TinyHD: Efficient Video Saliency Prediction with Heterogeneous Decoders using Hierarchical Maps Distillation

Feiyan Hu[1], Simone Palazzo[2], Federica Proietto Salanitri[2], Giovanni Bellitto[2], Morteza Moradi[2], Concetto Spampinato[2], Kevin McGuinness[1]
[1] Insight SFI Research Centre for Data Analytics, Dublin City University, Dublin, Ireland
{feiyan.hu, kevin.mcguinness}@dcu.ie
[2] PeRCeiVe Lab, University of Catania, Catania, Italy
{simone.palazzo, concetto.spampinato}@unict.it

Abstract

Video saliency prediction has recently attracted attention of the research community, as it is an upstream task for several practical applications. However, current solutions are particurly computationally demanding, especially due to the wide usage of spatio-temporal 3D convolutions. We observe that, while different model architectures achieve similar performance on benchmarks, visual variations between predicted saliency maps are still significant. Inspired by this intuition, we propose a lightweight model that employs multiple simple heterogeneous decoders and adopts several practical approaches to improve accuracy while keeping computational costs low, such as hierarchical multi-map knowledge distillation, multi-output saliency prediction, unlabeled auxiliary datasets and channel reduction with teacher assistant supervision. Our approach achieves saliency prediction accuracy on par or better than state-of-the-art methods on DFH1K, UCF-Sports and Hollywood2 benchmarks, while enhancing significantly the efficiency of the model.

1. Introduction

Video saliency prediction aims at estimating patterns of human attention during free-viewing of dynamic scenes, to emulate the capabilities of the human visual system of quickly analyzing and interpreting the surrounding environment. Due to its several practical applications [7, 5, 28, 33, 9, 40, 12, 27], it is an active area of research in computer vision. However, the solution to this problem is not trivial, for several reasons. First, attention mechanisms in the human visual system are not fully known, so it is not clear how to emulate them. Also, it requires complex modeling of both visual features and their motion and interaction: an object with striking visual patterns may be shadowed by a bland element of the scene that starts moving in a peculiar way. Finally, modeling the temporal dimension may become computationally expensive, especially with current deep learning methods based on spatio-temporal 3D convolutions, thus limiting the applicability to low-power devices.

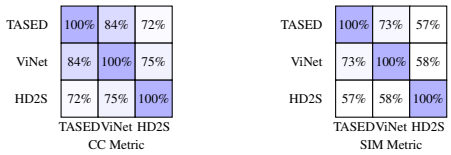


Figure 1: Measuring prediction similarity among video saliency prediction models TASED, ViNet and HD2S on DHF1K validation set.

Many solutions have been proposed, based on different assumptions on how to capture video saliency. It is interesting to note that, in spite of the remarkably different research directions followed by the variety of works in the literature, top results over video saliency prediction benchmarks are very close [1, 3, 36], suggesting that predictions of different models are similar. We assessed the validity of this conclusion by comparing three of the best performing methods on the DHF1K dataset [36] — TASED [25], HD2S [1] and ViNet [16] — not in terms of their scores on summary metrics, but in terms of the relative similarity of the predicted saliency maps. To illustrate our findings, Fig. 1 shows pairwise similarities between predicted maps over two common metrics, Linear Correlation Coefficient (CC) and Similarity (SIM). Although the three approaches achieve similar scores on both metrics on DHF1K (between 0.470 and 0.511 for CC, and between 0.361 and 0.406 for SIM), the same metrics computed between each other are relatively low, compared to what one would expect given their similarity to the saliency ground truth. A visual inspection of the saliency maps generated by methods under

comparison confirms this behavior: Fig. 2 shows that it is common to find cases where each approach produces remarkably different saliency maps.

Notably, all three methods — TASED, HD2S and ViNet — are encoder-decoder networks and share the same encoder, S3D [38], while employing different decoding strategies (a U-Net–like approach for TASED and ViNet, a hierarchical map aggregation for HD2S). This suggests that a key factor underlying differences between current video saliency prediction approaches lies in the way encoded features are processed in the decoding path, leading to models that learn specific (and often exclusive) representations. We strengthen this hypothesis by experimenting with another decoding strategy, exemplified by DLA [39], which combines hierarchical decoding with complex feature interactions: the results, also included in Fig. 2, show yet another saliency prediction pattern, while using the same encoder network, S3D. These results lead us to hypothesize that different model architectures introduce different inductive biases, which are more suitable to recognize certain patterns more than others, thus requiring to increase model capacity in order to generalize well to multiple saliency dynamics. Indeed, the size of weights of models in our analysis range between 82 MB and 116 MB, and the size of the top ten models in the DHF1K leaderboard[1] is on average 238 MB.

Given these premises, instead of increasing the complexity of a single decoding strategy, it may be more efficient to employ multiple simpler architectures with fewer parameters, relying on each architecture's capability to attend to different salient regions and combining their results. Hence, we propose **TinyHD**, a *lightweight, efficient and heterogeneous multi-decoder architecture* for video saliency prediction. The proposed method is inspired by encoder-decoder architectures, but introduces the adoption of heterogeneous decoding strategies in order to reduce the complexity of each decoder, increasing efficiency (the weights of the resulting model take only 16 MB) and improving the accuracy of predictions, as we show in our experiments. Furthermore, along the direction of reducing computational costs while retaining high accuracy, we also introduce a novel knowledge distillation approach, based on exploiting a teacher with multiple hierarchical predictions: this allows the model to freely learn its own features, since no explicit conditioning on representations is enforced, while at the same time receiving a supervision signal that encodes information at different layers of abstraction.

