# Model-guided Multi-path Knowledge Aggregation for Aerial Saliency Prediction

Kui Fu, Jia Li, Senior Member, IEEE, Yu Zhang, Member, IEEE, Hongze Shen and Yonghong Tian, Senior Member, IEEE

Abstract—As an emerging vision platform, a drone can look from many abnormal viewpoints which brings many new challenges into the classic vision task of video saliency prediction. To investigate these challenges, this paper proposes a largescale video dataset for aerial saliency prediction, which consists of ground-truth salient object regions of 1,000 aerial videos, annotated by 24 subjects. To the best of our knowledge, it is the first large-scale video dataset that focuses on visual saliency prediction on drones. Based on this dataset, we propose a Modelguided Multi-path Network (MM-Net) that serves as a baseline model for aerial video saliency prediction. Inspired by the annotation process in eye-tracking experiments, MM-Net adopts multiple information paths, each of which is initialized under the guidance of a classic saliency model. After that, the visual saliency knowledge encoded in the most representative paths is selected and aggregated to improve the capability of MM-Net in predicting spatial saliency in aerial scenarios. Finally, these spatial predictions are adaptively combined with the temporal saliency predictions via a spatiotemporal optimization algorithm. Experimental results show that MM-Net outperforms ten stateof-the-art models in predicting aerial video saliency.

Index Terms—Multi-path CNNs, knowledge transfer, visual saliency, aerial video, eye-tracking.

#### I. INTRODUCTION

ISUAL saliency prediction is one of the fundamental vision problems that has been extensively studied for several decades [1], [2], [3]. With the proposal of comprehensive rules [4], [5], large training datasets [6], [7], [8] and deep learning algorithms [9], [10], [11], [12], the performance of saliency models has been improving steadily. Meanwhile, many saliency-based attentive systems have achieved impressive performance in image recognition [13], video compression [14], content-based adverting [15], robot interaction [16] and navigation [17]. Despite the success of various existing

Kui Fu is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, and also with the Peng Cheng Laboratory, Shenzhen 518000, China.

Jia Li is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, also with the Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China, and also with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: jiali@buaa.edu.cn).

Yu Zhang and Hongze Shen are with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China.

Yonghong Tian is with the National Engineering Laboratory for Video Technology, Department of Computer Science and Technology, Peking University, Beijing 100871, China, and also with the Peng Cheng Laboratory, Shenzhen 518000, China.

J. Li is the corresponding author. URL: http://cvteam.net.

models, an important concern still remains in the literature: whether visual knowledge extracted from existing saliency models can boost the performance of saliency prediction in various scenarios, especially the newly emerging ones?

In most existing works, this question is explored using plenty of images [7] and videos [18] collected from Internet to cover various daily scenarios taken by digit cameras and mobile phones. In this paper, we conduct research on an emerging but less studied domain, the aerial videos captured by drones. A drone can observe the world from many different viewpoints, providing us an opportunity to revisit the problem from new perspectives. In particular, from these aerial videos, we wish to explore reliable answers to two questions in visual saliency estimation:

- 1. Whether previous ground-level saliency models still work well in processing aerial videos?
- 2. How to transfer the ground-level knowledge to aerial platforms to develop an aerial saliency model?

To answer these two questions, we collect 1,000 videos captured by drones, which are then free-viewed by 24 subjects in eye-tracking experiments to collect dense, accurate fixations. In this manner, the ground-truth salient regions in aerial videos can be well annotated, based on which we can benchmark the performance of various visual saliency models deployed on drones. By testing the performance of ten ground-level saliency models, we find these models still work impressively in capturing salient regions in aerial videos. However, there still exist large gap between their predictions and the ground-truth maps. Therefore, it is necessary to explore the way that drones look and develop a saliency model suitable for the saliency prediction task on aerial videos.

Based on this dataset, we also develop a Model-guided Multi-path Network (MM-Net) that can be used as a baseline model for aerial video saliency prediction. The design of MM-Net is motivated by the settings of eye-tracking experiments (see Fig. 1) and aims to aggregate the visual saliency knowledge from many classic models into a unified deep network. As shown in Fig. 2, common low-level features are first extracted in a low-level module of MM-Net, which are then fed into a multi-path module. This module contains multiple information paths, each of which is initialized under the guidance of a classic saliency model. After that, redundant paths are identified and removed via a path selection algorithm that jointly considers path diversity, representativeness and the overall complexity of the multi-path module. The selected paths then enter the fusion module, and the visual saliency knowledge encoded in the selected paths are aggregated to

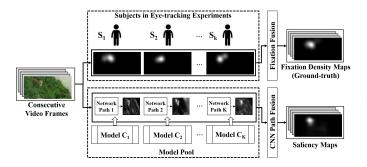


Fig. 1: In eye-tracking experiments, saliency maps can be generated by fusing fixations of multiple subjects. This process motivates the design of a multi-path network architecture, in which different paths are guided by different classic saliency models to encode different knowledge about visual saliency prediction.

predict spatial saliency [19] in aerial scenarios. Finally, spatial saliency maps can be efficiently predicted, which are then adaptively fused with the temporal saliency predictions to obtain clean and accurate saliency maps for various types of aerial videos. Experimental results show that MM-Net outperforms ten state-of-the-art models on aerial videos.

The main contributions of this paper include: 1) We propose, to the best of our knowledge, the first large-scale video dataset for aerial saliency prediction; 2) We propose a Model-guided Multi-path Network that provides a way to transfer the knowledge from multiple classic models into a single deep model; 3) We propose an effective path selection algorithm, which can be used to balance the complexity and effectiveness of the multi-path network.

The rest of this paper is organized as follows: Section II reviews related works and Section III presents the dataset. Our approach is presented in Section IV and tested in Section V. Finally, Section VI concludes the paper.

#### II. RELATED WORKS

In this section, we present a brief review of computational saliency models from three aspects: heuristic models, non-deep learning models and deep learning models.

Heuristic saliency models can be roughly categorized into bottom-up [1], [20] and top-down categories [21], [22]. The bottom-up models are stimulus-driven and infer saliency from visual stimuli themselves with hand-crafted features (e.g., direction, color and intensity) and/or limited human knowledge (e.g., center-bias). Due to the imperfect hand-crafted features or heuristic fusion strategies, bottom-up models may have some difficulties in suppressing background distractors. To address this problem, some top-down models heuristically incorporate high-level factors. For example, Borji et al. [23] proposed an unified Bayesian approach to integrate global context of a scene, previous attended locations and previous motor actions over time to predict the next attending locations. Chen et al. [24] proposed a video saliency model that predicted video saliency by combining the top-down saliency maps with the bottom-up ones through point-wise multiplication.

