SEOUL NATIONAL UNIV.
VISION & LEARNING

AAAI-18:
Thirty-Second AAAI Conference on Artificial Intelligence
February 2-7, 2018, Hilton New Orleans Riverside, New Orleans, Louisiana, USA

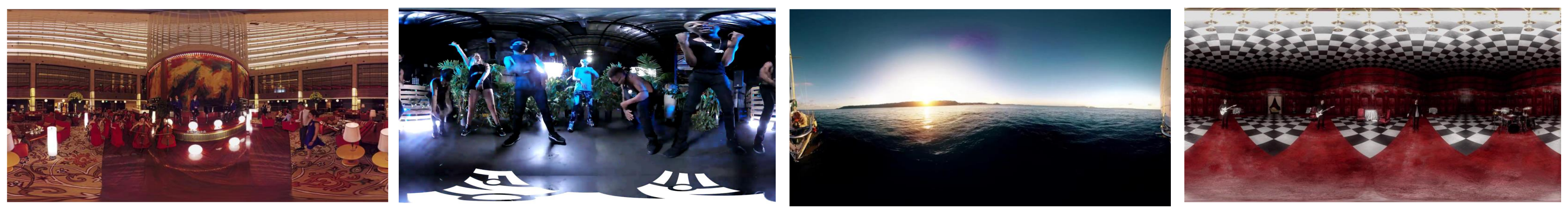
A Deep Ranking Model for Spatio-Temporal Highlight Detection from a 360º Video

Youngjae Yu Sangho Lee Joonil Na Jaeyun Kang Gunhee Kim

Seoul National University, Seoul, Korea

Motivation

360º video provides panoramic view of the entire scene.



☹ However, viewer experience can be severely handicapped due to the limited human's field-of-view.

Objective

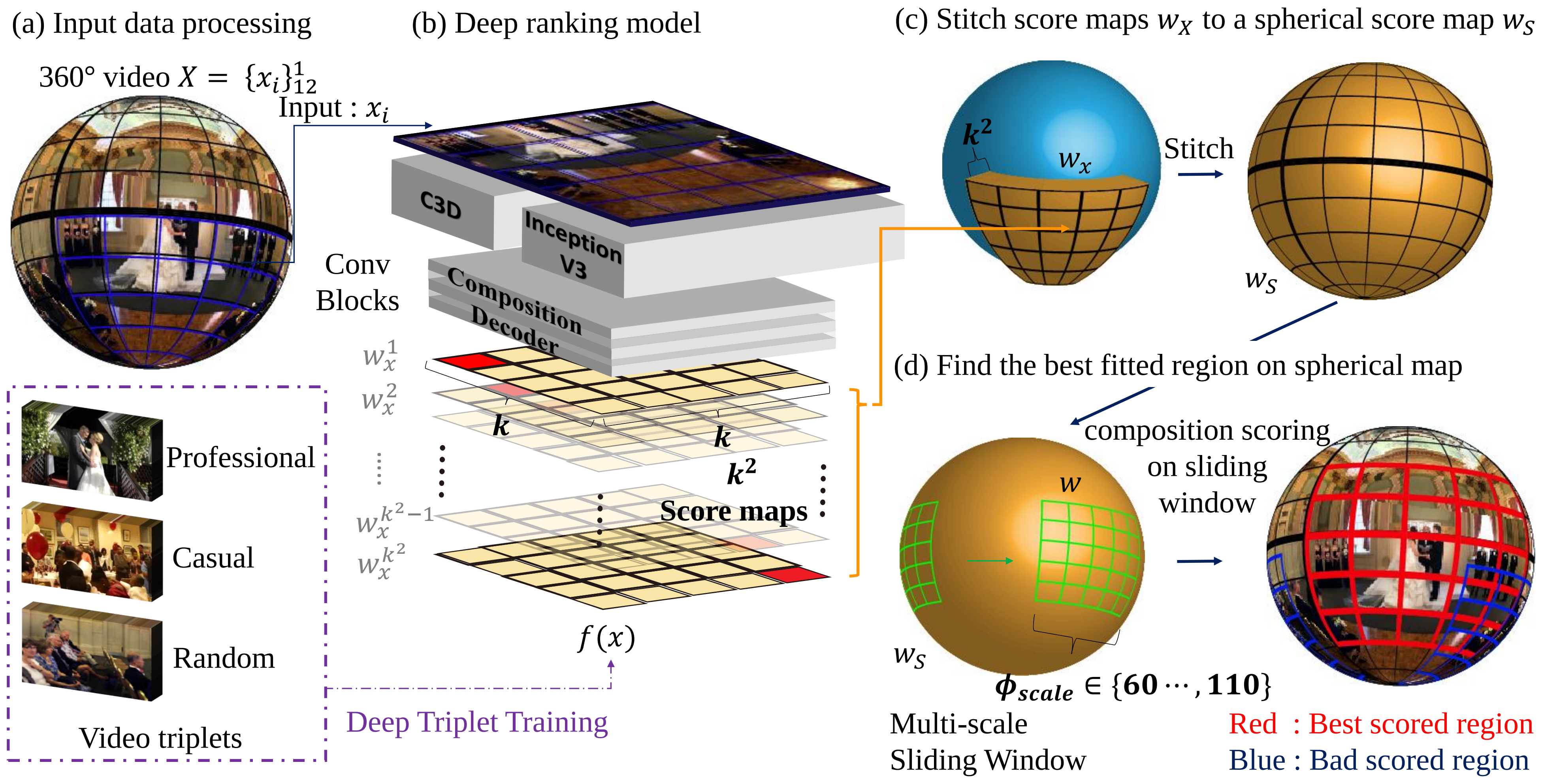
Summarize the 360º video spatially and temporally.

