IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 30, 2019, accepted December 12, 2019, date of publication December 23, 2019, date of current version January 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2961835

Supersaliency: A Novel Pipeline for Predicting Smooth Pursuit-Based Attention Improves Generalisability of Video Saliency

MIKHAIL STARTSEV AND MICHAEL DORR
Chair of Human-Machine Communication, Department of Electrical and Computer Engineering, Technical University of Munich, 80333 München, Germany
Corresponding author: Mikhail Startsev (mikhail.startsev@tum.de)

This work was supported in part by the Elite Network Bavaria, funded by the Bavarian State Ministry for Research and Education.

ABSTRACT Predicting attention is a popular topic at the intersection of human and computer vision. However, even though most of the available video saliency data sets and models claim to target human observers' fixations, they fail to differentiate them from smooth pursuits (SPs), a major eye movement type that is unique to perception of dynamic scenes. In this work, we strive for a more meaningful prediction and conceptual understanding of saliency in general. Because of the higher attentional selectivity of smooth pursuit compared to fixations modelled in traditional saliency research, we refer to the problem of SP prediction as ''supersaliency''. To make this distinction explicit, we (i) use algorithmic and manual annotations of SPs and fixations for two well-established video saliency data sets, (ii) train Slicing Convolutional Neural Networks for saliency prediction on either fixation- or SP-salient locations, and (iii) evaluate our and 26 publicly available dynamic saliency models on three data sets against traditional saliency and supersaliency ground truth. Overall, our models outperform the state of the art in both the new supersaliency and the traditional saliency problem settings, for which literature models are optimised. Importantly, on two independent data sets, our supersaliency model shows greater generalisation ability than its counterpart saliency model and outperforms all other models, even for fixation prediction. Furthermore, we tested an end-to-end video saliency model, which also showed systematic improvements when smooth pursuit was predicted either exclusively or together with fixations, with the best performance achieved when the model was trained for the supersaliency problem. This demonstrates the practical benefits and the potential of principled training data selection based on eye movement analysis.

INDEX TERMS Eye movements, saliency, smooth pursuit prediction.

I. INTRODUCTION

Saliency prediction has a wide variety of applications, be it in computer vision, robotics, or art [1], ranging from image and video compression [2], [3] to such high-level tasks as video summarisation [4], scene recognition [5], or human-robot interaction [6]. Its underlying idea is that in order to efficiently use the limited neural bandwidth, humans sequentially sample informative parts of the visual input with the high-resolution centre of the retina, the *fovea*. The prediction of gaze should thus be related to the classification of informative and uninformative video regions. However, humans use two different processes to foveate visual content. During fixations, the eyes remain mostly stationary; during smooth pursuit (SP), in contrast, a moving target is tracked by the eyes to maintain foveation. Notably, SP is impossible without such a target, and it needs to be actively initiated and maintained. For models of attention, this is a critical distinction: Because the eyes are stationary (''fixating'') in their default state, ''spurious'' fixations may be detected even if a subject is not attentively looking at the input; SP, however, always co-occurs with attention. In addition, visual sensitivity seems to be improved during SP (e.g. higher chromatic contrast sensitivity [7] and enhanced visual motion prediction [8]).

The ultimate goal of all eye movements and perception is to facilitate action in the real world. In a seminal paper [11], Land showed that gaze strategies, and SP in particular, play a critical role during many everyday activities. Similar results have been found for driving scenarios, where attention is crucial. Studies show that tangential [12] and target [13]

The associate editor coordinating the review of this manuscript and approving it for publication was Jianqing Zhu.

FIGURE 1. Empirically observed neurological differences between fixation and smooth pursuit: Large brain areas (highlighted) show significantly increased activation levels during pursuits compared to fixations (detected by [9]) in the *studyforrest* data set [10]; none demonstrate the inverse effects. A set of representative slices along orthogonal planes for a model brain is presented in this figure (slice numbers labelled on the figure) for the visualisation of the differences between fixation and pursuit conditions. Significance was determined via analysing the "standard score", or "Z-score" values.

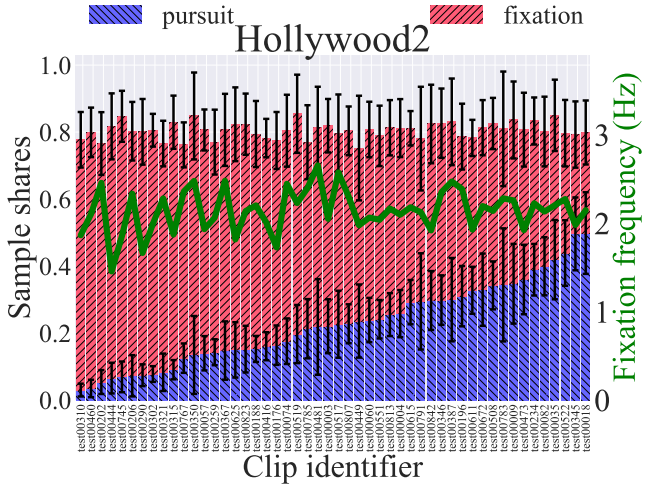


FIGURE 2. Behavioural differences between fixation and smooth pursuit: Saliency metrics typically evaluate against fixation onsets, which, as detected by a traditional approach [21] (green line), are roughly equally frequent across videos. However, applying a more principled approach to separating smooth pursuit from fixations [9] reveals great variation in the number of fixation (red bars) and pursuit (blue bars) samples (remaining samples are saccades, as well as blinks and other unreliably tracked samples).

locations during curve driving are "fixated" with what actually consists, in part, of SP. In natural driving, roadside objects are often followed with pure SP, without head motion [14]. Following objects that are moving relative to the car with gaze (by turning the head, via an SP eye movement, or a combination of both) is a clearer sign of attentive viewing, compared to the objects of interest crossing the line of sight.

In practice, it is difficult to segment the – often noisy – eye tracking signal into fixations and SPs, and thus many researchers combine all intervals where the eyes are keeping track of a point or an object into "fixations" [15]. Nevertheless, it is well established that e.g. individuals with schizophrenia show altered SP behaviour [16], [17], and recently new methods for gaze-controlled user interfaces based on SP have been presented [18]–[20]. This demonstrates some of the practical benefits of carefully separating the eye movements that make up the human gaze behaviour.

FIGURE 1 and FIGURE 2 show two analyses corroborating the importance of SP for models of attention in the context of a more tractable task of video watching. In FIGURE 1, data from the publicly available *studyforrest* data set[1] [10], which combine functional brain imaging and eye tracking during prolonged movie watching, were comparatively evaluated for SP vs. fixation episodes in a preliminary study. The highlighted voxels show that large brain areas are more active during SP compared to fixations; notably, no brain areas were more active during fixation than during SP. In other words, SP is representative of greater neurological engagement. The sparser selectivity of SP is demonstrated in FIGURE 2, where the relative share of SP and fixation gaze samples is plotted for 50 randomly selected clips from Hollywood2 [22]. Even though the number of traditionally detected fixations (but not their duration) is roughly the same for all clips, the amount of SP ranges from almost zero to half of the viewing time.

Taken together, these observations let us hypothesise that SP is used to selectively foveate video regions that demand greater cognitive resources, i.e. contain more information. In practice, automatic pursuit classification as applied to the *studyforrest* and Hollywood2 data sets may not be perfect, but the results in FIGURE 1 corroborate that even with potentially noisy detections, SP corresponds to higher brain activity, and thus to more meaningful saliency.

Therefore, explicitly modelling SP in a saliency pipeline should benefit the classification of informative video regions. Beyond a better understanding of attention, there might also be direct applications of SP prediction itself, e.g. in semi-autonomous driving (verification of attentive supervision), telemedicine (monitoring of SP impairment as a vulnerability marker for schizotypal personality disorder [23], e.g. during TV or movie watching [17]), or gaze-based interaction (analysis of potential distractors in user interfaces for AR/VR).

Despite the fundamental differences between SP and fixations, however, available saliency data sets ignore this distinction, and the computational models naturally follow suit [24], [25]. In fact, not one of the video saliency models we came across mentions the tracking of objects performed via SP, and the only data set we found to purposefully attempt separating SP from fixations is GazeCom [21], which simply discarded likely pursuits in order to achieve cleaner fixation detection.

We argue that processing the eye tracking recordings in a systematic and comprehensively described way in order

[1]These data were obtained from the OpenfMRI database. Its accession number is ds000113d.

