Supplementary material
STAViS: Spatio-Temporal AudioVisual Saliency Network

Antigoni Tsiami, Petros Koutras and Petros Maragos
School of E.C.E., National Technical University of Athens, Greece
`{antsiami, pkoutras, maragos}@cs.ntua.gr`

1. Configuration and parameters of the employed CNN architectures

In Tables 1, 2 we present the configuration and the parameters of the employed visual and audio sub-networks based on the 3D-ResNet-50 [5] and SounNet [1] architectures respectively. Figure 1 depicts the architecture of a 3D Bottleneck Block that constitutes the main building unit of the spatio-temporal 3D ResNet.

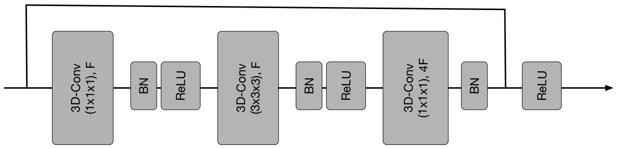

Figure 1. Architecture of the 3D Bottleneck Block. F denotes the number of feature maps of the 3D convolutional filter while BN refers to batch normalization.

2. Data splits and training

2.1. Data splits

As mentioned in the main body of the paper, training has been performed by combining data from all datasets: For DIEM, the standard split from literature has been employed [2]. For the other 5 databases, where there is no particular split, we created 3 different splits of the data, in the sense of 3-fold cross-validation, with non-overlapping videos between train, validation and test sets for each split, uniformly split among datasets. Among these 3 different splits, different videos were placed in each split, namely, if video1 of dataset1 was placed in test set of split1, then it would not appear in any other test set. We ensured that all videos of each dataset would appear once in the test set of one split. The models' final performance was obtained by taking the average among all 3 splits. The same procedure was carried out both for our audiovisual and visual variants, in order to ensure a fair comparison. In Table 3, the detailed list of the video contents of each split is depicted for completeness.

2.2. Training

In this subsection, the training process is described in more detail. First the visual-only network is trained, using as starting point the pretrained model in the Kinetics 400 database [5, 3] and employing DSAM skip connections. Afterwards, the whole audio-visual network is trained end-to-end, using the previously trained visual-only network as starting point for the visual path, while for the audio representation path we use as starting point the pretrained model from Flickr [1]. The network was trained either for 100 epochs or until loss did not further improve during 10 epochs, which usually happened around 60 epochs. Also, multi-step learning rate has been employed. The hyperparameters listed in the main body of the paper have mainly been decided upon experimentation.

3. Ablation study per database

Due to space restrictions, in the main body of the paper paper, the ablation study was presented as a table summarizing the results for all the videos contained in the employed databases. Here, in Tables 4,5 the ablation study results and the state-of-the-art evaluation results have been concatenated and presented per database for completeness. For details about the state-of-the-art methods please refer to the main paper.

We notice that except for a few cases, audiovisual combinations outperform all other visual-only methods, including our visual only variant. Sometimes, the performance is better by a large margin, as for example in Coutrot2, which is the most "audiovisual" database. In Coutrot2, all audiovisual combinations outperform the visual-only by far, indicating that the network indeed learns to fuse auditory saliency in order to predict fixations closer to the human ones.

An interesting remark concerns SumMe database, which contains the most unedited, user-made videos. Many of its videos include footages, GoPro cameras, videos with artificial sound, etc., thus, the majority does not contain many actual audiovisual events. However, audiovisual

Layer	3D-Conv1	3D Max Pool	3D Conv2 Block	3D Conv3 Block	3D Conv4 Block
Number of feature maps (F)	64	64	64	128	256
Filter Kernel	$7 \times 7 \times 7$	$3 \times 3 \times 3$	Bottleneck	Bottleneck	Bottleneck
Stride	$7 \times 7 \times 1$	$2 \times 2 \times 2$	$1 \times 1 \times 1$	$2 \times 2 \times 2$	$2 \times 2 \times 2$
Number of Blocks	-	-	3	4	6

Table 1. Configuration and parameters of the 3D ResNet architecture that has been employed in the patio-temporal network for visual saliency.

