TSP: Temporally-Sensitive Pretraining of Video Encoders for Localization Tasks –Supplementary Material–

Humam Alwassel Silvio Giancola Bernard Ghanem
King Abdullah University of Science and Technology (KAUST)
{humam.alwassel,silvio.giancola,bernard.ghanem}@kaust.edu.sa
http://humamalwassel.com/publication/tsp

A. Pooling Function for GVF

Table 1 compares between the performance of TSP with max-pooled *vs*. average-pooled GVF on the three target tasks: TAL, Proposals, and Dense-Captioning. TSP with max-pooled GVF offers better performance across all the tasks.

Table 1: **Effects of GVF pooling function on target tasks.** We compare features pretrained with TSP using average-pooled *vs*. max-pooled GVF. We use R(2+1)D-34 encoders and pretrain on ActivityNet. We use G-TAD, BMN, and BMT as the methods for the ActivityNet TAL, Proposals, and Dense-Captioning tasks, respectively. TSP with max-pooled GVF is better on all tasks.

Video Task	Temporal Action Localization				Action Proposal Generation				Dense Video Captioning		
Feature Pretraining	0.5	0.75	0.95	Avg.	AR@1	AR@10	AR@100	AUC	BLEU@3	BLEU@4	METEOR
TSP (avg GVF)	51.04	37.07	**9.54**	35.74	34.88	**59.32**	76.40	68.85	3.56	1.71	8.17
TSP (max GVF)	**51.26**	**37.12**	9.29	**35.81**	**34.99**	58.96	**76.63**	**69.04**	**4.16**	**2.02**	**8.75**

B. Extended Ablation Study Results

In this section, we provide further statistical analysis for our studies on TSP. Statistical analyses are of great importance when comparing the performances of different algorithms. In particular, it helps us to understand whether a given improvement is significant or it is within the noise range. For each ablation study in the main paper, we reported the *maximum* value over 5 runs, following common practice in the field. However, such practice might be misleading under certain scenarios, in particular if a proposed approach has high performance variance but has on average worse performance. To alleviate any doubts, and in an effort of transparency, we share here the *mean* and *standard deviation* performances for our proposed approach along with those of the TAC-pretrained baselines. Refer to Tables 2, 3, 4, and 5 for the extended statistical results of Study 1, 2, 3, and 4, respectively. These tables report the performance on *all* tIoUs as well. The extended results show that our improvements are consistent with those reported in the main paper, and that such improvements do *not* lie within the noise range. With such analysis, we can confidently say that TSP is *statistically* better than the TAC-pretrained baselines.

C. Extended State-of-the-Art Comparison

Tables 6, 7, 8, and 9 present extended SOTA comparison with more methods and all tIoUs for TAL on ActivityNet, TAL on THUMOS14, Dense-Captioning on ActivityNet Captions, and Proposals on ActivityNet, respectively. For the Dense-Captioning task, we report additional results for captioning ground truth proposals.

D. Extended Feature Analysis Study

Here, we extended the feature similarity study to compare with *TAC on ActivityNet*. Figure 1 provides more examples comparing TSP features with those of *TAC on Kinetics* and *TAC on ActivityNet*. Not only does TSP show better temporal sensitivity compared to *TAC on Kinetics* (as we have shown in the main paper), but it also presents a better distinguishing of background *vs*. foreground representation compared to *TAC on ActivityNet*.

Table 2: **Effects of TSP on target tasks (extended results).** Each experiment in Study 1 is repeated five times, and we report the the mean, standard deviation (std), and max values over those five runs. Each table entry is given by ***mean*** $\pm$ ***std (max)***. The row/column corresponding to the main evaluation metric for each task is highlighted in grey and the best (*mean*) performance is in bold.