- Select pleasant looking normal field-of-view within 360º field-of-view.
- Produce a concise highlight at the same time.

(Input) 360º Video

(Output) Highlighted Projection

Composition View Score (CVS) Model



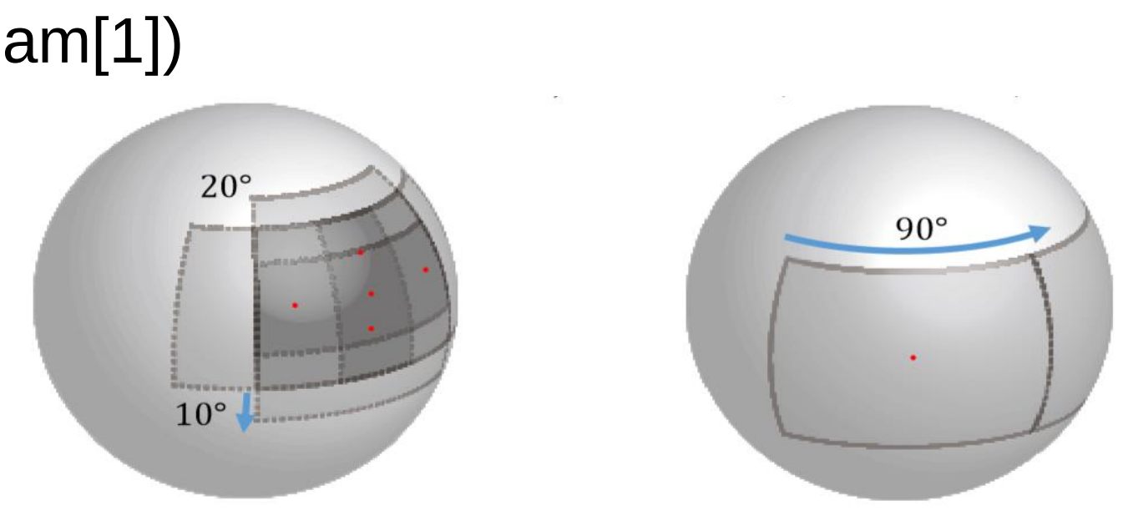
(1) Fully convolutional CVS generates a layered spherical score maps.

