Supplementary Material

1. Introduction

In our work, we propose two versions of a video saliency model that employs lightweight multiple heterogeneous decoders. **TinyHD-S** generates a single saliency map from given 16 input video frames, while **TinyHD-M** generates in a single pass 16 saliency maps, one for each input frame. We support open research. If the paper is accepted, we will release code and/or weights for models and training on Github, so that the experiment results can be reproduced.

In the supplementary materials, we provide the following additional information:

- Decoder architectural details.
- Additional experiments on knowledge distillation (training without ground-truth labels).
- Additional ablation studies on MIMO configuration (usage of homogeneous decoders; prediction of shorter output windows).
- Dataset sampling procedures.
- Output videos for qualitative comparison.

2. Decoder architectures

Fig. 1, 2, 3 respectively show the architecture of the employed decoders in the MIMO configuration. Fig. 5, 6, 7 respectively show the architecture of the employed decoders in the MISO configuration. Please note that the input temporal sizes of MISO is 8, 4, 2, 2 in order to match that of HD2S teacher. However the input temporal sizes of MIMO is 8, 4, 2, 1 to minimize FLOPs during encoding process.

It should be noted that D1 (inspired by hierarchical map aggregation), even in the MIMO configuration, produces a single saliency map, that is then replicated to obtain 16 intermediate maps. We denote this version of D1 as $D1^{S16}$. This procedure has been applied to compute results presented in the paper, and allows to significantly reduce computational, complexity. Indeed, the variant of D1 where 16 distinct saliency maps are computed, named $D1^{M}$ (shown in Fig. 4), introduces a large computational overhead due to temporal interpolation needed in order to double temporal at each decoding layer, and to finally generate homogeneous temporal output for each hierarchical level. For the sake of efficiency, we therefore employ $D1^{S16}$ as default version of D1.

Table 1 compares the two variants of the resulting model: prediction accuracy are similar, but $D1^{S16}$ is significantly more efficient, hence our choice of including it in the final version of our model.

3. Additional experiments on knowledge distillation

3.1. Training without ground-truth labels

In the main manuscript, results are reported assuming that ground-truth labels are always available, and that knowledge distillation loss terms are added on top of standard supervision. We hereby assess the impact of introducing ground-truth supervision on top of knowledge distillation. In this case, the baseline is training with teacher supervision only; hence, we are also able to present results computed on Kinetics-400 for comparison. Table 2 reports the results of these experiments on the TinyHD-S model. As we can observe, using ground truth label always helps to increase prediction accuracy.

4. Additional experiments on MIMO configuration

4.1. Usage of homogeneous decoders

As done in the manuscript for the TinyHD-S variant, we hereby present results when using homogenous decoders in the MIMO setting. Results in Table 3 demonstrate that combining different decoder strategy achieves a generally higher accuracy. A similar pattern is observed in the analogous experiment for TineyHD-S: in particular, metrics CC and NSS are better than other configurations, with the latter showing a significant improvement in accuracy.

4.2. Prediction of shorter output windows

Finally, we investigate whether selecting certain output saliency maps from the MIMO model can increase predition accuracy. This is based on the intuition that later frames in an input sequence can benefit from a larger temporal context than previous ones. In Table 4, our default TinyHD-M variant is shown as the baseline, where 16 frames are predicted by the model and they are all employed as predictions for the corresponding input. We compare the baseline to two variants: in the first, the model produces 16 output saliency

Table 1: Comparing $D1^{S16}$ and $D1^{M}$ designs.

Decoder	AUC-J	AUC-B	CC	NSS	SIM	GMACs	#params
TinyHD-M ($D1^{M}$)	**0.9053**	**0.8245**	0.4855	2.8068	0.3818	22.03G×1	3.92M
TinyHD-M ($D1^{S16}$)	0.9049	0.8237	**0.4877**	**2.8154**	**0.3841**	**7.95G×1**	3.92M

