Spatiotemporal Knowledge Distillation for Efficient Estimation of Aerial Video Saliency

Jia Li, *Senior Member, IEEE*, Kui Fu, Shengwei Zhao, and Shiming Ge, *Senior Member, IEEE*

***Abstract*— The performance of video saliency estimation techniques has achieved significant advances along with the rapid development of Convolutional Neural Networks (CNNs). However, devices like cameras and drones may have limited computational capability and storage space so that the direct deployment of complex deep saliency models becomes infeasible. To address this problem, this paper proposes a dynamic saliency estimation approach for aerial videos via spatiotemporal knowledge distillation. In this approach, five components are involved, including two teachers, two students and the desired spatiotemporal model. The knowledge of spatial and temporal saliency is first separately transferred from the two complex and redundant teachers to their simple and compact students, while the input scenes are also degraded from high-resolution to low-resolution to remove the probable data redundancy so as to greatly speed up the feature extraction process. After that, the desired spatiotemporal model is further trained by distilling and encoding the spatial and temporal saliency knowledge of two students into a unified network. In this manner, the inter-model redundancy can be removed for the effective estimation of dynamic saliency on aerial videos. Experimental results show that the proposed approach is comparable to 11 state-of-the-art models in estimating visual saliency on aerial videos, while its speed reaches up to 28,738 FPS and 1,490.5 FPS on the GPU and CPU platforms, respectively.**

***Index Terms*— Spatiotemporal knowledge distillation, visual saliency estimation, aerial video.**

Manuscript received January 15, 2019; revised June 24, 2019 and September 24, 2019; accepted October 2, 2019. Date of publication October 14, 2019; date of current version November 27, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61672072, Grant 61922006, and Grant 61532003, in part by the Beijing Nova Program under Grant Z181100006218063, and in part by the project from the Beijing Municipal Science and Technology Commission under Grant Z191100007119002. The work of S. Ge was supported by the Open Projects Program of the National Laboratory of Pattern Recognition, Ant Financial through the Ant Financial Science Funds for Security Research, and the Youth Innovation Promotion Association, Chinese Academy of Sciences. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Joao M. Ascenso. *(Corresponding author: Shiming Ge.)*

J. Li is with the State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China, with the Peng Cheng Laboratory, Shenzhen 518000, China, and also with the Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100191, China.

K. 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.

S. Zhao is with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100095, China, and also with the School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100095, China.

S. Ge is with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100095, China (e-mail: geshiming@iie.ac.cn).

I. Introduction

THE rapid development of mobile devices further emphasizes the importance of effectively and efficiently estimating dynamic visual saliency on videos. For example, a drone, one of the most popular mobile devices in recent years, is capable of collecting high-resolution aerial videos in various scenarios due to its flexible operability. To analyze these high-resolution videos on the drone with limited memory and computational capability, a highly efficient and accurate saliency model is required so that the limited resources can be spent on the attractive visual content with a high priority. By understanding the human attentional behavior to aerial data, the visual saliency models have the ability to automatically detect, locate and mine the most important part of massive visual information and can facilitate subsequent complex drone vision tasks in both speed and accuracy, such as drone event understanding [1], navigation [2], target tracking [3], obstacle avoidance [4], and object detection [5].

In the past decades, many models have been constructed in visual saliency estimation by defining comprehensive rules [6]–[9] or using deep learning frameworks [10]–[12]. In particular, the deep models have achieved impressive performance along with the development of large-scale benchmark datasets [13]–[15] but at the cost of huge memories and massive computations. However, these "ground-level" deep models may have difficulties to be directly deployed on drones for processing high-resolution aerial videos. The main reasons are two-fold: 1) the limited computational resource on drones is far from sufficient to meet the requirement of complex deep models; and 2) the ground-level saliency models may have difficulties to handle aerial videos since the data distributions change remarkably. Additionally, most existing video saliency models are suffering from slow computation and low estimation accuracy, since they rely on motion information from time-consuming optical flow and are designed without taking the spatiotemporal consistency in the inference process into consideration. As a result, to deploy these complex saliency models on resource-limited drones, two issues should be addressed first: 1) How to reduce the computational cost and memory footprint of deep saliency models without remarkable loss of accuracy? and 2) How to fuse both spatial and temporal cues to extract powerful features that apply to aerial videos?

To address these issues, we inspect existing deep models and find two major factors that restrict the computational cost: the model redundancy and the data redundancy.

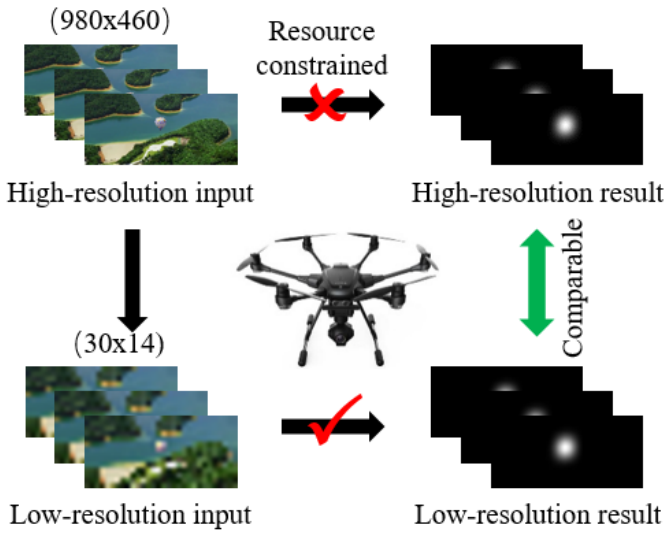

Fig. 1. Low-resolution frames lack details but salient targets can still be easily localized by the human-being. Since the directly deployment of complex saliency models on resource-limited drones may have difficulty in processing high-resolution aerial videos, a feasible solution is to distill the knowledge from complex models into a simple and compact model and remove the data redundancy (*e.g.*, resolution degradation) to ensure highly efficient aerial video processing.

Typically, some researchers tend to train deeper networks from large-scale datasets for dynamic saliency estimation. However, the models may become highly redundant for such a low-level vision task, leading to the dramatic increase of computation and memory costs. This fact facilitates the studies of model compression [16]–[19] that aim to remove the redundancy to generate compact models by parameter pruning and quantization. In this way, the resulting compact models usually have low computational cost and memory footprint but often encounter a sharp accuracy drop. To address this issue, some works [20]–[22] proposed to distill the knowledge from complex models and then transfer to simple ones without a significant performance drop.

Beyond the model redundancy, the data redundancy is less considered. Actually, saliency estimation on aerial scenarios is a low-level vision task that does not need so many details represented by high-resolution frame sequences. As shown in Fig. 1, salient targets in the heavily blurred low-resolution frames can still be easily localized by the human-being without such details. This implies that there exist strong data redundancy that is not necessary for the saliency estimation task. As a consequence, removing such data redundancy may be another way to further reduce the computational cost.

Inspired by these two findings, this paper proposes a spatiotemporal knowledge distillation approach. As shown in Fig. 2, the framework of this approach consists of five components, including two teachers, two students and the desired spatiotemporal model. The knowledge of visual saliency is transferred from the spatial and temporal teachers to the final spatiotemporal model by using the spatial and temporal students as the bridges. In this process, the spatial and temporal knowledge is first separately extracted from complex and redundant teachers and then transferred into simple and compact students to remove the intra-model redundancy. Meanwhile, the input scenes are also degraded from high-resolution to low-resolution to remove the data redundancy. After that, the desired spatiotemporal model is trained by distilling and encoding the spatial and temporal knowledge of the two students into a unified network to further remove the inter-model redundancy. By step-wisely removing the intra-model, data and inter-model redundancies, the dynamic saliency of aerial videos can be effectively estimated with an extremely high speed. Experimental results show that the proposed approach outperforms ten state-of-the-art models. In particular, its speed can reach up to 28,738 FPS and 1,490.5 FPS on the GPU and CPU platforms, respectively.

Our main contributions are summarized as follows: 1) We propose a two-step knowledge distillation framework that can greatly reduce the computational cost with little accuracy drop in aerial video saliency estimation; 2) We design a lightweight spatiotemporal network that can extract and fuse both spatial and temporal saliency cues; and 3) we conduct extensive experiments and prove that our approach achieves an ultrafast speed and is comparable to 11 state-of-the-art models.

The rest of this paper is organized as follows: Section II reviews related works and Section III presents the spatiotemporal knowledge distillation. Section IV benchmarks the proposed model. Finally, Section V concludes the paper.

II. Related Works

In this paper, we aim to distill knowledge from well pretrained saliency models that serve as teachers and transfer their knowledge to the students to facilitate efficient saliency estimation. Therefore, we present a brief review of visual saliency models and knowledge distillation studies.

A. Visual Saliency Models

The recent advances in the field of saliency estimation from videos result in many visual saliency models [23]. These models can be roughly grouped into three categories according to their features and frameworks, including heuristic, shallow-learning and deep-learning saliency models.

