Video Summarization with a Dual Attention Capsule Network

Hao Fu[1], Hongxing Wang[1], Jianyu Yang[2]

[1]Chongqing University, China
[2]Soochow University, China

Why Need Video Summarization?

Properties of a Good Summary

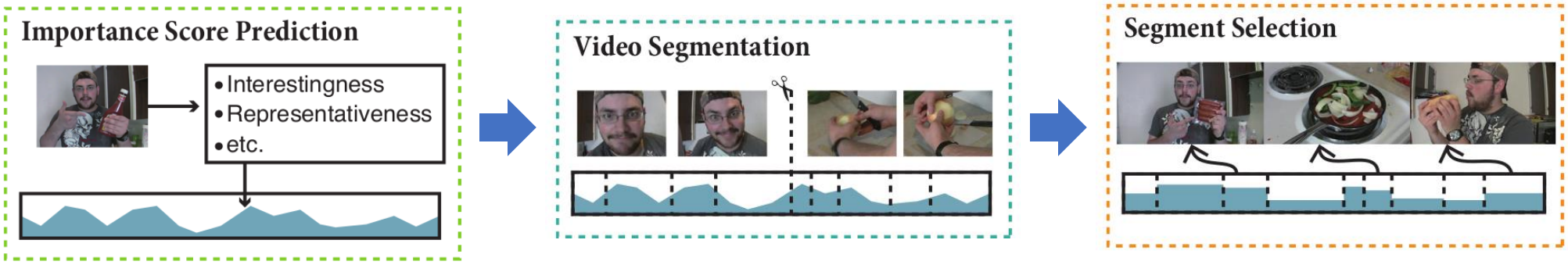
- Representative
- Diverse
- Coherent/has a sense of story


Case 2


Case 1


Case 3

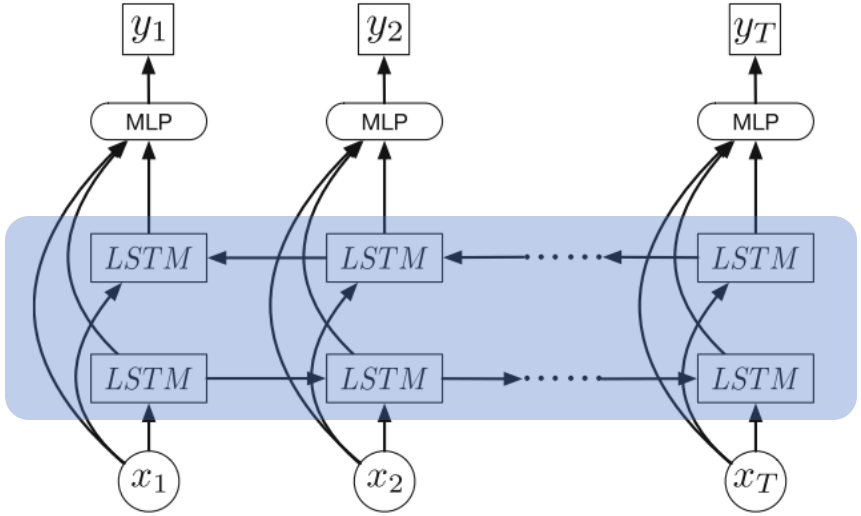
Common Video Summarization Pipeline



Modified From: Otani et al., CVPR 2019

- Gygli et al., Creating summaries from user videos, ECCV 2014.
- Song et al., TVSum: Summarizing web videos using titles, CVPR 2015
- …