(a) **TAL on ActivityNet using G-TAD with R(2+1)D-34.**

mAP@tIoU	Feature Pretraining			
	TAC on Kinetics	TAC on ActivityNet	TSP w/o GVF	TSP on ActivityNet
0.50	48.269 $\pm$ 0.241 (48.538)	49.223 $\pm$ 0.349 (49.761)	**51.389** $\pm$ **0.145 (51.445)**	51.206 $\pm$ 0.162 (51.263)
0.55	45.535 $\pm$ 0.239 (45.900)	46.588 $\pm$ 0.346 (46.919)	**48.532** $\pm$ **0.107 (48.641)**	48.472 $\pm$ 0.124 (48.551)
0.60	42.824 $\pm$ 0.284 (43.262)	43.790 $\pm$ 0.316 (44.131)	**45.683** $\pm$ **0.154 (45.872)**	45.637 $\pm$ 0.109 (45.723)
0.65	40.185 $\pm$ 0.309 (40.617)	41.111 $\pm$ 0.297 (41.510)	43.182 $\pm$ 0.112 (43.212)	**43.239** $\pm$ **0.108 (43.327)**
0.70	37.487 $\pm$ 0.285 (37.853)	38.366 $\pm$ 0.245 (38.631)	40.451 $\pm$ 0.135 (40.673)	**40.534** $\pm$ **0.169 (40.834)**
0.75	33.977 $\pm$ 0.232 (34.241)	34.780 $\pm$ 0.154 (34.865)	36.816 $\pm$ 0.069 (36.865)	**36.845** $\pm$ **0.164 (37.123)**
0.80	30.191 $\pm$ 0.158 (30.434)	30.856 $\pm$ 0.197 (30.874)	**32.756** $\pm$ **0.068 (32.865)**	32.678 $\pm$ 0.118 (32.772)
0.85	25.119 $\pm$ 0.157 (25.299)	25.820 $\pm$ 0.155 (25.849)	27.478 $\pm$ 0.092 (27.620)	**27.690** $\pm$ **0.126 (27.712)**
0.90	19.152 $\pm$ 0.164 (19.157)	19.569 $\pm$ 0.232 (19.617)	20.998 $\pm$ 0.095 (21.181)	**21.375** $\pm$ **0.152 (21.487)**
0.95	08.028 $\pm$ 0.290 (07.847)	08.274 $\pm$ 0.429 (08.647)	09.429 $\pm$ 0.241 (09.109)	**09.440** $\pm$ **0.243 (09.286)**
Average	33.077 $\pm$ 0.186 (33.315)	33.837 $\pm$ 0.189 (34.080)	35.671 $\pm$ 0.066 (35.748)	**35.712** $\pm$ **0.062 (35.808)**

(b) **Proposals on ActivityNet using BMN with R(2+1)D-34.**

Metric	Feature Pretraining			
	TAC on Kinetics	TAC on ActivityNet	TSP w/o GVF	TSP on ActivityNet
AR@1	34.002 $\pm$ 0.251 (34.185)	34.452 $\pm$ 0.152 (34.667)	34.961 $\pm$ 0.475 (34.971)	**35.011** $\pm$ **0.089 (34.991)**
AR@10	57.194 $\pm$ 0.501 (57.520)	57.772 $\pm$ 0.149 (57.892)	58.831 $\pm$ 0.640 (59.346)	**59.126** $\pm$ **0.101 (58.961)**
AR@100	75.415 $\pm$ 0.305 (75.561)	75.654 $\pm$ 0.067 (75.648)	76.212 $\pm$ 0.490 (76.469)	**76.539** $\pm$ **0.108 (76.627)**
AUC	67.637 $\pm$ 0.277 (67.912)	67.959 $\pm$ 0.087 (68.075)	68.572 $\pm$ 0.532 (68.875)	**68.906** $\pm$ **0.119 (69.035)**

(c) **Dense-Captioning on ActivityNet Captions using BMT with R(2+1)D-34.**

Feature Pretraining	Ground Truth Proposals			Learned Proposals		
	BLEU@3	BLEU@4	METEOR	BLEU@3	BLEU@4	METEOR
TAC on Kinetics	4.32	1.76	10.93	3.42	1.58	8.17
TAC on ActivityNet	4.64	1.94	10.99	3.63	1.74	8.21
TSP w/o GVF	**4.88**	**2.09**	11.29	3.75	1.83	8.42
TSP on ActivityNet	4.76	1.99	**11.31**	**4.16**	**2.02**	**8.75**

Table 3: **TSP for different video encoders (extended results).** Each experiment in Study 2 is repeated five times, and we report the the mean, standard deviation (std), and max values over those five runs. Each table entry is given by ***mean*** $\pm$ ***std (max)***.