The heuristic saliency models [24]–[29] generally use hand-crafted features and design heuristic rules to perform visual saliency estimation in a bottom-up or top-down manner. The bottom-up models are stimulus-driven and compete fairly to pop-out conspicuous visual signals. In these models, hand-crafted features such as directions, colors and intensities as well as heuristic fusion rules are widely used. For example, Fang *et al.* [30] proposed a spatiotemporal framework to separately detect the spatial and temporal saliency cues. These cues were then fused according to the spatial compactness and the temporal motion contrast. Later, they (Fang *et al.* [31]) proposed uncertainty weighting to fuse the spatial and temporal saliency results. However, such unbiased heuristic fusion strategies may have difficulties in suppressing background distractors. To alleviate this issue, some task-driven models heuristically incorporate high-level factors in a top-down manner. For example, Borji *et al.* [32] modeled the task-driven

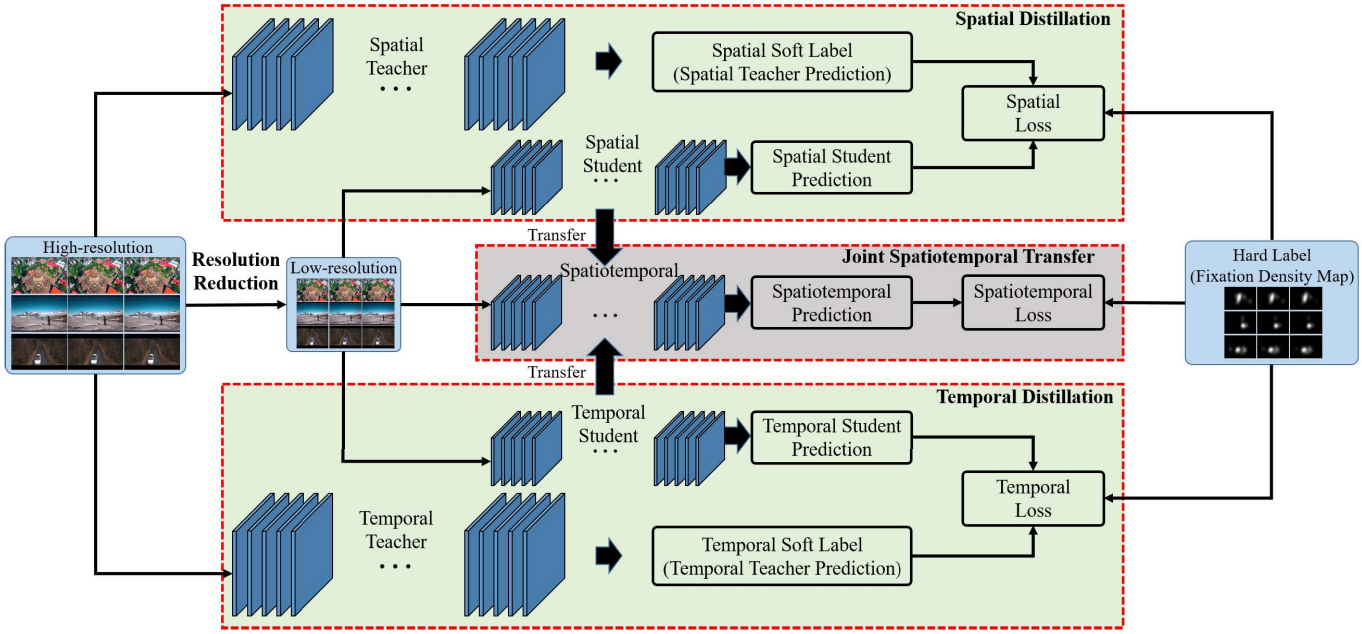

Fig. 2. System framework. The framework consists of five components: two teachers, two students and the desired spatiotemporal model. The knowledge is transferred from teachers to the desired model via two steps: 1) distilling the knowledge separately from the spatial teacher $\mathbb{T}_s$ and the temporal teacher $\mathbb{T}_t$ to their students $\mathbb{S}_s$ and $\mathbb{S}_t$, respectively. The distillation is conducted along with resolution degradation to remove data redundancy. 2) transferring and fusing the knowledge of $\mathbb{S}_s$ and $\mathbb{S}_t$ into a unified spatiotemporal model $\mathbb{S}_{st}$ to improve the accuracy and speed of dynamic saliency estimation on aerial videos.

visual attention with a unified Bayesian approach by integrating global scene context, previous attention locations and motion actions to predict the next attention locations. Chen *et al.* [33] predicted video saliency by combining the bottom-up saliency maps and the top-down ones through point-wise multiplication. Generally speaking, these heuristic saliency models often perform efficiently in estimating saliency but may suffer from poor accuracy and low robustness since the hand-crafted features and heuristic fusion strategies may be not optimal for all scenarios.

Inspired by the pros and cons of heuristic models, the shallow-learning saliency models [34]–[37] aim to directly learn an optimal fusion strategy of hand-crafted features from data. For example, the saliency model proposed by Vig *et al.* [36] used supervised learning to fine-tune the free parameters in dynamic scenarios. Fang *et al.* [38] proposed an optimization framework with pairwise binary terms by learning a set of discriminative subspaces to pop out targets and suppress distractors. Moreover, some works [35], [39] proposed to apply learning algorithms to combine multi-level features into the saliency estimation processes. For example, Song *et al.* [39] estimated saliency by fusing the low-level and high-level features as well as the center-bias priors. Due to optimized fusion strategy, learning-based saliency models generally achieve promising results. However, these models inherently share an upper bound in the performance since the hand-crafted features used may be also not optimal.

To address the feature issue, deep saliency models [40]–[43] proposed to use feature representation learned from data by using Convolutional Neural Networks (CNNs) [44], [45]. Some of these models directly employ the state-of-the-art deep models pretrained in large-scale visual tasks as feature extractors. For example, Kümmerer *et al.* reused AlexNet [46] and VGG-19 [47] to generate high-dimensional features for fixation prediction in [48] and [49], respectively. In contrast, Pan *et al.* [50] proposed to train a shallow CNN and a deep CNN in an end-to-end manner for fixation prediction. In addition, some deep saliency models focus on designing specific architectures or loss functions. For example, Imamoglu *et al.* [51] utilized the objectiveness scores predicted by the features selected from CNNs to detect conspicuous regions. Due to the rich knowledge extracted by the complex deep saliency models, the deep models usually outperform heuristic and shallow-learning models in accuracy. However, the computational cost often increase remarkably due to the rich redundancy in both models (*e.g.*, unnecessary computations) and data (*e.g.*, unnecessary high-resolution inputs), which prevents them from being directly deployed on mobile devices such as drones and cameras. Therefore, it is necessary to compress or distill these saliency models to greatly reduce the computational cost without remarkable performance drop.

B. Knowledge Distillation

Knowledge distillation [21] is a specific model compression technique that distills the inherent knowledge from a complex teacher model to a simple student one so as to greatly reduce the model redundancy and maintain a comparable performance. To this end, the student model is trained under the supervision of the teacher model in many existing works. For example, Hinton *et al.* [21] introduced the soft labels generated by a teacher model as an extra supervision, which was combined with the hard supervision defined by data labels. There also exist some other forms of supervision

such as classification probabilities [21], feature representations [22], [52], and inter-layer flows (the inner product of feature maps) [53]. Zhang *et al.* [54] proposed deep mutual learning, which conducted online distillation in one-phase training between two peer student models. Rusu *et al.* [55] proposed a multi-teacher single-student policy to distill knowledge from multiple teachers into a single student.

Generally speaking, these knowledge distillation approaches provide a powerful way to reduce the model redundancy, which is efficient in dealing with high-resolution static saliency estimation. However, the data redundancy is less considered, especially in low-level vision tasks like visual saliency estimation. Actually, saliency estimation is a low-level vision task that does not need so many details represented by high-resolution frame sequences. By removing such data redundancy hidden in the high-resolution as well as the consecutive video frames, the speed of a dynamic saliency model can be greatly boosted. To this end, we propose a spatiotemporal knowledge distillation approach to simultaneously reduce both the model and data redundancies while the model accuracy can be well maintained. Note that it is an approach of inductive transfer similar to [56] which uses relevant tasks to improve the generalization of the main task.

III. The Proposed Approach

In this section, we present a Spatiotemporal Knowledge Distillation (SKD) approach for dynamic saliency estimation in aerial videos. The proposed approach operates in two major steps: the separate spatial/temporal knowledge distillation, and the joint spatiotemporal knowledge transfer. Here we start with a brief overview of the approach and then elaborate on these two steps as well as their implementation details.