Layer	Conv1	Pool1	Conv2	Pool2	Conv3	Conv4	Conv5	Pool5	Conv6	Conv7
Number of feature channels	16	16	32	32	64	128	256	256	512	1024
Filter Kernel	64	8	32	8	16	8	4	4	4	4
Stride	2	1	2	1	2	2	2	1	2	2
Padding	32	0	16	0	8	4	2	0	2	2

Table 2. Configuration and parameters of the Audio Representation Network.

methods have performed quite well. Surprisingly though, the best performance in 3 out of 5 metrics is achieved by an audio-only method, that localizes the sound source among the other redundant information included in SumMe videos. Regarding the rest 2 metrics, the best performance is achieved by two different state-of-the-art spatial-only visual saliency methods.

4. Video demos of the proposed STAViS Network

In the supplementary material folder we have also included 4 demo videos (which correspond to the figures of the main body of the paper) that demonstrate better our approach over several video from all the employed databases.

References

[1] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In *Proc. Advances in Neural Information Processing Systems (NIPS)*, pages 892–900, 2016.

[2] A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Trans. Image Process.*, 22(1):55–69, 2013.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[4] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *IEEE Trans. Image Process.*, 27(10):5142–5154, 2018.

[5] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang. Deepvs: A deep learning based video saliency prediction approach. In *Proc. European Conf. on Computer Vision (ECCV)*. 2018.

[7] Kyle Min and Jason J. Corso. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, pages 2394–2403, 2019.

[8] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *Computer Vision and Image Understanding*, 2018.

[9] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E. O'Connor. Shallow and deep convolutional networks for saliency prediction. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 598–606, 2016.

[10] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE Trans. Image Process.*, 27(5):2368–2378, 2018.

[11] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting video saliency: A large-scale benchmark and a new model. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibin Ling, and Ali Borji. Revisiting video saliency prediction in the deep learning era. *IEEE Trans. Pattern Anal. Mach. Intell. (PAMI)*, pages 1–17, 2019.

	Test split 1	Test split 2	Test split 3
Database	video	video	video
AVAD	V32_Dancers	V22_Tap2	V40_Guitar5
AVAD	V6_Basketball1	V34_Beat	V16_Drummer2
AVAD	V33_Harp	V35_Squirrel	V30_Dog3
AVAD	V17_Soccer1	V3_Speech3	V7_Basketball2
AVAD	V41_Violin1	V11_News4	V42_Violin2
AVAD	V29_Dog2	V45_Darbuka2	V28_Dog1
AVAD	V25_Piano1	V38_Guitar3	V5_Interview2
AVAD	V31_Bird	V14_Conservation3	V21_Tap1
AVAD	V27_Piano3	V8_News1	V37_Guitar2
AVAD	V15_Drummer1	V44_Darbuka1	V26_Piano2
AVAD	V23_Tap3	V1_Speech1	V43_Violin3
AVAD	V24_Tap4	V36_Guitar1	V19_Singing1
AVAD	V13_Conservation2	V2_Speech2	V4_Interview1
AVAD	V9_News2	V18_Soccer2	V20_Singing2
AVAD	V10_News3	V39_Guitar4	V12_Conservation1
Coutrot 1	clip4	clip1	clip2
Coutrot 1	clip5	clip3	clip6
Coutrot 1	clip10	clip7	clip9
Coutrot 1	clip14	clip8	clip12
Coutrot 1	clip15	clip11	clip13
Coutrot 1	clip16	clip25	clip18
Coutrot 1	clip17	clip27	clip19
Coutrot 1	clip22	clip28	clip20
Coutrot 1	clip24	clip29	clip21
Coutrot 1	clip26	clip30	clip23
Coutrot 1	clip32	clip31	clip34
Coutrot 1	clip37	clip33	clip35
Coutrot 1	clip38	clip36	clip40
Coutrot 1	clip39	clip42	clip41
Coutrot 1	clip44	clip43	clip45
Coutrot 1	clip47	clip49	clip46
Coutrot 1	clip48	clip52	clip51
Coutrot 1	clip50	clip54	clip53
Coutrot 1	clip56	clip55	clip59
Coutrot 1	clip58	clip57	clip60
Coutrot2	clip4	clip12	clip3
Coutrot2	clip15	clip13	clip1
Coutrot2	clip11	clip10	clip9
Coutrot2	clip14	clip7	clip6
Coutrot2	clip2	clip5	clip8
DIEM	BBC_life_in_cold_blood	BBC_life_in_cold_blood	BBC_life_in_cold_blood
DIEM	BBC_wildlife_serpent	BBC_wildlife_serpent	BBC_wildlife_serpent
DIEM	DIY_SOS	DIY_SOS	DIY_SOS
DIEM	advert_bbc4_bees	advert_bbc4_bees	advert_bbc4_bees
DIEM	advert_bbc4_library	advert_bbc4_library	advert_bbc4_library
DIEM	advert_iphone	advert_iphone	advert_iphone
DIEM	ami_ib4010_closeup	ami_ib4010_closeup	ami_ib4010_closeup
DIEM	ami_ib4010_left	ami_ib4010_left	ami_ib4010_left
DIEM	harry_potter_6_trailer	harry_potter_6_trailer	harry_potter_6_trailer
DIEM	music_gummybear	music_gummybear	music_gummybear
DIEM	music_trailer_nine_inch_nails	music_trailer_nine_inch_nails	music_trailer_nine_inch_nails
DIEM	news_tony_blair_resignation	news_tony_blair_resignation	news_tony_blair_resignation
DIEM	nightlife_in_mozambique	nightlife_in_mozambique	nightlife_in_mozambique