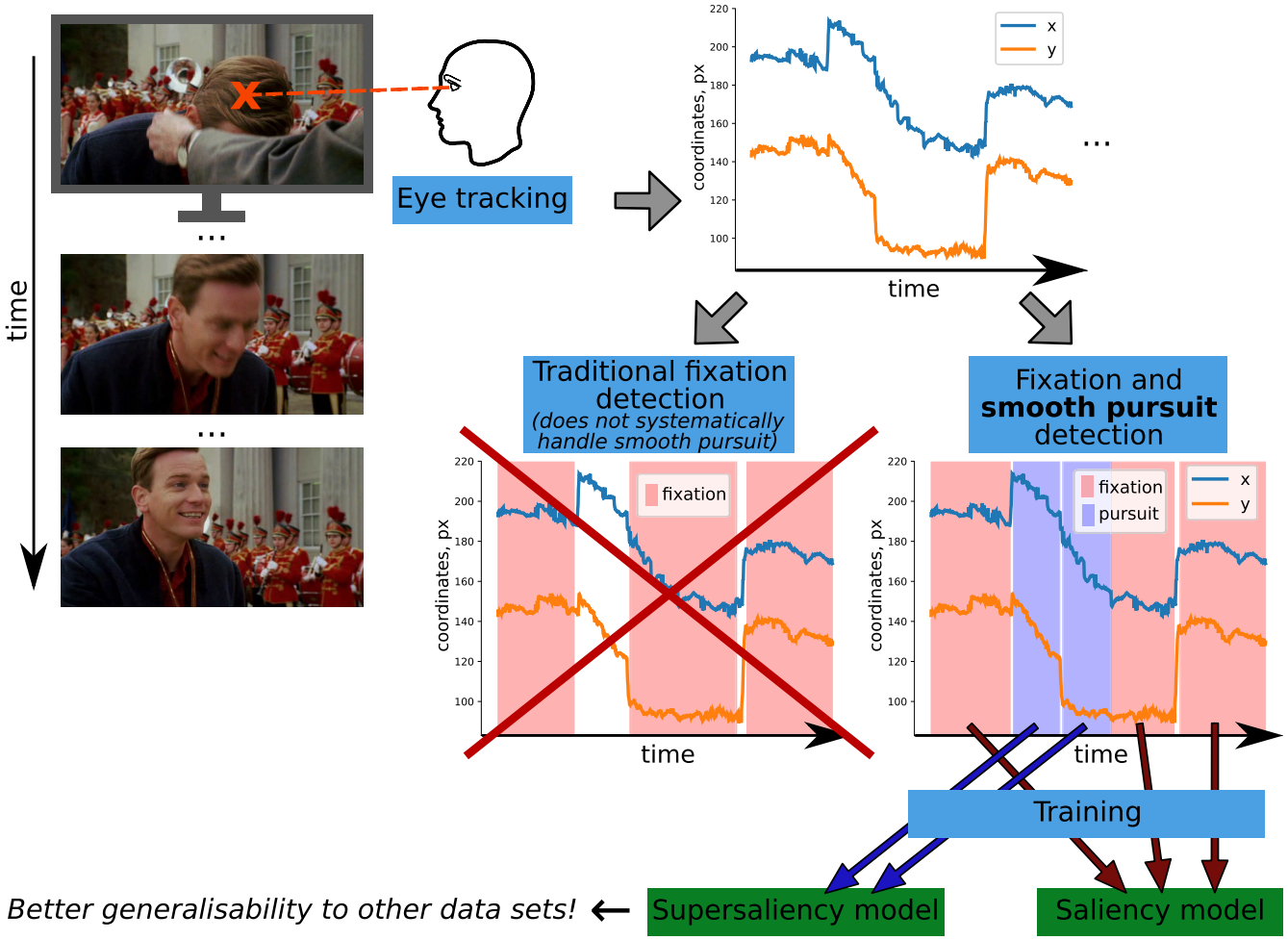


FIGURE 3. Overview of the proposed pipeline.

to extract moments of attention, be that fixations or smooth pursuits, is a vital first step in any pipeline of modelling human attention. This would allow for saliency to be treated not as a purely computational challenge of predicting some heat map frames for a video input, but as a task that could help us better understand human perception and attention.

In this manuscript, we extend our previous work [26] and make the following contributions: First, we introduce the problem of smooth pursuit prediction – *supersaliency*, so named due to the properties separating it from traditional, fixation-based saliency (e.g. see FIGURE 1 and FIGURE 2). In this problem setting, the saliency map values correspond to how likely a certain input video location is to induce SP. We then provide automatically labelled [9], large-scale training and test sets for this problem (building on the Hollywood2 data set [22]), as well as a manually labelled, smaller-scale test set of more complex scenes in order to test the generalisability of saliency models (building on the GazeCom data set [21], [27]). For both, we provide SP-only and fixation-only ground truth saliency maps. We also discuss the necessary adjustments to the evaluation of supersaliency (and video saliency in general) due to its high inter-video variance, introducing weighted averaging of individual clip scores.

Furthermore, we propose a deep dynamic saliency model for (super)saliency prediction, which is based on the slicing convolutional neural network (S-CNN) architecture [28]. After training our proposed model for both saliency and supersaliency prediction on the same overall data set, we demonstrate that our models excel at their respective problems in the test subset of the large-scale data set, compared to over two dozen literature models. Finally, we show that training for predicting smooth pursuit reduces data set bias: The supersaliency-trained model better generalises to two independent sets (without any additional training) and performs best even for traditional saliency prediction. We demonstrate the same pattern with an additional, end-to-end video saliency model we introduce in this work.

The overview of the (super)saliency modelling pipeline we are proposing in this work can be seen in Figure 3.

II. RELATED WORK

Predicting saliency for images has been a very active research field. A widely accepted benchmark is represented by the

MIT300 data set [29], [30], which is currently dominated by deep learning solutions. Saliency prediction for videos, however, lacks an established benchmark. It generally is a challenging problem, because, in addition to larger computational cost, objects of interest in a dynamic scene may be displayed only for a limited time and in different positions and contexts, so attention prioritisation is more crucial.

Taking this prioritisation principle to the extreme, works on salient object *detection* typically attempt to identify an object of interest in each frame (usually the same throughout a processed video clip). Saliency *prediction*, on the contrary, is not attempting to identify a single attention centre in the video, but aims at predicting the overall distribution of attention in the video as a heat map sequence. The salient object detection task is, therefore, much closer to segmentation at its core, with the added aspect of automatically selecting the dominant object in the scene. Despite the difference in problem formulations, both video saliency prediction and salient object detection essentially belong to the class of video-to-video transformation tasks, so some methodology can be shared between the two. We therefore include several works on both problems in our literature overview, when the methods are either directly or potentially applicable to the problem posed in this study.

Somewhat bridging these two saliency-related areas, [31] enabled attention shifting in the domain of salient object detection. That work directly tied the annotated objects of interest to human gaze directions, and therefore allowed for the objects to become or stop being salient as the scene unfolds.

A. SALIENCY PREDICTION

A variety of algorithms has been introduced to deal with human attention prediction [1]. Video saliency approaches broadly fall into two groups: Published algorithms mostly operate either in the original pixel domain [2], [24], [32], [33] and its derivatives (such as optic flow [34] or other motion representations [35]), or in the compression domain [25], [36], [37]. Transferring expert knowledge from images to videos in terms of saliency prediction is consistent with pixel-domain approaches, and the mounting evidence that motion attracts our eyes contributed to the development of compression-domain algorithms.

Traditionally, from the standpoint of perception, saliency models are also separated into two categories based on the nature of the features and information they employ. Bottom-up models focus their attention (and assume human observers do the same) on low-level features such as luminance, contrast, or edges. For videos, local motion can also be added to the list, together with the video encoding information. Hence, all the currently available compression-domain saliency predictors are effectively bottom-up.

Top-down models, on the contrary, use high-level, semantic information, such as concepts of objects, faces, etc. These are notoriously hard to formalise. One way to do so would be to detect certain objects in the video scenes, as was done in [22], where whole human figures, faces, and cars were detected. Another way would be to rely on developments in deep learning and the field's endeavour to implicitly learn important semantic concepts from data. In [38], either RGB space or contrast features are augmented with residual motion information to account for the dynamic aspect of the scenes (i.e. motion is processed before the CNN stage in a hand-crafted fashion). The work in [39] uses a 3D CNN to extract features, plus an LSTM network to expand the temporal span of the analysis. Other researchers use further additional modules, such as the attention mechanism [40] or object-to-motion sub-network [41]. In [42], a modified convolutional LSTM (using multi-scale dilations) is employed to accurately detect salient objects in video sequences. In a similar vein of research, [43] also modified the typical convolutional LSTM structure for video-to-video prediction by developing a parallel multi-dimensional extension of this structure. This modification allows for a much more complete utilisation of the relevant past information for each pixel. While our work does not focus on the architecture design, it would doubtlessly be interesting to explore the effects of systematically differentiating between fixations and smooth pursuits in the context of saliency prediction with a wider spectrum of computational models (our work tested two different approaches).

Whereas using a convolutional neural network in itself does not guarantee the top-down nature of the resulting model, its multilayer structure fits the idea of hierarchical computation of low-, mid-, and high-level features. A work by Krizhevsky *et al.* [44] pointed out that while the first convolutional layers learned fairly simplistic kernels that target frequency, orientation, and colour of the input signal, the activations in the last layer of the network corresponded to a feature space, in which conceptually similar images are close, regardless of the distance in the low-level representation space. Another study [45] concluded that, just like certain neural populations of a primate brain, deep networks trained for object classification create such internal representation spaces, where images of objects in the same category get similar responses, and images of differing categories get dissimilar ones. Other properties of the networks discussed in that work indicate potential insights into the visual processing system that can be gained from them.

B. VIDEO SALIENCY DATA SETS

A broad overview of existing data sets is given in [46]. Here, we dive into the aspect particularly relevant to this study – the identification of "salient" locations of the videos, i.e. how did the authors deal with dynamic eye movements. For the most part, this question is addressed inconsistently. The majority of the data sets either make no explicit mention of separating smooth pursuit from fixations (ASCMN [47], SFU [48], two Hollywood2-based sets [22], [49], DHF1K [40]) or rely on the event detection built into the eye tracker, which in turn does not differentiate SP from fixations (TUD [50], USC CRCNS [51], CITIUS [24], LEDOV [41]). IRCCyN/IVC (Video 1) [52] does not mention any eye movement types at

all, whereas IRCCyN/IVC (Video 2) [53] only names SP in passing.

There are two notable exceptions from this logic. First, DIEM [54], which comprises video clips from a rich spectrum of sources, including amateur footage, TV programs, and movie trailers, so one would expect a hugely varying fixation–pursuit balance. The respective paper touches on the properties of SP that separate it from fixations, but in the end only distinguishes between blinks, saccades, and non-saccadic eye movements, referring to the latter as generic *foveations*, which combine fixations and SPs.

GazeCom [21], on the other hand, explicitly acknowledges the difficulty of distinguishing between fixations and smooth pursuits in dynamic scenes. The used fixation detection algorithm employed a dual criterion based on gaze speed and dispersion. However, the recently published manually annotated ground truth data [27] show that these coarse thresholds are insufficient to parse out SP episodes.

Part of this work's contribution is, therefore, to provide a large-scale supersaliency (SP) and saliency (fixations) data set based on Hollywood2, as well as establishing a pipeline for (super)saliency evaluation.

III. SALIENCY AND SUPERSALIENCY

In this section, we describe the methodology behind the (super)saliency prediction in this work. Our approach relies on two main components: A large-scale data set of human video free-viewing, where the raw eye tracking data are available, and a computational model. Such data set would allow us to analyse the gaze recordings to parse out the episodes of either fixations or smooth pursuits. The detected samples of the two eye movements can be then directly used to train the proposed model.