Non-deep learning saliency models propose to learn the fusion strategies of various heuristic saliency cues [25], [26], [27]. For example, Vig et al. [26] proposed a simple bottom-up model for dynamic scenarios with the aim of keeping the number of salient regions to a minimum. Recently, Fang et al. [10] proposed an image saliency model by learning a set of discriminative subspaces that perform the best in popping out targets and suppressing distractors. Li et al. [28] proposed a saliency model that measured the joint visual surprise from intrinsic and extrinsic contexts. However, the hand-crafted features used in these models may inherently set an upper bound for the final performance.

**Deep learning saliency models** have great advantages in learning feature representations [29], [30], [31]. For example, Kümmerer *et al.* [32] presented a Convolutional Neural Network (CNN) that reused AlexNet [33] to generate high-dimensional features. Pan *et al.* [9] proposed two designs, a shallow network and a deeper network, that can be trained end-to-end for fixation prediction. Lahiri *et al.* [34] proposed a saliency model which used a two step learning strategy. These deep learning models usually have high computational efficiency and impressive performance in daily scenarios. However, it is unclear whether these saliency models can be reused in aerial platforms, which has remarkable viewpoint and depth changes. Therefore, it is necessary to construct a large-scale aerial saliency dataset to benchmark saliency models.

### III. THE AERIAL VIDEO SALIENCY DATASET

In this section, we follow [35] to label the salient regions in the videos by eye-tracking and present a large-scale dataset for aerial video saliency. Note that such an annotation process is a regular process in the field of visual video attention prediction and minimizes the bias from different subjects via a voting manner, leading to accurate annotations. We also benchmark classic ground-level models to show the difference and correlation between aerial and ground-level saliency prediction.

#### A. Dataset Construction

To construct the dataset, we download hundreds of long aerial videos from Internet that are captured by drones. We manually divide these long videos into shots and randomly sample 1,000 shots with a total length of 1.6 hours (*i.e.*, 177,664 frames at 30 FPS). We find that the dataset mainly covers videos from four genres: building, human, vehicle, and others. Thus the dataset, denoted as AVS1K<sup>1</sup>, contains four subsets that are denoted as AVS1K-B, AVS1K-H, AVS1K-V and AVS1K-O, respectively.

To annotate the ground-truth salient regions, we conduct massive eye-tracking experiments involved with 24 subjects. For each video in the dataset, we collect eye fixations on each frame and convert them to fixation density maps following the setting of [35], making the whole dataset densely annotated. Note that each video is free-viewed by 17-20 randomly selected subjects. All subjects have normal or corrected to

<sup>&</sup>lt;sup>1</sup>Dataset is publicly available at http://cvteam.net

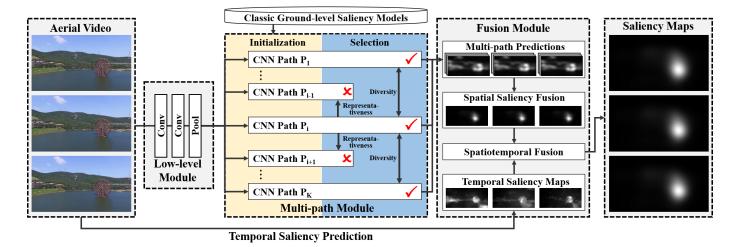


Fig. 2: System framework of our proposed baseline model MM-Net. The low-level module extracts low-level features. After that, these features are delivered into a multi-path module, in which each path is a sub-network pre-trained on massive aerial scenes under the guidance of a classic saliency model. In this manner, various visual saliency knowledge can be encoded in multiple paths to enhance the capability of MM-Net in processing aerial videos. To reduce computational cost, we further propose a selection algorithm to remove redundant paths according to path representativeness, diversity as well as module complexity. After that, the selected paths are fine-tuned on aerial scenarios to generate spatial saliency maps. They are then adaptively fused with the temporal saliency predictions to obtain clean and accurate saliency prediction results.

normal vision, and they have never seen these videos before. Note that we do not provide any prior information to the subjects and let them watch videos in a free-viewing manner in the eye-tracking experiments. In experiments, the videos are displayed on a 22-inch color monitor with the resolution of  $1680 \times 1050$ . A chin set is adopted to eliminate the error caused by the head wobble and fix the monitor viewing distance to 75cm. Other experimental conditions such as illumination and noise are set to constant for all subjects.

Given the fixation data, we can compute a fixation density map for each frame to annotate the ground-truth salient regions, *i.e.*, the salient regions that a drone should look at from the perspective of human-being. Let  $\mathcal{I}_t \in \mathcal{V}$  be a frame presented at time t, we measure the fixation density map of  $\mathcal{I}_t$ , denoted as  $S_t$  as in [35]. The value of  $S_t$  at pixel p can be computed as

$$S_t(p) = \sum_{f \in \mathbb{F}_{\mathcal{V}}} \delta(t_f \ge t) \cdot D_{spa}(f, p) \cdot D_{tem}(f, p), \quad (1)$$

where  $D_{spa}(f,p)$  and  $D_{tem}(f,p)$  measure the spatial and temporal influences of the fixation f to the pixel p, respectively. By using an indicator function  $\delta(t_f \geq t)$  that equals 1 if  $t_f \geq t$  and 0 otherwise, we only consider the influence of fixations in a short period after t. Let  $(x_p,y_p)$  be the coordinate of p, the values of  $D_{spa}(f,p)$  and  $D_{tem}(f,p)$  can be computed as

$$D_{spa}(f, p) = \exp\left(-\frac{(x_f - x_p)^2 + (y_f - y_p)^2}{2\sigma_D^2}\right),$$

$$D_{tem}(f, p) = \exp\left(-\frac{(t_f - t)^2}{2\sigma_T^2}\right),$$
(2)

where  $\sigma_D$  and  $\sigma_T$  are two constants to control the spatial and temporal influences of fixations, which are empirically set to 3% of video width (or video height if it is larger) and 0.1s,

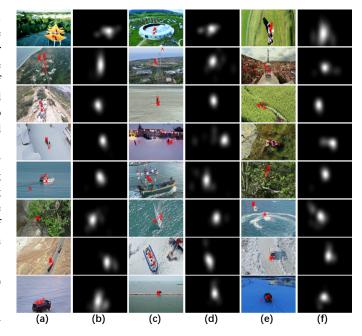


Fig. 3: Sample aerial frames (a, c, e), fixations (red dots) and ground-truth saliency maps (b, d, f) from AVS1K.

respectively. Some representative frames, recorded fixations and generated ground-truth saliency maps can be found in Fig. 3. Dataset statistics can be found in Tab. I.

# B. Model Performance in Aerial Scenarios

As it stands, aerial videos often have higher viewpoints, a wider field of vision and smaller targets. In other words, the visual patterns in aerial videos may be remarkably different

TABLE I: Dataset statistics of the AVS1K dataset.