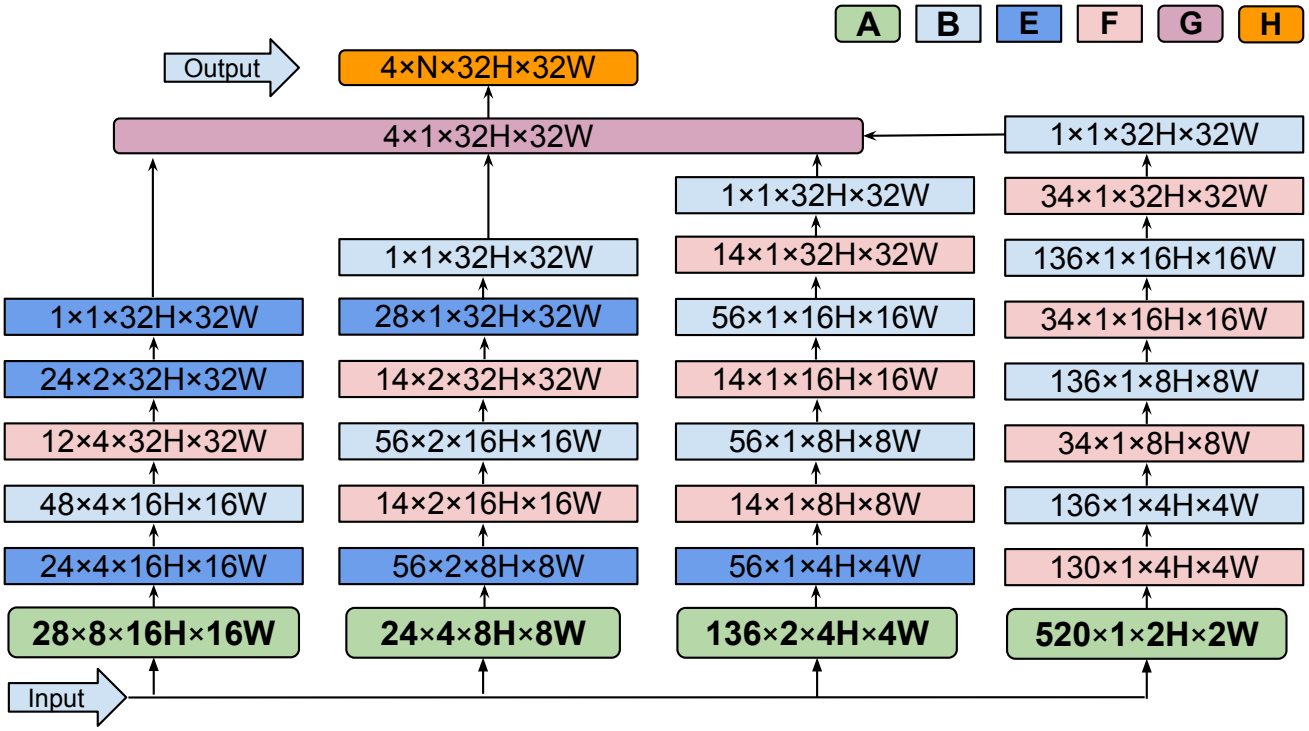

Figure 1: Decoder 1 with single saliency map output using temporal expansion ($D1^{S16}$). *Legend:* **A**: Input features; **B**: inverted residual module with expansion ratio of 2 and temporal stride of 1; **E**: same as **B**, but with temporal stride of 2; **F**: sub-pixel convolution layer or pixel shuffle layer of factor 2 for spatial dimension; **G**: concatenation on temporal dimension; **H**: copy or repeat on temporal dimension such that the output is matching number of output salieny maps. All sizes are with format $C \times T \times H \times W$. For input size 192×256, we have $H = 6$ and $W = 8$.

maps, but the first 8 are discarded; in the second variant, the model produces 8 saliency maps only. Note that all variants receive 16 frames as input.

It can be noticed that reducing the number of predicted frames improves accuracy, since the model can focus on a smaller set of predictions. However, this unavoidably leads to about doubling computational costs, since two forward passes by the model are required in order to predict 16 output saliency maps.

5. Dataset sampling

When training on DHF1K with Kinetics-400 as auxiliary dataset, at each iteration one video clip is uniformly sampled both datasets. When no auxiliary dataset is used, two video clips are uniformly samples from the dataset: one is used for ground-truth supervision, the other for teacher supervision. When only Kinetics-400 is used (Table 2 in this document), 600×2 clips are sampled from the full dataset in order to match the number of clips used in DHF1K. The same sampling strategy is applied when using Hollywood2 as the main dataset, with Kinetics-400 as auxiliary. When training our model on UCF-Sports as main dataset and Kinetics-400 as auxiliary, for every clip sampled in UCF-Sports we sample 10 clips from Kinetics-400, due to the limited size of UCF-Sports.