Experiments confirm that our model can generate high-quality predictions with low computational costs and model size (only 16 MB). We assess the impact of our heterogeneous multi-decoder strategy by carrying out extensive ablation studies and comparing alternative architectures. We also demonstrate the effectiveness of our knowledge distillation strategy, compared to the employment of a non-hierarchical teacher. To summarize our contributions:

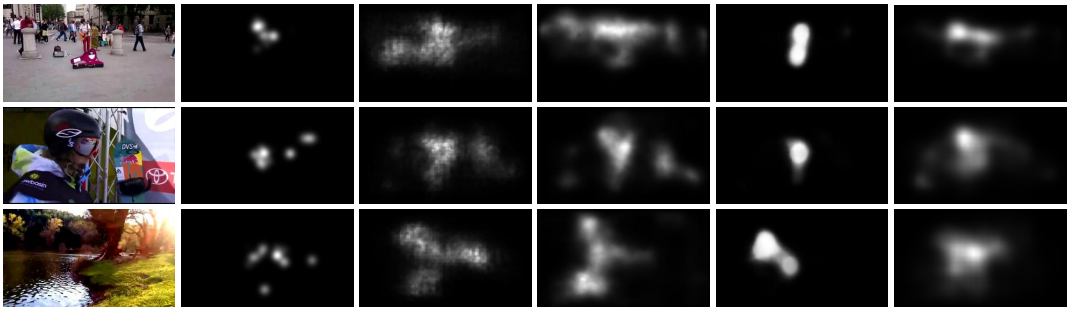
- We propose a decoding strategy for video saliency prediction which combines heterogeneous decoders to exploit the specific pattern analysis capabilities, while reducing the overall model complexity. To our knowledge, we are first to propose multiple saliency maps output using 3D CNN to improve model efficiency.
- We employ a knowledge distillation approach based on a hierarchical teacher, providing saliency maps estimated from different abstraction layers.
- Extensive experiments show that our model achieves state-of-the-art performance on the DHF1K benchmark, at lower computational costs of current methods. Ablation studies support the motivations for our decoding and knowledge distillation strategies.

2. Related Work

The main contributions of the proposed approach consist of a novel heterogeneous multi-decoder scheme, which combines lightweight versions of common decoding strategies, and a multi-objective knowledge distillation approach. In this section, we briefly present the state-of-the-art on these topics.

Decoding strategies for video saliency prediction. Among recent methods from the state-of-the-art for video saliency prediction, leveraging encoder-decoder networks can be considered a mainstream approach; however, several architectural variations have been proposed for feature sharing between encoder and decoder and for output reconstruction. As shown in the taxonomy presented in Fig. 3, a simpler class of approaches employs independent encoder and decoder, with no feature sharing between the two paths. Among these, approaches based on recurrent layers typically model temporal dynamics at the bottleneck of the architecture [18, 37, 22]. Non-recurrent architectures, instead, model time by means of 3D convolutions [42, 38, 4], Other approaches employ architectures similar to U-Net by introducing skip connections that encourage feature sharing between encoder and decoder. TASED [25] aggregates spatio-temporal features through the use of auxiliary pooling for reducing the temporal dimension. ViNet [16] integrates S3D features from multiple hierarchical levels by employing trilinear interpolation and 3D convolutions. UNISAL [6] proposes a multi-objective unified framework for both 2D and 3D saliency with domain-specific modules and a lightweight recurrent architecture to handle temporal dynamics; While single-decoder approaches are common, multi-decoder output integration has recently attracted interest. DVA [35] and HD2S [1] fuse maps predicted by independent decoders operating at different abstraction levels. RecSal [30] predicts multiple saliency maps in a multi-objective training framework. Recent works introduce more

[1] https://mmcheng.net/videosal/



(a) Frame (b) GT (c) TASED (d) ViNet (e) HD2S (f) DLA

Figure 2: Examples of video saliency maps from state-of-the-art methods. Although they achieve very similar performance on popular metrics, remarkable differences can be seem in the learned saliency patterns.

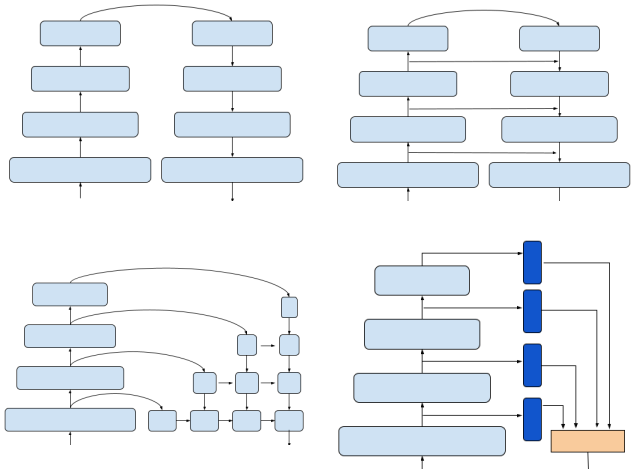


Figure 3: A taxonomy of decoding strategies commonly employed in video saliency prediction. Subfigures(top-left, top-right, bottom-left, bottom-right): Independent encoder and decoder, with no feature sharing between the two paths [4, 18, 22, 37, 42]; U-Net–like architecture, with features sharing between encoder and decoder [6, 16, 19, 25]; Deep Layer Aggregation [39]; Hierarchical intermediate map aggregation [1, 30, 35].

complex feature interactions among decoding paths, where high-resolution features are affected by deeper high-level features, as in DLA [39] and TSFP-Net [3]. All of the approaches presented employ either a single-decoder architecture or a homogeneous multi-decoder one, where differences between decoders lie in the number of layers rather than in their structure. In our work, we propose an architecture which combines heterogeneous decoder structures, in order to better exploit their distinctive saliency prediction properties and thus increase computational efficiency.