A. DATA SETS AND THEIR ANALYSIS

GazeCom [21], which we used because it is the only saliency data set that also provides full manual annotation of eye movement events [27], [55], contains eye tracking data for 54 subjects, with 18 dynamic natural scenes used as stimuli, around 20 seconds each. At over 4.5 total hours of viewing time, this is the largest manually annotated eye tracking data set that accounts for SP. A high number of observers and the hand-labelled eye movement type information make this a suitable benchmark set. FIGURE 4a displays an example scene, together with its empirical saliency maps for both fixations and smooth pursuits, and the same frames in saliency maps predicted by different models.

Hollywood2 [22], selected for its diversity and the sheer amount of eye tracking recordings, contains about 5.5 hours of video (1707 clips, split into training and test sets), viewed by 16 subjects. The movies have all types of camera movement, including translation and zoom, as well as scene cuts. While the full training subset was used, we randomly selected 50 clips from the test subset (same as in FIGURE 2) for testing all the models. Example frames and respective (super)saliency maps can be seen in FIGURE 4b. Since manual labelling is impractical due to the data set size (over 70 h of total viewing time), we used our publicly available toolbox [27] implementing a state-of-the-art SP and fixation detection algorithm [9], [55]. A large-scale evaluation of this toolbox was performed in [56], where it demonstrated excellent performance when compared to the GazeCom ground truth data, and generalised well to an independent data set.

CITIUS [24] was recently used for a large-scale evaluation of the state of the art in connection with a novel model (AWS-D). It contains both real-life and synthetic video sequences, split into subcategories of static and moving camera. For our evaluation, we used the real-life part, **CITIUS-R** (22 clips totalling ca. 7 minutes, 45 observers). Only fixation onset and duration data are provided by the authors, so SP analysis was impossible.

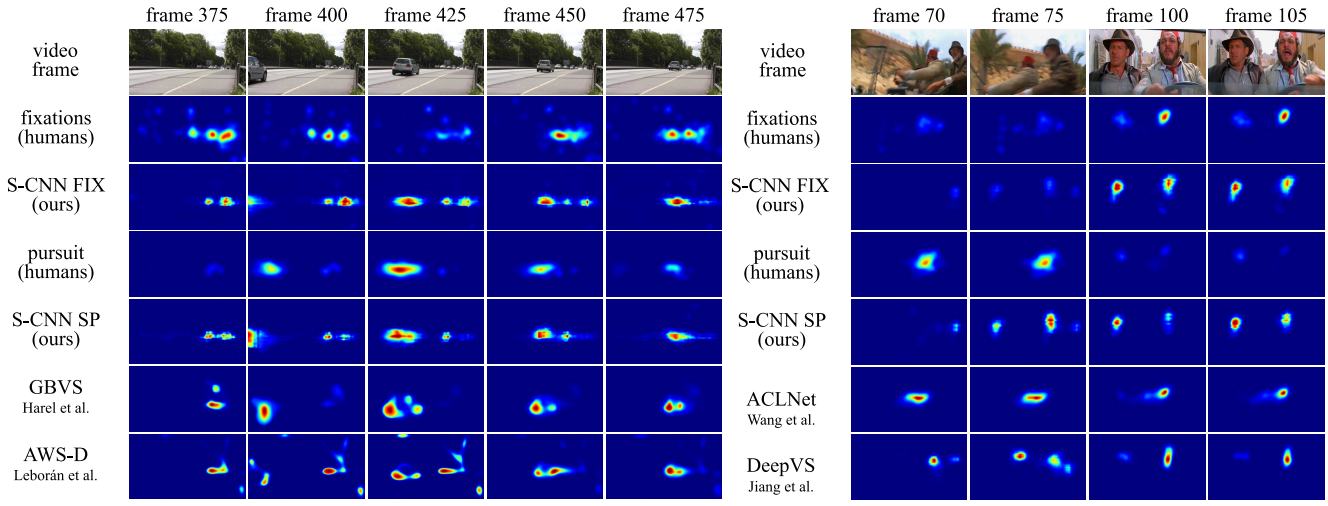
By definition, fixations are almost stationary, so that a single point (usually, mean gaze position placed at temporal onset) sufficiently describes an entire fixation. In line with the literature, we evaluated the prediction of such fixation onsets in the ''onset'' condition (detected by a standard algorithm [21] for GazeCom and Hollywood2, provided with the data set for CITIUS-R). Notably, the reference models are likely optimised for this problem setting.

To describe the trajectory of an SP episode, however, all its gaze samples need to be taken into account. Accordingly, both the GazeCom ground truth and the toolbox [27] we used for Hollywood2 provide sample-level annotations. These annotations were used for evaluating the prediction of pursuit-based attention in the ''SP'' condition, i.e. model predictions were tested against the set of individual pursuit gaze samples. The ''FIX'' condition utilised individual fixation samples as well (similar to [54]), and is, in principle, not very different from ''onset''. By directly mirroring the implementation of the ''SP'' condition, however, it allowed for a fairer comparison between the two.

B. SLICING CNN SALIENCY MODEL

We adopted the slicing convolutional neural network (S-CNN) architecture [28]. To achieve saliency prediction, we extended patch-based image analysis (e.g. [57] for image saliency, and [38] for individual video frames) to subvolume-based video processing. This way, we are still able to capture motion patterns, while maintaining a relatively straightforward binary classification-based architecture – (super)salient vs. non-salient subvolumes. Initially, we did not use more complex end-to-end approaches in order to keep the proof-of-concept implementation of fixation- and pursuit-based training as straightforward as possible, without intermediate steps of having to convert locations of corresponding samples into continuous saliency maps. These steps would introduce additional data parametrisation and, potentially, biases into the pipeline. However, we additionally validate the idea of supersaliency prediction with an end-to-end model in Section IV.

S-CNN [28] takes an alternative approach to extracting motion information from a video sequence. Instead of



(a) GazeCom ("street" clip) (b) Hollywood2 ("actioncliptest00416" clip)

FIGURE 4. Frame examples from GazeCom (a) and Hollywood2 (b) videos (first row), with their respective empirical ground truth fixation-based saliency (second row) and smooth pursuit-based supersaliency (fourth row) ground truth maps. Algorithmic predictions (all identically histogram-equalised, for fair visual comparison) occupy the rest of the rows. The choice of saliency models for visual comparison was based on best average performance on the respective data set.

handcrafted motion descriptors [38], 3D convolutions [58], or recurrent structures [39], S-CNN achieves temporal integration by rotating the feature tensors after initial individual frame-based feature extraction. This way, time (frame index) is one of the axes of the subsequent convolutions. The architecture is based on VGG-16 [59], with the addition of dimension swapping operations and temporal pooling. The whole network would consist of three branches, in each of which the performed rotation is different, and the ensuing convolutions are performed in the planes *xy* (equivalent to no rotation), *xt*, or *yt* (branches are named respectively). Due to the size of the complete model, only one branch could be trained at a time. We decided to use the *xt*-branch for our experiments (see FIGURE 5), since it yielded the best individual results in [28], and the horizontal axis seems to be more important for human vision [60] and SP in particular [61]. We also tested the other branches separately and the late fusion of their results, but the *xt* branch was the best individual performer,

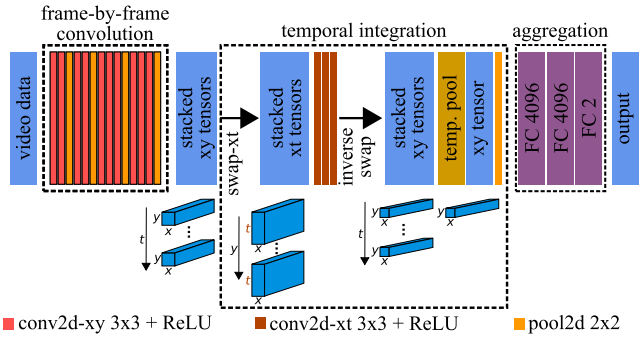


FIGURE 5. The *xt* branch of the S-CNN architecture for binary salient vs. non-salient video subvolume classification. Temporal integration is performed after the *swap-xt* operation via the convolutions operating in the *xt* plane and temporal pooling.

and the fusion did not produce sufficient performance gains to justify the tripled computation time. Therefore, we do not report these results in this paper. Similarly, our preliminary tests with 3D-CNN architectures, similar to results in [28], led us to opt for the better-performing S-CNNs instead.

As input to our model, we used RGB video subvolumes $128\,\mathrm{px} \times 128\,\mathrm{px} \times 15\,\mathrm{frames}$ (px denoting pixels) around the pixel to be classified. Similar subvolumes were used in [62] for unsupervised feature learning. Unlike [38], we did not extract motion information explicitly, but relied on the network architecture entirely without any further input manipulations in order to achieve a simpler data processing pipeline.

To go from binary classification to generating a continuous (super)saliency map, we took the probability for the positive class at the soft-max layer of the network (for each respective surrounding subvolume of each video pixel). To reduce computation time, we only did this for every 10^{th} pixel along both spatial axes. We then upscaled the resulting low-resolution map to the desired dimensions. For GazeCom and Hollywood2, we generated saliency maps at $640 \times 360\,\mathrm{px}$, whereas for CITIUS-R, the original resolution of 320×240 was used.

C. TRAINING DETAILS

Out of 823 training videos in Hollywood2, 90% (741 clips) were used for training and 10% for validation. Before extracting the subvolumes centred around positive or negative locations of our videos, these were rescaled to 640×360 pixels size and mirror-padded to reduce boundary effects. In total, the 823 clips contain 4,520,813 unique SP and 10,448,307 unique fixated locations. To assess the influence of the eye movement type in the training data, we fitted the same model twice for two different purposes. First,

we trained the *S-CNN SP* model for predicting *supersaliency*, so the positive locations were those where SP had occurred. Analogously, for the *S-CNN FIX* model predicting purely fixation-based (i.e. excluding SP) *saliency*, the input video subvolumes where observers had fixated were labelled as positive.

