上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Saliency Prediction on Omnidirectional Images with Brain-Like Shallow Neural Network

Dandan Zhu
Artificial Intelligence Institute, Shanghai Jiaotong University

Collaborated with Yongqing Chen, Xiongkuo Min, Defang Zhao, Yucheng Zhu, Qiangqiang Zhou, Tian Han, Guangtao Zhai and Xiaokang Yang

SJTU

Outline

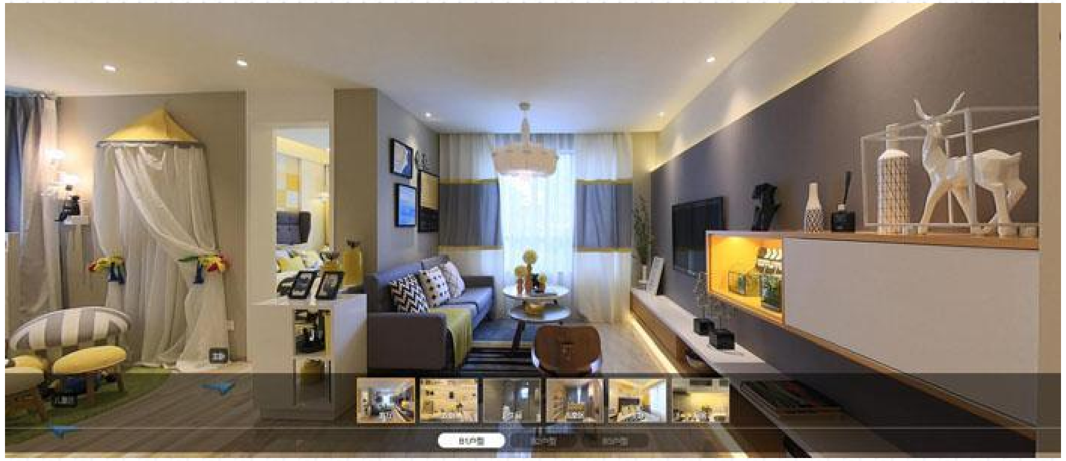
- **Background**
- **Related Works**
- **Proposed Method**
- **Experimental Results**

Background

◆ **Most recently, Omnidirectional image (ODI) has become part of our daily life.**

Live streaming **of CCTV award presentation ceremony** in 360° format

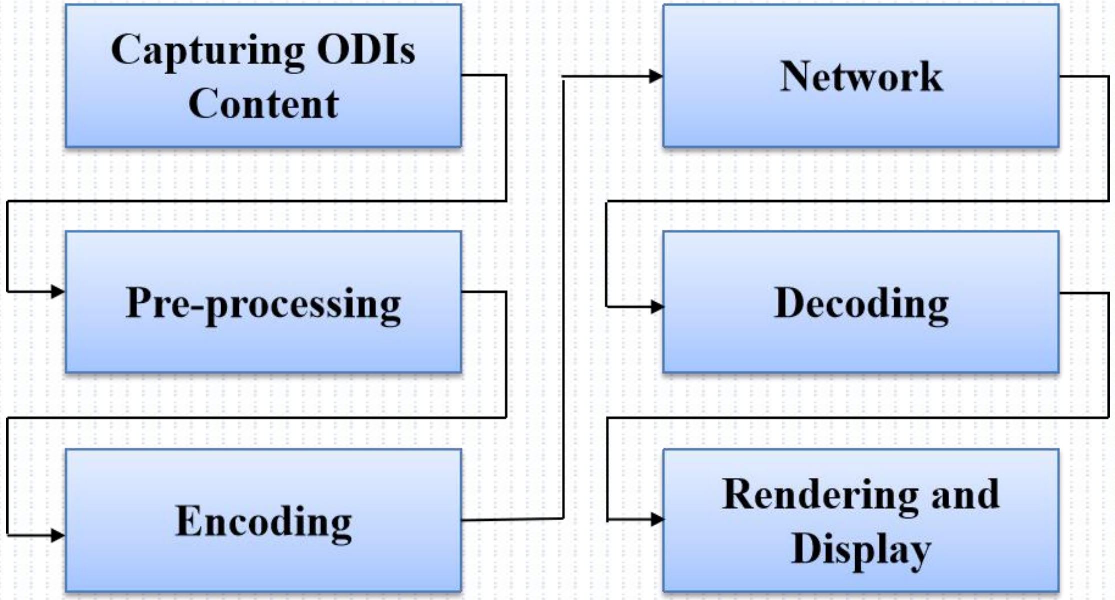
Immersive images (ODIs) for education



VR viewing house (in the form of ODI)