6. Output videos for qualitative comparison

Supplementary materials also include videos for the entire DHF1K validation set. Each video shows an input sequence with corresponding ground truth and saliency maps

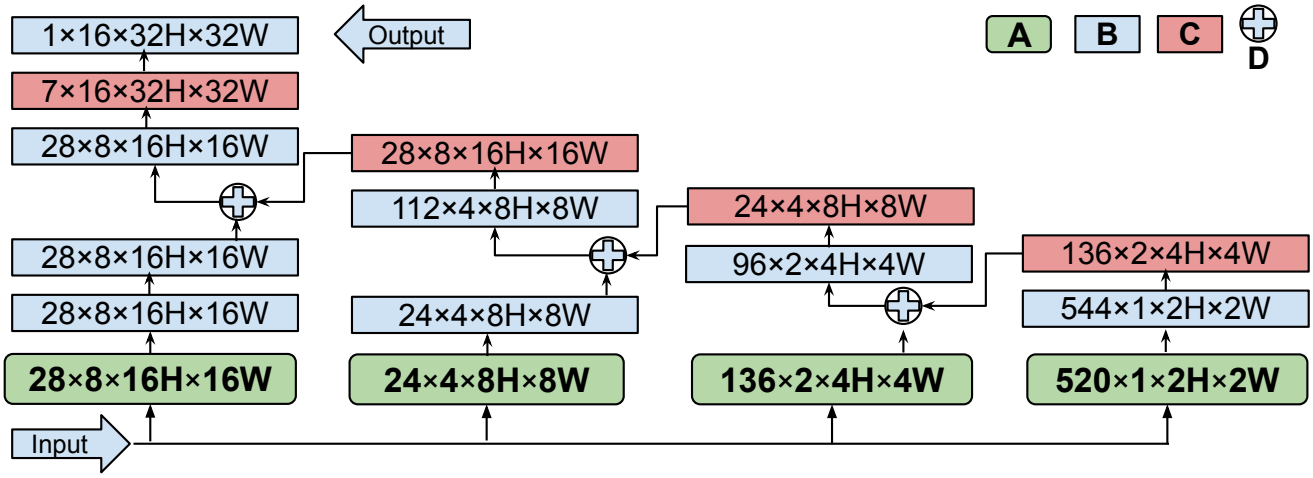

Figure 2: Decoder 2 with multiple output. *Legend:* **A**: Input features; **B**: inverted residual module with expansion ratio of 2 and temporal stride of 1; **C**: sub-pixel convolution layer or pixel shuffle layer of factor 2 for spatial dimension followed by a trilinear up-sampling on temporal dimension of factor 2; **D**: fusion by addition. All sizes are with format $C \times T \times H \times W$. For input size 192×256, we have $H = 6$ and $W = 8$.