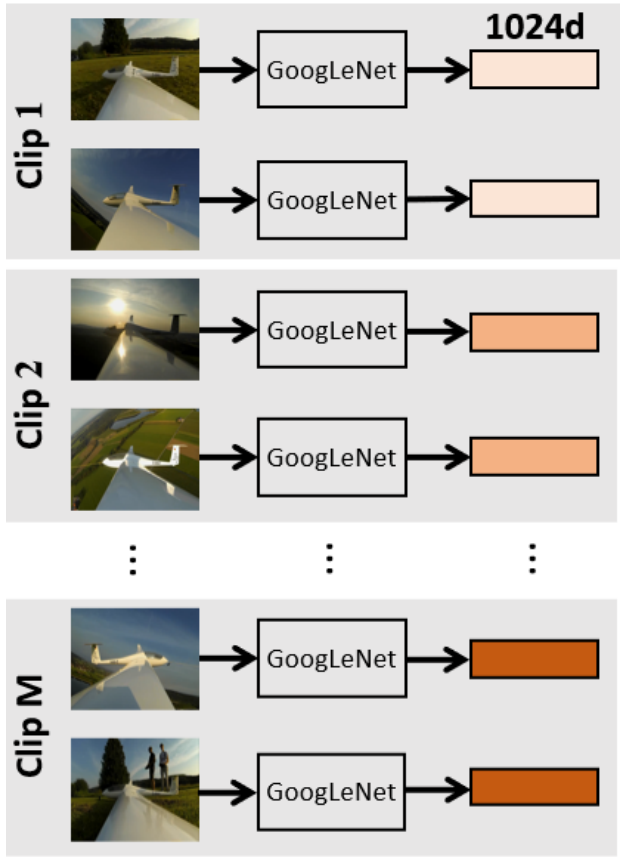
Popular Solution and Existing Challenges



Modified From: Zhang et al., ECCV 2016

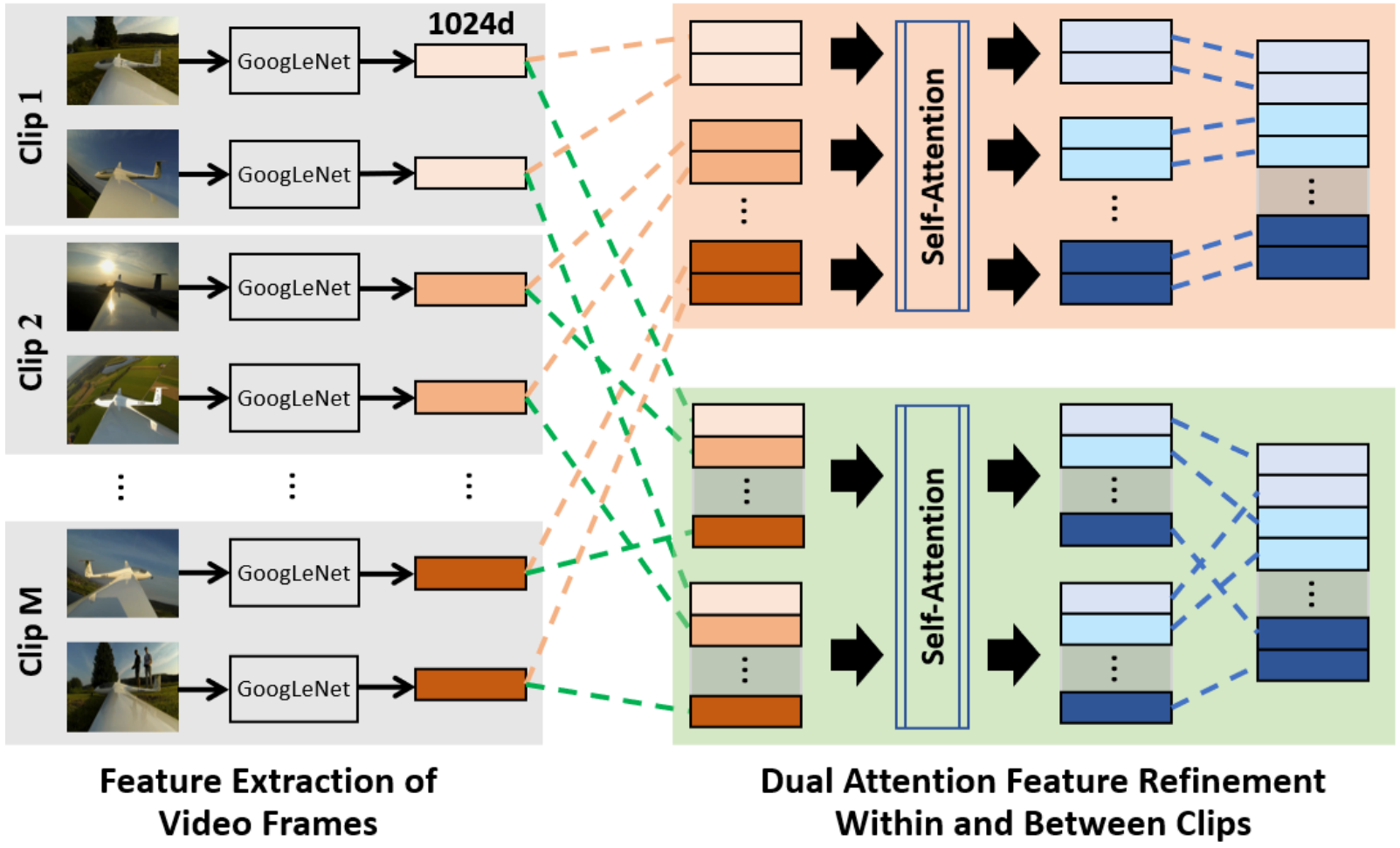
- Unfriendliness for parallel processing
- Considerable computational burdens
- Not capable enough to capture long-term dependencies