Dataset	Video	Max Res.	Frames	Avg. Len. (s)
AVS1K-B	240	1280×720	41,471	5.76
AVS1K-H	210	$1280 \times 720$	31,699	5.03
AVS1K-V	200	$1280 \times 720$	27,092	4.52
AVS1K-O	350	$1280 \times 720$	77,402	7.37
AVS1K	1000	$1280 \times 720$	177,664	5.92

TABLE II: Performance comparisons on AVS1K and DHF1K.

Method	AUC	AVS1K sAUC	NSS	AUC	DHF1K sAUC	NSS
AIM	0.644	0.652	0.944	0.713	0.576	0.812
AWS	0.681	0.728	1.391	0.693	0.571	0.854
BMS	0.710	0.746	1.660	0.743	0.580	1.031
GB	0.612	0.650	0.984	0.631	0.558	0.599
HFT	0.789	0.715	1.671	0.806	0.579	1.323
ICL	0.698	0.651	1.252	0.712	0.535	0.697
IT	0.540	0.572	0.844	0.559	0.517	0.387
QDCT	0.689	0.696	1.302	0.720	0.569	0.909
SP	0.781	0.706	1.602	0.811	0.581	1.400
SUN	0.587	0.603	0.672	0.628	0.546	0.540

from those on the ground. Thus, it is worth exploring the performance of classic saliency models in aerial scenarios. To address this concern, we test ten classic saliency models on AVS1K. These models include AIM [36], AWS [37], BMS [38], GB [39], HFT [40], ICL [41], IT [1], QDCT [42], SP [43] and SUN [44]. Note that they are not learning-based and thus less sensitive to dataset bias. As an intuitive comparisons, we also show the performance of these models over DHF1K [45], the latest large-scale video saliency dataset fulfilled with daily videos collected by digital cameras and mobile phones.

To measure the performance of these models, we select three widely adopted metrics according to [46], [47], [48], including the traditional Area Under the ROC Curve (AUC), the shuffled AUC (sAUC) and the Normalized Scanpath Saliency (NSS). Typically, AUC may assign high scores to a fuzzy saliency map if it correctly predicts the orders of salient and less-salient locations, while sAUC and NSS prefer clean saliency maps that only pop-out the most salient locations and suppress all the other regions.

Based on the three metrics, Tab. II shows the performance of ten classic models on AVS1K and DHF1K. From Tab. II, we find that the AUC scores of all models on AVS1K are lower than those on DHF1K. Such inferior performance intuitively demonstrates that it is challenging for classic saliency models in dealing with aerial scenarios since they are designed based on human visual mechanisms on the ground-level but can not directly fit to aerial scenarios. Note that it is not affected by the bias of ground-truths labeling. This may be caused by the fact that aerial videos often have wider field of vision and thus contain richer contents. Surprisingly, sAUC and NSS, which focus on the saliency amplitude, achieve even higher scores on AVS1K than on DHF1K. This implies that the salient targets in aerial videos, which are usually very small, demonstrate impressive capability to pop-out from its surroundings from the higher viewpoints. These results imply that visual saliency knowledge encoded in classic models can be reused in aerial saliency prediction after certain domain adaptation operations.

### IV. THE MODEL-GUIDED MULTI-PATH NETWORK

#### A. Path Initialization

Motivated by above observations, we present MM-Net, a baseline model that absorbs the visual saliency knowledge in classic models and evolves to handle aerial video saliency like the human being does. As shown in Fig. 2, MM-Net starts with a low-level module that consists of two convolution layers and one max pooling layer. We initialize the parameters of the low-level module with the first two convolution layers of VGG16 [49]. Given these low-level features, there exist many ways in classic models to extract and fuse saliency cues from them. To make use of the knowledge in these models, we select the ten classic models we have tested in Section III, each of which is used to guide the initialization process of a network path in Fig. 2. In the initialization, we first obtain the saliency maps of a classic model on the training set (500 videos) and validation set (250 videos) of AVS1K. These model-estimated saliency maps are then used as "ground-truth" to fine-tune the layers in each MM-Net path. Note that the multiple paths and fusion modules are independently trained. The selected three paths are fixed in the fusion process. Each path is initialized with Xavier's algorithm and the input resolution is  $320 \times 320$ . In this process, the low-level module is fixed so that the parameters of each network path are independently updated. By minimizing the cross entropy loss between path outputs and classic model predictions, each path is forced to behave like a classic model so as to distillate its knowledge of saliency prediction.

After initialization, MM-Net inherently learns how to extract and fuse various saliency-related features. However, the knowledge encoded in different paths is highly redundant, and how to remove such redundancy to reduce model complexity is the next issue to be addressed.

#### B. Path Selection

To remove the path redundancy, we propose a path selection algorithm that jointly considers path diversity, representativeness and the overall complexity of the multi-path module. Formally, a binary column vector  $\boldsymbol{\alpha}$  with M binary components (M=10 in this study) is adopted, which indicates the ith path is selected if  $\boldsymbol{\alpha}_i=1$ , or discarded otherwise. For all the K frames from the 250 validation videos of AVS1K, we denote the saliency map predicted by the ith path on the kth frame as  $S_{k_i}$ . As a result, the path selection process can be solved by optimizing

$$\alpha^* = \arg \max_{\alpha} \Omega_r + \lambda_d \Omega_d$$
s.t.  $1 \le \|\alpha\|_0 \le M$ , and  $\alpha_i \in \{0, 1\}, \forall i$  (3)

where  $\|\alpha\|_0$  denotes the number of non-zero components in  $\alpha$  and thus reflects the complexity of the multi-path module. The terms  $\Omega_r$  and  $\Omega_d$  denote the representativeness and diversity to be maximized, respectively. The  $\lambda_d$  is a weight parameter to balance the representativeness and diversity, which is empirically set to 0.2 (its influence on final results will be discussed in experiments).

The term  $\Omega_r$  is defined according to path similarities. That is, the unselected paths should be highly similar to selected ones that are considered to be representative. This term can be defined as

$$\Omega_r = \frac{\sum_{i=1}^{M} (1 - \alpha_i) \cdot \max\{\alpha_j \cdot \operatorname{Sim}_{ij} | \forall \ j \neq i\}}{\sum_{i=1}^{M} (1 - \alpha_i) + \epsilon},$$
 (4)

where  $\epsilon$  is a small value to avoid dividing by zero. The term  $Sim_{ij}$  measures the similarity between the *i*th and *j*th paths that can be measured in a data-driven manner:

$$\operatorname{Sim}_{ij} = \frac{1}{K} \sum_{k=1}^{K} \sum_{x=1}^{W} \sum_{y=1}^{H} \min(S_{k_i}(x, y), S_{k_j}(x, y)), \quad (5)$$

where W and H are the width and height of the input images, respectively. By resizing  $S_{k_i}$  and  $S_{k_j}$  to the input image resolution and normalizing them into probability distributions,  $\operatorname{Sim}_{ij}$  measures the average histogram interactions between the saliency maps estimated by two paths. By maximizing  $\Omega_r$ , the similarity between selected and unselected paths can become very high, leading to a less-redundant multi-path module.