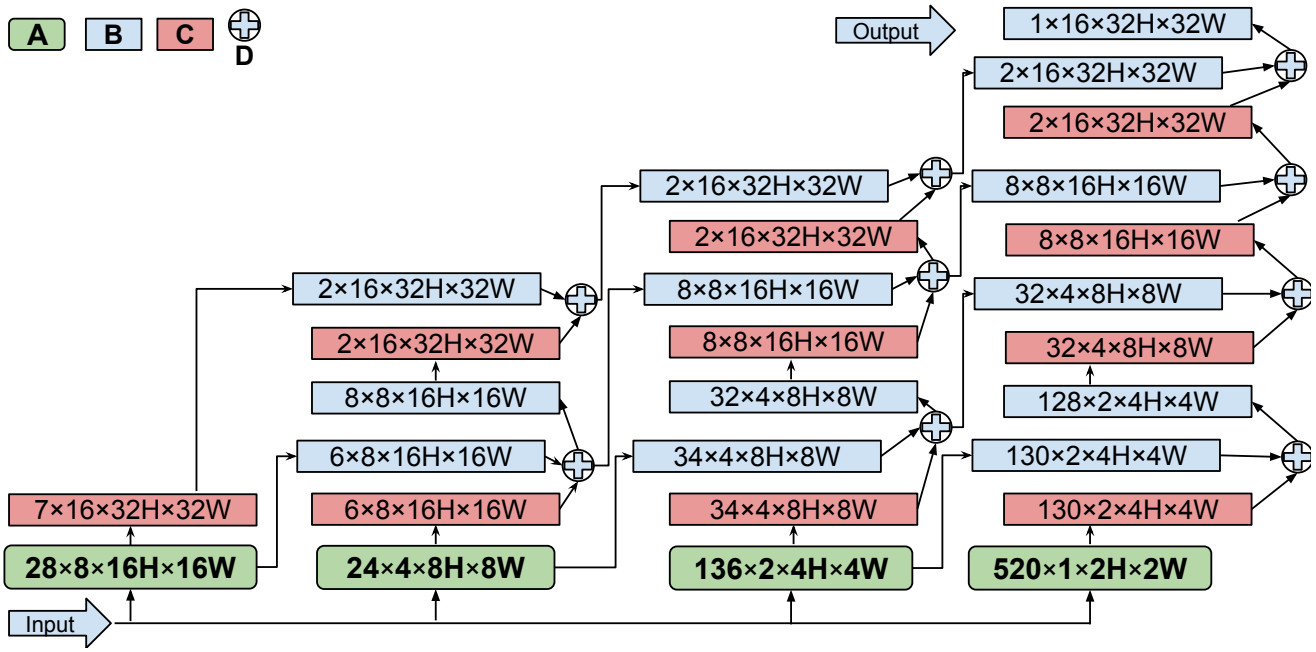

Figure 3: Decoder 3 with multiple output. *Legend:* **A**: Input features; **B**: inverted residual module with expansion ratio of 2 and temporal stride of 1; **C**: sub-pixel convolution layer or pixel shuffle layer of factor 2 for spatial dimension followed by a trilinear up-sampling on temporal dimension of factor 2; **D**: fusion by addition. All sizes are with format $C \times T \times H \times W$. For input size 192×256, we have $H = 6$ and $W = 8$.

obtained by TinyHD-S, TinyHD-M, HD2S, TASED, ViNet. Due to size limitations, videos are provided at low resolution.