Knowledge distillation for visual saliency prediction. Knowledge distillation [13, 11] is commonly employed to train an efficient *student* model from a more complex *teacher* model, with higher accuracy than when training the student directly from dataset labels. Several knowledge distillation approaches have been recently proposed for video saliency prediction. SKD-DVA [20] proposes spatio-temporal knowledge distillation with two teachers and two students, with each pair focusing on either spatial or temporal transfer. SV2T-SS [41] distills corresponding features of teacher and student (implemented as encoder-decoder networks), based on first- and second-order feature statistics transfer. UVA-DVA [10] employs separate spatial and temporal teachers, whose knowledge is transferred to a single student model, which then fuses the resulting features in the final saliency prediction, achieving reasonable accuracy at impressive speed. Leveraging knowledge distillation for video salient object detection is the main theme of the work in [34]. The knowledge distillation setting proposed in our work differs from existing techniques in two main aspects: 1) we define a multi-objective distillation target on saliency maps directly; 2) we employ a hierarchical model as a teacher in order to further capture differences in saliency patterns extracted at multiple scales.

3. Methodology

3.1. Overview

The overall architecture of the proposed saliency prediction network with knowledge distillation is shown in Fig. 4. Following the taxonomy introduced in Sect. 2, a shared encoder extracts multi-level features that are then processed by three parallel decoding architectures: **decoder 1 (D1)** implements hierarchical intermediate maps aggregation (inspired by HD2S); **decoder 2 (D2)** employs a U-Net–like approach; **decoder 3 (D3)** is based on deep layer aggregation concepts (as in DLA [39]).

The hierarchical aggregation decoder (i.e., *decoder 1* in Fig. 4) produces four intermediate saliency maps from features extracted at different encoder layers; then, the set of

predictions from all decoders are fused into the final prediction. At training time, we compute a supervised loss by comparing the final prediction to the ground-truth map, and a knowledge distillation loss on the final prediction and the intermediate maps extracted by D1 (all losses are based on Kullback-Leibler divergence between saliency maps; see Sect. 3.3). In order to have a correspondence between intermediate maps produced by D1 and teacher maps, we employ HD2S as a teacher, since it naturally and semantically matches the decoder's hierarchical structure.

3.2. Encoder structure

Depthwise separable convolutions are widely used for efficient network design, as in MobileNetV2 [31], commonly pre-trained on ImageNet and used as a backbone for lightweight models. In order to adapt it as a 3D video feature extractor, we follow the *kernel inflation* approach introduced in [2] and already employed for static [14] and dynamic [6] saliency prediction. A 2D convolutional kernel of size $C_{\text{in}} \times C_{\text{out}} \times H \times W$ can be inflated into a 3D kernel of size $C_{in} \times C_{out} \times T \times H \times W$ by replicating its weights along the temporal dimension T. This simple trick provides a convenient initialization that responds to common spatial patterns and can be gradually adapted to temporal dynamics during training, eliminating the burden of learning basic spatial structures from scratch. Given the inflated MobileNetV2 encoder, we follow the approach in FastSal [14] to extract four blocks of concatenated feature from the whole set of layers.

3.2.1 Decoder structure and multiple prediction

The set of heterogeneous decoders, employed in our model, includes **D1** (hierarchical map aggregation), **D2**(U-Net–like) and **D3** (deep layer aggregation). Our realizations of each of these approaches are designed to process the four input streams of features extracted by the encoder. D1 produces four intermediate saliency maps, while D2 and D3 produce a map each. The fusion layer that computes the final output map is implemented as a 1×1 convolution of the predicted maps. Architectural details are reported in the supplementary materials. As an additional efficiency consideration, we note that the high computation cost of many state-of-the-art approaches due to multiple-input/single-output (MISO) prediction, where a sequence of frames is used to predict a single saliency map, usually referring to the last frame. This provides a full context of previous frames to the model, but also means that, in order to predict N saliency maps (without interpolation), N forward passes are also required, with a proportional increase of computational power. In order to further improve efficiency, we implement a multiple-input/multiple-output (MIMO) schema for output generation, by designing decoders that predict a number of saliency maps equal to the number of frames provided to the encoder. MIMO decoders can intrinsically make use of the similarity between consecutive saliency maps, and employ this information to reduce the computational power required to generate the same number of saliency maps by MISO decoders. Of course, the downside is that each frame has a different amount of surrounding context; however, in our experiments this has little impact on our model's accuracy.

3.3. Knowledge distillation

Given the presence of multiple decoders in our model, one of which also produces intermediate saliency maps, choosing a distillation approach to supervise the student's training is not trivial. As illustrated in Fig. 4, we carry out knowledge distillation by extracting intermediate and final outputs from a hierarchical teacher to supervise intermediate of one student decoder and final student outputs. Our design of the distillation process is guided by several observations. First and foremost, it is necessary to provide a training signal at the very output of the model, in order to train the final fusion layer. Second, carrying out distillation at the representation level, by enforcing similarity between teacher and student features, defeats the purpose of having multiple decoders that are meant to recognize their own distinctive saliency patterns and should therefore be free to independently learn their own features. Also, using saliency maps directly ensures that the output and target have the same size, so that the use of adaptation layers to match feature size of student's and teacher's can be avoided. We therefore choose to use intermediate saliency maps from a hierarchical teacher, HD2S [1], since this makes it possible in a natural way to affect the model at different depths of the encoder, without providing as strong a training signal as internal features.