The representativeness term  $\Omega_r$  is defined between selected and unselected paths, while the diversity term  $\Omega_d$  is defined only on the selected ones that aims to maximize their difference

$$\Omega_{d} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{M} \delta(i \neq j) \cdot \alpha_{i} \cdot \alpha_{j} \cdot (1 - \operatorname{Sim}_{ij})}{\sum_{i=1}^{M} \sum_{j=1}^{M} \delta(i \neq j) \cdot \alpha_{i} \cdot \alpha_{j} + \epsilon}.$$
 (6)

We can see that this term will penalize the co-selection of two highly similar paths.

By incorporating (4) and (6) into the optimization objective (3), we can obtain a binary optimization problem with quadratic terms. As M is relatively small, enumeration can be adopted in ideal case. For large M, set optimization methods (e.g. submodular optimization) can quickly find a good local minimum.

#### C. Spatial Prediction and Spatiotemporal Fusion

Since the regular GPU with 11G memory (e.g., GTX 1080Ti) can not afford the training of the proposed model with a large number of classic methods, we optimize the path settings within the acceptable range of hardware to seek a complexity-accuracy trade-off. By solving (3) defined over the similarity matrix of ten paths, we obtain three representative paths (i.e., paths pre-trained by IT, QDCT, and SUN in the given parameters). After that, these selected paths are fused to output the spatial saliency maps. The overall structure of MM-Net can be found in Fig. 4. In training MM-Net, the parameters of the fusion module are randomly initialized and then optimized with a learning rate of  $5 \times 10^{-6}$ .

Beyond the spatial fusion of paths, another necessary fusion is the spatial and temporal saliency maps. Let  $S_k$  be the spatial saliency map given by MM-Net and  $T_k$  be the temporal saliency map given by an existing temporal saliency model

(e.g., [50]). Inspired by [51], we propose to spatiotemporally fuse  $S_k$  and  $T_k$  in adaptive fashion:

$$S_k^* = \lambda \cdot S_k^{int} + (1 - \lambda) \cdot S_k^{sel}, \tag{7}$$

where  $S_k^*$  is the refined saliency map,  $S_k^{int}$  is the collaborative interaction of  $S_k$  and  $T_k$ , and  $S_k^{sel}$  is the selected spatial or temporal saliency map according to a heuristic rule.  $\lambda$  is a positive scalar/weight to balance  $S_k^{int}$  and  $S_k^{sel}$ . We first compute the spatial-to-temporal consistency score  $C_{s2t}$  and temporal-to-spatial consistency score  $C_{t2s}$ :

$$C_{s2t} = e(S_k \odot T_k)/e(T_k), \quad C_{t2s} = e(S_k \odot T_k)/e(S_k), \quad (8)$$

where  $e(\cdot)$  is the entropy function and  $\odot$  indicates the perpixel multiplication. We can see that the spatial-to-temporal consistency  $C_{s2t}$  will be higher than the temporal-to-spatial consistency  $C_{t2s}$  if the temporal saliency map is cleaner, and vice versa. As a result, the collaboration interaction map can be computed by emphasizing the cleaner map:

$$S_k^{int} = \frac{C_{t2s} \cdot T_k + C_{s2t} \cdot S_k}{C_{t2s} + C_{s2t}}.$$
 (9)

Let  $d_k^S$  and  $d_k^T$  be the weighted average of distances of salient pixels to their gravity centers in  $S_k$  and  $T_k$ , respectively. The  $S_k^{sel}$  is defined as the map with more compact salient regions:

$$S_k^{sel} = \begin{cases} S_k & \text{if } d_k^S \le d_k^T \\ T_k & \text{otherwise} \end{cases}$$
 (10)

Intuitively, we can trust  $S_k^{int}$  if the spatial and temporal saliency maps are highly consistent. If not, we can select the most compact map as the final prediction. Let  $d_k^{int}$  be the average weighted distances of salient pixels to their gravity center in  $S_k^{int}$ , the parameter  $\lambda$  can be computed as

$$\lambda = \begin{cases} \min(C_{t2s}, C_{s2t}) & \text{if } d_k^{int} < \omega \cdot \min(d_k^S, d_k^T) \\ 0 & \text{otherwise} \end{cases}$$
 (11)

where  $\omega$  is a predefined weight that is empirically set to 2.1 (its influence will be discussed in experiments). By incorporating (9), (10) and (11) into (7), we can adaptively fuse the spatial and temporal saliency maps.

# V. EXPERIMENTS

We test our approach using the proposed aerial video saliency dataset AVS1K as well as the latest video saliency dataset DHF1K [52] that are fulfilled with daily scenarios captured by digital cameras and mobile phones. Both datasets, to the best of our knowledge, are currently the largest in their own domain. For DHF1K we use the official split, which contains 600, 100 and 300 videos for training, validation and testing, respectively.

On these two datasets, we compare MM-Net and two variants: MM-Net+ (with spatiotemporal refinement) and MM-Net- (without spatiotemporal refinement and the guidance of classic models). We also make comparisons with ten state-of-the-art models, including the heuristic group (H Group) HFT [40], SP [43] and PNSP [50], the Non-Deep Learning Group (NL Group) SSD [5] and LDS [10], and the Deep Learning

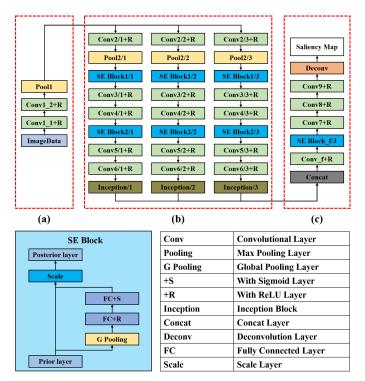


Fig. 4: The structure of MM-Net, including (a) low-level module, (b) multi-path module and (c) fusion module.

Group (DL Group) eDN [27], iSEEL [53], SalNet [9], DVA [52] and STS [54]. Among these models, iSEEL and eDN are built on pre-extracted features and thus cannot be re-trained. For the other three models, we fine-tune them on the two datasets and use a mark \* to indicate the retrained model.

In the comparisons, we adopt five evaluation metrics, including the aforementioned AUC, sAUC and NSS as well as the Similarity Metric (SIM) [55] and Correlation Coefficient (CC) [56]. SIM is computed to measure the similarity of two saliency maps as probability distributions, while CC is computed as the linear correlation between the estimated and ground-truth saliency maps. Values of all the five metrics are positively correlated with the performance.

# A. Comparison with the State-of-the-art Models

Performance of 13 evaluated models on the AVS1K dataset are shown in Tab. III. We also illustrate the Receiver Operating characteristics Curves (ROC) in Fig. 5 and several representative results of these models in Fig. 6.