A. The Framework

As shown in Fig. 2, the proposed SKD approach consists of five major components, including a spatial teacher, a temporal teacher, a spatial student, a temporal student and the desired spatiotemporal model. Note that the spatial/temporal teachers can be any complex spatial/temporal saliency models pretrained on massive high-resolution data. Typically, such teacher models give impressive performances in predicting spatial saliency (*e.g.*, DVA [10], SalNet [50] and SSNet [57]) or temporal saliency (*e.g.*, TSNet [57]) at the expense of high computational cost. As a result, the objective of the proposed distillation framework is to distill their knowledge into much simpler student models and finally the desired spatiotemporal saliency model by removing the model and data redundancy in two consecutive steps.

In the first step, we separately distill the knowledge from spatial and temporal teachers to the two students to reduce intra-model redundancy, respectively. Meanwhile, the high-resolution inputs, which often contain unnecessary details for low-level vision tasks such as saliency estimation, are degraded into low-resolution ones to remove data redundancy.

In the second step, the knowledge in the spatial and temporal students is jointly transferred and encoded into the desired spatiotemporal model. In this manner, the inter-model redundancy of two students in extracting common visual features can be also removed. As a result, the intra-model, inter-model and data redundancies are step-wisely removed, leading to a model with high accuracy and extremely low computational cost.

B. Separate Spatial and Temporal Knowledge Distillation

The separate spatial and temporal distillation operations in the first step force two simple students to mimic the behavior of two complex teachers in a spatial high-resolution frame and a temporal consecutive frame pair, respectively. Let $\mathcal{D} = \{I_n, Y_n\}_{n=1}^N$ be the training dataset containing N samples and I_n be an high-resolution frame with the ground-truth saliency map Y_n. Then, we can easily compute their resolution-degraded version as $\hat{\mathcal{D}} = \{\hat{I}_n, \hat{Y}_n\}_{n=1}^N$ by using a resolution-reduction operation $\mathcal{R}$ so that $\hat{I}_n = \mathcal{R}(I_n)$ and $\hat{Y}_n = \mathcal{R}(Y_n)$. For the sake of simplification, we denote the spatial and temporal teachers as $\mathbb{T}_s$ and $\mathbb{T}_t$, respectively. Similarly, the spatial and temporal students are denoted as $\mathbb{S}_s$ and $\mathbb{S}_t$, respectively. Note that $\mathbb{T}_s$ and $\mathbb{S}_s$ take a single frame as the input, while $\mathbb{T}_t$ and $\mathbb{S}_t$ use a pair of consecutive frames.

Inspired by [21], we first generate the high-resolution soft labels of teachers and use their resolution-degraded versions to supervise the training process of the two students. In this sense, the spatial and temporal students are trained to optimize the following spatial and temporal losses, respectively.

$$\begin{aligned}\mathcal{L}_s = \mu \cdot \mathcal{L}_{soft}\left(\mathbb{S}_s(\hat{I}_n), \mathcal{R}(\mathbb{T}_s(I_n))\right) \\ + (1-\mu)\cdot \mathcal{L}_{hard}\left(\mathbb{S}_s(\hat{I}_n), \hat{Y}_n\right),\end{aligned} \tag{1}$$

$$\begin{aligned}\mathcal{L}_t = \mu \cdot \mathcal{L}_{soft}\left(\mathbb{S}_t(\hat{I}_n, \hat{I}_{n+1}), \mathcal{R}\left(\mathbb{T}_t\left(I_n, I_{n+1}\right)\right)\right) \\ + (1-\mu)\cdot \mathcal{L}_{hard}\left(\mathbb{S}_t(\hat{I}_n, \hat{I}_{n+1}), \hat{Y}_n\right),\end{aligned} \tag{2}$$

where the scale parameter μ is used to balance the soft loss $\mathcal{L}_{soft}$ and hard loss $\mathcal{L}_{hard}$ (we empirically set $\mu = 0.5$). The $\mathcal{L}_{soft}$ is used to measure the difference between the resolution-degraded predictions of teachers and their students, while $\mathcal{L}_{hard}$ is computed between the resolution-degraded ground-truth maps and the student predictions. Both of the two losses use normalized $\mathcal{L}_2$ loss:

$$\mathcal{L}_{soft}(\mathbb{S}, \mathbb{T}) = \frac{1}{w \cdot h} \cdot \|\mathbb{S} - \mathcal{R}(\mathbb{T})\|_2^2, \tag{3}$$

$$\mathcal{L}_{hard}(\mathbb{S}, Y) = \frac{1}{w \cdot h} \cdot \|\mathbb{S} - \mathcal{R}(Y)\|_2^2, \tag{4}$$

where $\mathbb{S}$ and $\mathbb{T}$ denote the predictions of spatial/temporal student and teacher, respectively. The w and h are the width and height of low-resolution video frames. The distillation flow in the first step is shown in Fig. 3, where the high-resolution teacher knowledge is distilled into low-resolution students. We can see that this distillation flow seeks a balance of generalization ability and prediction accuracy. The soft labels given by the complex teacher models pre-trained on large-scale public or private datasets reflect a probabilistic understanding of the input scenes. With the supervision of the soft labels, the generalization ability of the student models can be

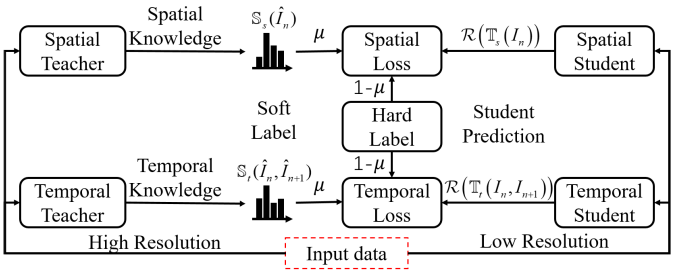

Fig. 3. Distillation flow. The spatial/temporal student networks are trained under the supervision of hard labels as well as soft labels generated by spatial/temporal teacher networks. By this way, the private knowledge inherited in the spatial and temporal teacher networks can be transfer into the spatial and temporal student networks.

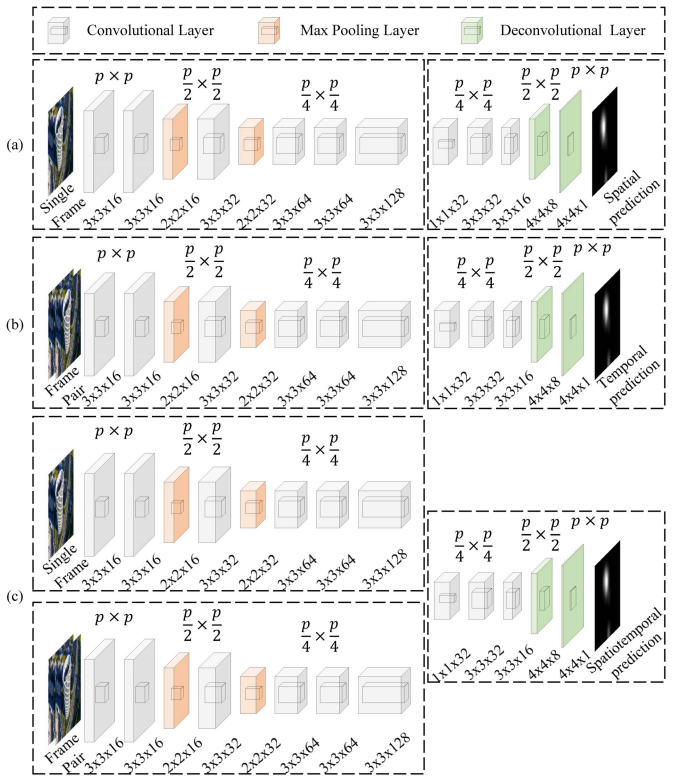

Fig. 4. Networks. (a) Spatial student network. (b) Temporal student network. (c) Spatiotemporal network. The spatial network takes a single frame as the input, while the temporal network takes a pair of successive frames as the input.

enhanced. In this way, the computational cost can be greatly reduced without remarkable loss of accuracy.

The detail structures of students are shown in Fig. 4 (a) and (b). The $\mathbb{S}_s$ is a fully convolutional network (FCN) which takes a single low-resolution frame $\hat{I}_n$ as input. Inspired by SalNet [50], $\mathbb{S}_s$ contains 13 layers. Aiming at dealing the small targets in aerial videos, we use the majority 3×3 convolutional kernels to enhance the local information extraction ability. In order to gradually expand the receptive fields, we adopt two pooling layers, in the 3rd and 5th layers of each path, respectively. Convolutional layers with 1×1 kernel size are adopted in the 9th layer to reduce the dimension of the feature maps while maintaining the diversity and effectiveness of the feature maps. A Rectified Linear Unit (ReLU) layer is adopted after every convolutional layer to improve feature representation capability. In this manner, we can obtain a low-level and mid-level feature extractor with good performance. After that, we use two convolutional layers with kernel size 3×3 to extract high-level saliency cues. In addition, we design a decoder network which contains two deconvolutional layers to upsample feature maps and constructs an output that maintains the original resolution of the input.