For both *S-CNN SP* and *S-CNN FIX*, the training set consisted of 100,000 subvolumes, half of which were positives (as described above, randomly sampled from the respective eye movement locations in the training videos), half negatives (randomly selected in a uniform fashion to match the number of positive samples per video, excluding the subvolumes already in the positive set). For validation, 10,000 subvolumes were used, same sampling procedure as for the training set.

Convolutional layers were initialised with pre-trained VGG-16 weights, fully-connected layers were initialised randomly. We used a batch size of 5, and trained both models for 50,000 iterations with stochastic gradient descent (with momentum of 0.9, learning rate starting at 10^{-4} and decreasing 10-fold after every 20,000 iterations), at which point both loss and accuracy levelled out.

D. ADAPTIVE CENTRE BIAS

Since our model is inherently spatial bias-free, as it deals purely with individual subvolumes of the input video, we applied an adaptive solution to each frame – the gravity centre bias approach of Wu *et al.* [34], which emphasises not the centre of the frame, but the centre of mass in the saliency distribution. At this location, a single unit pixel is placed on the bias map, which is then blurred with a Gaussian filter (σ equivalent to three degrees of the visual field was chosen) and normalised to contain values ranging from 0 to the highest saliency value of the currently processed frame. Each frame of the video saliency map was then linearly mixed with its respective bias map (with a weight of 0.4 for the bias, and 0.6 for the original frame, as in [34]).

IV. VALIDATION WITH A MORE COMPLEX MODEL

As discussed in Section II-A, more sophisticated architectures have been developed over time to better handle both the spatial and the temporal aspects of deep video processing. While the slicing CNN model we used in Section III-B allowed us to avoid any additional steps when going from concept to implementation, end-to-end architectures provide a more modern and efficient tool for saliency prediction.

In order to investigate whether the benefits of supersaliency hold for an end-to-end model, we implemented an architecture combining two recent works: (i) the fully-convolutional deep DenseNet from [63] for efficient information extraction from each 2D frame, and (ii) the introduction of several convolutional LSTMs into an encoder-decoder network [64] for temporal integration. Thus, we replaced the encoder part of the network in [64] with a DenseNet structure as in [63], keeping the decoder simple. The dense blocks were modified to process the video frames in a time-distributed fashion (i.e. identical operations applied to all frames). The model is sketched in FIGURE 6. A detailed model description can be found in the supplementary material.

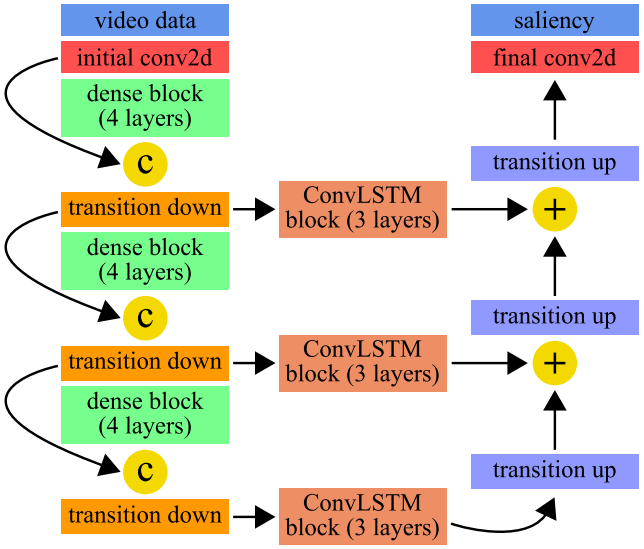


FIGURE 6. The outline of the end-to-end architecture we used for additional testing of our pipeline. In this scheme, "c" stands for the concatenation operation, "+" – for addition. In our experiments, ground truth saliency is provided as related to solely fixations, solely smooth pursuits, or both eye movements together.

A. TRAINING DETAILS

The Hollywood2 training set was randomly subdivided in the following way: 770 clips (ca. 200,000 frames) were used for training, 53 clips (ca. 15,000 frames) – for validation. The ground truth saliency map sequences were generated in the same way as for evaluation (see Section V-B). For this experiment, we trained the model to predict the saliency maps produced either for fixation or smooth pursuit samples only, or for the combination of both. The first two conditions correspond to purely fixation-based (traditional) saliency and purely pursuit-based supersaliency; the latter is very similar to only removing the saccades, and aggregating all the remaining gaze samples, as e.g. in [54].

We used Kullback-Leibler divergence as loss on the three-dimensional tensors of saliency (time $\times$ x $\times$ y), and trained the model for 10 epochs (500 iterations in each) with Adam optimiser [65] with default parameters (cf. Keras 2.2.4). The final model was selected based on the validation loss. Due to GPU memory constraints, we limited the input to this relatively large model to sequences of 12 frames (at $128 \times 72\,\mathrm{px}$) and used a batch size of 4. During training, the model produced sequences of 12 corresponding saliency frames for each input sequence. During testing, no video subdivision was performed.

Since this model operated on relatively low-resolution clips, we did not expect its saliency prediction to achieve benchmark-beating performance, but separately evaluated it and used its results to support our argument about the

potential of the supersaliency problem setting, and the importance of the smooth pursuit eye movement for saliency in general.

V. EVALUATION

A. REFERENCE MODELS

We compared our approach to a score of publicly available dynamic saliency models. For compression domain models, we followed the pipeline and provided source code of Khatoonabadi *et al.* [25], generating the saliency maps for all videos at 288 pixels in height, and proportionally scaled width for PMES [66], MAM [67], PIM-ZEN [68], PIM-MCS [69], MCSDM [70], MSM-SM [71], PNSP-CS [72], and a range of OBDL-models [25], as well as pixel-domain GBVS [32], [73] and STSD [74]. Instead of the static AWS [75] that was used in [25], we evaluated AWS-D [24], its recent extension to dynamic stimuli (for GazeCom, after downscaling to $640 \times 360\,\mathrm{px}$ due to memory constraints, other data sets – at their original resolution). We also computed the three invariants (H, S, and K) of the structure tensor [76] at fixed temporal (second) and spatial (third) scales. For Hollywood2, the approach of Mathe and Sminchisescu [77], combining static (low-, mid-, and high-level) and motion features, was evaluated as well.

Deep models for saliency prediction on videos are much scarcer than such models for static images. As of yet, the problem of finding reference models in this domain is further confounded by the absence of publicly available code or data of some approaches, e.g. [39], and the popularity of salient object detection approaches and data sets, e.g. [78]–[80]. Included in our set of reference models are two recent approaches: DeepVS (OMCNN-2CLSTM) [41] – code available via [81] – and ACLNet [40] – code available via [82]. We ran both with default parameters on all three data sets.

B. BASELINES

The set of baselines was inspired by the works of Judd *et al.* [29], [30]: *Chance*, *Permutation*, *Centre*, *One Human*, and *Infinite Humans* (as a limit). The latter two cannot be computed unless gaze data *for each individual observer* are available (i.e. not possible for CITIUS). All the random baselines were repeated five times per video of each data set. The *ground truth saliency maps* were obtained via superimposing spatio-temporal Gaussians at every attended location of all the considered observers. The two spatial sigmas were set to one degree of visual angle (commonly used in the literature as the approximate fovea size, e.g. [29], [83]; [77] uses 1.5°). The temporal sigma was set to a frame count equivalent of $1/3$ of a second (so that the effect would be mostly contained within one second's distance).

C. METRICS

For a thorough evaluation, we took a broad spectrum of metrics (all computed the same way for fixation samples and onsets – saliency – and smooth pursuit samples – supersaliency – for the data sets described in Section III-A), mostly based on [83]: AUC-Judd, AUC-Borji, shuffled AUC (sAUC), normalised scanpath saliency (NSS), histogram similarity (SIM), correlation coefficient (CC), and Kullback-Leibler divergence (KLD), as well as Information Gain (IG) [84]. We additionally computed balanced accuracy (same positive and negative location sets as for AUC-Borji; accuracy at the equal error rate point).

In our implementation of sAUC and IG, in order to obtain salient locations of other clips, we first rescaled their temporal axes to fit the duration of the evaluated clip, and then sampled not just spatial (like e.g. [24]), but also temporal coordinates. This preserves the temporal structure of the stimulus-independent bias: E.g. the first fixations after stimulus display tend to have heavier centre bias than subsequent ones in both static images [85] and videos [86].

For GazeCom and Hollywood2, we fixed all saliency maps to $640 \times 360\,\mathrm{px}$ resolution during evaluation, either for memory constraints, or for symmetric evaluation in case of differently shaped videos. For CITIUS, the native resolution of $320 \times 240\,\mathrm{px}$ was maintained.

1) METRIC AVERAGING

Due to its selectivity (i.e. observers can decide not to pursue anything), SP is sparse and highly unbalanced between videos (see FIGURE 2). Simply averaging the performance scores across all videos of the data set could introduce artefacts for many metrics. For AUC-based metrics, for example, there exists a "perfect" aggregated score, which could be computed by combining the data over all the videos *before* computing the metric, i.e. merging all positives and all negatives beforehand. This is, however, not always possible, as many models use per-video or even per-frame normalisation as the final step, either to allow for easier visualisation, or to use the full spectrum of the 8-bit integer range, if the result is stored as a video. To demonstrate this averaging problem, we randomly sampled non-trivial subsets of video clips (100 times for all the possible subset sizes) of all three utilised test sets, and computed per-clip AUC-Borji and sAUC scores for our *S-CNN SP* model (without any normalisation of its outputs). We combined these via either regular or weighted (according to the number of SP- or fixation-salient locations samples, depending on the problem setting) averaging. This combination is then compared to the perfect score, as described above. We found that averaging per-video AUC scores is a significantly poorer approximation of the ideal score than their *weighted mean* ($p \ll 0.01$, for (super)saliency prediction on GazeCom and Hollywood2, see Table 1).