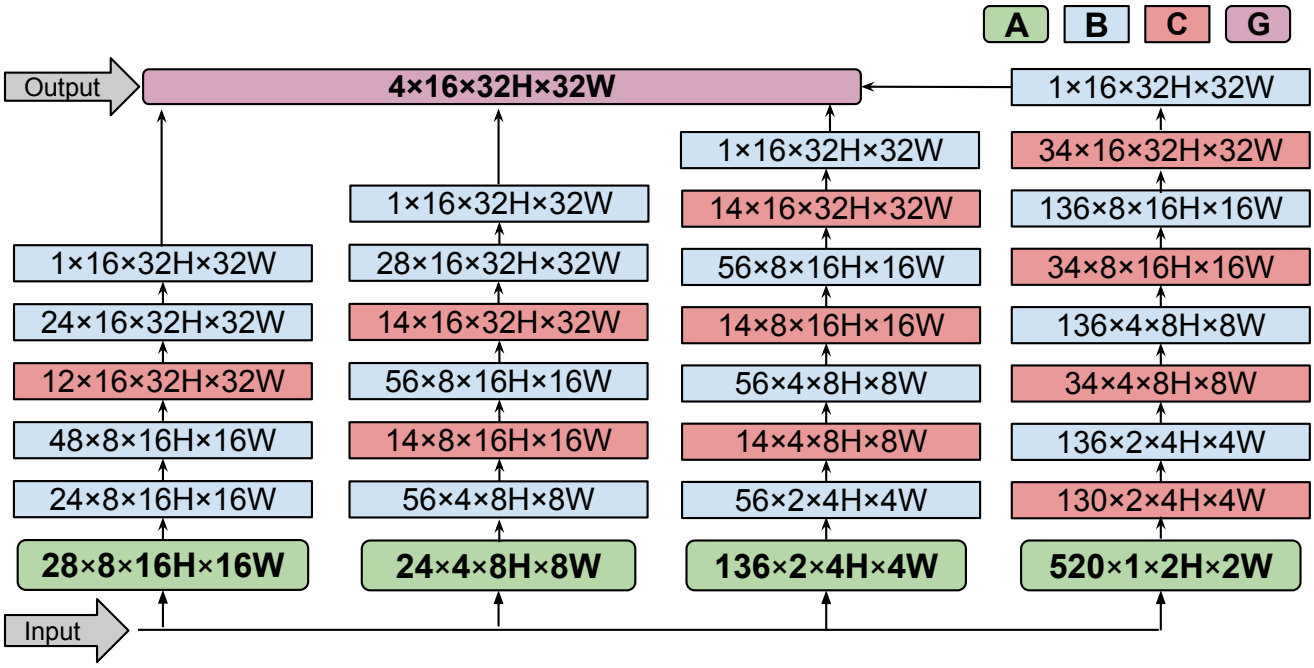

Figure 4: Decoder 1 with multiple saliency map output ($D1^M$). *Legend:* **A**: Input features; **B**: inverted residual module with expansion ratio of 2 and temporal stride of 1; **C**: sub-pixel convolution layer or pixel shuffle layer of factor 2 for spatial dimension followed by a trilinear up-sampling on temporal dimension of factor 2; **G**: concatenation on temporal dimension. For input size 192×256, we have $H = 6$ and $W = 8$.

Table 2: Results reported on the DHF1K validation set using the TinyHD-S model. Teacher is HD2S. "GT" denotes usage of ground-truth maps.

Dataset	Losses	AUC-J	AUC-B	CC	NSS	SIM
Kinetic	Baseline (teacher only)	0.8980	0.8104	0.4613	2.7082	0.3657
DHF1K	Baseline (teacher only)	0.8943	0.8018	0.4527	2.6486	0.3645
	+ GT	0.9058	0.8237	0.4875	2.8182	0.3846
	+ Aux. dataset	0.8982	0.8070	0.4651	2.7383	0.3723
	+ GT	**0.9075**	**0.8245**	**0.4946**	**2.8742**	**0.3890**

Table 3: Performance achieved when using homogeneous decoders in the MIMO configuration, on the DHF1K validation set.

Decoder	AUC-J	AUC-B	CC	NSS	SIM	GMACs	#params
D1×1	0.8982	0.8241	0.4791	2.7367	0.3849	3.66G×1	2.50M
D2×1	0.9037	**0.8294**	0.4815	2.7547	0.3763	4.00G×1	5.54M
D3×1	0.9041	0.8265	0.4814	2.7600	0.3772	3.43G×1	2.48M
D1×2	0.8985	0.8218	0.4794	2.7614	0.3835	5.75G×1	2.71M
D2×2	0.9039	0.8268	0.4825	2.7606	0.3752	6.43G×1	4.78M
D3×2	0.9029	0.8270	0.4784	2.7395	0.3717	5.29G×1	2.66M
D1×3	0.8999	0.8244	0.4853	2.8018	**0.3858**	7.84G×1	2.91M
D2×3	**0.9052**	0.8259	0.4842	2.7768	0.3773	8.86G×1	6.01M
D3×3	0.9028	0.8221	0.4769	2.7345	0.3756	7.15G×1	2.84M
TinyHD-M	0.9049	0.8237	**0.4877**	**2.8154**	0.3841	7.95G×1	3.92M