We formalize our knowledge distillation procedure as follows. Let $\mathcal{V}$ be the space of video sequences and $\mathcal{S}$ be the space of saliency maps (whether for the entire sequence or for a single frame); let $\mathcal{M}$ be a family of models such that each element in $\mathcal{M}$ is a function $M : \mathcal{V} \to \mathcal{S}^{n+1}$, which provides n intermediate and one output saliency maps. We thus define a teacher $T \in \mathcal{M}$ and student $S \in \mathcal{M}$. For simplicity, the notations S_i and T_i will indicate the i-th map generated by, respectively, the student and teacher; indexes from 1 to n will denote intermediate maps, while index $n+1$ will refer to the final output. Saliency map distance is measured by Kullback-Leibler (KL) divergence:

$$\mathcal{L}_{\text{KL}}(\mathbf{x}, \mathbf{y}) = \sum_i y_i \log \frac{y_i}{x_i}, \tag{1}$$

with i iterating over spatial locations of the saliency maps.

At each training iteration, we sample a video sequence $\mathbf{v} \in \mathcal{V}$ and its ground-truth saliency $\mathbf{s} \in \mathcal{S}$. The em-

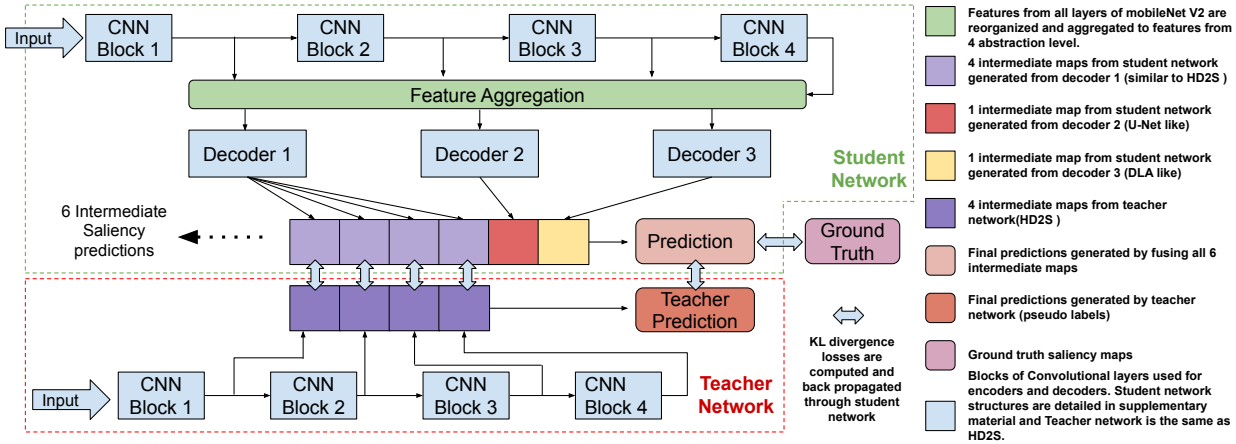


Figure 4: Overview of the proposed multi-decoder architecture with hierarchical knowledge distillation.

ployed loss function aims at minimizing the KL divergence between student and teacher maps (both intermediate and final) and between (final) student and ground-truth maps:

$$\mathcal{L} = \sum_{i=1}^{n+1} \mathcal{L}_{\mathrm{KL}}\big(S_i\left(\mathbf{v}\right), T_i\left(\mathbf{v}\right)\big) + \mathcal{L}_{\mathrm{KL}}\big(S_{n+1}\left(\mathbf{v}\right), \mathbf{s}\big). \quad (2)$$

3.3.1 Training with auxiliary dataset

The usage of unlabeled auxiliary datasets in a knowledge distillation setting has been shown to help boost performance [21, 32, 15]. Following this approach, we introduce a new video distribution $\mathcal{W}$, and extend the loss function with a term that measures the distance between student's predicted saliency maps and "pseudo-labels" (which are, in fact, also maps) provided by the teacher. As a result, given a pair of input videos $\mathbf{v} \in \mathcal{V}$ and $\mathbf{w} \in \mathcal{W}$, the new loss function becomes:

$$\mathcal{L} = \sum_{i=1}^{n+1} \mathcal{L}_{\mathrm{KL}}\big(S_i\left(\mathbf{v}\right), T_i\left(\mathbf{v}\right)\big) + \sum_{i=1}^{n+1} \mathcal{L}_{\mathrm{KL}}\big(S_i\left(\mathbf{w}\right), T_i\left(\mathbf{w}\right)\big) + \mathcal{L}_{\mathrm{KL}}\big(S_{n+1}\left(\mathbf{v}\right), \mathbf{s}\big). \quad (3)$$

3.3.2 Channel reduction with teacher assistant

Previous works have shown that, with a suitable network design, it is possible to decrease the number of channels in the encoder's layers, in order to reduce the computational cost, without an excessive loss in accuracy [8]. Our channel reduction strategy applies multiple knowledge distillation iterations: at each of them, a new student is initialized by averaging the weights of each pair of consecutive kernels into a new kernel. Although kernel ordering is essentially random, this approach has been shown to provide a meaningful initialization to the new student. Additionally, we also explore the "teacher assistant" [26] distillation strategy: rather than using the original teacher to perform knowledge distillation on reduced-channel students, we employ the full-capacity student (i.e., before any channel reduction) as a new teacher. As a result, by combining the channel reduction and teacher assistant, we encourage the model to distill more information while reducing computational cost.

4. Experiments

4.1. Datasets and Metrics

We conduct experiments on DHF1K [36], UCF-Sports [29, 24] and Hollywood2 [23, 24] datasets, commonly employed to evaluate video saliency prediction.