(a) **TAL on ActivityNet using G-TAD with ResNet3D-18.**

mAP@tIoU	Feature Pretraining TAC on Kinetics	 TAC on ActivityNet	 TSP on ActivityNet
0.50	47.514 $\pm$ 0.310 (47.970)	48.351 $\pm$ 0.188 (48.708)	**49.182 $\pm$ 0.305 (49.806)**
0.55	44.629 $\pm$ 0.323 (45.207)	45.598 $\pm$ 0.170 (45.926)	**46.278 $\pm$ 0.251 (46.683)**
0.60	42.005 $\pm$ 0.377 (42.582)	42.977 $\pm$ 0.146 (43.178)	**43.724 $\pm$ 0.209 (44.069)**
0.65	39.312 $\pm$ 0.359 (39.895)	40.306 $\pm$ 0.153 (40.339)	**41.094 $\pm$ 0.203 (41.374)**
0.70	36.515 $\pm$ 0.329 (36.990)	37.587 $\pm$ 0.110 (37.627)	**38.298 $\pm$ 0.213 (38.541)**
0.75	32.878 $\pm$ 0.253 (33.206)	34.085 $\pm$ 0.129 (34.217)	**34.654 $\pm$ 0.166 (34.814)**
0.80	29.014 $\pm$ 0.207 (29.314)	30.121 $\pm$ 0.116 (30.205)	**30.697 $\pm$ 0.076 (30.702)**
0.85	24.459 $\pm$ 0.111 (24.517)	25.507 $\pm$ 0.189 (25.602)	**26.176 $\pm$ 0.075 (26.182)**
0.90	18.808 $\pm$ 0.229 (19.197)	19.402 $\pm$ 0.123 (19.427)	**20.268 $\pm$ 0.121 (20.156)**
0.95	08.306 $\pm$ 0.516 (08.955)	08.424 $\pm$ 0.270 (08.816)	**08.661 $\pm$ 0.435 (08.625)**
Average	32.344 $\pm$ 0.273 (32.783)	33.235 $\pm$ 0.089 (33.404)	**33.903 $\pm$ 0.147 (34.095)**

(b) **TAL on ActivityNet using G-TAD with R(2+1)D-18.**

mAP@tIoU	Feature Pretraining TAC on Kinetics	 TAC on ActivityNet	 TSP on ActivityNet
0.50	47.218 $\pm$ 0.297 (47.573)	48.701 $\pm$ 0.170 (49.003)	**49.883 $\pm$ 0.187 (50.069)**
0.55	44.445 $\pm$ 0.289 (44.827)	45.846 $\pm$ 0.164 (46.157)	**47.079 $\pm$ 0.203 (47.226)**
0.60	41.715 $\pm$ 0.297 (42.055)	43.275 $\pm$ 0.133 (43.475)	**44.483 $\pm$ 0.179 (44.433)**
0.65	38.992 $\pm$ 0.273 (39.367)	40.816 $\pm$ 0.141 (40.977)	**41.983 $\pm$ 0.164 (41.909)**
0.70	36.233 $\pm$ 0.270 (36.653)	38.037 $\pm$ 0.126 (38.202)	**39.245 $\pm$ 0.101 (39.243)**
0.75	32.762 $\pm$ 0.216 (33.113)	34.273 $\pm$ 0.205 (34.562)	**35.568 $\pm$ 0.091 (35.608)**
0.80	28.919 $\pm$ 0.242 (29.301)	30.609 $\pm$ 0.224 (30.743)	**31.595 $\pm$ 0.191 (31.987)**
0.85	24.481 $\pm$ 0.163 (24.745)	25.837 $\pm$ 0.157 (25.993)	**26.677 $\pm$ 0.146 (26.885)**
0.90	18.691 $\pm$ 0.170 (18.839)	19.920 $\pm$ 0.198 (20.119)	**20.464 $\pm$ 0.167 (20.773)**
0.95	08.111 $\pm$ 0.574 (08.099)	08.971 $\pm$ 0.447 (09.424)	**09.072 $\pm$ 0.434 (08.958)**
Average	32.157 $\pm$ 0.220 (32.457)	33.629 $\pm$ 0.161 (33.865)	**34.605 $\pm$ 0.101 (34.709)**