From Tab. III, we find that our fundamental multi-path network MM-Net-, in which the spatiotemporal refinement and the guidance of classic models are not used, still outperforms the other ten state-of-the-art models in terms of NSS and CC and ranks the second place in terms of AUC (worse than DVA), sAUC (worse than SalNet) and SIM (worse than DVA). Note that NSS is the primary metric recommended by many surveys on saliency evaluation metrics [52], [57]. The impressive performance of MM-Net- can be explained by its multi-path structure. The low-level module of MM-Net- can extract many low-level preattentive features, based

TABLE III: Benchmarking results on AVS1K. The best and runner-up models of each column are marked with bold and underline, respectively. Except our models, the other deep models fine-tuned on AVS1K are marked with \*.

	Models	AUC	sAUC	NSS	SIM	CC
Н	HFT SP PNSP	0.789 0.781 0.787	0.715 0.706 0.634	1.671 1.602 1.140	0.408 0.422 0.321	0.539 0.520 0.370
NL	SSD LDS	0.737 0.808	0.692 0.720	1.564 1.743	0.404 0.452	0.503 0.565
DL	eDN iSEEL SalNet* DVA* STS* MM-Net MM-Net-	0.855 0.801 0.797 <u>0.864</u> 0.804 0.858 0.860 <b>0.869</b>	0.732 0.767 0.769 0.761 0.732 <u>0.771</u> 0.768 <b>0.784</b>	1.262 1.974 1.835 2.044 1.821 2.110 2.087 2.133	0.289 0.458 0.410 <u>0.544</u> 0.472 <b>0.547</b> 0.541 0.532	0.417 0.636 0.593 0.658 0.578 0.673 0.666 <b>0.682</b>

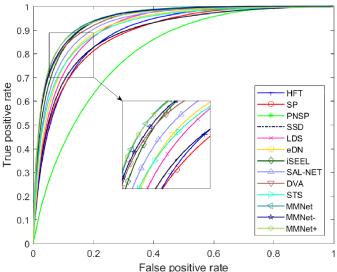


Fig. 5: ROC curves of 13 models on AVS1K.

on which the multi-path module can further extract saliency cues at higher levels from different perspectives. These highlevel saliency cues are then fused to obtain the final saliency map. In this way, MM-Net- has better representation capability when compared with traditional single path network (such as SalNet) and classic two-stream network for video (such as STS). Although DVA also adopts a multi-stream structure that directly fed supervisions into multi-layers, the MM-Net- still performs better than DVA in terms of sAUC, NSS and CC.

In Tab. III, we also find that MM-Net outperforms MM-Netin terms of all metrics except AUC. This can be explained by the model guidance strategy adopted in training multiple paths. After initializing different network paths under the guidance of selected models with heuristically designed saliency features and rules, MM-Net inherently learns how to extract and fuse saliency-related features. In this manner, the biases of classic models can be further investigated and utilized to improve the saliency prediction accuracy. As a result, the effectiveness of the model guidance strategy can be well justified.

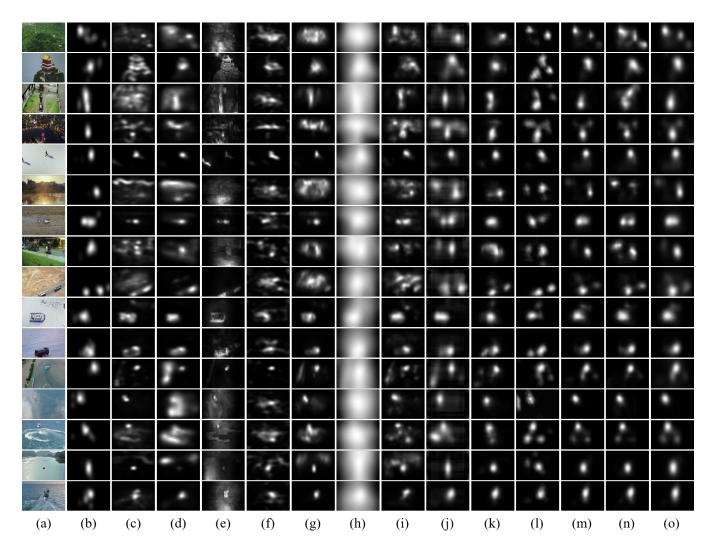


Fig. 6: Representative frames of state-of-the-art models on AVS1K. (a) Video frame, (b) Ground truth, (c) HFT, (d) SP, (e) PNSP, (f) SSD, (g) LDS, (h) eDN, (i) iSEEL, (j) SalNet, (k) DVA, (l) STS, (m) MM-Net, (n) MM-Net-, (o) MM-Net+.

From Tab. III, we also observe that MM-Net+ outperforms all the other models in terms of all metrics except SIM. This may be caused by the proposed spatiotemporal optimization algorithm. Based on the mutual consistency and weighted spatiotemporal saliency, the optimization algorithm tends to generate cleaner saliency maps with more compact salient regions (see Fig. 6). As a result, the most salient locations can pop-out in the saliency maps predicted by MM-Net+, leading to high NSS and sAUC scores.

We also find that the heuristic models in the H Group perform worse than the models in the NL Group and the DL Group. For H Group, the key issue here is that hand-crafted features designed for daily scenarios may be no longer suitable for the aerial scenarios. In other words, there may exist many irregular saliency visual patterns in aerial videos, which should be learned from data. This also explains the impressive performance of models in the DL Group since they can benefit from the powerful capabilities of CNNs in extracting hierarchical feature representations.

From these results, we conclude that, in aerial scenarios, the

salient visual patterns as well as the feature fusion strategies may become remarkably different. As a result, it is necessary to learn the saliency cues and their fusion strategies that best characterize the salient visual patterns from the aerial perspective. In addition, there exist some inherent correlations between the daily and aerial scenarios, implying that a drone can also benefit from the knowledge encoded in previous models in learning how to look. By transferring such knowledge, a drone can gain a better capability of handling various visual patterns.

## B. Performance Analysis

In this section, we conduct several experiments to analyze the performance of MM-Net+ (and MM-Net) from multiple perspectives, including parameter influences, generalization ability and performance on four subsets of AVS1K.