Background

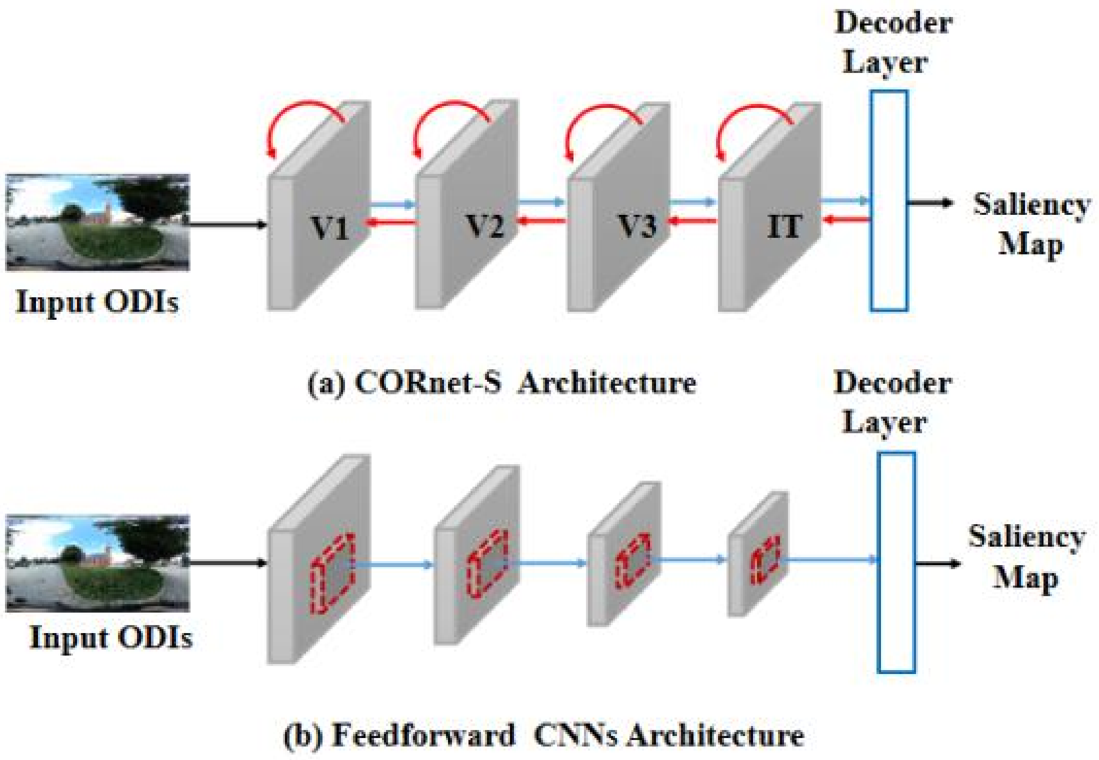
- However, ODIs have higher resolution, making it difficult for streaming and rendering.



- Meanwhile, deep feedforward CNNs are too complex in design and contain a vast number of layers, which is difficult to map to the ventral stream structure of the brain visual system.

Background

- They lack biologically-important brain structures (i.e. recurrence connectivity), which is difficult to match the complex neurons states in the brain.



Related Works

- **Saliency prediction of eye fixations on ODIs**

Non-Deep-Learning : VCTs (De Abreu *et al.,* 2017), Feature-based estimation (Battist *et al.,* 2018), Color dictionary-based (Ling *et al.,* 2018).

These works are all based on heuristics (hand-crafted features) .

Deep-Learning: SalNet360 (Monroy *et al.,* 2018), SaltiNet (Reina *et al.,* 2017).

Saliency prediction accuracy of these methods is sub-optimal.

- **Saliency prediction of head fixations on ODIs**

Non-Deep-Learning : BMS360 and GBVS360 (Lebreton *et al.,* 2018), SJTU (Zhu *et al.,* 2018).