DHF1K contains 1000 videos split into 600/100/300 for training, validation, and test (unreleased). Eye fixations are collected from 17 participants in free-viewing experiments. **UCF-Sports** is a task-driven dataset that includes 150 videos (103 for training, 47 for test) covering 9 sport activities. Participants were asked to identify the activity in each video sequence. **Hollywood2** includes 1707 videos extracted from 69 movies and categorized between 12 action classes. At data collection, 3 observers are free-viewing, 12 observers are asked recognize the action, and 4 observers are asked to recognize the scene. 823 videos are used for training and 884 for test. We also employ the Kinetic-400 [17] action recognition benchmark as auxiliary dataset, used by the teacher to generate additional training inputs with pseudo-labels. For evaluation purposes, we report results in terms of the standard metrics for video saliency prediction [35]: AUC-Judd (AUC-J), AUC-Borji (AUC-B), Linear Correlation Coefficient (CC), Normalized Scanpath Saliency (NSS), and Similarity Metric (SIM).

Table 1: Comparison with SoA on the DHF1K and Hollywood2 test set in both the MISO and MIMO settings.

(a) Prediction accuracy and computational cost on the DHF1K test set. GMACs are estimated for 16 frames, hence the $\times 16$ multiplication for MISO models. Models marked with a * are image saliency models.

Models	AUC-J	SIM	sAUC	CC	NSS	GMACs	#params
Multi-input/single-output (MISO) prediction							
SalGAN*	0.866	0.262	0.709	0.370	2.043	45.73×16	31.92M
FastSal*	0.887	0.293	0.712	0.426	2.330	2.64×16	2.47M
3DSal	0.850	0.321	0.623	0.356	1.996	136.45×16	46.15M
TASED	0.895	0.361	0.712	0.470	2.667	91.75×16	21.26M
ViNet	0.908	0.381	**0.729**	**0.511**	2.872	115.28×16	31.1M
HD2S	0.908	**0.406**	0.700	0.503	2.812	11.08×16	29.8M
TinyHD-S	**0.909**	0.396	0.714	0.505	**2.921**	5.57×16	3.94M
Multi-input/multi-output (MIMO) prediction							
SalEMA	0.890	0.466	0.667	0.449	2.574	640.16×1	31.79M
STRA-Net	0.895	0.355	0.663	0.458	2.558	266.01×3	168.02M
UNISAL	0.901	**0.390**	0.691	0.490	2.776	19.42×1	3.71M
TinyHD-M	**0.905**	0.387	**0.707**	**0.493**	**2.819**	7.95×1	3.92M

(b) Prediction accuracy on Hollywood2

Models	AUC-J	SIM	CC	NSS
Multi-input/single-output prediction				
ACLNet	0.913	0.757	0.623	3.086
SalSAC	0.931	0.529	0.670	3.356
TASED	0.918	0.507	0.646	3.302
ViNet	0.930	0.550	**0.693**	3.730
HD2S	**0.936**	0.551	0.670	3.352
TinyHD-S	0.935	**0.561**	0.690	**3.815**
Multi-input/multi-output prediction				
SalEMA	0.919	0.487	0.613	3.186
STRA-Net	0.923	0.536	0.662	3.478
UNISAL	**0.934**	0.542	0.673	**3.901**
TinyHD-M	**0.934**	**0.553**	**0.686**	3.744

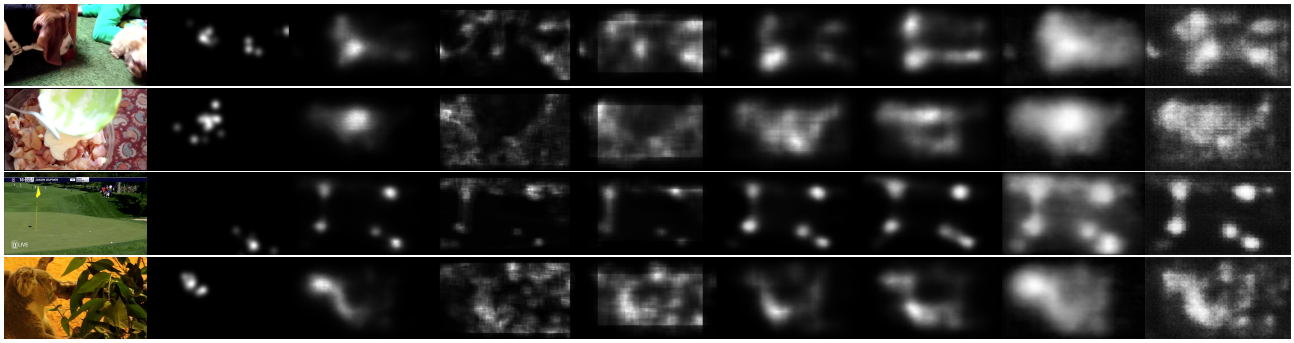


Figure 5: Examples of video saliency maps predicted by the proposed model, as well as intermediate maps by multiple decoders. Values between parentheses indicate one of the intermediate saliency maps by decoder D1.

4.2. Training procedure

Models are trained for 200 epochs using mini-batch stochastic gradient descent, with a mini-batch size of 12. The initial learning rate is 0.01, and it is reduced by a factor of 0.1 at epochs 100, 150, and 180. Input sequence length is 16 frames, spatially resized to 192×256. We carry out data augmentation by means of random horizontal flips; in our experiments, spatial resize and cropping do not lead to significant benefits. When the teacher assistant strategy is employed for channel reduction, we perform two additional knowledge distillation, each time training a new student network whose encoder contains, respectively, half and a quarter of the original number of channel at each encoder layer.