(2) Position-wise composition score learns fidelity of views and determines which view is suitable for highlight.

$$f(x) = \sum_{i,j}\sum_{l,m} \kappa(l-i)\kappa(m-j)\mathbf{w}_x^{c(k,l,m)}(i,j|\mathcal{M}) \quad \text{where } \kappa(u) = \frac{exp(-u^2/2h^2)}{\sqrt{2\pi}h}, c(k,l,m) = k \times l + m$$

(3) CVS model significantly reduces a burden of normal field-of-view projection

	Processing time	# of ST-glimpses
AutoCam[1]	178 min	198
CVS	11 min	12



(a) Dense Projection (TS-DCNN[3], RankNet[2], AutoCam[1])

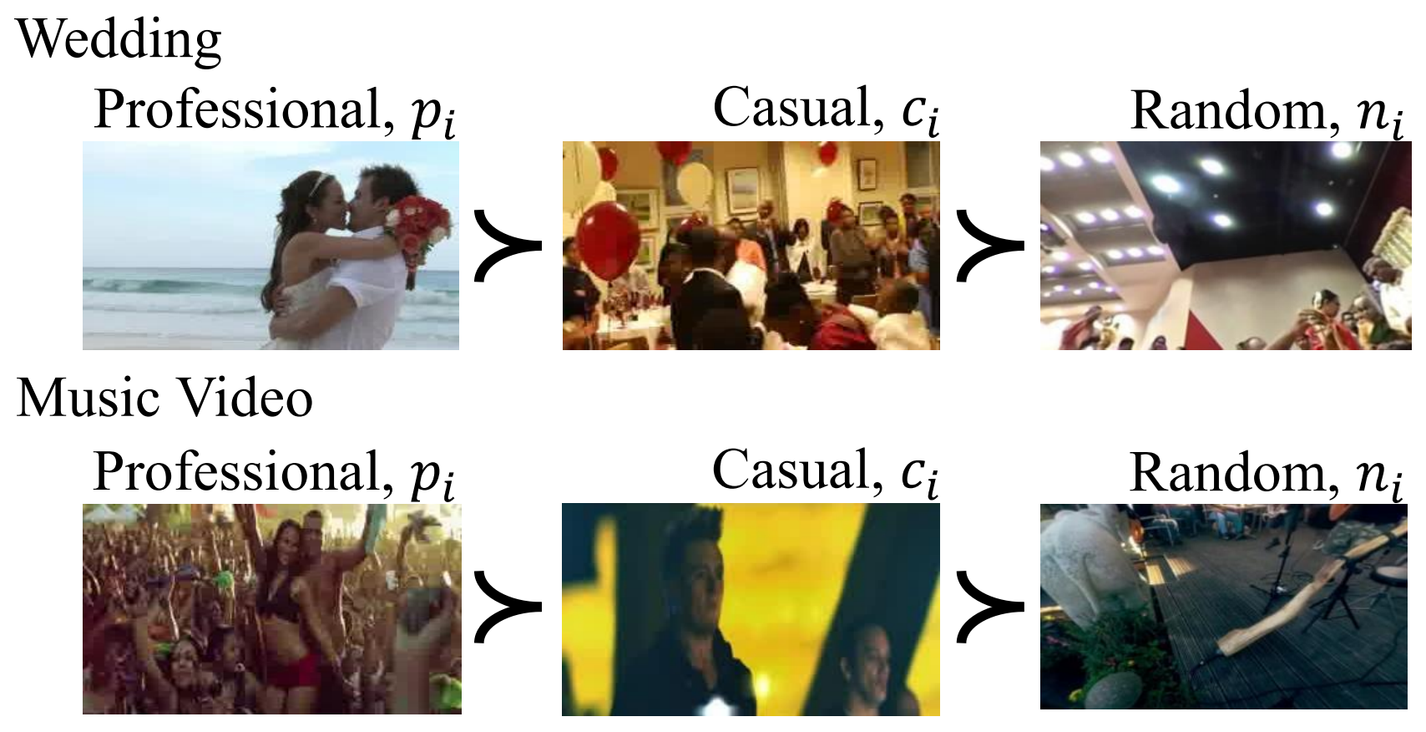
(b) Sparse Projection (Our CVS model)

Triplet Ranking in order

- Rank quality of visual composition in video.