The $\mathbb{S}_t$ has a similar structure as $\mathbb{S}_s$ while takes a pair of successive low-resolution frames $(\hat{I}_n, \hat{I}_{n+1})$ as input. It avoids the heuristic, time-consuming optical flow calculation used in traditional methods, and instead uses learnable parameters to directly obtain the inner motion correlation between frames. In this manner, the $\mathbb{S}_t$ can calculate the temporal saliency on a low computational cost. In practice, we concatenate the current frame $\hat{I}_n$ and the next frame $\hat{I}_{n+1}$ to an input tensor with the size of $h \times w \times 6$. Note that the teacher models can be any classic deep models trained in existing public or private datasets, and we fine-tune them on the high-resolution aerial videos so that they can adapt to the specific visual attributes of aerial videos such as large-scale scenarios, small targets and vertical viewpoints.

C. Joint Spatiotemporal Knowledge Transfer

After the separate spatial and temporal knowledge distillation, the teacher knowledge has been distilled into the corresponding students. Considering that the spatial student takes one frame as the input while the temporal student takes a pair of frames, there surely exist some redundancy in these two student models, especially in the feature extraction. To further remove such inter-model redundancy, we conduct a joint spatiotemporal knowledge transfer step to extract compact and powerful spatiotemporal saliency features with the desired spatiotemporal model $\mathbb{S}_{st}$.

The network architecture of $\mathbb{S}_{st}$ is shown in Fig. 4(c), which has two input information streams. The spatial and temporal input streams share the same structure as the first eight layers of the spatial and temporal students to extract the spatial features $\mathcal{F}_s$ and the temporal features $\mathcal{F}_t$, respectively. After that, these two streams are combined into a fusion sub-network which takes a similar structure as the last four layers to the students. The input $\mathcal{F}$ of the fusion sub-network is the concatenation of $\mathcal{F}_s$ and $\mathcal{F}_t$. In this manner, the $\mathbb{S}_{st}$ takes the spatiotemporal consistency into consideration, which can effectively utilize consistent features between the spatial domain and the temporal domain to pop out foreground regions and suppress background regions.

During training $\mathbb{S}_{st}$, we initialize its first eight layers with the spatial and temporal student models and the fusion sub-network with a truncated random normal distribution. Then the training process is performed by optimizing the following hard spatiotemporal loss $\mathcal{L}_{st}$:

$$\mathcal{L}_{st} = \mathcal{L}_{hard}(\mathbb{S}_{st}(\hat{I}_n, \hat{I}_{n+1}), \hat{Y}_n). \tag{5}$$

The knowledge transfer process in training the desired spatiotemporal model is shown in Fig. 5. We can see that the

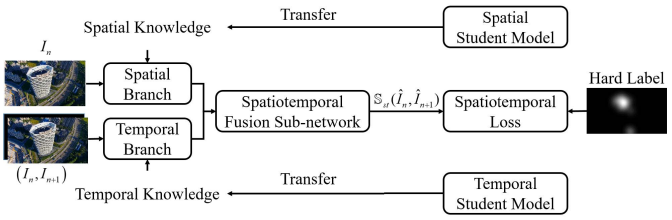

Fig. 5. The knowledge transfer process in training the desired spatiotemporal saliency model. The spatial and temporal knowledge learned by the spatial and temporal student networks is transferred into the spatiotemporal network for extracting spatial feature $\mathcal{F}_s$ and temporal feature $\mathcal{F}_t$. Then the spatiotemporal network fuses them for extracting powerful spatiotemporal features for better performance.

knowledge is first transferred from the two students into the two streams of the desired spatiotemporal model, and the features extracted by the two streams are further fine-tuned on low-resolution aerial videos in a fully supervised manner to remove the redundancy in the spatial and temporal students.

In the implementation, the model adopts the Tensorflow platform [58] on NVIDIA GPU 1080Ti and a single core Intel CPU 3.4GHz. The learning rate and batch size are set as 1×10^{-3} and 128, respectively. The Optimizer adopts Adam algorithm [59]. After training, the learned spatiotemporal model is deployed for processing aerial videos collected by drones. It takes a successive low-resolution frame pair $(\hat{I}_n, \hat{I}_{n+1})$ as the input. Then, the spatial stream receives $\hat{I}_n$ to generate spatial features, while the temporal branch takes $(\hat{I}_n, \hat{I}_{n+1})$ to output the temporal features.

IV. Experiments

To verify the effectiveness and efficiency of the proposed SKD approach, we conduct the experiments on a large-scale aerial video benchmark dataset AVS1K [60] and study the scalability of SKD on DHF1K [61]. We first introduce the experimental settings and then benchmark with ten state-of-the-art models. Finally, we conduct several diagnostics experiments to give an insight analyze of SKD approach.

A. Experimental Setting

The main experiments are conducted on AVS1K [60], a largest aerial video dataset for saliency estimation. AVS1K contains $1,000$ aerial videos and $177,644$ frames. Its maximal video resolution and average video length are 1280×720 and 5.92s, respectively. According to the salient targets, the videos in AVS1K can be divided into four categories: Building (AVS1K-B), Human (AVS1K-H), Vehicle (AVS1K-V) and Others (AVS1K-O):

- **AVS1K-B** contains 240 aerial videos with $41,471$ frames, and the average video length is 5.76s.
- **AVS1K-H** contains 210 aerial videos with $31,699$ frames, and the average video length is 5.03s.
- **AVS1K-V** contains 200 aerial videos with $27,092$ frames, and the average video length is 4.52s.
- **AVS1K-O** contains 240 aerial videos with $77,402$ frames, and the average video length is 7.37s.

Additionally, we conduct a scalability experiment on DHF1K [61], which is the current largest ground-level video visual attention dataset and has made a significant leap in terms of scalability, diversity, and difficulty when compared with conventional ground-level datasets.

To evaluate the proposed approach, we quantitatively compare its performance against that of 11 state-of-the-art models from three category groups:

1) The Heuristic Group (denoted as H Group) contains three heuristic models, including HFT [62], SP [63] and PNSP [30].

2) The Shallow Learning Group (denoted as S Group) contains two shallow-learning models, including SSD [7] and LDS [38].

3) The Deep Learning Group (denoted as D Group) contains six deep-learning models, including eDN [37], iSEEL [64], DVA [10], SalNet [50], STS [57] and ACLNet [61].

Based on the investigation in [65]–[67], we report quantitative evaluation results on five widely used evaluation metrics, including the traditional Area Under the ROC Curve (AUC), the shuffled AUC (sAUC), the Normalized Scanpath Saliency (NSS), the Similarity Metric (SIM) [68] and Correlation Coefficient (CC) [69]. AUC intuitively reflects the classification ability of ROC curve, which is generated by enumerating all probable thresholds of true positive rate versus false positive rate. Different from AUC, sAUC takes the fixations shuffled from other frames as negatives in generating the curve. NSS measures the average response at the eye fixation locations and normalizes the estimated saliency maps to zero mean and unit standard deviation. In this paper, the implementation in [70] is adopted, which efficiently computes NSS via element-wise multiplication of the estimated and ground-truth saliency maps. SIM is computed to measure the similarity between the estimated and ground saliency maps, while CC is computed as the linear correlation between them. Noting that the values of all metric are positively correlated with the model performance. However, individual metric can not perfectly indicate whether the model is efficient or not. For example, AUC prefers to assign high score to a saliency map if it correctly predicts the order of saliency and less-salient locations, even if it is fuzzy. While sAUC and NSS trend to clean saliency maps that only pop-out the most salient locations and suppress all the distractors. Particularly, we take NSS as the primary metric according to the surveys on saliency evaluation metrics [10], [71].

B. Performance Evaluation

For simplicity, we use TSNet [72] as the fixed temporal teacher and denote our models as SKD-$\mathbb{T}_s$-R where $\mathbb{T}_s$ is the spatial teacher model and R indicates the input resolution. The performance of ten state-of-the-art and our two models on the AVS1K is presented in Tab. I. Here, SKD-DVA-32 and SKD-DVA-64 use DVA as the spatial teacher and take the input resolution of 32×32 and 64×64, respectively. Moreover, the ROC Curves are given in Fig. 6. Some representative results of these models are shown in Fig. 7.