4.3. Performance comparison with state-of-the-art models

In these experiments, we report results of our model in both the MISO and MIMO configurations (respectively, **TinyHD-S** and **TinyHD-M**), trained with the auxiliary unlabeled dataset but *without* channel reduction that using the teacher assistant strategy (which introduces trade-offs between accuracy and computational costs that will be discussed later). We also report the number of multiply-accumulate operations (MAC) carried out by each method[2] to generate a 16-frame saliency sequence. Results on

[2]Values are computed from official implementations when available and from our own implementations otherwise.

DHF1K are shown in Table 1a. In the MISO configuration, our model is on par with state-of-the-art methods (and even better on NSS), but only employs a fraction of the their computational cost. In the MIMO configuration, our method sets a new state of the art, outperforming (on four metrics out of five) also UNISAL, which has a similar number of parameters but is about twice as demanding in terms of GMACs. Fig. 5 presents a few examples of saliency predictions by our model[3]. For each example, we also show the intermediate maps provided by each decoder. Qualitatively, our model predicts reasonable saliency regions, sometimes identifying additional elements not included in the ground truth (e.g., the third example). Intermediate maps also exhibit a certain variability, although similar patterns can be found in pairs (e.g., maps 1-2 and maps 3-4 from D1, and maps from D2 and D3). In general, the highest-level map from D1 (the fourth) mostly affects the output prediction: this is expected, since the corresponding architecture matches the teacher's. However, the fusion layer includes all information from intermediate maps, as shown in the last example, where two salient areas identified by the highest-level map from D1 are discarded.

Table 1b and 2 report results on Hollywood2 and UCF-Sports. While the model performs very well on the former, especially in the more efficient MIMO setting, ViNet and UNISAL achieve higher accuracy on UCF-Sports. This may be due to the lower performance of the HD2S teacher on that specific dataset, and to the arguable suitability of UCF-Sports as a video saliency prediction benchmark: the vast majority of its videos has fewer than 100 frames, and user fixations are driven by action classification, rather than free-viewing saliency [1].

4.4. Ablation studies

In order to experimentally substantiate our architectural and methodological choices, we carry out a set of ablation studies on each component of the model. the results of these experiments are reported on the DHF1K validation set, since testing set is not publicly available. First, we assess the effect of our heterogeneous multi-decoder strategy, evaluating the model's performance under several decoder configurations. We carry out this experiment in the MISO configuration, which achieves higher accuracy, as shown in Table 1a. In order to demonstrate the importance of combining different decoder architectures, Table 3 reports results when using homogeneous decoders in our architecture. Table 3 show that the heterogeneous approach generally performs better than configurations with a single decoder type, most remarkably in the NSS metric. For the sake of completeness, we also show configurations where a smaller number of homogeneous decoders are employed; these setups are, of course, more computationally efficient, but exhibit lower performance on average in the accuracy metrics.

In the second part of our ablation study, we evaluate of the impact of our knowledge distillation strategy. Table 4 reports the results obtained by the proposed model, in MISO configuration, when trained on ground-truth maps only, and when gradually adding knowledge distillation terms on DHF1K and on Kinetics-400, using HD2S as teacher. The full loss setting achieves better performance on average — as previously. This is most evident in the NSS metric.

[3]More examples are provided in the supplementary materials, as well as a visual comparison with state-of-the-art models.

Table 2: Performance comparison on UCF-Sports in both the MISO and MIMO settings.

Models	AUC-J	SIM	CC	NSS
Multi-input/single-output prediction				
ACLNet	0.897	0.406	0.510	2.567
3DSal	0.881	0.478	0.590	2.802
TASED	0.899	0.469	0.582	2.920
ViNet	**0.924**	**0.522**	**0.673**	**3.620**
HD2S	0.904	0.507	0.604	3.114
TinyHD-S	0.918	0.510	0.624	3.280
Multi-input/multi-output prediction				
SalEMA	0.906	0.431	0.544	2.638
STRA-Net	0.910	0.479	0.593	3.018
UNISAL	**0.918**	**0.523**	**0.644**	**3.381**
TinyHD-M	0.911	0.499	0.609	3.234

Table 3: Performance of our architecture with homogeneous decoders, on the DHF1K validation set. Number of parameters of models with homogeneous decoders are: D1×1 (2.55M), D2×1 (3.57M). D3×1 (2.53M); D1×2 (2.75M), D2×2 (4.78M). D3×2 (2.70M); D1×3 (2.95M), D2×3 (6.00M). D3×3 (2.88M); TinyHD-S (3.94M).

Decoder	AUC-J	AUC-B	CC	NSS	SIM	GMACs
D1×1	0.8993	0.8210	0.4881	2.8163	**0.3939**	3.55×16
D2×1	0.9040	0.8235	0.4837	2.7976	0.3820	2.88×16
D3×1	0.9034	0.8248	0.4836	2.7851	0.3794	2.45×16
D1×2	0.8998	0.8195	0.4882	2.8256	0.3928	5.45×16
D2×2	0.9046	0.8251	0.4855	2.8117	0.3806	4.11×16
D3×2	0.9046	0.8239	0.4864	2.8095	0.3819	3.24×16
D1×3	0.9013	0.8253	0.4922	2.8420	0.3924	7.35×16
D2×3	0.9049	**0.8266**	0.4847	2.8042	0.3774	5.33×16
D3×3	0.9047	0.8242	0.4845	2.7967	0.3799	4.03×16
TinyHD-S	**0.9075**	0.8244	**0.4945**	**2.8735**	0.3887	5.57×16

Table 4: Impact of loss terms on our model in the MISO configuration, starting from training on ground-truth (GT) maps only, and gradually adding knowledge distillation terms on DHF1K (target dataset or TD) and on Kinetics-400 (auxiliary dataset or AD), using HD2S as a teacher.