(c) **TAL on ActivityNet using G-TAD with R(2+1)D-34.**

mAP@tIoU	Feature Pretraining TAC on Kinetics	 TAC on ActivityNet	 TSP on ActivityNet
0.50	48.269 $\pm$ 0.241 (48.538)	49.223 $\pm$ 0.349 (49.761)	**51.206 $\pm$ 0.162 (51.263)**
0.55	45.535 $\pm$ 0.239 (45.900)	46.588 $\pm$ 0.346 (46.919)	**48.472 $\pm$ 0.124 (48.551)**
0.60	42.824 $\pm$ 0.284 (43.262)	43.790 $\pm$ 0.316 (44.131)	**45.637 $\pm$ 0.109 (45.723)**
0.65	40.185 $\pm$ 0.309 (40.617)	41.111 $\pm$ 0.297 (41.510)	**43.239 $\pm$ 0.108 (43.327)**
0.70	37.487 $\pm$ 0.285 (37.853)	38.366 $\pm$ 0.245 (38.631)	**40.534 $\pm$ 0.169 (40.834)**
0.75	33.977 $\pm$ 0.232 (34.241)	34.780 $\pm$ 0.154 (34.865)	**36.845 $\pm$ 0.164 (37.123)**
0.80	30.191 $\pm$ 0.158 (30.434)	30.856 $\pm$ 0.197 (30.874)	**32.678 $\pm$ 0.118 (32.772)**
0.85	25.119 $\pm$ 0.157 (25.299)	25.820 $\pm$ 0.155 (25.849)	**27.690 $\pm$ 0.126 (27.712)**
0.90	19.152 $\pm$ 0.164 (19.157)	19.569 $\pm$ 0.232 (19.617)	**21.375 $\pm$ 0.152 (21.487)**
0.95	08.028 $\pm$ 0.290 (07.847)	08.274 $\pm$ 0.429 (08.647)	**09.440 $\pm$ 0.243 (09.286)**
Average	33.077 $\pm$ 0.186 (33.315)	33.837 $\pm$ 0.189 (34.080)	**35.712 $\pm$ 0.062 (35.808)**

Table 4: **TSP with other localization algorithms (extended results).** Each experiment in Study 3 is repeated five times, and we report the the mean, standard deviation (std), and max values over those five runs. Each table entry is given by ***mean*** $\pm$ ***std (max)***.

(a) **TAL on ActivityNet using BMN with R(2+1)D-18.**

mAP@tIoU	Feature Pretraining TAC on Kinetics	TAC on ActivityNet	TSP on ActivityNet
0.50	49.798 $\pm$ 0.253 (49.951)	50.339 $\pm$ 0.270 (50.775)	**51.283** $\pm$ **0.206 (51.228)**
0.55	47.127 $\pm$ 0.253 (47.391)	47.731 $\pm$ 0.289 (48.239)	**48.665** $\pm$ **0.209 (48.712)**
0.60	44.230 $\pm$ 0.341 (44.621)	44.760 $\pm$ 0.252 (45.173)	**45.759** $\pm$ **0.176 (45.741)**
0.65	41.588 $\pm$ 0.280 (41.905)	42.027 $\pm$ 0.252 (42.471)	**43.310** $\pm$ **0.166 (43.319)**
0.70	38.727 $\pm$ 0.343 (39.078)	39.016 $\pm$ 0.232 (39.427)	**40.346** $\pm$ **0.122 (40.442)**
0.75	35.020 $\pm$ 0.351 (35.306)	35.201 $\pm$ 0.178 (35.397)	**36.577** $\pm$ **0.165 (36.782)**
0.80	31.149 $\pm$ 0.313 (31.521)	31.494 $\pm$ 0.190 (31.542)	**32.609** $\pm$ **0.186 (32.803)**
0.85	25.893 $\pm$ 0.227 (26.186)	26.315 $\pm$ 0.217 (26.394)	**27.398** $\pm$ **0.108 (27.333)**
0.90	19.568 $\pm$ 0.208 (19.980)	19.991 $\pm$ 0.276 (20.070)	**20.825** $\pm$ **0.196 (20.813)**
0.95	07.651 $\pm$ 1.041 (08.613)	**08.614** $\pm$ **0.531 (07.963)**	08.420 $\pm$ 0.557 (09.504)
Average	34.075 $\pm$ 0.304 (34.455)	34.549 $\pm$ 0.169 (34.745)	**35.519** $\pm$ **0.129 (35.668)**

Table 5: **TSP on different pretraining datasets (extended results).** Each experiment in Study 4 is repeated five times, and we report the the mean, standard deviation (std), and max values over those five runs. Each table entry is given by ***mean*** $\pm$ ***std (max)***.