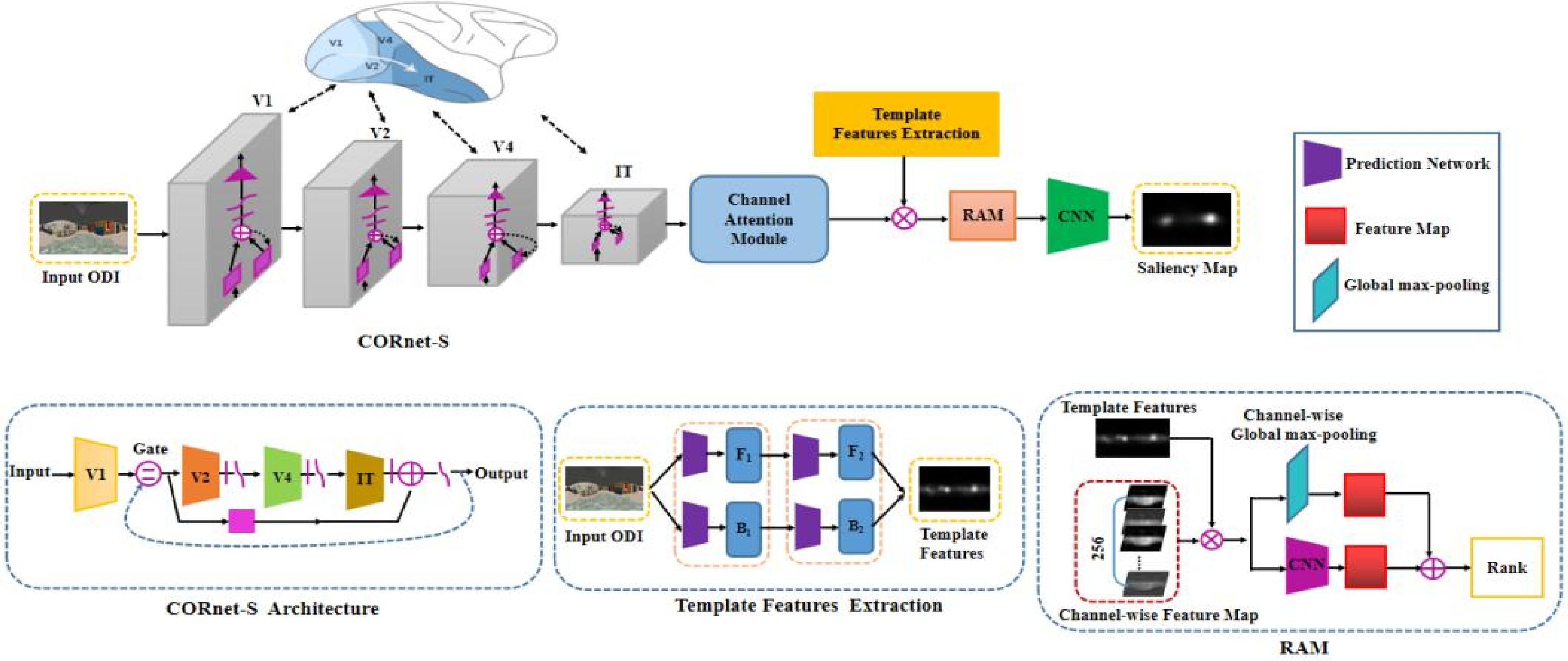
Existing traditional works are all based on heuristics.

Deep-Learning: SalGAN360 (Chao *et al.,* 2018).

No existing saliency prediction work on ODIs benefits from the brain-like neural network.

◆ Model Architecture

Our proposed model consists of three modules: CORnet-S module, template features extraction module, Ranking Attention Module (RAM) .



Proposed Method

CORnet-S module

- The CORnet-S module is a lightweight ANN with four computational areas, conceptualized as analogous to the ventral visual areas (V1, V2, V4 and IT) and recurrent connections.
- In order to implement recurrence, the output of an area is passed through that area several times.
- It is worth noting that we modify the original CORnet-S structure and add a channel attention module behind the IT area, which can further improve the feature extraction capability of the model. Specifically, the channel attention maps are calculated as follows:

$$F_c = \boldsymbol{\sigma}(MLP(Avgpool(f)) + MLP(Maxpool(f)))$$

$$= \boldsymbol{\sigma}(w_1(w_0(f_{avg}^c)) + w_1(w_0(f_{\mathbf{max}}^c))),$$

where $\boldsymbol{\sigma}$ is the sigmoid function, w_0 and w_1 are the MLP weights.