We will, therefore, present the weighted averaging results for supersaliency prediction. Since fixations suffer from this problem to a lesser extent, this adjustment is not essential there. However, in the data sets with great variation of fixation samples' share (e.g. Hollywood2: 30% to 78% in our 50-clip subset), we would generally recommend using weighting for fixation prediction evaluation as well. Conventional mean

TABLE 1. Means and standard deviations of the absolute error of "perfect AUC" estimation with *regular* and *weighted* averaging, as well as one-sided two-sample Kolmogorov-Smirnov test p-values (with the null hypothesis that regular averaging, as a way to estimate the perfect AUC score, produces absolute errors that are smaller than or equal to those of weighted averaging). Except for CITIUS-R, weighted averaging always demonstrates a statistically significant ($p \ll 0.01$) advantage over regular averaging.

Statistic	Absolute error properties	GazeCom			Hollywood2 (50 clips)			CITIUS-R
		SP	FIX	onsets	SP	FIX	onsets	onsets
AUC-Borji	mean (*regular* averaging)	0.038	0.011	0.012	0.022	0.017	0.018	0.0125
	mean (*weighted* averaging)	0.011	0.01	0.01	0.008	0.012	0.009	0.0135
	SD (*regular* averaging)	0.03	0.007	0.008	0.009	0.009	0.008	0.0079
	SD (*weighted* averaging)	0.01	0.007	0.007	0.004	0.004	0.004	0.0075
	p-value	9e-205	4e-16	8e-16	0e+00	0e+00	0e+00	0.92
sAUC	mean (*regular* averaging)	0.039	0.013	0.014	0.038	0.029	0.031	0.0173
	mean (*weighted* averaging)	0.015	0.011	0.011	0.011	0.015	0.012	0.0169
	SD (*regular* averaging)	0.031	0.008	0.009	0.016	0.014	0.013	0.0092
	SD (*weighted* averaging)	0.014	0.008	0.008	0.006	0.005	0.005	0.0089
	p-value	1e-137	4e-20	2e-25	0e+00	0e+00	0e+00	0.02

results for fixations are, nevertheless, presented for comparability with the literature (weighted results reveal a similar picture).

2) CROSS-AUC

Another point we raise in our evaluation is directly distinguishing SP-salient from fixation-salient pixels based on the saliency maps. To this end, we introduced *cross-AUC (xAUC)*: The AUC is computed for the positive samples' set of all pursuit-salient locations, with an equal number of randomly selected fixation-salient locations for the same stimulus used as negatives. The baselines' performance on this metric will be indicative of how well the targets for these two eye movements can be separated (in comparison to the separation of salient and non-salient locations). If a model scores above 50% on this metric, it on average favours (i.e. assigns higher saliency scores to) pursuit-salient locations over fixation-salient ones (since SP is chosen as the positive class). For the purpose of distinguishing the two eye movement types, the scores of 70% and 30% are, however, equivalent: Such scores would reveal that a model favours either SPs over fixations, or vice versa, respectively, with the same bias from not displaying any preference whatsoever (and the corresponding xAUC of 50%).

VI. RESULTS AND DISCUSSION

A. SLICING CNN RESULTS

We tested the outputs of 26 published dynamic saliency models, including two deep learning-based solutions, as well as our own S-CNN models – SP and fixation predictors both with and without the additional post-processing step of gravity centre bias. For brevity and because there is no principled way of averaging different metrics numerically, we present the results as average ranks (over the 9 metrics we used – see Section V-C) in Table 2. Complete tables of all metric scores for all 7 data types (corresponding to the columns of Table 2) and 35 baselines and models can be found in the supplementary material.

Traditional saliency prediction commonly evaluates only one sample per fixation, as we did in the "onset" condition. For supersaliency, however, all gaze samples need to be predicted individually, and for consistency we did the same for fixations in the "FIX" condition. In principle, this should give greater weight to longer fixations with more samples, but our results show that differences between evaluating in the "FIX" and "onset" conditions are small in practice (cf. respective columns in Table 2).

On average, our pursuit prediction model, combined with adaptive centre bias (*S-CNN SP + Gravity CB*), performs best, almost always making it to the first or the second position (and always in the top-4). Remarkably, this holds true both for the prediction of smooth pursuits and the prediction of fixations, despite training exclusively on SP-salient locations as positive examples. The success of our pursuit prediction approach in predicting fixations can be potentially attributed to humans pursuing and fixating similar targets, but the relative selectivity of SP allows the model to focus on the particularly interesting objects in the scene. Even without the gravity centre bias, both our saliency *S-CNN FIX* and supersaliency *S-CNN SP* models outperform the models from the literature on the whole, with their average rank at least two positions better than that of the next best model (ACLNet).

The fact that all our *S-CNN* models consistently outperform the traditional "shallow" reference models for both saliency and supersaliency prediction on all data sets demonstrates the potential of deep video saliency models. This is in line with the findings in e.g. [39], [87], where a deep architecture has shown superior fixation prediction performance, compared to non-CNN models. On Hollywood2, due to the very centre biased nature of the gaze locations [21], for example, only the deep learning models (*S-CNN*, ACLNet, and DeepVS) rank higher than the Centre Baseline or achieve non-negative information gain scores (cf. Table 2 and the tables in the supplementary material).

Only in the fixation prediction task on the Hollywood2 data set, the results of our best model are inferior to the two deep reference approaches (and only to those) – DeepVS and ACLNet. On both other data sets (GazeCom and CITIUS-R), as well as for supersaliency prediction on Hollywood2, our model is outperforming all reference algorithms. The two evaluated deep literature approaches are particularly weak on the GazeCom data set, and especially in the task of predicting

TABLE 2. Evaluation results, presented as the mean of rank values for all the metrics we compute (except for xAUC). "Onset" refers to evaluation against fixation onsets ("traditional" saliency). Where marked with *, ranking was computed for the weighted average of the scores. The rows with gray background correspond to baselines. Top-3 non-baseline results in each category are **boldified**.

Model	GazeCom			Hollywood2 (50 clips)			CITIUS-R	average rank
	SP*	FIX	onset	SP*	FIX	onset	onset	
Infinite Humans	1.0	1.0	1.0	1.0	1.0	1.0	–	1.0
S-CNN SP + Gravity CB	**4.9**	**2.9**	**2.9**	**4.0**	5.1	5.0	**3.3**	**4.0**
S-CNN FIX + Gravity CB	12.2	**2.8**	**2.8**	5.3	**4.6**	**4.1**	**3.9**	**5.1**
S-CNN SP	**3.0**	**4.4**	**4.1**	6.2	7.6	7.4	4.8	**5.4**
S-CNN FIX	**9.1**	4.6	4.8	7.7	6.7	6.8	5.6	6.4
ACLNet	24.3	11.0	10.7	**4.3**	**2.9**	**3.4**	**3.3**	8.6
DeepVS (OMCNN-2CLSTM)	25.4	9.8	11.0	**5.0**	**4.7**	**4.7**	8.2	9.8
GBVS	11.1	11.2	10.1	11.6	11.3	11.1	7.6	10.6
OBDL-MRF-O	13.8	12.2	11.9	13.7	13.3	11.8	9.9	12.4
OBDL-MRF-OC	15.1	13.9	13.7	14.8	14.2	12.9	11.2	13.7
AWS-D	14.9	7.9	7.4	24.0	18.0	18.2	7.2	14.0
OBDL-MRF-TO	18.6	13.9	14.8	12.6	14.3	15.2	13.3	14.7
OBDL-MRF	18.9	16.0	16.3	13.8	11.3	12.8	13.7	14.7
Centre	29.4	16.8	16.4	9.6	10.3	10.1	10.6	14.7
OBDL-MRF-T	23.1	13.2	14.9	12.6	12.2	15.1	15.3	15.2
One Human	18.7	19.2	22.6	11.1	9.8	10.6	–	15.3
OBDL-T	13.7	15.1	12.4	17.9	18.7	18.6	11.1	15.3
OBDL-MRF-C	16.3	16.1	16.0	15.9	15.9	14.2	13.1	15.4
OBDL-MRF-TC	20.6	12.9	14.2	12.8	16.8	17.1	15.6	15.7
OBDL-S	14.7	19.8	18.4	19.1	20.2	19.9	17.3	18.5
Mathe	–	–	–	20.7	21.7	21.9	–	21.4
Invariant-K	11.3	20.8	19.6	30.2	25.0	25.0	20.8	21.8
STSD	18.0	21.6	21.6	27.4	25.4	25.1	18.0	22.4
OBDL	22.4	23.2	22.4	22.9	22.9	22.8	20.8	22.5
PMES	11.8	27.8	27.0	22.0	27.0	27.4	27.0	24.3
PIM-ZEN	13.6	26.4	26.3	24.0	26.6	27.1	26.8	24.4
PIM-MCS	14.3	25.9	26.1	25.8	26.7	27.2	26.2	24.6
Invariant-S	28.6	21.9	22.2	32.0	27.7	27.2	22.8	26.0
MSM-SM	16.8	33.1	32.3	21.7	28.4	27.2	25.4	26.4
PNSP-CS	13.9	28.7	28.7	28.1	30.1	30.1	27.2	26.7
Permutation	33.4	29.3	29.4	27.7	23.4	22.3	24.8	27.2
MCSDM	13.4	28.4	28.3	30.7	31.0	31.6	28.1	27.4
Invariant-H	29.8	23.7	24.4	33.3	30.9	30.9	25.8	28.4
Chance	31.0	28.0	28.0	32.4	31.8	31.9	28.2	30.2
MAM	27.9	31.6	32.1	28.3	32.6	32.2	31.1	30.8

pursuit-based supersaliency. Qualitatively, we observed that their predicted saliency distributions tend to miss moving salient targets, unless these are close to the centre of the frame.