(a) **TAL on THUMOS14 using P-GCN with R(2+1)D-34.**

mAP@tIoU	Feature Pretraining TAC on Kinetics	TSP on ActivityNet	TAC on THUMOS14	TSP on THUMOS14
0.1	70.978 $\pm$ 0.157 (71.215)	72.202 $\pm$ 0.115 (72.193)	71.713 $\pm$ 0.090 (71.611)	**73.889** $\pm$ **0.116 (74.023)**
0.2	68.720 $\pm$ 0.115 (68.640)	69.836 $\pm$ 0.139 (69.739)	69.416 $\pm$ 0.071 (69.362)	**72.172** $\pm$ **0.139 (72.286)**
0.3	65.876 $\pm$ 0.087 (65.867)	65.412 $\pm$ 0.178 (65.403)	66.289 $\pm$ 0.080 (66.418)	**68.840** $\pm$ **0.165 (69.057)**
0.4	60.043 $\pm$ 0.114 (60.048)	59.844 $\pm$ 0.165 (59.979)	60.228 $\pm$ 0.165 (60.302)	**63.290** $\pm$ **0.175 (63.314)**
0.5	48.763 $\pm$ 0.348 (49.007)	50.331 $\pm$ 0.394 (51.038)	49.879 $\pm$ 0.421 (50.028)	**52.901** $\pm$ **0.337 (53.545)**
0.6	36.313 $\pm$ 0.519 (37.048)	36.339 $\pm$ 0.298 (36.732)	36.744 $\pm$ 0.309 (36.559)	**40.092** $\pm$ **0.295 (40.445)**
0.7	22.380 $\pm$ 0.386 (22.892)	22.191 $\pm$ 0.255 (22.221)	22.769 $\pm$ 0.302 (23.327)	**25.691** $\pm$ **0.210 (26.009)**
0.8	09.325 $\pm$ 0.199 (09.126)	09.270 $\pm$ 0.203 (09.285)	09.508 $\pm$ 0.157 (09.713)	**10.619** $\pm$ **0.229 (10.469)**
0.9	01.413 $\pm$ 0.071 (01.409)	01.399 $\pm$ 0.067 (01.393)	01.435 $\pm$ 0.103 (01.484)	**01.615** $\pm$ **0.071 (01.674)**

(b) **TAL on THUMOS14 using G-TAD with R(2+1)D-34.**

mAP@tIoU	Feature Pretraining TAC on Kinetics	TSP on ActivityNet	TAC on THUMOS14	TSP on THUMOS14
0.1	58.311 $\pm$ 0.553 (58.934)	60.747 $\pm$ 1.114 (62.106)	59.747 $\pm$ 0.655 (60.546)	**67.605** $\pm$ **1.096 (68.498)**
0.2	54.909 $\pm$ 0.383 (55.446)	57.173 $\pm$ 1.144 (58.991)	56.756 $\pm$ 0.662 (57.738)	**64.542** $\pm$ **1.106 (65.279)**
0.3	49.728 $\pm$ 0.685 (50.590)	51.622 $\pm$ 1.136 (53.449)	51.202 $\pm$ 0.717 (52.608)	**58.205** $\pm$ **1.236 (59.628)**
0.4	42.405 $\pm$ 0.591 (43.232)	43.945 $\pm$ 1.163 (45.924)	43.999 $\pm$ 0.708 (45.538)	**50.853** $\pm$ **1.243 (51.987)**
0.5	33.255 $\pm$ 0.760 (34.521)	35.089 $\pm$ 1.066 (37.034)	34.797 $\pm$ 0.558 (35.823)	**41.500** $\pm$ **1.118 (43.232)**
0.6	23.618 $\pm$ 0.615 (24.080)	24.865 $\pm$ 1.027 (26.734)	25.024 $\pm$ 0.747 (26.194)	**30.196** $\pm$ **1.422 (32.201)**
0.7	14.467 $\pm$ 0.918 (15.467)	14.771 $\pm$ 0.789 (16.128)	15.536 $\pm$ 0.557 (15.565)	**18.446** $\pm$ **1.318 (21.052)**
0.8	06.763 $\pm$ 0.694 (07.254)	06.479 $\pm$ 0.498 (07.403)	07.264 $\pm$ 0.401 (07.231)	**08.836** $\pm$ **0.873 (10.592)**
0.9	01.224 $\pm$ 0.157 (01.313)	01.086 $\pm$ 0.155 (01.351)	01.271 $\pm$ 0.113 (01.355)	**01.491** $\pm$ **0.133 (01.721)**

Table 6: **SOTA comparison for TAL on ActivityNet (extended results).** We use G-TAD as the algorithms atop our features. TSP achieves SOTA performance.