TABLE I

PERFORMANCE COMPARISON OF 11 STATE-OF-THE-ART AND OUR TWO MODELS ON AVS1K. THE BEST AND RUNNER-UP MODELS OF EACH COLUMN ARE MARKED WITH BOLD AND UNDERLINE, RESPECTIVELY. THE MODELS FINE-TUNED ON AVS1K ARE MARKED WITH *

	Models	AUC	sAUC	NSS	SIM	CC	Parameters (M)	Memory Footprint (MB)	Speed (FPS) GPU (NVIDIA 1080Ti)	Speed (FPS) CPU (Intel 3.4GHz)
H	HFT [62]	0.789	0.715	1.671	0.408	0.539	—	—	—	7.6
	SP [63]	0.781	0.706	1.602	0.422	0.520	—	—	—	3.6
	PNSP [30]	0.787	0.634	1.140	0.321	0.370	—	—	—	—
S	SSD [7]	0.737	0.692	1.564	0.404	0.503	—	—	—	32.2
	LDS [38]	0.808	0.720	1.743	0.452	0.565	—	—	—	4.6
D	eDN [37]	0.855	0.732	1.262	0.289	0.417	—	—	—	0.2
	iSEEL [64]	0.801	0.767	1.974	0.458	0.636	—	—	—	—
	DVA* [10]	0.864	0.761	2.044	0.544	0.658	25.07	59.01	49	2.5
	SalNet* [50]	0.797	0.769	1.835	0.410	0.593	25.81	43.22	28	1.5
	STS* [57]	0.804	0.732	1.821	0.472	0.578	41.25	86.94	17	0.9
	ACLNet [61]	**0.891**	**0.809**	**2.361**	**0.611**	**0.753**	—	—	43	3.1
	SKD-DVA-32	0.859	0.760	2.040	0.527	0.657	**0.30**	**0.14**	**28,738**	**1,490.5**
	SKD-DVA-64	0.867	0.770	2.100	0.534	0.674	0.30	0.58	8,522	411.7

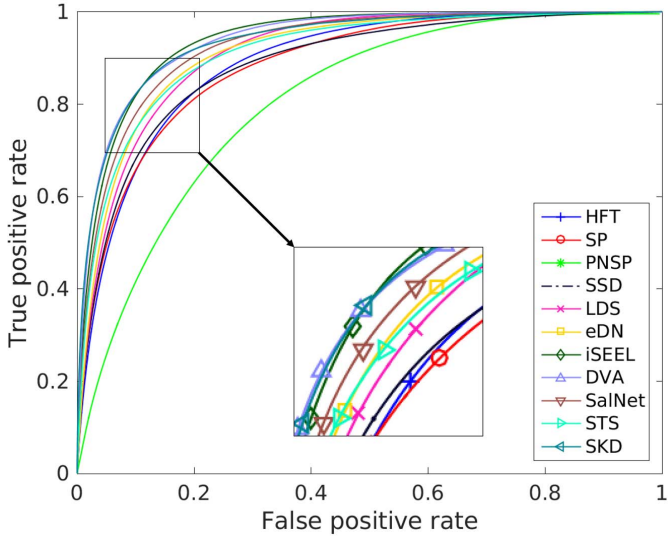

Fig. 6. ROC curves of 11 models on AVS1K.

From Tab. I, we observe that both the SKD-DVA-32 and SKD-DVA-64 are comparable to the 11 state-of-the-art models. Particularly, the SKD-DVA-64 ranks the second place in terms of AUC, sAUC, NSS and CC, while in the third place in term of SIM. Such performance improvement can be attributed to the spatiotemporal distillation framework. It distills the spatial and temporal knowledge inherited in the teachers into students in the separate spatial/temporal distillation step. Then the framework transfers such spatial and temporal knowledge into a desired spatiotemporal model and fine-tunes it for better estimation accuracy. Experimental results reveal that the SKD-DVA-64 has better representation capability when compared with traditional single stream networks (*e.g.*, SalNet), classic two-stream networks for dynamical scenarios (*e.g.*, STS) as well as multi-stream networks (*e.g.*, DVA). In term of NSS, SKD-DVA-64 achieves 2.7%, 14.4% and 15.3% performance gain to DVA, SalNet and STS, respectively. It is worth to note that the continued decrease in resolution results in a performance attenuation to some extent. The fly in the ointment is that the SKD approach still underperforms ACLNet, which employ attentive CNN-LSTM architecture and focus on learning more flexing temporal saliency representation across successive frames. Intuitively, the SKD-DVA-32 has a 2.9% accuracy drop to SKD-DVA-64 in term of NSS.

We also find that the proposed approach can achieve an ultrafast speed in aerial video saliency estimation, which can be explained by the extremely low computational cost. Our spatiotemporal network has only 0.30M parameters, namely with a 98.8% reduction to DVA. Benefiting from the combined effect of reduced parameters and input resolution, the computational cost and the memory footprint of the proposed approach are compressed into a extremely low extent. The SKD-DVA-32 and SKD-DVA-64 can achieve 421.5× and 101.7× memory reduction to DVA, respectively. In summary, the SKD-DVA-32 can achieve an ultrafast speed (28, 738 FPS) with comparable performance to 11 state-of-the-art models, while the SKD-DVA-64 can achieve a very fast speed (8, 522 FPS) and performs better than SKD-DVA-32.

Additionally, we can observe the difference lies among different categories. The heuristic models in the H group have the poorest performance. The reason may be that these heuristic models usually rely on low-level hand-crafted features and predefined rules for feature fusion. Thus these models may encounter huge challenges when infer saliency cues in unknown scenarios. By adopting learnable fusion strategies, the models in S group can achieve slightly better performance but still far from satisfactory. The key issue is that the hand-crafted features adopted in H group and S group are designed for ground-level scenarios, which may not be applicable to aerial videos. This also indicates that there may exits some unconventional visual patterns in such aerial scenarios, which should be learned from the data. Table I reveals that the models in D group generally exceed that of the H group and S group, which can be attributed to the powerful capabilities of CNNs in extracting hierarchical feature representations.

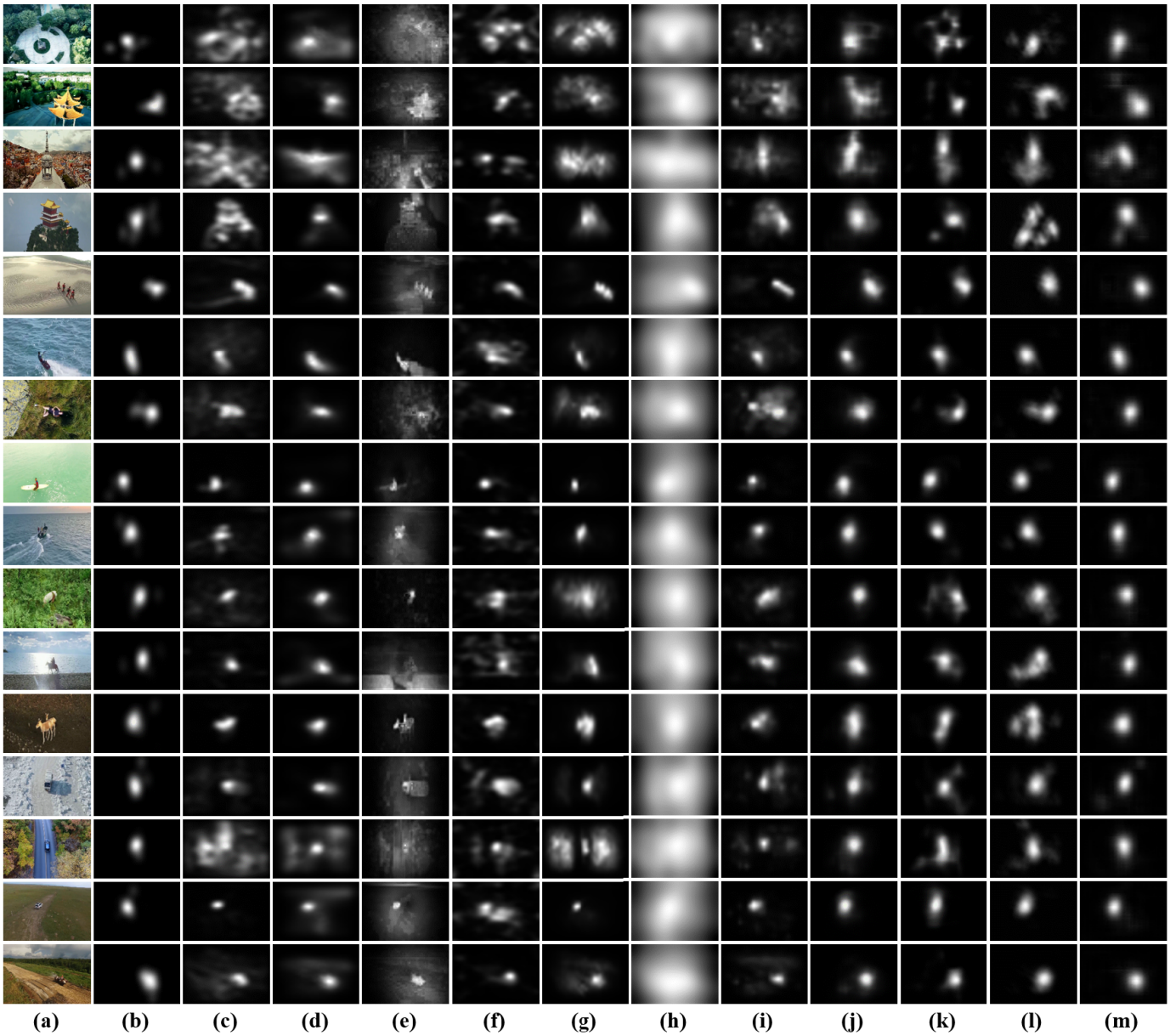

Fig. 7. Representative frames of the models on AVS1K. (a) Video frame, (b) Ground truth, (c) HFT, (d) SP, (e) PNSP, (f) SSD, (g) LDS, (h) eDN, (i) iSEEL, (j) DVA, (k) SalNet, (l) STS, (m) SKD.