In the first experiment, we analyze the parameter  $\lambda_d$  in (3) that is used to balance the representativeness and diversity in MM-Net+. The NSS curve of MM-Net+ on AVS1K with different  $\lambda_d$  is shown in Fig. 7. From Fig. 7, we find that when

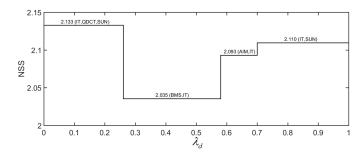


Fig. 7: Performance of MM-Net+ on AVS1K with the paths selected under different  $\lambda_d$ . In a wide range of  $\lambda_d$ , the selected paths stay almost stable (*i.e.*, 2 or 3 paths)

 $\lambda_d$  falls in [0, 0.24], MM-Net+ achieves the best performance (NSS=2.133) with three representative paths (IT, QDCT and SUN). When  $\lambda_d$  grows, the number of selected path decreases. When  $\lambda_d$  falls in [0.26, 0.56], MM-Net+ has lower complexity (only two selected paths, BMS and IT) as well as the lowest performance (NSS=2.035). When  $\lambda_d$  grows larger, only two paths keep on being selected but they may be guided by two different models. For example, when  $\lambda_d$  falls in [0.58, 0.68], MM-Net+ selects AIM and IT with NSS=2.093. When  $\lambda_d$  falls in [0.70, 1.00], MM-Net+ selects IT and SUN with NSS=2.093. To sum up, in a wide range of  $\lambda_d$ , the path selection algorithm tends to select two or three paths to reduce the model complexity. Therefore, we select  $\lambda_d=0.2$  in all experiments for pursuing better performance at an acceptable performance.

In the second experiment, we analyze the parameter  $\omega$  in (11) that serves as a threshold in computing  $\lambda$  and further balance the fusion of spatial and temporal saliency maps in MM-Net+. The curves of AUC, sAUC, NSS, SIM and CC scores on AVS1K with different  $\omega$  are shown in Fig. 8. We find that the AUC, NSS and CC curves are convex, the sAUC curve is monotonically increasing, and the SIM curve is monotonically decreasing. The overall performance is generally stable when  $\omega$  falls between [1.8, 2.3]. With a small  $\omega$ , many saliency maps are refined with  $\lambda=0$  (see (11)), implying that the spatial and temporal saliency maps cannot be adaptively fused in (7). On the contrary, a large  $\omega$  may generate non-zero  $\lambda$  in most cases and thus lead to noisy saliency maps due to the additive fusion strategy of (7). Therefore, we select  $\omega=2.1$  in all experiments.

In the third experiment, we do some ablation studies to show how the path selection and spatiotemporal fusion affects or contributes to the saliency estimation. We present the performance of different settings of path selection in Tab. IV. From this table, we find that both the path selection and spatiotemporal fusion boost the performance of saliency prediction on aerial videos. The model with three paths is superior to those with two paths ones. This can be interpreted as the three paths that can encode more complete visual knowledge into the overall model, leading to more powerful intermediate representations. Meanwhile, we find that performance gaps exist between the different selection of classic models. The overall performance not necessarily positively relates to the

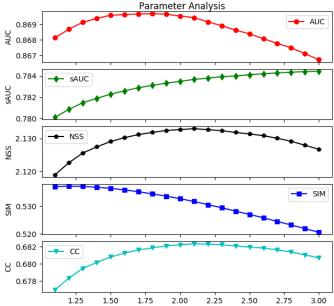


Fig. 8: Parameter analysis on AVS1K with different  $\omega$  in the interval [1.1, 3.0].

TABLE IV: Performance of different settings of path selection and spatiotemporal fusion on AVS1K. The best and runner-up models of each column are marked with bold and underline, respectively.

Init	Fusion	Nums	Paths	AUC	sAUC	NSS	SIM	CC
		3	_	0.860	0.768	2.087	0.541	0.666
$\checkmark$		3	IT,QDCT,SUN	0.858	0.771	2.110	0.547	0.673
$\checkmark$	$\checkmark$	3	IT,QDCT,SUN	0.869	0.784	2.133	0.532	0.682
$\checkmark$	$\checkmark$	2	AIM,IT	0.867	0.778	2.093	0.527	0.670
$\checkmark$	$\checkmark$	2	BMS,IT	0.861	0.769	2.035	0.514	0.650
✓	✓	2	IT,SUN	0.872	<u>0.781</u>	<u>2.110</u>	0.528	<u>0.674</u>

performance of the classic models but depends on the fitness of the encoded knowledge.

In the fourth experiment, we compare MM-Net with 16 state-of-the-art models on the latest video saliency dataset DHF1K that are fulfilled with videos captured by digital cameras and mobile phones. The main objective of this experiment is to verify the generalization ability of MM-Net on various scenarios. Quantitative results of these two models, after being fine-tuned on DHF1K, are shown in Tab. V. We can observe the proposed MM-Net outperforms the DVA on DHF1K in terms of all of metrics except for NSS. This proves the generalization ability of MM-Net, implying that the multi-path network architecture can be used for the saliency prediction task in both aerial and daily scenarios.

To verify the mutual generalization ability of saliency models in the aerial and general scenarios, we present the performance of two general models in Tab. VI. This table demonstrates the poor mutual generalization ability of saliency models in the aerial and general models.

Furthermore, to verify the performance of MM-Net+ on different scenarios, we show its performance on the four subsets of AVS1K in Tab. VII. MM-Net+ have relatively better

TABLE VI: Performance comparison of 16 state-of-the-art mod- TABLE VII: Performance of MM-Net+ on subsets of AVS1K. els on DHF1K. The best and runner-up models of each column The best and runner-up models of each column are marked are marked with bold and underline, respectively.

	Models	AUC	sAUC	NSS	SIM	CC
S	ITTI [1]	0.774	0.553	1.207	0.162	0.233
models	GBVS [39]	0.828	0.554	1.474	0.186	0.283
Ĕ	SALICON [58]	0.857	0.590	1.901	0.232	0.327
	Shallow-Net [9]	0.833	0.529	1.509	0.182	0.295
Static	Deep-Net [9]	0.855	0.592	1.775	0.201	0.331
Š	DVA [52]	0.860	0.595	2.013	0.262	0.358
	PQFT [59]	0.699	0.562	0.749	0.139	0.137
re	Seo et al. [20]	0.635	0.499	0.334	0.142	0.070
er	Rudoy et al. [60]	0.769	0.501	1.498	0.214	0.285
models	Hou et al. [41]	0.726	0.545	0.847	0.167	0.150
	Fang et al. [50]	0.819	0.537	1.539	0.198	0.273
ĕ	OBDL [61]	0.638	0.500	0.495	0.171	0.117
<u> </u>	AWS-D [62]	0.703	0.513	0.940	0.157	0.174
Dynamic	OM-CNN [63]	0.856	0.583	1.911	0.256	0.344
_	Two-stream [54]	0.834	0.581	1.632	0.197	0.325
	Ours*	0.875	0.627	1.972	0.271	0.389

Tested on the validation set of DHF1K.

TABLE VI: Performance comparison of two general models on AVS1K and DHF1K.

Method	AUC	sAUC	NSS	SIM	CC
DVA*	0.807	0.610	1.071	0.354	0.351
SalNet*	0.754	0.617	0.980	0.295	0.319
DVA <sup>+</sup>	0.822	0.574	1.430	0.240	0.283
SalNet <sup>+</sup>	0.799	0.574	1.312	0.189	0.261

<sup>\*</sup> Trained on DHF1K and tested on AVS1K.

performance on AVS1K-H and AVS1K-V than on AVS1K-B and AVS1K-O. In most aerial videos, humans and vehicles usually have relatively appropriate sizes and significant motions, which makes them easier to pop-out from the local context. On the contrary, buildings are static and usually have big sizes, making both the spatial and the temporal saliency prediction challenging. Similarly, AVS1K-O contains many diverse scenarios about planes, boats and animals. In these cases, the appearances and motion patterns of salient targets may change remarkably, making it difficult to separate them from distractors.