Method	0.5	0.75	0.95	Avg.
R-C3D [32]	26.80	–	–	–
TAL-Net [6]	38.23	18.30	1.30	20.22
SCC [5]	40.00	17.90	4.70	21.70
TCN [7]	37.49	23.47	4.47	23.58
CDC [28]	45.30	26.00	0.20	23.80
BSN [21]	46.45	29.96	8.02	30.03
Zhao *et al*. [35]	43.47	33.91	9.21	30.12
C-TCN [17]	47.60	31.90	6.20	31.10
P-GCN [34]	48.26	33.16	3.27	31.11
BMN [20]	50.07	34.78	8.29	33.85
GTAN [24]	52.61	34.14	8.91	34.31
PBRNet [22]	**53.96**	34.97	8.98	35.01
G-TAD [33]	50.36	34.60	9.02	34.09
TSP (ours)	51.26	**37.12**	**9.29**	**35.81**

Table 7: **SOTA comparison for TAL on THUMOS14 (extended results).** We use P-GCN as the algorithms atop our features. TSP achieves SOTA performance.

Method	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Hou *et al*. [13]	51.3	–	43.7	–	22.0	–	–	–	–
SST [4]	–	–	37.8	–	23.0	–	–	–	–
CDC [28]	–	–	40.1	29.4	23.3	13.1	7.9	–	–
TCN [7]	–	–	–	33.3	25.6	15.9	9.0	–	–
TURN-TAP [10]	54.0	50.9	44.1	34.9	25.6	–	–	–	–
R-C3D [32]	54.5	51.5	44.8	35.6	28.9	–	–	–	–
SS-TAD [29]	–	–	45.7	–	29.2	–	9.6	–	–
SSN [36]	66.0	59.4	51.9	41.0	29.8	–	–	–	–
CTAP [8]	–	–	–	–	29.9	–	–	–	–
Action Search [1]	–	–	51.8	42.4	30.8	20.2	11.1	–	–
CBR [11]	60.1	56.7	50.1	41.3	31.0	19.1	9.9	–	–
ETP [26]	–	–	48.2	42.4	34.2	23.4	13.9	–	–
BSN [21]	–	–	53.5	45.0	36.9	28.4	20.0	–	–
MGG[23]	–	–	53.9	46.8	37.4	29.5	21.3	–	–
GTAN [24]	–	–	57.8	47.2	38.8	–	–	–	–
BMN [20]	–	–	56.0	47.4	38.8	29.7	20.5	–	–
DBG [19]	–	–	57.8	49.4	39.8	30.2	21.7	–	–
CMS-RC3D [3]	61.6	59.3	54.7	48.2	40.0	–	–	–	–
G-TAD [33]	–	–	54.5	47.6	40.2	30.8	23.4	–	–
TAL-Net [6]	59.8	57.1	53.2	48.5	42.8	33.8	20.8	–	–
Zhao *et al*. [35]	–	–	53.9	50.7	45.4	38.0	28.5	–	–
PBRNet [22]	–	–	58.5	54.6	51.3	41.8	**29.5**	–	–
C-TCN [17]	72.2	71.4	68.0	62.3	52.1	–	–	–	–
TSA-Net [12]	–	–	65.6	61.4	53.0	**42.4**	28.8	–	–
P-GCN [34]	69.5	67.8	63.6	57.8	49.1	–	–	–	–
TSP (ours)	**74.0**	**72.3**	**69.1**	**63.3**	**53.5**	40.4	26.0	**10.5**	**1.7**

Table 8: **SOTA comparison for Dense-Captioning on ActivityNet Captions (extended results).** We use BMT as the algorithms atop our features. TSP achieves SOTA performance in terms of average BLEU and is competitive in terms of average METEOR. The best numbers are highlighted in bold and the second best is underlined.

Method	Ground Truth Proposals			Learned Proposals		
	BLEU@3	BLEU@4	METEOR	BLEU@3	BLEU@4	METEOR
Rahman *et al*. [27]	3.04	1.46	7.23	1.85	0.90	4.93
Krishna *et al*. [16]	4.09	1.60	8.88	1.90	0.71	5.69
Bi-SST [30]	–	–	10.89	2.27	1.13	6.10
Masked Transformer [37]	**5.76**	**2.71**	11.16	2.91	1.44	6.91
DVC [18]	4.55	1.62	10.33	2.27	0.73	6.93
MFT [31]	–	–	–	2.82	1.24	7.08
MDVC [15]	4.52	1.98	11.07	2.53	1.01	7.46
SDVC [25]	4.41	1.28	**13.07**	2.94	0.93	**8.82**
BMT [14]	4.63	<u>1.99</u>	10.90	<u>3.84</u>	<u>1.88</u>	8.44
TSP (ours)	<u>4.76</u>	<u>1.99</u>	<u>11.31</u>	**4.16**	**2.02**	<u>8.75</u>

Table 9: **SOTA comparison for Proposals on ActivityNet (extended results)**. We use BMN atop our features. TSP significantly improves over BMN original performance, and is competitive with SOTA.