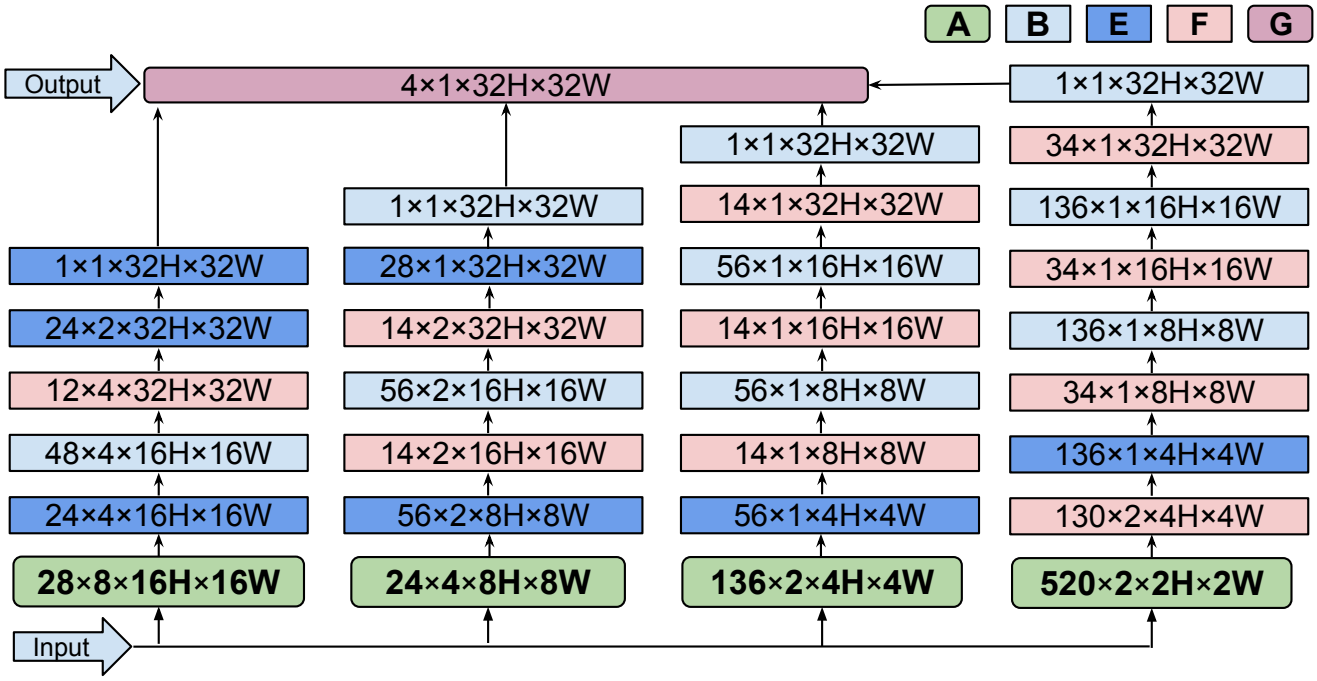

Figure 5: Decoder 1 with single output. *Legend:* **A**: Input features; **B**: inverted residual module with expansion ratio of 2 and temporal stride of 1; **E**: same as **B**, but with temporal stride of 2; **F**: sub-pixel convolution layer or pixel shuffle layer of factor 2 for spatial dimension; **G**: concatenation on temporal dimension. All sizes are with format $C \times T \times H \times W$. For input size 192×256, we have $H = 6$ and $W = 8$.

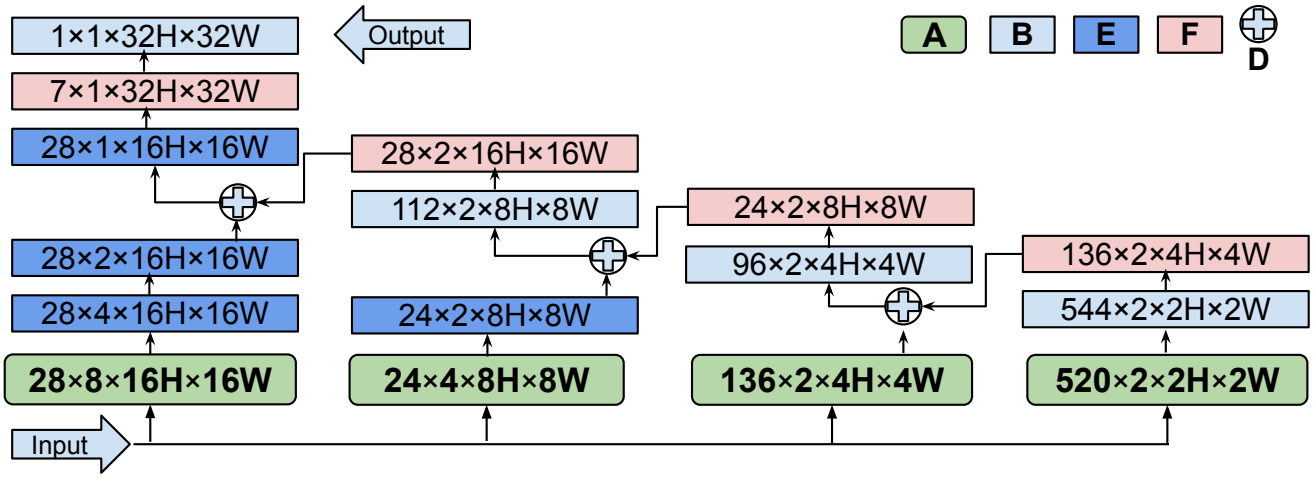

Figure 6: Decoder 2 with single output. *Legend:* **A**: Input features; **B**: inverted residual module with expansion ratio of 2 and temporal stride of 1; **E**: same as **B**, but with temporal stride of 2; **F**: sub-pixel convolution layer or pixel shuffle layer of factor 2 for spatial dimension; **D**: fusion by addition. All sizes are with format $C \times T \times H \times W$. For input size 192×256, we have $H = 6$ and $W = 8$.