### VI. CONCLUSION

In this work, we introduce a large-scale video dataset for aerial saliency prediction. Based on this dataset, we propose MM-Net, a baseline model for aerial saliency prediction, which adopts a multi-path network structure and a modelguided training strategy to transfer human knowledge from classic models into the network paths. A spatiotemporal optimization algorithm is also proposed to fuse the spatial and temporal saliency maps. Experimental results demonstrate the superior performance of our proposed models.

In the future work, we will explore the feasibility of learning a saliency model with few or none domain-specific training data. Such one-shot or zero-shot learning may further help the deployment of visual saliency models in many unknown scenarios.

with bold and underline, respectively.

Subset	AUC	sAUC	NSS	SIM	CC
AVS1K-B	0.872	0.780	1.974	0.543	0.681
AVS1K-H	0.892	0.808	2.482	0.548	0.729
AVS1K-V	0.865	0.826	2.497	0.566	0.759
AVS1K-O	0.859	0.757	1.926	0.503	0.630

#### ACKNOWLEDGMENT

This work was supported in part by National Natural Science Foundation of China (No. 61922006, No. 61825101 and No. 61532003), and the Beijing Nova Program (No. Z181100006218063).

#### REFERENCES

- [1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 11, pp. 1254-1259, Nov. 1998.
- [2] 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.
- [3] T. V. Nguyen, Q. Zhao, and S. Yan, "Attentive systems: A survey," Int. J. Comput. Vis., vol. 126, no. 1, pp. 86-110, 2017.
- [4] S. Goferman, L. Zelnikmanor, and A. Tal, "Context-aware saliency detection," IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 10, pp. 1915-1926, Dec. 2012.
- [5] J. Li, L. Y. Duan, X. Chen, T. Huang, and Y. Tian, "Finding the secret of image saliency in the frequency domain," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 12, pp. 2428-2440, Apr. 2015.
- [6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in Proc. CVPR, 2009, pp. 248-255.
- [7] A. Borji and L. Itti, "Cat2000: A large scale fixation dataset for boosting saliency research," CVPR workshop on "Future of Datasets", 2015, arXiv preprint arXiv:1505.03581.
- [8] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," arXiv:1609.08675. [Online]. Available: http://arxiv.org/abs/1609.08675, 2016.
- [9] J. Pan, E. Sayrol, X. Giroinieto, K. Mcguinness, and N. E. Oconnor, 'Shallow and deep convolutional networks for saliency prediction," in Proc. CVPR, 2016, pp. 598-606.
- [10] S. Fang, J. Li, Y. Tian, T. Huang, and X. Chen, "Learning discriminative subspaces on random contrasts for image saliency analysis," IEEE TNNLS, vol. 28, no. 5, pp. 1095-1108, May 2017.
- [11] K. Zhang, W. Luo, L. Ma, W. Liu, and H. Li, "Learning joint gait representation via quintuplet loss minimization," in Proc. of IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun 2019, pp. 4700-4709.
- [12] K. Zhang, W. Luo, L. Ma, and H. Li, "Cousin network guided sketch recognition via latent attribute warehouse," in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 9203-9210.
- [13] B. Zhou, A. L. Garcia, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," Proc. Adv. Neural Inf. Process. Syst., pp. 487-495, 2014.
- [14] R. Gupta, M. T. Khanna, and S. Chaudhury, "Visual saliency guided video compression algorithm," Signal Process., Image Commun, vol. 28, no. 9, pp. 1006-1022, Oct. 2013.
- C. Shen and Q. Zhao, "Webpage saliency," Proc. ECCV, Springer, pp. 33-46, 2014.
- [16] C. Muhl, Y. Nagai, and G. Sagerer, "On constructing a communicative space in hri," in Proc. German Conf. on Adv. Artif. Intell., 2007, pp. 264 - 278
- [17] D. W. Gage, "Saliency detection and model-based tracking: a two part vision system for small robot navigation in forested environment," Proc. SPIE, vol. 8387, no. 15, p. 27, 2012.
- [18] Y. Fang, C. Zhang, J. Li, J. Lei, M. Perreira Da Silva, and P. Le Callet, "Visual attention modeling for stereoscopic video: A benchmark and computational model," IEEE Trans. Image Process., vol. 26, no. 10, pp. 4684–4696, 2017.

<sup>+</sup> Trained on AVS1K and tested on the validation set of DHF1K.

- [19] X. Ding and Z. Chen, "Improving saliency detection based on modeling photographer's intention," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 124–134. Jan. 2018.
- [20] 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.
- [21] P. Jiang, H. Ling, J. Yu, and J. Peng, "Salient region detection by ufo: Uniqueness, focusness and objectness," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1976–1983.
- [22] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. CVPR*, 2014, pp. 2806–2813.
- [23] A. Borji, D. N. Sihite, and L. Itti, "Probabilistic learning of task-specific visual attention," in *Proc. CVPR*, 2012, pp. 470–477.
- [24] Y. Chen, G. Tao, Q. Xie, and M. Song, "Video attention prediction using gaze saliency," *Multimedia Tools and Appl.*, vol. 78, no. 19, pp. 26867–26884, 2016.
- [25] J. Li, Y. Tian, T. Huang, and W. Gao, "Probabilistic multi-task learning for visual saliency estimation in video," *Int. J. Comput. Vis.*, vol. 90, no. 2, pp. 150–165, Nov. 2010.
- [26] 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.
- [27] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in CVPR, 2014.
- [28] J. Li, Y. Tian, X. Chen, and T. Huang, "Measuring visual surprise jointly from intrinsic and extrinsic contexts for image saliency estimation," *Int. J. Comput. Vis.*, vol. 120, no. 1, pp. 44–60, 2016.
- [29] P. Hu, B. Shuai, J. Liu, and G. Wang, "Deep level sets for salient object detection," in CVPR, 2017, pp. 2300–2309.
- [30] J. Kuen, Z. Wang, and G. Wang, "Recurrent attentional networks for saliency detection," in *Proc. CVPR*, 2016, pp. 3668–3677.
- [31] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3919–3930, Aug. 2016.
- [32] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet," arXiv:1411.1045. [Online]. Available: http://arxiv.org/abs/1411.1045, 2014.
- [33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1097–1105.
- [34] A. Lahiri, S. Roy, A. Santara, P. Mitra, and P. K. Biswas, "Wepsam: Weakly pre-learnt saliency model," arXiv preprint arXiv:1605.01101, 2016.
- [35] J. Li, C. Xia, and X. Chen, "A benchmark dataset and saliency-guided stacked autoencoders for video-based salient object detection," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 349–364, Jan. 2018.
- [36] N. Bruce and J. Tsotsos, "Attention based on information maximization," J. Vis., vol. 7, no. 9, p. 950, 2007.
- [37] A. Garcia-Diaz, V. Leboran, X. R. Fdez-Vidal, and X. M. Pardo, "On the relationship between optical variability, visual saliency, and eye fixations: A computational approach," J. Vis., vol. 12, no. 6, p. 17, 2012.
- [38] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," in *Proc. ICCV*, 2013, pp. 153–160.
- [39] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. NIPS*, 2007, pp. 545–552.
- [40] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 996–1010, Apr. 2013.
- [41] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Proc. NIPS*, 2009.
- [42] B. Schauerte and R. Stiefelhagen, "Quaternion-based spectral saliency detection for eye fixation prediction," in *Proc. ECCV*, 2012, pp. 116– 129.
- [43] J. Li, Y. Tian, and T. Huang, "Visual saliency with statistical priors," Int. J. Comput. Vis., vol. 107, no. 3, pp. 239–253, May 2014.
- [44] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, p. 32, Dec. 2008.
- [45] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," 2018, pp. 4894–4903.
- [46] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: state-of-the-art and study of comparison metrics," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1153–1160.
- [47] J. Li, C. Xia, Y. Song, S. Fang, and X. Chen, "A data-driven metric for comprehensive evaluation of saliency models," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 190–198.