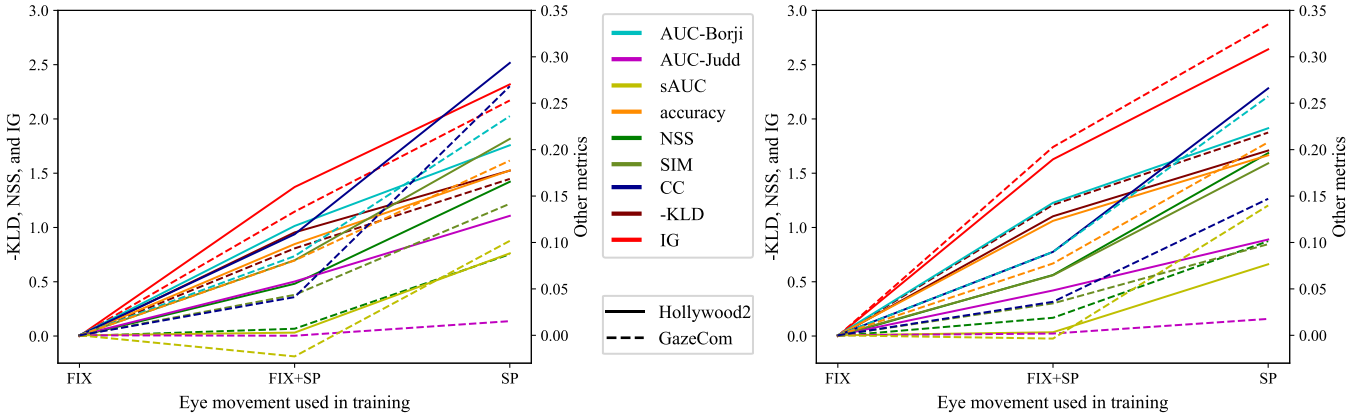
Both with and without the gravity centre bias, our supersaliency *S-CNN SP* models perform better than our respective saliency *S-CNN FIX* models (with the difference in average rank values of ca. one position). We emphasise that these models were only trained on the Hollywood2 training set. On the Hollywood2 test set, maybe not surprisingly, the fixation-predicting models perform better for fixation-based saliency and SP-predicting models perform better for pursuit-based supersaliency. On the two other data sets, however, the models that were trained for SP prediction generally perform better than their fixation-trained counterparts, indicating their greater generalisation capability.

To find informative video regions, we use humans as a yardstick, since they clearly excel at real-world tasks despite their limited perceptual throughput. Smooth pursuit is more selective than fixations and thus likely restricted to particularly interesting objects. The use of such sparser (yet more densely concentrated [27]), higher-quality training data could explain the superior generalisability of the supersaliency models to independent data sets.

For visual comparison, example saliency map sequences are presented in FIGURE 4a and FIGURE 4b for select GazeCom and Hollywood2 clips, respectively. It can be seen, for example, that our *S-CNN FIX* model differentiates well between fixation-rich and SP-rich frames in an example Hollywood2 clip.

B. END-TO-END VALIDATION

To additionally highlight the importance of pursuit and supersaliency in the context of a more state-of-the-art-like architecture, we trained a model encompassing both DenseNet and convolutional LSTM elements (see Section IV) in several set-ups: While keeping the training pipeline the same, we differently generated the ground truth saliency maps. We examined three conditions: (i) fixation-only attention, (ii) fixation- and pursuit-based attention, and (iii) pursuit-only attention. Taking the performance in the first condition as a baseline, we plot the absolute improvements of the saliency metrics in other conditions in



(a) Fixation prediction scores improvement

(b) Pursuit prediction scores improvement

FIGURE 7. Absolute improvement in the scores of our end-to-end saliency prediction model (see Section IV) due to the type of training data used (see *x* axis). FIGURE 7a reports the improvements of fixation-based saliency prediction, while FIGURE 7b depicts the same improvements for pursuit-based supersaliency prediction. Including pursuit-based attention into training (FIX+SP condition) is beneficial for the vast majority of metrics, compared to training for predicting purely fixation-based attention (FIX condition). Notably, training the model for the supersaliency problem directly (SP condition) always benefited our model, when tested for both the traditional saliency (7a) and supersaliency (7b) tasks.

FIGURE 7 (full absolute performance scores can be found in the supplementary material).

In these plots, the models trained in one of the three conditions are tested on the task of either fixation prediction (FIGURE 7a) or smooth pursuit prediction (FIGURE 7b). For both of the tasks, on GazeCom and Hollywood2 data sets alike, the values of performance measures are almost always improved when pursuit samples are added to fixation-only attention modelling (transition from "FIX" to "FIX+SP" conditions in the figures). Most importantly, performance of the model is invariably and noticeably improved when only pursuit samples are used for training. Only AUC-Judd on the GazeCom data set is just slightly improving between these conditions, because the metric is not class-balanced and is saturating at high saliency map resolutions. Evaluation at a lower resolution of saliency maps yields much more noticeable performance improvements for AUC-Judd as well (data not shown).

The results on CITIUS-R are qualitatively and quantitatively similar, and are not depicted for better figure readability. This again points to the greater across-data set generalisation capability of a model that was trained to predict supersaliency maps, compared to an identically trained model for saliency map prediction.

C. DISTINGUISHING FIXATION AND PURSUIT TARGETS

In the task of separating SP- and fixation-salient locations (the xAUC metric), most models yield a result above 0.5 on GazeCom, which means that they still, by chance or by design, assign higher saliency values to SP locations (unlike e.g. the centre baseline with xAUC score of 0.44, which implies that fixations on this data set are more centre biased than pursuits). Probably due to their emphasis on motion information, the top of the chart with respect to this metric is heavily dominated by compression-domain approaches (top-7 non-baseline models for GazeCom, top-4 for Hollywood2, cf. tables in the supplementary material). Even though in the limit (*Infinite Humans* baseline) this metric's weighted average can be confidently above 0.9, the best model's (MSM-SM [71]) result is just below 0.74 for GazeCom, and below 0.6 for Hollywood2. This particular aspect needs more investigation and, possibly, dedicated training: Notably, the models proposed in this work were not trained to maximise xAUC, but rather to achieve better general-purpose saliency prediction, conditioned on one eye movement type or the other.

D. GENERAL IMPLICATIONS

The work presented here points out a major methodological concern: While smooth pursuit comprises a significant part of the viewing behaviour, is has never been systematically analysed in the context of saliency prediction. This lack of specialised analysis means that the gaze samples corresponding to the form of attention expressed as smooth pursuit will be either discarded in the analysis, or labelled inconsistently.

Analysing attention means analysing both fixations *and* smooth pursuits, with a caveat: Fixations are not always intentional and can correspond to inattentive viewing or mind wandering [88], [89]. Typical works on saliency prediction only talk about fixations, never accounting for what can be called their attentiveness. Our work, on the contrary, demonstrates that using only smooth pursuit gaze samples – i.e. those when the eye movements reveal attentive viewing by following a moving target – can help improve on traditional saliency approaches.

This, however, is not the end of the story: We only consider pure eye movement information to uncover something about the observer's attention. Instead, e.g. pupil size can be used to infer attention (see e.g. [90] for a review of the works on connecting pupil size dynamics to a variety of perception aspects; [91], [92]), though the analysis might be more complex. If an

EEG signal is recorded simultaneously with eye tracking data, this can be analysed to infer periods of attentive viewing as well [93], [94]. Recent technological advancements have enabled simultaneous fMRI and eye tracking recording [10], [95], which could open the next frontier for analysing attention allocation with the help of brain imaging.

Directly tying together pursuit, saliency, and brain activity, albeit with synthetic stimuli and single-neuron recordings, [96] examined neuron spiking in monkeys, comparing the extent to which different regions in the brain encode visual saliency (in a low-level sense). Generalising such conclusions to more naturalistic [97] and realistic visual stimuli would require a better method to analyse naturally occurring smooth pursuit, and could further our understanding of what exactly contemporary saliency models learn.

VII. CONCLUSION

In this paper, we introduced the concept of *supersaliency* – smooth pursuit-based attention prediction. We argue that pursuit exhibits properties that set it apart from fixations in terms of perception and behavioural consequences, and that predicting smooth pursuit should thus be studied separately from fixation prediction. To this end, we provide our pipeline and the ground truth for saliency and supersaliency problems for the large-scale Hollywood2, as well as for the manually annotated GazeCom at `https://gin.g-node.org/MikhailStartsev/supersaliency`.

To better understand a model's behaviour on supersaliency data, we introduced the cross-AUC metric that assesses an algorithm's preference for pursuit vs. fixation locations, thus describing its ability to distinguish between the two. Whereas the human data showed that there are clear systematic differences between the two target types, it remains an open question how to reliably capture these differences with video-based saliency models.

Finally, we proposed and evaluated a deep saliency model with the slicing CNN architecture, which we trained for both smooth pursuit and fixation-based attention prediction. In both settings, our model outperformed all 26 tested dynamic reference models. Importantly, training for supersaliency yielded better results even for traditional fixation-based saliency prediction on two additional independent data sets. The same trend was observed with an additionally introduced deep end-to-end saliency model, further validating our conclusions that supersaliency demonstrates better generalisability. These findings demonstrate the potential of smooth pursuit modelling and prediction.

REFERENCES

[1] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

[2] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[3] H. Hadizadeh and I. V. Bajić, "Saliency-aware video compression," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 19–33, Jan. 2014.

[4] S. Marat, M. Guironnet, and D. Pellerin, "Video summarization using a visual attention model," in *Proc. 15th Eur. Signal Process. Conf.*, Sep. 2007, pp. 1784–1788.

[5] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 300–312, Feb. 2007.

[6] Y. Nagai and K. J. Rohlfing, "Computational analysis of motionese toward scaffolding robot action learning," *IEEE Trans. Auton. Mental Develop.*, vol. 1, no. 1, pp. 44–54, May 2009.

[7] A. C. Schütz, D. I. Braun, and K. R. Gegenfurtner, "Eye movements and perception: A selective review," *J. Vis.*, vol. 11, no. 5, p. 9, 2011, doi: 10.1167/11.5.9.

[8] M. Spering, A. C. Schütz, D. I. Braun, and K. R. Gegenfurtner, "Keep your eyes on the ball: Smooth pursuit eye movements enhance prediction of visual motion," *J. Neurophysiol.*, vol. 105, no. 4, pp. 1756–1767, 2011. [Online]. Available: http://jn.physiology.org/content/105/4/1756

[9] I. Agtzidis, M. Startsev, and M. Dorr, "Smooth pursuit detection based on multiple observers," in *Proc. 9th Biennial ACM Symp. Eye Tracking Res. Appl. (ETRA)*, New York, NY, USA, 2016, pp. 303–306, doi: 10.1145/2857491.2857521.