DIEM	one_show	one_show	one_show
DIEM	pingpong_angle_shot	pingpong_angle_shot	pingpong_angle_shot
DIEM	pingpong_no_bodies	pingpong_no_bodies	pingpong_no_bodies
DIEM	sport_scramblers	sport_scramblers	sport_scramblers
DIEM	sport_wimbledon_federer_final	sport_wimbledon_federer_final	sport_wimbledon_federer_final
DIEM	tv_uni_challenge_final	tv_uni_challenge_final	tv_uni_challenge_final
DIEM	university_forum_construction_ionic	university_forum_construction_ionic	university_forum_construction_ionic
SumMe	Valparaiso_Downhill	Cooking	Base_jumping
SumMe	Car_railcrossing	Bus_in_Rock_Tunnel	Bike_Polo
SumMe	Bearpark_climbing	Uncut_Evening_Flight	Scuba
SumMe	Playing_on_water_slide	playing_ball	paluma_jump
SumMe	Fire_Domino	St_Maarten_Landing	Kids_playing_in_leaves
SumMe	Cockpit_Landing	Paintball	Eiffel_Tower
SumMe	Saving_dolphines	Air_Force_One	Statue_of_Liberty
SumMe	Notre_Dame	car_over_camera	Excavators_river_crossing
SumMe			Jumps
ETMD	CHI_1_color	CRA_1_color	DEP_1_color
ETMD	CHI_2_color	CRA_2_color	DEP_2_color
ETMD	GLA_1_color	FNE_1_color	LOR_1_color
ETMD	GLA_2_color	FNE_2_color	LOR_2_color

Table 3. Detailed list of the video contents of each one of the three test splits. Note that regarding DIEM, the same videos are contained in every split, because there is a specific train, validation and test split from the literature.