It is no doubt that the SKD-DVA-64 has an impressive performance on the aerial dataset, which usually has abnormal viewpoints and small targets, but there arose another concern about its scalability to a ground-level dataset with normal viewpoints and target scales. To validate this point, we conduct a scalability experiment on DHF1K [61] and present its performance in Tab. II. We find that the performance of our model is not satisfactory, which may be caused by the fact that our model adopts small receptive field to deal with small targets in aerial scenarios, and such a design is hard to meet the requirements of the normal-scale targets in ground-level scenarios.

To follow this assumption, we present the performance of two modified versions in Tab. II. The **Modify-A** gets a slightly larger receptive field via modifying the kernel sizes of the first two convolutional layers to 5×5, while the **Modify-B** has an even larger receptive field since it modifies the kernel sizes in the first two convolutional layers to 7×7 and 5×5, receptively. Obviously, the performance of the two modified versions is comparable to the state-of-the-art models. Particularly, the **Modify-B** ranks the seventh place in term of NSS and is superior to **Modify-A**, which reveals the correctness of our assumption. To sum up, the proposed method with appropriate modification is an scalable model that can be generalized to ground-level scenarios without remarkable performance drop.

C. Diagnostics Experiments

After the promising performance is achieved, we further conduct six diagnostics experiments on AVS1K to delve into our SKD framework. In the experiments, SKD-DVA-64 is taken as the baseline model.

TABLE II

PERFORMANCE COMPARISON OF 16 STATE-OF-THE-ART MODELS ON DHF1K. THE BEST AND RUNNER-UP MODELS OF EACH COLUMN ARE MARKED WITH BOLD AND UNDERLINE, RESPECTIVELY

	Models	AUC	sAUC	NSS	SIM	CC
Static models	ITTI [24]	0.774	0.553	1.207	0.162	0.233
	GBVS [73]	0.828	0.554	1.474	0.186	0.283
	SALICON [74]	0.857	0.590	1.901	0.232	0.327
	Shallow-Net [50]	0.833	0.529	1.509	0.182	0.295
	Deep-Net [50]	0.855	0.592	1.775	0.201	0.331
	DVA [10]	0.860	0.595	2.013	0.262	0.358
Dynamic models	PQFT [75]	0.699	0.562	0.749	0.139	0.137
	Seo *et al.* [26]	0.635	0.499	0.334	0.142	0.070
	Rudoy *et al.* [76]	0.769	0.501	1.498	0.214	0.285
	Hou *et al.* [77]	0.726	0.545	0.847	0.167	0.150
	Fang *et al.* [30]	0.819	0.537	1.539	0.198	0.273
	OBDL [78]	0.638	0.500	0.495	0.171	0.117
	AWS-D [79]	0.703	0.513	0.940	0.157	0.174
	OM-CNN [80]	0.856	0.583	1.911	0.256	0.344
	Two-stream [57]	0.834	0.581	1.632	0.197	0.325
	ACLNet [61]	**0.890**	**0.601**	**2.354**	**0.315**	**0.434**
	Ours	0.831	0.502	1.511	0.199	0.294
	Modify-A	0.832	0.503	1.557	0.189	0.305
	Modify-B	0.830	0.506	1.566	0.194	0.303

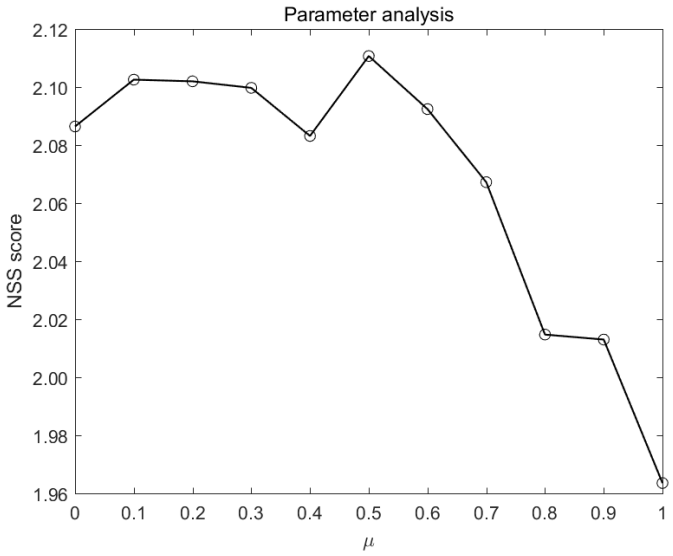

Fig. 8. Parameter analysis on AVS1K with different μ in the interval [0.0, 1.0].

1) Parameter Analysis: In the first experiment, we analyze the parameter μ in (1) and (2) that is served as a scale parameter in computing $\mathcal{L}_s$ and $\mathcal{L}_t$. The curve of NSS scores on AVS1K with different μ is shown in Fig. 8, which is computed as the mean performance value in three tests. We find that the average NSS is relatively high (greater than 2.08) when μ falls between [0.0, 0.6]. Particularly, the desired spatiotemporal model achieves the best performance when the μ is set to 0.5. However, when the μ continues to grow, the performance drops sharply. This can be interpreted as it is difficult for soft labels to accurately represent the true distribution of data, and the supervision of hard labels is indispensable. In other words, the soft labels provide an opportunity to improve the generalization ability, while the hard labels emphasize only the accuracy. When both the generalization ability and accuracy are taken into consideration, the overall performance on the testing set can become better.

TABLE III

PERFORMANCE OF SKD-DVA-64 ON FOUR SUBSETS OF AVS1K. THE BEST AND RUNNER-UP MODELS OF EACH COLUMN ARE MARKED WITH BOLD AND UNDERLINE, RESPECTIVELY

Subset	AUC	sAUC	NSS	SIM	CC
AVS1K-B	0.858	0.762	1.901	0.535	0.663
AVS1K-H	**0.903**	0.795	**2.465**	0.551	0.719
AVS1K-V	0.883	**0.804**	2.396	**0.551**	**0.728**
AVS1K-O	0.849	0.804	1.934	0.519	0.637

2) Generalization Analysis: The second experiment presents the performance of SKD-DVA-64 on four subsets to verify its generalization, as shown in Tab. III. We can find that SKD-DVA-64 achieves better performance on AVS1K-H and AVS1K-V than that on AVS1K-B and AVS1K-O. The reason arises from that the targets like human and vehicles have appropriate sizes and conspicuous motion patterns on most aerial videos so that the spatiotemporal feature extraction can be easier. By contrast, AVS1K-B and AVS1K-O have relative static or larger targets, making the saliency model difficult to separate the targets from the distractors. This result implies that our model can generalize the ability in estimating the salient targets.

3) Teacher and Resolution: The third experiment aims to assess the effect of the distilled teacher and input resolution. Without loss of generality, we fix the temporal teacher model as TSNet, and consider three candidate spatial teachers including DVA, SalNet and SSNet [72], that is $\mathbb{T}_s \in$ {DVA, SalNet, SSNet}. The performance of all the teacher models is presented in Tab. V. We find that for the spatial teachers, the DVA ranks the first place which is followed by the SalNet, and the SSNet is the worst. While for the temporal teacher, the performance of the TSNet is similar to the SalNet and the SSNet. A possible explanation for this is that the DVA has a multi-stream structure, which allows it has stronger feature representation ability than single stream networks (*e.g.*, SalNet, SSNet), can learn higher level semantic knowledge and decrease the redundancy. Meanwhile, we also check four input resolutions, having R$\in$ $\{256, 128, 64, 32\}$. The performance of our different models is presented in Tab. IV, which shows some observations. First, under the same teacher, the input resolution has a remarkable effect on the model performance in terms of all metrics. The best and runner-up performance are achieved when the input resolutions are 64×64 and 32×32, respectively. It reveals that the data redundancy could be efficiently removed by using our framework, leading to better performance. Second, the performance is consistent under the same input resolution. For example, SKD-$\mathbb{T}_s$-64 models ranks top-1 no matter what spatial teacher model it adopts. It indicates that our framework provides a general way to distill teacher knowledge for improving saliency estimation. Third, the results generated by the models in a low resolution of 32×32 still have competitive performance. Fig. 9 shows some representative results, where the results tend to be clearer

Fig. 9. Representative frames of the proposed model in various resolutions on AVS1K. (a) Video frame, (b) Ground truth, (c) SKD-DVA-256, (d) SKD-DVA-128, (e) SKD-DVA-64, (f) SKD-DVA-32.