Proposed Method



Feature Extraction of Video Frames

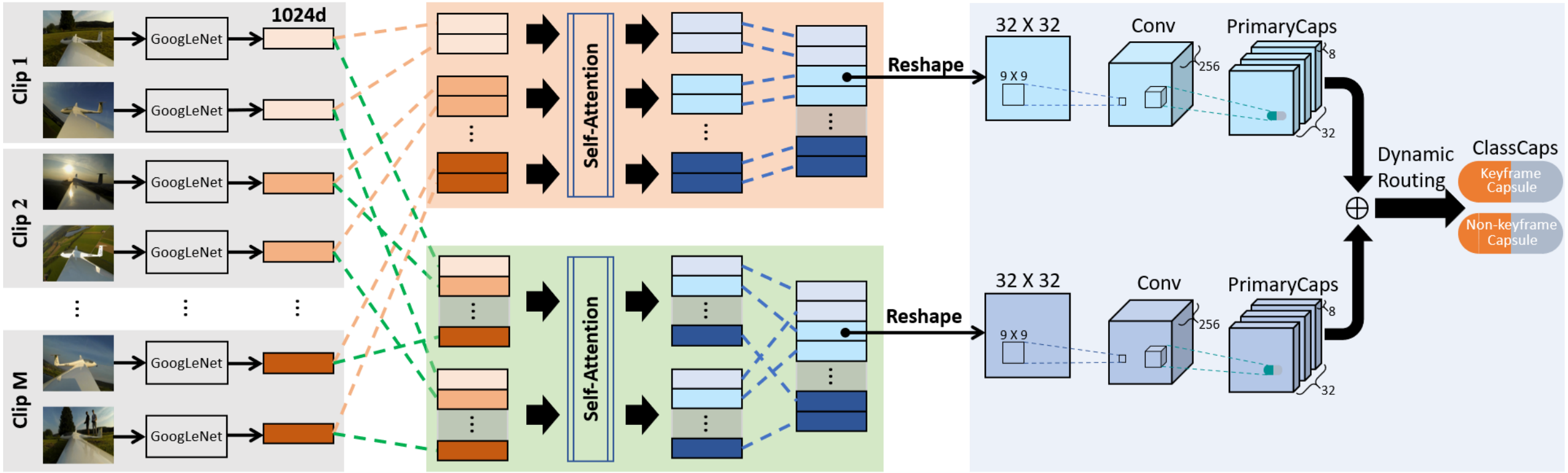
Proposed Method



Local self-attention network

Global self-attention network

Proposed Method



Feature Extraction of Video Frames

Dual Attention Feature Refinement Within and Between Clips

Two-Stream Capsule Network for Feature Fusion and Frame Labeling

Experiments

Datasets

- SumMe: 25 videos with various events such as travel, cooking, and sports
- TVSum: 50 videos with various genres such as news, how-to, and egocentric

Evaluation metrics

- F-score
- Kendall's τ
- Spearman's ρ

Comparison with SOTA (F-score)

Method	SumMe			TVSum		
-	C	A	T	C	A	T
dppLSTM [5]	38.6	41.6	40.7	54.2	57.9	56.9
SUM-GAN$_{sup}$ [6]	41.7	43.6	-	56.3	**61.2**	-
DR-DSN$_{sup}$ [7]	42.1	43.9	42.6	58.1	59.8	58.9
SASUP [35]	45.3	-	-	58.2	-	-
CSNet$_{sup}$ [21]	**48.6**	48.7	44.1	58.5	57.1	57.4
Ours	47.5	**49.3**	**45.2**	**59.4**	59.8	**59.2**