Dataset	DIEM					Coutrot1					Coutrot2				
Method	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑
Visual	0.5665	2.19	0.8792	0.6648	0.4719	0.4587	1.99	0.8617	0.5764	0.3842	0.6529	4.19	0.9405	0.6895	0.4470
L_1_AudioOnly	0.5362	2.05	0.8719	0.6596	0.4573	0.4444	1.93	0.8605	0.5789	0.3813	0.6917	4.67	0.9519	0.7013	0.4970
L_2_AudioOnly	0.5458	2.10	0.8737	0.6601	0.4569	0.4687	2.04	0.8669	0.584	0.3880	0.7223	4.99	0.9572	0.7054	0.5000
L_3_AudioOnly	0.5407	2.10	0.8719	0.6594	0.4552	0.4491	1.99	0.8618	0.5799	0.3745	0.7126	4.96	0.9568	0.7023	0.4849
$L_1\_S_1^{av}$	0.5489	2.13	0.8743	0.6582	0.4623	0.4516	2.01	0.8612	0.5803	0.3808	0.7102	5.01	0.9523	0.7023	0.4911
$L_2\_S_1^{av}$	0.5731	2.25	0.8839	0.6701	0.4847	0.4577	2.04	0.8602	0.5770	0.3892	0.7040	5.02	0.9480	0.6954	0.5017
$L_3\_S_1^{av}$	0.5712	2.25	0.8825	0.6693	0.4848	0.4623	2.07	0.8649	0.5820	0.3934	0.7076	5.08	0.9499	0.7003	0.5112
$L_1\_S_2^{av}$	0.525	2.04	0.8367	0.661	0.3519	0.4437	1.91	0.8354	0.5787	0.2994	0.6250	3.90	0.9168	0.6958	0.2785
$L_2\_S_2^{av}$	0.5259	2.04	0.8376	0.6607	0.3527	0.4448	1.92	0.8374	0.5789	0.3005	0.6257	3.90	0.9171	0.6952	0.2801
$L_3\_S_2^{av}$	0.5309	2.07	0.8344	0.6648	0.3553	0.4364	1.90	0.8218	0.5782	0.2924	0.6309	4.07	0.9161	0.6977	0.2883
$L_3^{mul}\_\tilde{S}_2^{av}$	0.5594	2.14	0.8785	0.6681	0.4694	0.4588	2.00	0.8633	0.5823	0.3872	0.6983	4.70	0.9513	0.7035	0.4762
$\mathbf{L_3^{mul}\_S_3^{av}}$ (proposed)	**0.5795**	**2.26**	0.8838	**0.6741**	0.4824	0.4722	2.11	**0.8686**	**0.5847**	0.3935	**0.7349**	**5.28**	**0.9581**	**0.7106**	0.5111
$L_3^{mul}\_S_{fus}^{av}$	**0.5795**	2.25	**0.8843**	0.6727	**0.4877**	0.4679	2.08	0.8670	0.5835	**0.3954**	0.7215	5.10	0.9551	0.7083	**0.5137**
DeepNet [9]	0.4075	1.52	0.8321	0.6227	0.3183	0.3402	1.41	0.8248	0.5597	0.2732	0.3012	1.82	0.8966	0.6000	0.2019
DVA [10]	0.4779	1.97	0.8547	0.641	0.3785	0.4306	2.07	0.8531	0.5783	0.3324	0.4634	3.45	0.9328	0.6324	0.2742
SAM [4]	0.4930	2.05	0.8592	0.6446	0.4261	0.4329	2.11	0.8571	0.5768	0.3672	0.4194	3.02	0.9320	0.6152	0.3041
SalGAN [8]	0.4868	1.89	0.8570	0.6609	0.3931	0.4161	1.85	0.8536	0.5799	0.3321	0.4398	2.96	0.9331	0.6183	0.2909
ACLNet [11, 12]	0.5229	2.02	0.8690	0.6221	0.4279	0.4253	1.92	0.8502	0.5429	0.3612	0.4485	3.16	0.9267	0.5943	0.3229
DeepVS [6]	0.4523	1.86	0.8406	0.6256	0.3923	0.3595	1.77	0.8306	0.5617	0.3174	0.4494	3.79	0.9255	0.6469	0.2590
TASED [7]	0.5579	2.16	0.8812	0.6579	0.4615	**0.4799**	**2.18**	0.8676	0.5808	0.3884	0.4375	3.17	0.9216	0.6118	0.3142

Table 4. Ablation study and state-of-the-art evaluation for databases DIEM, Coutrot1 and Coutrot2.