⇨ Professional videos > casual 360º videos > Random regions.

$$f(p_i) \succ f(c_i) \succ f(n_i), \qquad \forall (p_i, c_i, n_i) \in \mathcal{D}$$



Video	Topic	Type	# video	Total (hour)	mean (minute)
360 ° video	wedding	360°	62	54.8	53.1
	MV		53	17.2	19.5
NFOV video	wedding	professional	755	87.1	6.9
		casual	664	104.5	9.4
	MV	professional	333	3.3	4.4
		casual	654	47.5	0.6

- Our model correctly quantify the quality differences among the samples with ranking loss.

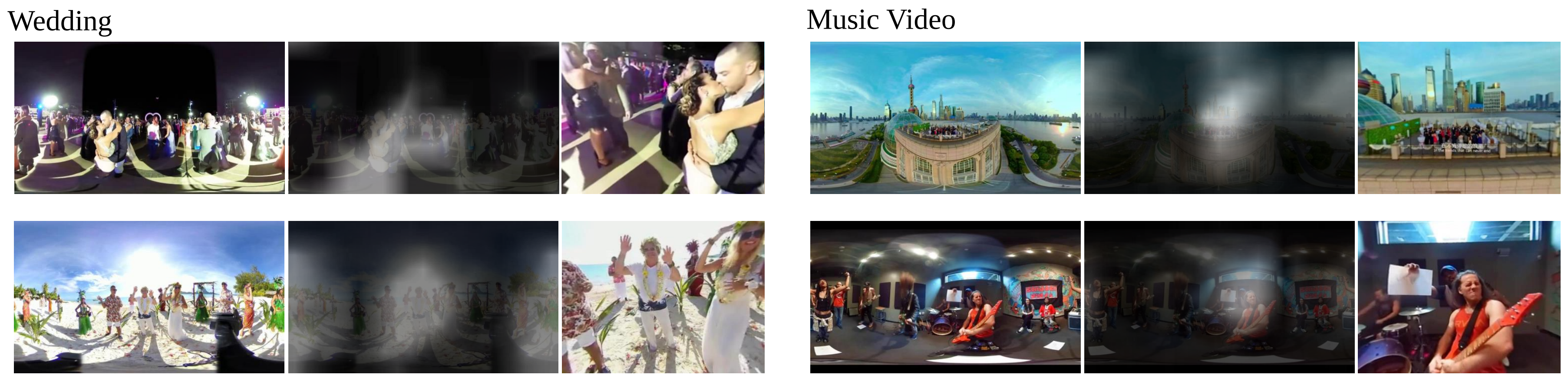
$$\mathcal{L}_i = \alpha \max(0, f(c_i) - f(p_i) + 1) + (1-\alpha)\max(0, f(n_i) - f(c_i) + 1)$$

$$\mathcal{L} = \sum_i \mathcal{L}_i + \lambda \|\mathcal{M}\|_F^2$$

where $\mathcal{M}$ denotes CVS model parameters, i denotes index of minibatch.