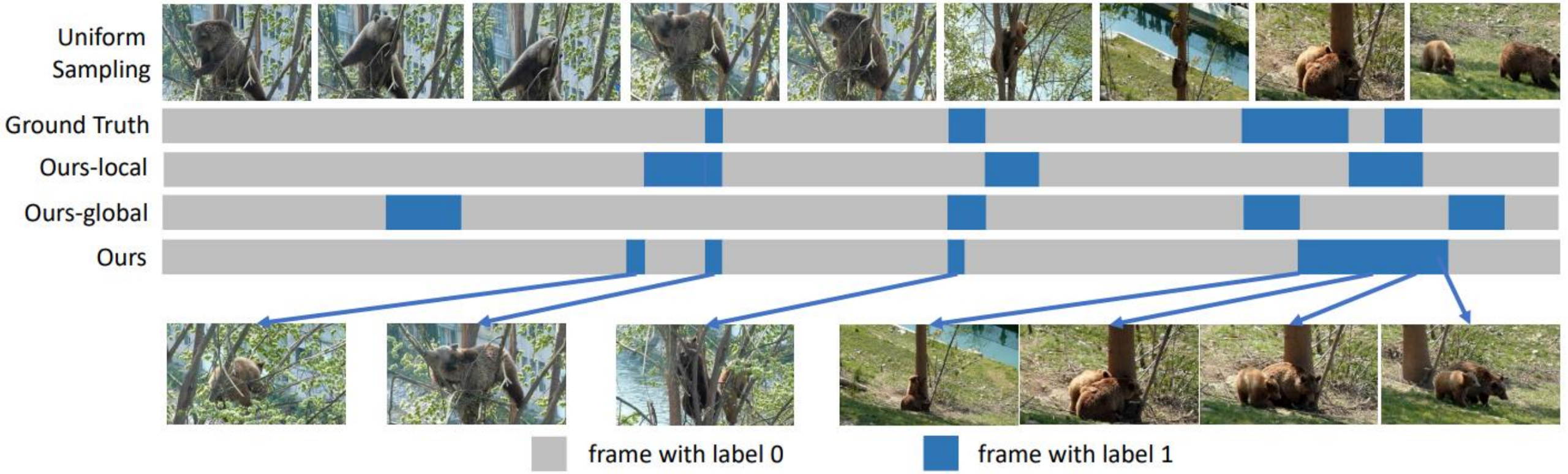
Comparison with SOTA (τ and ρ)

Dataset	SumMe		TVSum	
Metric	Kendall's τ	Spearman's ρ	Kendall's τ	Spearman's ρ
Random	0.000	0.000	0.000	0.000
DR-DSN$_{sup}$ [7]	0.034	0.041	0.025	0.039
dppLSTM [5]	0.040	0.049	0.042	0.055
Ours	**0.063**	**0.059**	**0.058**	**0.065**

Results under Unbalanced Train-Test Duration

Dataset	SumMe		TVSum	
Metric	F-score	Kendall's τ	F-score	Kendall's τ
dppLSTM [5]	36.6(-2.0)	0.036(-0.004)	51.5(-3.2)	0.032(-0.010)
DR-DSN$_{sup}$ [7]	39.7(-2.4)	0.029(-0.005)	56.2(-1.9)	0.019(-0.006)
Ours	**46.1(-1.4)**	**0.076(+0.013)**	**59.0(-0.4)**	**0.056(-0.002)**

Qualitative Results



Ablation Study

Method	SumMe	TVSum
Ours-local	45.2	58.3
Ours-global	45.4	58.5
Ours-fc	46.6	58.7
Ours	**47.5**	**59.4**

Conclusions

- We propose a novel dual attention capsule network model, which can effectively incorporate the short- and long-term temporal dependencies among video frames for summarization.
- Our proposed video summarization is parallelizable, which can easily handle longer-term dependencies among video frames than the RNN/LSTM-based approaches.
- Experimental results show that our proposed method owns stronger learning ability, and is competitive with existing state-of-the-art methods.

Thanks

20151757@cqu.edu.cn