[10] M. Hanke, N. Adelhöfer, D. Kottke, V. Iacovella, A. Sengupta, F. R. Kaule, R. Nigbur, A. Q. Waite, F. Baumgartner, and J. Stadler, "A studyforrest extension, simultaneous fMRI and eye gaze recordings during prolonged natural stimulation," *Sci. Data*, vol. 3, Oct. 2016, Art. no. 160092.

[11] M. F. Land, "Eye movements and the control of actions in everyday life," *Prog. Retinal Eye Res.*, vol. 25, no. 3, pp. 296–324, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1350946206000036

[12] C. N. Authié and D. R. Mestre, "Optokinetic nystagmus is elicited by curvilinear optic flow during high speed curve driving," *Vis. Res.*, vol. 51, no. 16, pp. 1791–1800, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0042698911002173

[13] O. Lappi and E. Lehtonen, "Eye-movements in real curve driving: Pursuit-like optokinesis in vehicle frame of reference, stability in an allocentric reference coordinate system," *J. Eye Movement Res.*, vol. 6, no. 1, pp. 1–13, 2013. [Online]. Available: https://bop.unibe.ch/JEMR/article/view/2352

[14] O. Lappi, P. Rinkkala, and J. Pekkanen, "Systematic observation of an expert driver's gaze strategy—An on-road case study," *Frontiers Psychol.*, vol. 8, p. 620, Apr. 2017. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2017.00620

[15] J. B. Pelz and R. Canosa, "Oculomotor behavior and perceptual strategies in complex tasks," *Vis. Res.*, vol. 41, no. 25, pp. 3587–3596, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0042698901002450

[16] A. B. Sereno and P. S. Holzman, "Antisaccades and smooth pursuit eye movements in schizophrenia," *Biol. Psychiatry*, vol. 37, no. 6, pp. 394–401, 1995.

[17] J. E. Silberg, I. Agtzidis, M. Startsev, T. Fasshauer, K. Silling, A. Sprenger, M. Dorr, and R. Lencer, "Free visual exploration of natural movies in schizophrenia," *Eur. Arch. Psychiatry Clin. Neurosci.*, vol. 269, pp. 407–418, Jan. 2018, doi: 10.1007/s00406-017-0863-1.

[18] M. Vidal, A. Bulling, and H. Gellersen, "Pursuits: Spontaneous interaction with displays based on smooth pursuit eye movement and moving targets," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput. (UbiComp)*, New York, NY, USA, 2013, pp. 439–448. [Online]. Available: http://doi.acm.org/10.1145/2493432.2493477

[19] A. Esteves, E. Velloso, A. Bulling, and H. Gellersen, "Orbits: Gaze interaction for smart watches using smooth pursuit eye movements," in *Proc. 28th Annu. ACM Symp. User Interface Softw. Technol. (UIST)*, New York, NY, USA, 2015, pp. 457–466. [Online]. Available: http://doi.acm.org/10.1145/2807442.2807499

[20] S. Schenk, P. Tiefenbacher, G. Rigoll, and M. Dorr, "SPOCK: A smooth pursuit oculomotor control kit," in *Proc. CHI Conf. Extended Abstr. Hum. Factors Comput. Syst. (CHI EA)*, New York, NY, USA, 2016, pp. 2681–2687. [Online]. Available: http://doi.acm.org/10.1145/2851581.2892291

[21] M. Dorr, T. Martinetz, K. R. Gegenfurtner, and E. Barth, "Variability of eye movements when viewing dynamic natural scenes," *J. Vis.*, vol. 10, no. 10, p. 28, 2010, doi: 10.1167/10.10.28.

[22] S. Mathe and C. Sminchisescu, "Dynamic eye movement datasets and learnt saliency models for visual action recognition," in *Proc. 12th Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer-Verlag, 2012, pp. 842–856, doi: 10.1007/978-3-642-33709-3_60.

[23] L. J. Siever, R. D. Coursey, I. S. Alterman, M. S. Buchsbaum, and D. L. Murphy, "Impaired smooth pursuit eye movement: Vulnerability marker for schizotypal personality disorder in a normal volunteer population," *Amer. J. Psychiatry*, vol. 141, no. 12, pp. 1560–1566, 1984, doi: 10.1176/ajp.141.12.1560.

[24] V. Leborán, A. García-Díaz, X. R. Fdez-Vidal, and X. M. Pardo, "Dynamic whitening saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 5, pp. 893–907, May 2017.

[25] S. H. Khatoonabadi, N. Vasconcelos, I. V. Bajić, and Y. Shan, "How many bits does it take for a stimulus to be salient?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5501–5510.

[26] M. Startsev and M. Dorr, "Increasing video saliency model generalizability by training for smooth pursuit prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2018, pp. 2050–20503.

[27] M. Startsev, I. Agtzidis, and M. Dorr, "Characterizing and automatically detecting smooth pursuit in a large-scale ground-truth data set of dynamic natural scenes," *J. Vis.*, vol. 19, no. 14, p. 10, Dec. 2019, doi: 10.1167/19.14.10.

[28] J. Shao, C.-C. Loy, K. Kang, and X. Wang, "Slicing convolutional neural network for crowd video understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 5620–5628.

[29] T. Judd, F. Durand, and A. Torralba. (2012). *A Benchmark of Computational Models of Saliency to Predict Human Fixations*. [Online]. Available: http://hdl.handle.net/1721.1/68590

[30] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. *MIT Saliency Benchmark*. Accessed: Nov. 23, 2018. [Online]. Available: http://saliency.mit.edu/

[31] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8554–8564.

[32] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 545–552.

[33] J. Wang, H. R. Tavakoli, and J. Laaksonen, "Fixation prediction in videos using unsupervised hierarchical features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 2225–2232.

[34] Z. Wu, L. Su, Q. Huang, B. Wu, J. Li, and G. Li, "Video saliency prediction with optimized optical flow and gravity center bias," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2016, pp. 1–6.

[35] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. 14th ACM Int. Conf. Multimedia (MM)*, New York, NY, USA, 2006, pp. 815–824. [Online]. Available: http://doi.acm.org/10.1145/1180639.1180824

[36] Y. Li and Y. Li, "A fast and efficient saliency detection model in video compressed-domain for human fixations prediction," *Multimedia Tools Appl.*, vol. 76, pp. 1–23, Dec. 2016, doi: 10.1007/s11042-016-4118-3.

[37] M. Xu, L. Jiang, X. Sun, Z. Ye, and Z. Wang, "Learning to detect video saliency with hevc features," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 369–385, Jan. 2017.

[38] S. Chaabouni, J. Benois-Pineau, O. Hadar, and C. B. Amar, "Deep learning for saliency prediction in natural video," *CoRR*, vol. abs/1604.08010, pp. 1–34, Apr. 2016. [Online]. Available: https://dblp.uni-trier.de/rec/bibtex/journals/corr/ChaabouniBHA16

[39] L. Bazzani, H. Larochelle, and L. Torresani, "Recurrent mixture density network for spatiotemporal visual attention," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.

[40] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 4894–4903.

[41] L. Jiang, M. Xu, T. Liu, M. Qiao, and Z. Wang, "DeepVS: A deep learning based video saliency prediction approach," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018.

[42] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid dilated deeper ConvLSTM for video salient object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 715–731.

[43] W. Byeon, Q. Wang, R. K. Srivastava, and P. Koumoutsakos, "ContextVP: Fully context-aware video prediction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018.

[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Red Hook, NY, USA: Curran Associates, 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[45] C. F. Cadieu, H. Hong, D. L. K. Yamins, N. Pinto, D. Ardila, E. A. Solomon, N. J. Majaj, and J. J. DiCarlo, "Deep neural networks rival the representation of primate IT cortex for core visual object recognition," *PLoS Comput. Biol.*, vol. 10, no. 12, pp. 1–18, Dec. 2014, doi: 10.1371/journal.pcbi.1003963.

[46] S. Winkler and R. Subramanian, "Overview of eye tracking datasets," in *Proc. 5th Int. Workshop Qual. Multimedia Exper. (QoMEX)*, Jul. 2013, pp. 212–217.

[47] N. Riche, M. Mancas, D. Culibrk, V. Crnojevic, B. Gosselin, and T. Dutoit, "Dynamic saliency models and human attention: A comparative study on videos," in *Computer Vision*. Berlin, Germany: Springer, 2013, pp. 586–598, doi: 10.1007/978-3-642-37431-9_45.

[48] H. Hadizadeh, M. J. Enriquez, and I. V. Bajić, "Eye-tracking database for a set of standard video sequences," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 898–903, Feb. 2012.

[49] E. Vig, M. Dorr, and D. Cox, "Space-variant descriptor sampling for action recognition based on saliency and eye movements," in *Computer Vision*. Berlin, Germany: Springer, 2012, pp. 84–97, doi: 10.1007/978-3-642-33786-4_7.

[50] H. Alers, J. A. Redi, and I. Heynderickx, "Examining the effect of task on viewing behavior in videos using saliency maps," *Proc. SPIE, Hum. Vis. Electron. Imag.*, vol. 8291, Feb. 2012, Art. no. 82910X.

[51] R. Carmi and L. Itti, "The role of memory in guiding attention during natural vision," *J. Vis.*, vol. 6, no. 9, p. 4, 2006, doi: 10.1167/6.9.4.

[52] F. Boulos, W. Chen, B. Parrein, and P. L. Callet, "Region-of-interest intra prediction for H.264/AVC error resilience," in *Proc. 16th IEEE Int. Conf. Image Process. (ICIP)*, Nov. 2009, pp. 3109–3112.

[53] U. Engelke, R. Pepion, P. L. Callet, and H.-J. Zepernick, "Linking distortion perception and visual saliency in H.264/AVC coded video containing packet loss," *Proc. SPIE, Vis. Commun. Image Process.*, vol. 7744, Jul. 2010, Art. no. 774406.

[54] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson, "Clustering of gaze during dynamic scene viewing is predicted by motion," *Cogn. Comput.*, vol. 3, no. 1, pp. 5–24, 2011, doi: 10.1007/s12559-010-9074-z.