TABLE IV

THE PERFORMANCE OF THE MODELS UNDER DIFFERENT SETTINGS ON AVS1K DATASET. T_S: SPATIAL TEACHER MODEL, RES: INPUT RESOLUTION. THE BEST AND RUNNER-UP MODELS OF EACH COLUMN IN EACH SPATIAL TEACHER SIGNAL ARE MARKED WITH BOLD AND UNDERLINE, RESPECTIVELY

T_S	Res	AUC	sAUC	NSS	SIM	CC
DVA	256	0.800	0.708	1.624	0.391	0.514
	128	0.844	0.735	1.860	0.455	0.593
	64	**0.867**	**0.770**	**2.100**	**0.534**	**0.674**
	32	0.859	0.760	2.040	0.527	0.657
SalNet	256	0.806	0.713	1.664	0.397	0.527
	128	0.840	0.739	1.899	0.461	0.604
	64	**0.868**	**0.774**	**2.108**	**0.533**	**0.676**
	32	0.859	0.756	2.022	0.524	0.651
SSNet	256	0.807	0.712	1.650	0.403	0.523
	128	0.839	0.738	1.845	0.454	0.588
	64	**0.867**	**0.774**	**2.100**	**0.536**	**0.674**
	32	0.858	0.753	2.010	0.528	0.647

when the input resolution is reducing, which can be interpreted as the resolution reduction can effectively remove the data redundancy and make it easier for the networks to extract valuable information from the data.

4) Fusion Strategy Analysis: The fourth experiment aims to validate the effectiveness of the fusion strategy for the teacher models. The performance of four teacher models and three fusion models is presented in Tab. V. From this table, we find that the performance of all three fusion models is superior to the temporal teacher model (TSNet), but interestingly, when compared with their corresponding spatial teacher models, the SalNet-TSNet and SSNet-TSNet have performance gains while the DVA-TSNet has a performance drop. This can be interpreted as both the SalNet and SSNet have similar

TABLE V

PERFORMANCE COMPARISON OF FOUR TEACHER AND THREE FUSION MODELS ON AVS1K. THE BEST AND RUNNER-UP MODELS OF EACH COLUMN ARE MARKED WITH BOLD AND UNDERLINE, RESPECTIVELY

	Models	AUC	sAUC	NSS	SIM	CC
Teacher	DVA [10]	**0.864**	0.761	**2.044**	**0.544**	**0.658**
	SalNet [50]	0.797	**0.769**	1.835	0.410	0.593
	SSNet [57]	0.834	0.701	1.686	0.470	0.537
	TSNet [57]	0.843	0.719	1.754	0.479	0.561
Fusion	DVA-TSNet	0.850	0.722	1.839	0.503	0.591
	SalNet-TSNet	0.851	0.760	1.961	0.433	0.627
	SSNet-TSNet	0.804	0.732	1.821	0.472	0.578

TABLE VI

THE PERFORMANCE COMPARISONS OF THE FULL MODEL SKD-DVA-64 AND THREE BASELINE MODELS

Model	AUC	sAUC	NSS	SIM	CC
SKD-DVA-64	0.867	0.770	2.100	0.534	0.674
Ablation-S	0.865	0.769	2.085	0.536	0.669
Ablation-T	0.866	0.762	2.007	0.506	0.647
Ablation-O	0.845	0.751	1.943	0.489	0.621

structures to the TSNet, the spatial and temporal knowledge is synchronous and complementary, leads to a reasonable performance gain in SalNet-TSNet and SSNet-TSNet.

However, for DVA-TSNet, a huge difference exists in the backbone networks of its teachers, the learned temporal knowledge is redundant for the learned spatial knowledge. Even a powerful fusion sub-network cannot remove such redundancy and extract powerful spatiotemporal features, as a result, the DVA-TSNet has a performance drop. From this experiment, we can empirically prove the fusion strategy is effective for the teacher models with similar backbone networks, but not for those backbone networks with huge differences.

5) Ablation Analysis: The ablation analyses experiment is conducted to illustrate the contributions of the separate components. To this end, we further implement three ablation models. Two ablation models (Ablation-S and Ablation-T) are generated without the joint spatiotemporal transfer to distill only the spatial and temporal teacher knowledge, respectively. Similarly, the ablation model Ablation-O is implemented without the separate spatial/temporal distillation so that it can generate the spatiotemporal saliency maps without spatial and temporal teacher knowledge. The performance is presented in Tab. VI, where we can find that the baseline model SKD-DVA-64 achieves the best performance while all the three ablation models have somewhat performance degradation. A possible explanation is that Ablation-S or Ablation-T lacks the sufficient temporal or spatial information in generating aerial video saliency maps, leading to a performance drop. In addition, without the separate spatial/temporal distillation, Ablation-O may have trouble in extracting powerful spatiotemporal cues, resulting in the lowest performance.

6) Efficiency Analysis: Beyond the effectiveness, the efficiency of our approach is shown in Tab. I. After the step-wisely

TABLE VII

INFERENCE TIME AND MEMORY FOOTPRINT OF OUR APPROACH ON GPU (NVIDIA 1080Ti) AND CPU (INTEL 3.4GHz)

Model	GPU Time / #FPS	CPU Time / #FPS	Memory footprint (MB)
SKD-T_s-256	1.414 ms / 707	37.828 ms / 26.4	9.24
SKD-T_s-128	0.381 ms / 2,626	9.675 ms / 103.4	2.31
SKD-T_s-64	0.117 ms / 8,522	2.429 ms / 411.7	0.58
SKD-T_s-32	0.035 ms / 28,738	0.671 ms / 1,490.5	0.14

removing the redundancy in intra-model, data and inter-model, the resulting model contains 0.30M parameters, leading to a great reduction again 25.07M, 25.81M and 41.25M in DVA, SalNet and STS. Due to this reduction in model parameters as well as resolution, the model memory is greatly reduced, as shown in Tab. VII. Moreover, the inference speeds of our models in different input resolutions are all remarkably improved, as demonstrated in Tab. VII. The inference runtime on the GPU platform can be reduced to 1.414ms, 0.381ms, 0.117ms and 0.035ms in the resolution of 256×256, 128×128, 64×64 and 32×32, respectively. In particular, SKD-$\mathbb{T}_s$-32 achieves an extremely fast speed of 28,738 FPS and 1,490.5 FPS on the GPU and CPU platforms, respectively. These results indicate that our approach provides a practical solution to deploy existing complex deep saliency models on low-end mobile devices, such as drones.

7) Limitation Analysis: Despite their comparable performance and significant performance advantages, the proposed models have trouble to break through their performance upper limits, due to their limitations in design philosophy. 1) Complex training process. The models are trained in a two-step manner, which can not directly fuse the spatial and temporal cues, resulting in a redundancy training process. 2) Limited temporal cues. The temporal teacher adopted highly relies on optical flow, which is computationally expensive and can only yield the temporal cues between two frames. A cheaper and multi-frame temporal cues extractor will contribute to the further improvement of the final model performance. 3) Inadequate model presentation ability. The proposed spatial, temporal and spatiotemporal models are simple models without powerful presentation ability. This constrains their ability to yield powerful features.

V. CONCLUSION

At present, most deep models for dynamic saliency estimation suffer from heavy computational cost and memory footprint, which poses a dilemma for them to be deployed on devices with limited computational capability and memory space. To address this issue, this paper proposes a low-resolution dynamic saliency estimation approach via spatiotemporal knowledge distillation. By step-wisely removing the intra-model, inter-model and data redundancies, a compact and simple saliency model with impressive performance on aerial videos can be established. Experimental results show that the proposed approach is comparable to 11 state-of-the-art models in estimating visual saliency on aerial videos, while running at an extremely fast speed of 28,738 FPS and 1,490.5 FPS on the GPU and CPU platforms, respectively. Such a performance means the model can be easily deployed on drones.

In the future work, we will tentatively explore attention-assisted UVA video object detection, aiming at designing a robust detection model that can handle aerial scenes in complex weather environment.

REFERENCES

[1] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S. C. Zhu, "Joint inference of groups, events and human roles in aerial videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4576–4584.

[2] J. Zhang, Y. Wu, W. Liu, and X. Chen, "Novel approach to position and orientation estimation in vision-based UAV navigation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 46, no. 2, pp. 687–700, Apr. 2010.

[3] K. Boudjit and C. Larbes, "Detection and implementation autonomous target tracking with a quadrotor AR. Drone," in *Proc. 12th Int. Conf. Inform. Control, Automat. Robot. (ICINCO)*, vol. 2, Jul. 2015, pp. 223–230.

[4] A. C. Woods and H. M. La, "Dynamic target tracking and obstacle avoidance using a drone," in *International Symposium on Visual Computing*. Berlin, Germany: Springer, 2015, pp. 857–866.

[5] C. Aker and S. Kalkan, "Using deep networks for drone detection," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Sep. 2017, pp. 1–6.