- [48] 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.
- [49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556. [Online]. Available: http://arxiv.org/abs/1409.1556, 2014.
- [50] 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, 2014.
- [51] Z. Liu, X. Zhang, S. Luo, and O. Le Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1522–1540, Sep. 2014.
- [52] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.
- [53] H. R. Tavakoli, A. Borji, J. Laaksonen, and E. Rahtu, "Exploiting interimage similarity and ensemble of extreme learners for fixation prediction using deep features," *Neurocomputing*, vol. 244, pp. 10–18, Jun. 2017.
- [54] 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. 2017.
- [55] W. Hou, X. Gao, D. Tao, and X. Li, "Visual saliency detection using information divergence," *Pattern Recognit.*, vol. 46, no. 10, pp. 2658– 2669, Oct. 2013.
- [56] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proc. CVPR*, 2012, pp. 438–445.
- [57] N. Liu and J. Han, "A deep spatial contextual long-term recurrent convolutional network for saliency detection," arXiv:1610.01708. [Online]. Available: http://arxiv.org/abs/1610.01708, 2016.
- [58] X. Huang, C. Shen, X. Boix, and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *IEEE International Conference on Computer Vision*, Dec. 2015, pp. 262–270.
- [59] 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.
- [60] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor, "Learning video saliency from human gaze using candidate selection," in *IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2013, pp. 1147–1154.
- [61] S. Hossein Khatoonabadi, N. Vasconcelos, I. V. Bajic, and Y. Shan, "How many bits does it take for a stimulus to be salient?" in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2015, pp. 5501–5510.
- [62] V. Leboran, A. Garcia-Diaz, X. R. Fdez-Vidal, and X. M. Pardo, "Dynamic whitening saliency," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 5, pp. 893–907, May 2017.
- [63] L. Jiang, M. Xu, and Z. Wang, "Predicting video saliency with object-to-motion cnn and two-layer convolutional lstm," arXiv:1709.06316. [Online]. Available: http://arxiv.org/abs/1709.06316, 2017.



**Kui Fu** is currently pursuing the Ph.D. degree with the State Key Laboratory of Virtual Reality Technology and System, School of Computer Science and Engineering, Beihang University. His research interests include computer vision and image understanding.



Jia Li (M'12-SM'15) received the B.E. degree from Tsinghua University in 2005 and the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2011. He is currently a Full Professor with the School of Computer Science and Engineering, Beihang University, Beijing, China. Before he joined Beihang University in Jun. 2014, he used to conduct research in Nanyang Technological University, Peking University and Shanda Innovations. He is the author or coauthor of over 70 technical articles in refereed journals and con-

ferences such as TPAMI, IJCV, TIP, CVPR and ICCV. His research interests include computer vision and multimedia big data, especially the understanding and generation of visual contents. He is supported by the Research Funds for Excellent Young Researchers from National Nature Science Foundation of China since 2019. He was also selected into the Beijing Nova Program (2017) and ever received the Second-grade Science Award of Chinese Institute of Electronics (2018), two Excellent Doctoral Thesis Award from Chinese Academy of Sciences (2012) and the Beijing Municipal Education Commission (2012), and the First-Grade Science-Technology Progress Award from Ministry of Education, China (2010). He is a senior member of IEEE, CIE and CCF. More information can be found at http://cvteam.net.



(S'00-M'06-SM'10) Tian Yonghong Distinguished Professor currently Boya with the School of EECS, Peking University, China. and is also the deputy director Research of Artificial Intelligence Center. Laboratory, China. Peng Cheng Shenzhen, His include research interests computer multimedia big vision. data. and braininspired computation. He is the author or coauthor of over 180 technical articles in refereed iournals such IEEE

TPAMI/TNNLS/TIP/TMM/TCSVT/TKDE/TPDS, ACM CSUR/TOIS/TOMM and conferences such as NeurIPS/CVPR/ICCV/AAAI/ACMMM/WWW.

Prof. Tian was/is an Associate Editor of IEEE TCSVT (2018.1-), IEEE TMM (2014.8-2018.8), IEEE Multimedia Mag. (2018.1-), and IEEE Access (2017.1-). He co-initiated IEEE Int'l Conf. on Multimedia Big Data (BigMM) and served as the TPC Co-chair of BigMM 2015, and aslo served as the Technical Program Co-chair of IEEE ICME 2015, IEEE ISM 2015 and IEEE MIPR 2018/2019, and General Co-chair of IEEE MIPR 2020. He is the steering member of IEEE ICME (2018-) and IEEE BigMM (2015-), and is a TPC Member of more than ten conferences such as CVPR, ICCV, ACM KDD, AAAI, ACM MM and ECCV. He was the recipient of the Chinese National Science Foundation for Distinguished Young Scholars in 2018, two National Science and Technology Awards and three ministerial-level awards in China, and obtained the 2015 EURASIP Best Paper Award for Journal on Image and Video Processing, and the best paper award of IEEE BigMM 2018. He is a senior member of IEEE, CIE and CCF, a member of ACM.



Yu Zhang is currently a researcher with Sense-Time Research, Beijing China. He received his B.E. degree in 2012 and Ph.D. degree in 2018 from the School of Computer Science and Egnineering, Beihang University. His research interests include computer vision and image/video processing.



**Hongze Shen** received his Master degree in 2020 from the School of Computer Science and Engineering, Beihang University. His research interests include computer vision and image understanding.