[55] I. Agtzidis, M. Startsev, and M. Dorr, "In the pursuit of (ground) truth: A hand-labelling tool for eye movements recorded during dynamic scene viewing," in *Proc. IEEE 2nd Workshop Eye Tracking Vis. (ETVIS)*, Oct. 2016, pp. 65–68.

[56] M. Startsev, I. Agtzidis, and M. Dorr, "1D CNN with BLSTM for automated classification of fixations, saccades, and smooth pursuits," *Behav. Res. Methods*, vol. 51, no. 2, pp. 556–572, Apr. 2019, doi: 10.3758/s13428-018-1144-2.

[57] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2106–2113.

[58] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[59] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015, pp. 1–14. [Online]. Available: http://arxiv.org/abs/1409.1556

[60] E. Vig, M. Dorr, and E. Barth, "Contribution of spatio-temporal intensity variation to bottom-up saliency," in *Bio-Inspired Models of Network, Information, and Computing Systems*. Berlin, Germany: Springer, 2012, pp. 469–474, doi: 10.1007/978-3-642-32615-8_44.

[61] K. G. Rottach, A. Z. Zivotofsky, V. E. Das, L. Averbuch-Heller, A. O. Discenna, A. Poonyathalang, and R. Leigh, "Comparison of horizontal, vertical and diagonal smooth pursuit eye movements in normal human subjects," *Vis. Res.*, vol. 36, no. 14, pp. 2189–2195, 1996. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0042698995003029

[62] G. Chen, D. Clarke, M. Giuliani, A. Gaschler, D. Wu, D. Weikersdorfer, and A. Knoll, "Multi-modality gesture detection and recognition with un-supervision, randomization and discrimination," in *Computer Vision*, L. Agapito, M. M. Bronstein, and C. Rother, Eds. Cham, Switzerland: Springer, 2015, pp. 608–622, doi: 10.1007/978-3-319-16178-5_43.

[63] S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jul. 2017, pp. 1175–1183.

[64] S. S. Nabavi, M. Rochan, and Y. Wang, "Future semantic segmentation with convolutional LSTM," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*. Newcastle, U.K.: Northumbria Univ., Sep. 2018, p. 137. [Online]. Available: http://bmvc2018.org/contents/papers/0559.pdf

[65] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, Dec. 2015, pp. 1–15. [Online]. Available: http://arxiv.org/abs/1412.6980

[66] Y.-F. Ma and H.-J. Zhang, "A new perceived motion based shot content representation," in *Proc. Int. Conf. Image Process.*, vol. 3, 2001, pp. 426–429.

[67] Y.-F. Ma and H.-J. Zhang, "A model of motion attention for video skimming," in *Proc. Int. Conf. Image Process.*, vol. 1, 2002, pp. I-129–I-132.

[68] G. Agarwal, A. Anbu, and A. Sinha, "A fast algorithm to find the region-of-interest in the compressed MPEG domain," in *Proc. Int. Conf. Multimedia Expo (ICME)*, vol. 2, Jul. 2003, pp. II-133–II-136.

[69] A. Sinha, G. Agarwal, and A. Anbu, "Region-of-interest based compressed domain video transcoding scheme," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, May 2004, pp. iii-161-iii-164.

[70] Z. Liu, H. Yan, L. Shen, Y. Wang, and Z. Zhang, "A motion attention model based rate control algorithm for H. 264/AVC," in *Proc. 8th IEEE/ACIS Int. Conf. Comput. Inf. Sci.*, Jun. 2009, pp. 568–573.

[71] K. Muthuswamy and D. Rajan, "Salient motion detection in compressed domain," *IEEE Signal Process. Lett.*, vol. 20, no. 10, pp. 996–999, Oct. 2013.

[72] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, and C.-W. Lin, "A video saliency detection model in compressed domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 1, pp. 27–38, Jan. 2014.

[73] J. Harel. *A Saliency Implementation in MATLAB*. Accessed: Nov. 23, 2018. [Online]. Available: http://www.klab.caltech.edu/~harel/share/gbvs.php

[74] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, p. 15, 2009, doi: 10.1167/9.12.15.

[75] A. García-Díaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image Vis. Comput.*, vol. 30, no. 1, pp. 51–64, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885611001235

[76] E. Vig, M. Dorr, T. Martinetz, and E. Barth, "Intrinsic dimensionality predicts the saliency of natural dynamic scenes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1080–1091, Jun. 2012.

[77] S. Mathe and C. Sminchisescu, "Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 7, pp. 1408–1424, Jul. 2015.

[78] Y. Tang, W. Zou, Z. Jin, and X. Li, "Multi-scale spatiotemporal Conv-LSTM network for video saliency detection," in *Proc. ACM Int. Conf. Multimedia Retr.*, 2018, pp. 362–369.

[79] G. Ding and Y. Fang, "Video saliency detection by 3D convolutional neural networks," in *Digital TV and Wireless Multimedia Communication*, G. Zhai, J. Zhou, and X. Yang, Eds. Singapore: Springer, 2018, pp. 245–254.

[80] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 38–49, Jan. 2018.

[81] L. Jiang, M. Xu, and Z. Wang. (2017). *OMCNN_2CLSTM: The Model of DeepVS: A Deep Learning Based Video Saliency Prediction Approach (ECCV2018)*. [Online]. Available: https://github.com/remega/OMCNN_2CLSTM

[82] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji. (2017). *DHF1K: Revisiting Video Saliency: A Large-Scale Benchmark and a New Model (CVPR)*. [Online]. Available: https://github.com/wenguanwang/DHF1K

[83] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, Mar. 2019.

[84] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Information-theoretic model comparison unifies saliency metrics," *Proc. Nat. Acad. Sci. USA*, vol. 112, no. 52, pp. 16054–16059, 2015. [Online]. Available: http://www.pnas.org/content/112/52/16054.abstract

[85] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *J. Vis.*, vol. 7, no. 14, p. 4, 2007, doi: 10.1167/7.14.4.

[86] P.-H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *J. Vis.*, vol. 9, no. 7, p. 4, 2009, doi: 10.1167/9.7.4.

[87] C. Bak, A. Kocak, E. Erdem, and A. Erdem, "Spatio-temporal saliency networks for dynamic saliency prediction," *IEEE Trans. Multimedia*, vol. 20, no. 7, pp. 1688–1698, Jul. 2018.

[88] J. Smallwood, E. Beach, J. W. Schooler, and T. C. Handy, "Going AWOL in the brain: Mind wandering reduces cortical analysis of external events," *J. Cogn. Neurosci.*, vol. 20, no. 3, pp. 458–469, 2008, doi: 10.1162/jocn.2008.20037.

[89] J. Smallwood, "Mind-wandering while reading: Attentional decoupling, mindless reading and the cascade model of inattention," *Lang. Linguistics Compass*, vol. 5, no. 2, pp. 63–77, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-818X.2010.00263.x

[90] S. Mathôt and S. V. der Stigchel, "New light on the mind's eye: The pupillary light response as active vision," *Current Directions Psychol. Sci.*, vol. 24, no. 5, pp. 374–378, 2015, doi: 10.1177/0963721415593725.

[91] O. E. Kang, K. E. Huffer, and T. P. Wheatley, "Pupil dilation dynamics track attention to high-level information," *PLoS ONE*, vol. 9, no. 8, pp. 1–6, Aug. 2014, doi: 10.1371/journal.pone.0102463.

[92] S. M. Wierda, H. van Rijn, N. A. Taatgen, and S. Martens, "Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution," *Proc. Nat. Acad. Sci. USA*, vol. 109, no. 22, pp. 8456–8460, 2012. [Online]. Available: https://www.pnas.org/content/109/22/8456

[93] V. Balasubramanian, K. Adalarasu, and A. Gupta, "EEG based analysis of cognitive fatigue during simulated driving," *Int. J. Ind. Syst. Eng.*, vol. 7, no. 2, pp. 135–149, 2011. [Online]. Available: https://www.inderscienceonline.com/doi/abs/10.1504/IJISE.2011.038563

[94] N.-H. Liu, C.-Y. Chiang, and H.-C. Chu, "Recognizing the degree of human attention using EEG signals from mobile sensors," *Sensors*, vol. 13, no. 8, pp. 10273–10286, Aug. 2013, doi: 10.3390/s130810273.

[95] M. Kanowski, J. W. Rieger, T. Noesselt, C. Tempelmann, and H. Hinrichs, "Endoscopic eye tracking system for fMRI," *J. Neurosci. Methods*, vol. 160, no. 1, pp. 10–15, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165027006003931

[96] B. J. White, L. Itti, and D. P. Munoz, "Superior colliculus encodes visual saliency during smooth pursuit eye movements," *Eur. J. Neurosci.*, vol. 8, no. 14263, pp. 1–9, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/ejn.14432

[97] B. J. White, D. J. Berg, J. Y. Kan, R. A. Marino, L. Itti, and D. P. Munoz, "Superior colliculus neurons encode a visual saliency map during free viewing of natural dynamic video," *Nature Commun.*, vol. 8, pp. 1–9, Jan. 2017.

MIKHAIL STARTSEV received the Diplom degree in computational mathematics and informatics from the Lomonosov Moscow State University (LMSU), Russia, in 2015, where he was a member of the Graphics and Media Lab. He is currently pursuing the Ph.D. degree with the Chair of Human–Machine Communication, Department of Electrical and Computer Engineering, Technical University of Munich (TUM), Germany, as a part of an International Junior Research Group "Visual Efficient Sensing for the Perception-Action Loop" (VESPA). His researches focus on the human visual system, with an emphasis on the eye movements, and computer vision, in particular saliency.

MICHAEL DORR received the Dr.-Ing. degree in computer science from the University of Lübeck, Germany, in 2010. He completed his postdoctoral training at The Schepens Eye Research Institute, Harvard Medical School, before joining the Technical University of Munich (TUM) to lead the International Junior Research Group "Visual Efficient Sensing for the Perception-Action Loop" (VESPA), in 2014. He conducts research at the intersection of human and machine vision, with a focus on models of eye movements and selective attention in complex dynamic environments.