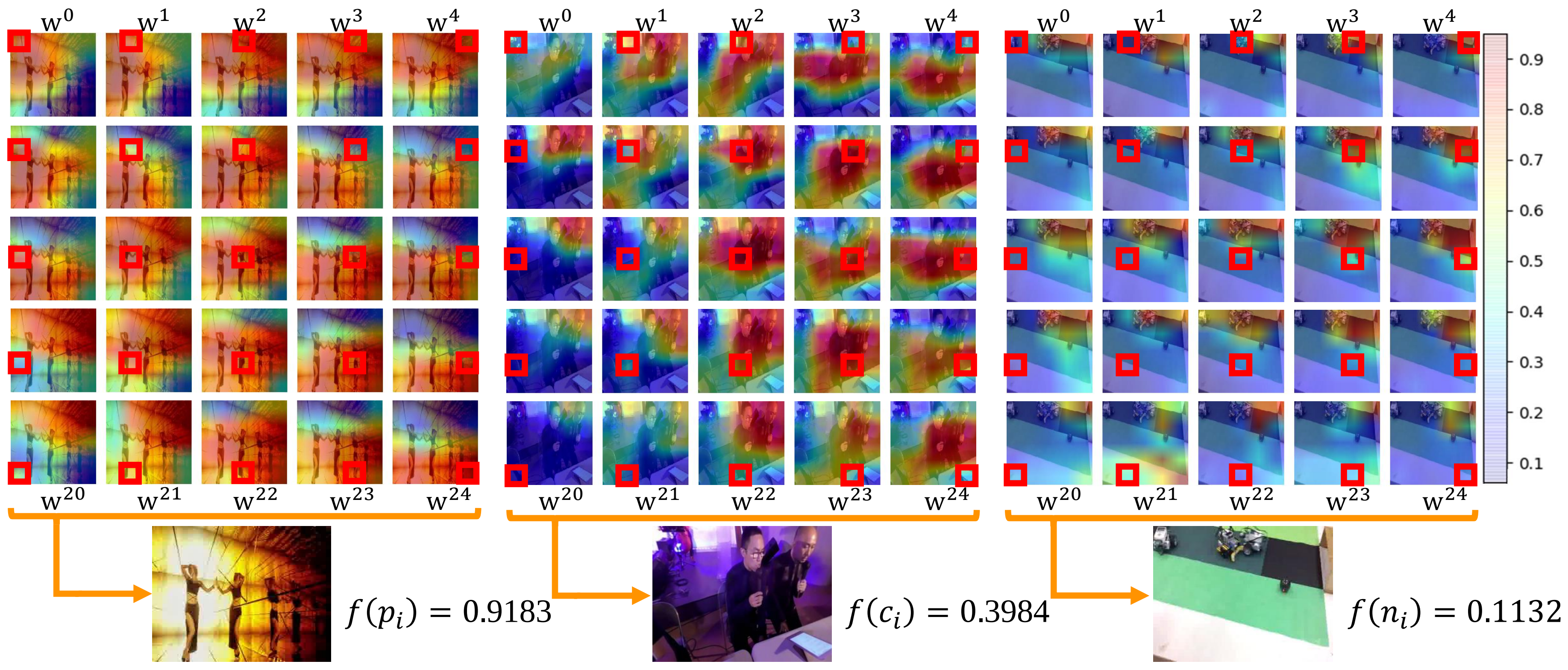
Experiments

- Examples of view selection.



Original Score map Projection Original Score map Projection

- Examples of position score maps $\mathbf{w}_x$.



(a) Professional (b) Casual (c) Random

- Results of spatial summarization on the Pano2Vid [1] dataset

Methods	Frame cosine sim	Frame overlap
Center	0.572	0.336
Eye-Level	0.575	0.392
Saliency [1]	0.387	0.188
AutoCam (w/o stitch) [1]	0.541	0.354
AutoCam-stitch [1]	0.581	0.389
RankNet [2]	0.562	0.398
TS-DCNN [3]	0.578	0.441
CVS-C3D	0.656	0.554
CVS-Inception	0.642	0.545
CVS-Fusion	0.701	0.590
CVS-C3D-stitch	0.774	0.646
CVS-Inception-stitch	0.768	0.666
CVS-Fusion-stitch	**0.800**	**0.677**

- Results of highlight detection on our newly collected dataset

Methods	Wedding	MV
Center	7.88	5.90
RankNet [2]	11.98	11.65
TS-DCNN [3]	13.23	12.28
CVS-C3D	16.32	12.15
CVS-Inception	16.13	12.38
CVS-Fusion (pairwise)	14.34	12.56
CVS-Fusion	**17.96**	**14.92**

- AMT results for our dataset

CVS-Fusion vs	Wedding	MV
Center	**68.0** % (117/150)	**57.3** % (86/150)
RankNet [2]	**67.3** % (101/150)	**65.3** % (98/150)
TS-DCNN [3]	**64.0** % (96/150)	**58.0** % (87/150)

References

[1] Su,Y.-C et al. 2016. Pano2Vid: Automatic Cinematography for Watching 360◦ Videos. In ACCV.
[2] Gygli, M et al. 2016. Video2GIF: Automatic Generation of Animated GIFs from Video. In CVPR.
[3] Yao,T et al. 2016. Highlight Detection with Pairwise Deep Ranking for First-Person Video Summarization. In CVPR.