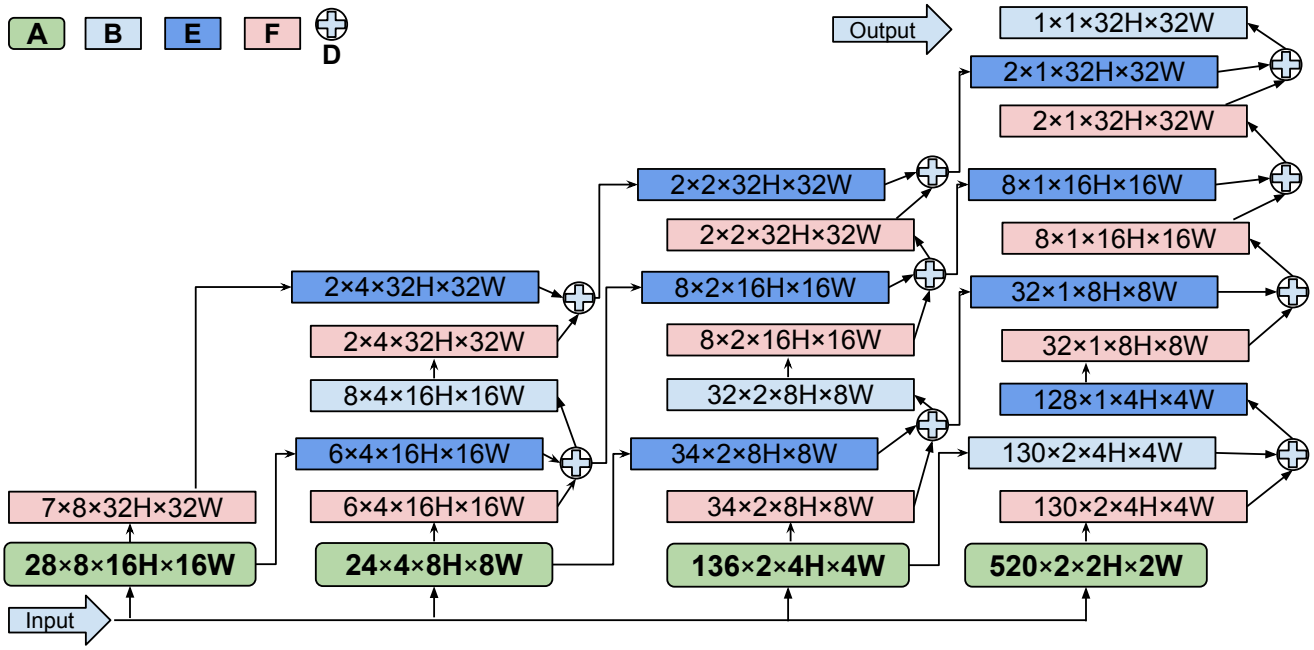

Figure 7: Decoder 3 with single output. *Legend:* **A**: Input features; **B**: inverted residual module with expansion ratio of 2 and temporal stride of 1; **E**: same as **B**, but with temporal stride of 2; **F**: sub-pixel convolution layer or pixel shuffle layer of factor 2 for spatial dimension; **D**: fusion by addition. All sizes are with format $C \times T \times H \times W$. For input size 192×256, we have $H = 6$ and $W = 8$.

Table 4: Results when producing or using a smaller number of output saliency maps, on the DHF1K validation set.

#output frames	#selected frames	AUC-J	AUC-B	CC	NSS	SIM	GMACs	#params
16	16	0.9049	0.8237	0.4877	2.8154	0.3841	**7.95G×1**	3.92M
16	8	0.9055	0.8219	0.4905	2.8442	0.3874	7.95G×2	3.92M
8	8	**0.9073**	**0.8259**	**0.4961**	**2.8686**	**0.3914**	6.91G×2	3.92M