Dataset	AVAD					SumMe					ETMD				
Method	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑	CC ↑	NSS ↑	AUC-J ↑	sAUC ↑	SIM ↑
Visual	0.6041	3.07	0.9157	0.5900	0.4431	0.4180	1.98	0.8848	0.6477	0.3325	0.5602	2.84	0.9290	0.7278	0.4121
L_1_AudioOnly	0.5836	2.89	0.9176	0.5949	0.4413	0.4072	1.93	0.8833	0.6520	0.3328	0.5407	2.71	0.9276	0.7298	0.4086
L_2_AudioOnly	0.6107	3.07	0.9199	0.5975	0.4475	**0.4413**	2.12	**0.8908**	0.6665	**0.3442**	0.5580	2.84	0.9302	0.7343	0.4087
L_3_AudioOnly	0.6009	3.07	0.9185	0.5956	0.4351	0.4097	1.95	0.8838	0.6527	0.3218	0.5504	2.81	0.9285	0.7301	0.3995
$L_1\_S_1^{av}$	0.6035	3.13	0.9166	**0.5984**	0.4463	0.4069	1.94	0.8826	0.6463	0.3271	0.5493	2.82	0.9277	0.7264	0.4073
$L_2\_S_1^{av}$	**0.6198**	**3.26**	0.9201	0.5949	**0.4700**	0.4146	2.01	0.8872	0.6514	0.3359	0.5599	2.90	0.9303	0.7256	0.4238
$L_3\_S_1^{av}$	0.6159	3.25	0.9195	0.5934	0.4690	0.4136	1.99	0.8856	0.6520	0.3391	0.5658	**2.93**	0.9312	0.7292	0.4296
$L_1\_S_2^{av}$	0.5886	2.96	0.9013	0.5951	0.3156	0.4035	1.91	0.8618	0.6540	0.2470	0.5418	2.74	0.911	0.7346	0.2773
$L_2\_S_2^{av}$	0.5912	2.98	0.9021	0.5957	0.3172	0.4038	1.91	0.8627	0.6542	0.2476	0.5423	2.74	0.9116	0.7346	0.2784
$L_3\_S_2^{av}$	0.5860	2.96	0.8899	0.5947	0.3036	0.4110	1.97	0.8710	0.6513	0.2620	0.5530	2.86	0.9161	0.7311	0.3070
$L_3^{mul}\_\tilde{S}_2^{av}$	0.5966	2.98	0.9179	0.5914	0.4449	0.4188	1.99	0.8853	0.6566	0.3358	0.5551	2.80	0.9292	0.7333	0.4136
$\mathbf{L_3^{mul}\_S_3^{av}}$ (proposed)	0.6086	3.18	0.9196	0.5936	0.4578	0.4220	2.03	0.8883	0.6562	0.3373	**0.5690**	**2.93**	0.9316	0.7317	0.4251
$L_3^{mul}\_S_{fus}^{av}$	0.6162	3.21	**0.9216**	0.5952	0.4690	0.4183	2.01	0.8871	0.6547	0.3403	0.5669	2.91	**0.9317**	0.7318	**0.4280**
DeepNet [9]	0.3831	1.85	0.8690	0.5616	0.2564	0.3320	1.55	0.8488	0.6451	0.2274	0.3879	1.90	0.8897	0.6992	0.2253
DVA [10]	0.5247	3.00	0.8887	0.5820	0.3633	0.3983	2.14	0.8681	0.6686	0.2811	0.4965	2.72	0.9039	0.7288	0.3165
SAM [4]	0.5279	2.99	0.9025	0.5777	0.4244	0.4041	**2.21**	0.8717	0.6728	0.3272	0.5068	2.78	0.9073	0.7310	0.3790
SalGAN [8]	0.4912	2.55	0.8865	0.5799	0.3608	0.3978	1.97	0.8754	**0.6882**	0.2897	0.4765	2.46	0.9035	**0.7463**	0.3117
ACLNet [11, 12]	0.5809	3.17	0.9053	0.5600	0.4463	0.3795	1.79	0.8687	0.6092	0.2965	0.4771	2.36	0.9152	0.6752	0.3290
DeepVS [6]	0.5281	3.01	0.8968	0.5858	0.3914	0.3172	1.62	0.8422	0.6120	0.2622	0.4616	2.48	0.9041	0.6861	0.3495
TASED [7]	0.6006	3.16	0.9146	0.5898	0.4395	0.4288	2.10	0.8840	0.6570	0.3337	0.5093	2.63	0.9164	0.7117	0.3660

Table 5. Ablation study and state-of-the-art evaluation for databases AVAD, SumMe and ETMD.