[6] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.

[7] 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, Dec. 2015.

[8] M. Zhang, Y. Pang, Y. Wu, Y. Du, H. Sun, and K. Zhang, "Saliency detection via local structure propagation," *J. Vis. Commun. Image Represent.*, vol. 52, no. 4, pp. 131–142, 2018.

[9] 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. 2019.

[10] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.

[11] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 20–33, Jan. 2018.

[12] 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.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[14] A. Borji and L. Itti, "Cat2000: A large scale fixation dataset for boosting saliency research," in *Proc. IEEE Comput. Vis. Pattern Recognit. Workshop Future Datasets*, May 2015.

[15] S. Abu-El-Haija *et al.*, "Youtube-8m: A large-scale video classification benchmark," 2016, *arXiv:1609.08675*. [Online]. Available: https://arxiv.org/abs/1609.08675

[16] M. Denil, B. Shakibi, L. Dinh, M. Ranzato, and N. de Freitas, "Predicting parameters in deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2148–2156.

[17] Y.-D. Kim, E. Park, S. Yoo, T. Choi, L. Yang, and D. Shin, "Compression of deep convolutional neural networks for fast and low power mobile applications," in *Proc. Int. Conf. Learn. Representations (ICLR)*, May 2016, pp. 1–16.

[18] X. Zhang, J. Zou, K. He, and J. Sun, "Accelerating very deep convolutional networks for classification and detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 1943–1955, Oct. 2016.

[19] X. Zhang, J. Zou, X. Ming, K. He, and J. Sun, "Efficient and accurate approximations of nonlinear convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1984–1992.

[20] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2006, pp. 535–541.

[21] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. Adv. Neural Inf. Process. Syst. Workshop*, Mar. 2015, p. 9.

[22] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: Hints for thin deep nets," in *Proc. Int. Conf. Learn. Representations(ICLR)*, Dec. 2015, pp. 1–12.

[23] W. Wang, Q. Lai, H. Fu, J. Shen, and H. Ling, "Salient object detection in the deep learning era: An in-depth survey," 2019, *arXiv:1904.09146*. [Online]. Available: https://arxiv.org/abs/1904.09146

[24] 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.

[25] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.

[26] 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.

[27] 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.*, Dec. 2013, pp. 1976–1983.

[28] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2806–2813.

[29] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4185–4196, Nov. 2015.

[30] 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.

[31] Y. Fang, C. Zhang, J. Li, J. Lei, M. P. 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, Oct. 2017.

[32] A. Borji, D. N. Sihite, and L. Itti, "Probabilistic learning of task-specific visual attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 470–477.

[33] Y. Chen, G. Tao, Q. Xie, and M. Song, "Video attention prediction using gaze saliency," *Multimedia Tools Appl.*, vol. 78, no. 19, pp. 26867–26884, Oct. 2019.

[34] 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, 2010.

[35] W.-F. Lee, T.-H. Huang, S.-S. Yeh, and H. H. Chen, "Learning-based prediction of visual attention for video signals," *IEEE Trans. Image Process.*, vol. 20, no. 11, pp. 3028–3038, Nov. 2011.

[36] 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.

[37] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2798–2805.

[38] S. Fang, J. Li, Y. Tian, T. Huang, and X. Chen, "Learning discriminative subspaces on random contrasts for image saliency analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 5, pp. 1095–1108, May 2017.

[39] M. Song, C. Chen, S. Wang, and Y. Yang, "Low-level and high-level prior learning for visual saliency estimation," *Inf. Sci.*, vol. 281, pp. 573–585, Oct. 2014.

[40] P. Hu, B. Shuai, J. Liu, and G. Wang, "Deep level sets for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2300–2309.

[41] J. Kuen, Z. Wang, and G. Wang, "Recurrent attentional networks for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3668–3677.

[42] X. Li *et al.*, "DeepSaliency: Multi-task deep neural network model for salient object detection," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3919–3930, Aug. 2016.

[43] 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.

[44] K. Zhang, W. Luo, Y. Zhong, L. Ma, W. Liu, and H. Li, "Adversarial Spatio-temporal learning for video deblurring," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 291–301, Jan. 2019.

[45] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017.

[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Sep. 2014, pp. 1–14.

[48] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze I: Boosting saliency prediction with feature maps trained on ImageNet," 2014, *arXiv:1411.1045*. [Online]. Available: https://arxiv.org/abs/1411.1045

[49] M. Kümmerer, T. S. Wallis, and M. Bethge, "DeepGaze II: Reading fixations from deep features trained on object recognition," 2016, *arXiv:1610.01563*. [Online]. Available: https://arxiv.org/abs/1610.01563

[50] J. Pan, E. Sayrol, X. Giro-i-Nieto, K. Mcguinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 598–606.

[51] N. Imamoğlu, C. Zhang, W. Shmoda, Y. Fang, and B. Shi, "Saliency detection by forward and backward cues in deep-CNN," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2017, pp. 430–434.

[52] J. Ba and R. Caruana, "Do deep nets really need to be deep," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2654–2662.

[53] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4133–4141.

[54] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4320–4328.

[55] A. A. Rusu *et al.*, "Policy distillation," in *Proc. Int. Conf. Learn. Representations*, Nov. 2015, pp. 1–12.

[56] S. Ruder, "An overview of multi-task learning in deep neural networks," 2017, *arXiv:1706.05098*. [Online]. Available: https://arxiv.org/abs/1706.05098

[57] 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.

[58] M. Abadi *et al.*, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement. (OSDI* , 2016, pp. 265–283.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations (ICLR)*, Dec. 2014, pp. –15.

[60] K. Fu, J. Li, H. Shen, and Y. Tian, "How drones look: Crowdsourced knowledge transfer for aerial video saliency prediction," 2018, *arXiv:1811.05625*. [Online]. Available: https://arxiv.org/abs/1811.05625

[61] 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.*, Jun. 2018, pp. 4894–4903.

[62] 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.

[63] 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.

[64] H. R. Tavakoli, A. Borji, J. Laaksonen, and E. Rahtu, "Exploiting inter-image similarity and ensemble of extreme learners for fixation prediction using deep features," *Neurocomputing*, vol. 244, pp. 10–18, Jun. 2017.

[65] 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.*, Dec. 2013, pp. 1153–1160.

[66] 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.*, Dec. 2015, pp. 190–198.

[67] 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.

[68] 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.

[69] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 438–445.

[70] S. Marat, T. H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Modelling spatio-temporal saliency to predict gaze direction for short videos," *Int. J. Comput. Vis.*, vol. 82, p. 231, May 2009.

[71] N. Liu and J. Han, "A deep spatial contextual long-term recurrent convolutional network for saliency detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3264–3274, Jul. 2018.

[72] C. C. Bak, A. Erdem, and E. Erdem, "Two-stream convolutional networks for dynamic saliency prediction," 2016, *arXiv:1607.04730*. [Online]. Available: https://arxiv.org/abs/1607.04730
[73] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 545–552.
[74] X. Huang, C. Shen, X. Boix, and Q. Zhao, "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 262–270.
[75] 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.
[76] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor, "Learning video saliency from human gaze using candidate selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1147–1154.
[77] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 681–688.
[78] S. H. Khatoonabadi, N. Vasconcelos, I. V. Bajic, and Y. Shan, "How many bits does it take for a stimulus to be salient," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5501–5510.
[79] 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.
[80] 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, pp. 602–617.

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 an Associate Professor with the School of Computer Science and Engineering, Beihang University, Beijing, China. Before he joined Beihang University in June 2014, he severed as a Researcher in several multimedia groups of Nanyang Technological University, Peking University, and Shanda Innovations. He has authored or coauthored over 60 technical articles in refereed journals and conferences, such as TPAMI, IJCV, TIP, CVPR, ICCV, and ACM MM. His research interests include computer vision and multimedia big data, especially learning-based visual content understanding. He is a Senior Member of CCF.

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. Part of this work was done during his short-term visit to Peng Cheng Laboratory. His research interests include computer vision and image understanding.

Shengwei Zhao received the B.S. degree from the School of Mathematics and Statistics, Wuhan University in 2017. He is currently pursuing the master's degree with the Institute of Information Engineering, Chinese Academy of Sciences, and the School of Cyber Security, University of Chinese Academy of Sciences. His major research interests include deep learning and computer vision, especially low-quality image analysis.

Shiming Ge (M'13–SM'15) received the B.S. and Ph.D. degrees in electronic engineering from the University of Science and Technology of China (USTC) in 2003 and 2008, respectively. He is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences. He is also a member of the Youth Innovation Promotion Association, Chinese Academy of Sciences. Prior to that, he was a Senior Researcher and a Project Manager with Shanda Innovations and a Researcher with Samsung Electronics and the Nokia Research Center. His research mainly focuses on computer vision, data analysis, machine learning, and AI security, especially efficient learning models and solutions toward scalable applications.