Method	[8]	[21]	[23]	[35]	[2]	[19]	[9]	BMN [20]	**TSP**
AR@100	73.17	74.16	74.54	75.27	76.73	76.65	**78.63**	75.01	76.63
AUC	65.72	66.17	66.43	66.51	68.05	68.23	**69.93**	67.10	69.04

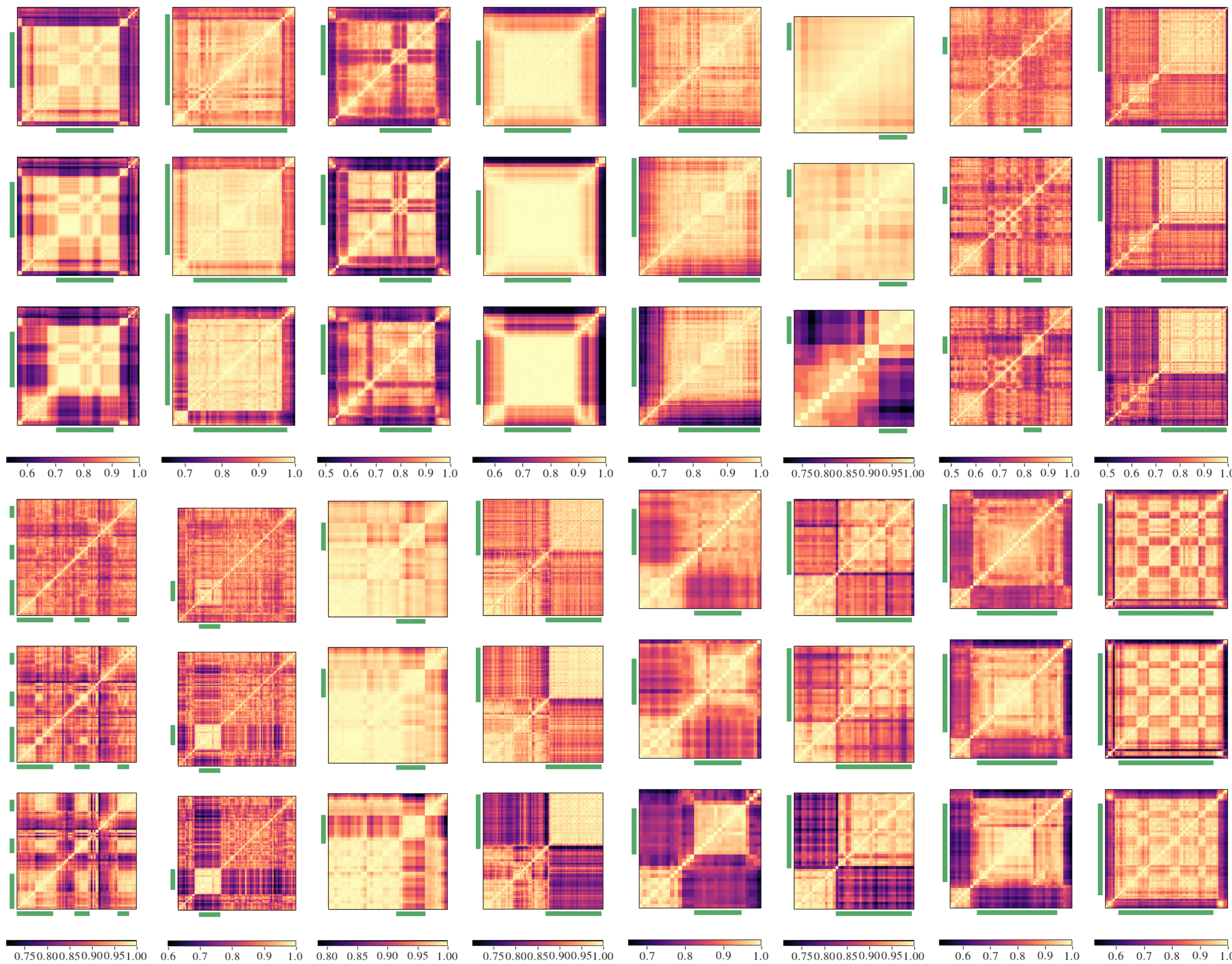

Figure 1: **Feature similarity (extended results)**. Each column (set of three matrices) shows the similarity matrices of one video using *TAC on Kinetics* (top), *TAC on ActivityNet* (middle), and TSP on ActivityNet (bottom) features. The green lines next to each matrix represent the temporal extent of ground truth actions. Better viewed in color.