Loss term	AUC-J	AUC-B	CC	NSS	SIM
GT maps	0.9033	**0.8286**	0.4864	2.7680	0.3765
+ K.D. on TD	0.9058	0.8237	0.4875	2.8182	0.3846
+ K.D. on AD	**0.9075**	0.8244	**0.4945**	**2.8735**	**0.3887**

4.5. Channel reduction with teacher assistant

Finally, we investigate further reducing computational costs by means of our channel reduction strategy: multiple distillation steps are carried out, with each student progressively halving its number of encoding and decoding features, as described in Sect. 3.3.2. We also evaluate the performance of this approach when training on the original teacher (HD2S) and when using the "teacher assistant" technique, with the full-capacity student used as a teacher. Table 5 reports results, on both MISO and MIMO settings, after one and two reduction steps steps, respectively resulting in models with half (marked as $\times\frac{1}{2}$) and a quarter (marked as $\times\frac{1}{4}$) of the original number of convolutional features (marked as $\times 1$). Rows with "+TA" denote the use of the full-capacity student as teacher for knowledge distillation, rather than HD2S. As expected, channel reduction introduces a trade-off between retaining the accuracy of the original model and reducing computational costs. As multiply-accumulate operations and model parameters are significantly reduced, accuracy also decreases, most evidently in the NSS and, to a smaller extent, in the SIM metrics. It is noteworthy that configurations employing a teacher assistant outperform the counterpart using HD2S.

5. Conclusions

In this work, starting from the observation that different encoder-decoder architectures recognize specific video saliency patterns, we propose a heterogeneous multi-decoder architecture that leverages simpler versions of state-of-the-art decoding strategies to achieve high prediction accuracy at a fraction of the computational cost. We train our model in a multi-target knowledge distillation setting, where a hierarchical decoder is used as a teacher to supervise a matching internal decoder in our model and the output prediction; additionally, we employ semi-supervised learning on an unlabeled auxiliary dataset to further improve model generalization. Our model sets new state-of-the-art performance when employed in a multi-input/multi-output setting, while being significantly more efficient in terms of floating-point operations and number of parameters. We further push the limits of our model by applying a channel reduction procedure through multiple distillation steps and using the full-capacity student as a teacher, according to the "teacher assistant" paradigm. In the resulting model, the number of floating-point operations is approximately halved compared to the full-capacity version, and the number of parameters becomes as small as about 500k, taking about 2.4 MB storage space without compression.

Table 5: Performance of the proposed model when employing channel reduction and teacher assistant distillation.

(a) Number of parameters of models with reduced channels and GMACs reported on generating 16 output saliency maps.

	GMACs			#params		
Models	$\times 1$	$\times\frac{1}{2}$	$\times\frac{1}{4}$	$\times 1$	$\times\frac{1}{2}$	$\times\frac{1}{4}$
TinyHD-S	89.12	59.52	37.44	3.94M	1.37M	513.1k
TinyHD-M	7.95	6.92	4.06	3.92M	1.37M	515.3k

(b) Performance of channel reduction reported on DHF1K validation set in both the MISO and MIMO settings.

Models	AUC-J	AUC-B	CC	NSS	SIM
Multi-input/single-output prediction					
TinyHD-S$\times 1$	**0.9075**	0.8244	**0.4945**	**2.8735**	**0.3887**
TinyHD-S$\times\frac{1}{2}$	0.9038	**0.8331**	0.4754	2.7194	0.3641
+TA	0.9052	0.8330	0.4805	2.7317	0.3684
TinyHD-S$\times\frac{1}{4}$	0.9005	0.8285	0.4560	2.5830	0.3514
+TA	0.9018	0.8318	0.4667	2.6329	0.3569
Multi-input/multi-output prediction					
TinyHD-M$\times 1$	**0.9050**	0.8239	**0.4880**	**2.8178**	**0.3844**
TinyHD-M$\times\frac{1}{2}$	0.9016	0.8272	0.4687	2.6718	0.3612
+TA	0.9021	0.8307	0.4718	2.6726	0.3630
TinyHD-M$\times\frac{1}{4}$	0.8980	0.8294	0.4487	2.5257	0.3438
+TA	0.8999	**0.8333**	0.4564	2.5581	0.3478

Acknowledgments

This publication has been financially supported by: Science Foundation Ireland (SFI) under grant number SFI/12/RC/2289_P2; Regione Sicilia, Italy, *RehaStart* project (grant identifier: PO FESR 2014/2020, Azione 1.1.5, N. 08ME6201000222, CUP G79J18000610007); University of Catania, *Piano della Ricerca di Ateneo*, 2020/2022, Linea 2D; MIUR, Italy, Azione 1.2 "Mobilità dei Ricercatori" (grant identifier: Asse I, PON R&I 2014-2020, id. AIM 1889410, CUP: E64I18002520007).

References

[1] Giovanni Bellitto, Federica Proietto Salanitri, Simone Palazzo, Francesco Rundo, Daniela Giordano, and Concetto Spampinato. Hierarchical domain-adapted feature learning for video saliency prediction. *International Journal of Computer Vision*, 129(12):3216–3232, 2021.

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[3] Qinyao Chang and Shiping Zhu. Temporal-spatial feature pyramid for video saliency detection. *arXiv preprint arXiv:2105.04213*, 2021.

[4] Yasser Abdelaziz Dahou Djilali, Mohamed Sayah, Kevin McGuinness, and Noel E O'Connor. 3dsal: An efficient 3d-cnn architecture for video saliency prediction. In *VISIGRAPP (4: VISAPP)*, pages 27–36, 2020.

[5] Richard Droste, Yifan Cai, Harshita Sharma, Pierre Chatelain, Aris T Papageorghiou, and J Alison Noble. Towards capturing sonographic experience: cognition-inspired ultrasound video saliency prediction. In *Annual Conference on Medical Image Understanding and Analysis*, pages 174–186. Springer, 2019.

[6] Richard Droste, Jianbo Jiao, and J Alison Noble. Unified image and video saliency modeling. In *European Conference on Computer Vision*, pages 419–435. Springer, 2020.

[7] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. Dada: Driver attention prediction in driving accident scenarios. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[8] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[9] João Filipe Ferreira and Jorge Dias. Attentional mechanisms for socially interactive robots–a survey. *IEEE Transactions on Autonomous Mental Development*, 6(2):110–125, 2014.

[10] Kui Fu, Peipei Shi, Yafei Song, Shiming Ge, Xiangju Lu, and Jia Li. Ultrafast video attention prediction with coupled knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10802–10809, 2020.

[11] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.

[12] Hadi Hadizadeh and Ivan V Bajić. Saliency-aware video compression. *IEEE Transactions on Image Processing*, 23(1):19–33, 2013.

[13] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

[14] Feiyan Hu and Kevin McGuinness. Fastsal: a computationally efficient network for visual saliency prediction. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9054–9061. IEEE, 2021.

[15] Sohei Itahara, Takayuki Nishio, Yusuke Koda, Masahiro Morikura, and Koji Yamamoto. Distillation-based semi-supervised federated learning for communication-efficient collaborative training with non-iid private data. *arXiv preprint arXiv:2008.06180*, 2020.

[16] Samyak Jain, Pradeep Yarlagadda, Shreyank Jyoti, Shyamgopal Karthik, Ramanathan Subramanian, and Vineet Gandhi. Vinet: Pushing the limits of visual modality for audio-visual saliency prediction. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3520–3527. IEEE, 2020.

[17] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[18] Qiuxia Lai, Wenguan Wang, Hanqiu Sun, and Jianbing Shen. Video saliency prediction using spatiotemporal residual attentive networks. *IEEE Transactions on Image Processing*, 29:1113–1126, 2019.

[19] Hao Li, Fei Qi, and Guangming Shi. A novel spatio-temporal 3d convolutional encoder-decoder network for dynamic saliency prediction. *IEEE Access*, 9:36328–36341, 2021.

[20] Jia Li, Kui Fu, Shengwei Zhao, and Shiming Ge. Spatiotemporal knowledge distillation for efficient estimation of aerial video saliency. *IEEE Transactions on Image Processing*, 29:1902–1914, 2019.

[21] Xuejun Liao, Ya Xue, and Lawrence Carin. Logistic regression with an auxiliary data source. In *Proceedings of the 22nd international conference on Machine learning*, pages 505–512, 2005.

[22] Panagiotis Linardos, Eva Mohedano, Juan José Nieto, Noel E. O'Connor, Xavier Giró-i-Nieto, and Kevin McGuinness. Simple vs complex temporal recurrences for video saliency prediction. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, 2019.

[23] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2929–2936. IEEE, 2009.

[24] Stefan Mathe and Cristian Sminchisescu. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(7):1408–1424, 2014.

[25] Kyle Min and Jason J Corso. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2394–2403, 2019.

[26] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5191–5198, 2020.

[27] K. L. Bhanu Moorthy, Moneish Kumar, Ramanathan Subramanian, and Vineet Gandhi. *GAZED– Gaze-Guided Cinematic Editing of Wide-Angle Monocular Video Recordings*,

page 1–11. Association for Computing Machinery, New York, NY, USA, 2020.

[28] Anne-Flore Perrin, Lu Zhang, and Olivier Le Meur. How well current saliency prediction models perform on uavs videos? In *International Conference on Computer Analysis of Images and Patterns*, pages 311–323. Springer, 2019.

[29] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2008.

[30] Oindrila Saha and Sandeep Mishra. Recsal : Deep recursive supervision for visual saliency prediction. In *31st British Machine Vision Conference 2020, BMVC 2020, Virtual Event, UK, September 7-10, 2020*, 2020.

[31] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[32] Felix Sattler, Tim Korjakow, Roman Rischke, and Wojciech Samek. Fedaux: Leveraging unlabeled auxiliary data in federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[33] Xiao Sun, Yuxing Hu, Luming Zhang, Yanxiang Chen, Ping Li, Zhao Xie, and Zhenguang Liu. Camera-assisted video saliency prediction and its applications. *IEEE transactions on cybernetics*, 48(9):2520–2530, 2017.

[34] Yi Tang, Yuanman Li, and Wenbin Zou. Fast video salient object detection via spatiotemporal knowledge distillation. *arXiv preprint arXiv:2010.10027*, 2020.

[35] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378, 2017.

[36] Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibing Ling, and Ali Borji. Revisiting video saliency prediction in the deep learning era. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[37] Xinyi Wu, Zhenyao Wu, Jinglin Zhang, Lili Ju, and Song Wang. Salsac: A video saliency prediction model with shuffled attentions and correlation-based convlstm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12410–12417, 2020.

[38] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018.

[39] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018.

[40] Geng Zhang, Zejian Yuan, Nanning Zheng, Xingdong Sheng, and Tie Liu. Visual saliency based object tracking. In *Asian conference on computer vision*, pages 193–203. Springer, 2009.

[41] Peng Zhang, Li Su, Liang Li, BingKun Bao, Pamela Cosman, GuoRong Li, and Qingming Huang. Training efficient saliency prediction models with knowledge distillation. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 512–520, 2019.

[42] Wenbin Zou, Shengkai Zhuo, Yi Tang, Shishun Tian, Xia Li, and Chen Xu. Sta3d: Spatiotemporally attentive 3d network for video saliency prediction. *Pattern Recognition Letters*, 147:78–84, 2021.