References

[1] Humam Alwassel, Fabian Caba Heilbron, and Bernard Ghanem. Action search: Spotting targets in videos and its application to temporal action localization. In *ECCV*, 2018. 5

[2] Yueran Bai, Yingying Wang, Yunhai Tong, Yang Yang, Qiyue Liu, and Junhui Liu. Boundary content graph neural network for temporal action proposal generation. In *ECCV*, 2020. 6

[3] Yancheng Bai, Huijuan Xu, Kate Saenko, and Bernard Ghanem. Contextual multi-scale region convolutional 3d network for activity detection. *arXiv preprint arXiv:1801.09184*, 2018. 5

[4] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. Sst: Single-stream temporal action proposals. In *CVPR*, 2017. 5

[5] Fabian Caba Heilbron, Wayner Barrios, Victor Escorcia, and Bernard Ghanem. Scc: Semantic context cascade for efficient action detection. In *CVPR*, 2017. 5

[6] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *CVPR*, 2018. 5

[7] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S Davis, and Yan Qiu Chen. Temporal context network for activity localization in videos. In *ICCV*, 2017. 5

[8] Jiyang Gao, Kan Chen, and Ram Nevatia. Ctap: Complementary temporal action proposal generation. In *ECCV*, 2018. 5, 6

[9] Jialin Gao, Zhixiang Shi, Guanshuo Wang, Jiani Li, Yufeng Yuan, Shiming Ge, and Xi Zhou. Accurate temporal action proposal generation with relation-aware pyramid network. In *AAAI*, 2020. 6

[10] Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, and Ram Nevatia. Turn tap: Temporal unit regression network for temporal action proposals. In *ICCV*, 2017. 5

[11] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Cascaded boundary regression for temporal action detection. In *BMVC*, 2017. 5

[12] Guoqiang Gong, Liangfeng Zheng, and Yadong Mu. Scale matters: Temporal scale aggregation network for precise action localization in untrimmed videos. In *ICME*, 2020. 5

[13] Rui Hou, Rahul Sukthankar, and Mubarak Shah. Real-time temporal action localization in untrimmed videos by sub-action discovery. In *BMVC*, 2017. 5

[14] Vladimir Iashin and Esa Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In *BMVC*, 2020. 6

[15] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *CVPRW*, 2020. 6

[16] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 6

[17] Xin Li, Tianwei Lin, Xiao Liu, Wangmeng Zuo, Chao Li, Xiang Long, Dongliang He, Fu Li, Shilei Wen, and Chuang Gan. Deep concept-wise temporal convolutional networks for action localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. 5

[18] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *CVPR*, 2018. 6

[19] Chuming Lin, Jian Li, Yabiao Wang, Ying Tai, Donghao Luo, Zhipeng Cui, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. Fast learning of temporal action proposal via dense boundary generator. In *AAAI*, 2020. 5, 6

[20] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *ICCV*, 2019. 5, 6

[21] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *ECCV*, 2018. 5, 6

[22] Qinying Liu and Zilei Wang. Progressive boundary refinement network for temporal action detection. In *AAAI*, 2020. 5

[23] Yuan Liu, Lin Ma, Yifeng Zhang, Wei Liu, and Shih-Fu Chang. Multi-granularity generator for temporal action proposal. In *CVPR*, 2019. 5, 6

[24] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In *CVPR*, 2019. 5

[25] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning. In *CVPR*, 2019. 6

[26] Haonan Qiu, Yingbin Zheng, Hao Ye, Yao Lu, Feng Wang, and Liang He. Precise temporal action localization by evolving temporal proposals. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 2018. 5

[27] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *ICCV*, 2019. 6

[28] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *CVPR*, 2017. 5

[29] Bernard Ghanem Shyamal Buch, Victor Escorcia and Juan Carlos Niebles. End-to-end, single-stream temporal action detection in untrimmed videos. In *BMVC*, 2017. 5

[30] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *CVPR*, 2018. 6

[31] Yilei Xiong, Bo Dai, and Dahua Lin. Move forward and tell: A progressive generator of video descriptions. In *ECCV*, 2018. 6

[32] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *ICCV*, 2017. 5

[33] Mengmeng Xu, Chen Zhao, David S. Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *CVPR*, 2020. 5

[34] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *ICCV*, 2019. 5

[35] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In *ECCV*, 2020. 5, 6

[36] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, 2017. 5

[37] Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *CVPR*, 2018. 6