Proposed Method

Template feature extraction module

In ODIs, template features represent the final attention map obtained by fusing the foreground attention map with the background attention map. Specifically, we employ a two-stage network to learn part attention maps. The first stage individually predicts foreground attention F^1 and background attention B^1 by two independent prediction networks:

$$F^1 = \varphi^1(F^{TF}), \quad B^1 = \phi^1(F^{TF}),$$

where F^{TF} is the feature map obtained by vgg16 network, φ^1 and ϕ^1 denote two prediction networks. In the second stage, the attention maps obtained by the first stage are further refined and the specific equations are expressed as follows:

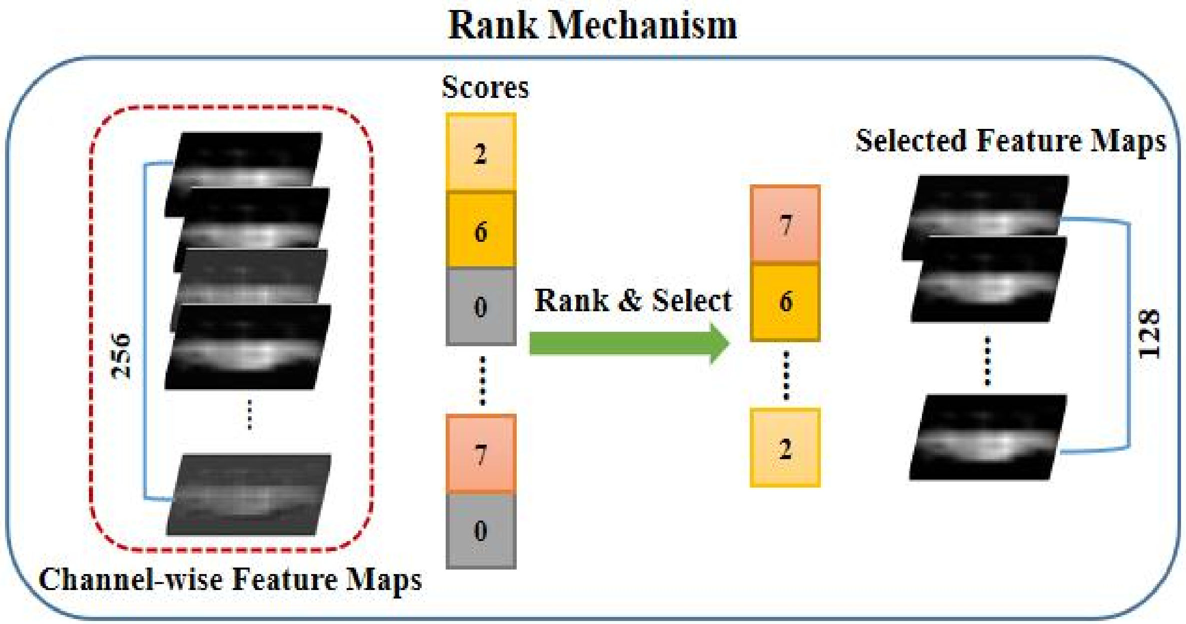
$$F^2 = \varphi^2(F \mid F^1, B^1), \quad B^2 = \phi^2(F \mid F^1, B^1),$$

where F^2 and B^2 are the foreground attention map and background attention map, respectively.

Proposed Method

Ranking attention module

In proposed RAM module, we can learn a scoring scheme for channel-wise feature maps and rank these feature maps based on such important scores. The rank mechanism of the proposed RAM is depicted as follows:



Proposed Method

◆ Ranking attention module

To calculate the ranking scores of the channel-wise feature maps, we utilize a two-layer network f_n refinement by summing with the channel-wise global max-pooling $f_{\max}$ of the tensor S_i in an element-wise manner:

$$r_i = f_n(S_i) + f_{\mathbf{max}}(S_i),$$

For the ranking scores of channel-wise feature maps in S_i, we need to rank these channel-wise feature maps according to the ranking score r_i :

$$S_i' = rank\ (S_i \mid r_i),$$

where S_i' represents the ordered channel-wise feature maps afeter rank. Then we need to select important features for the final fine-grained saliency prediction and discard redundant features.

Experimental Results

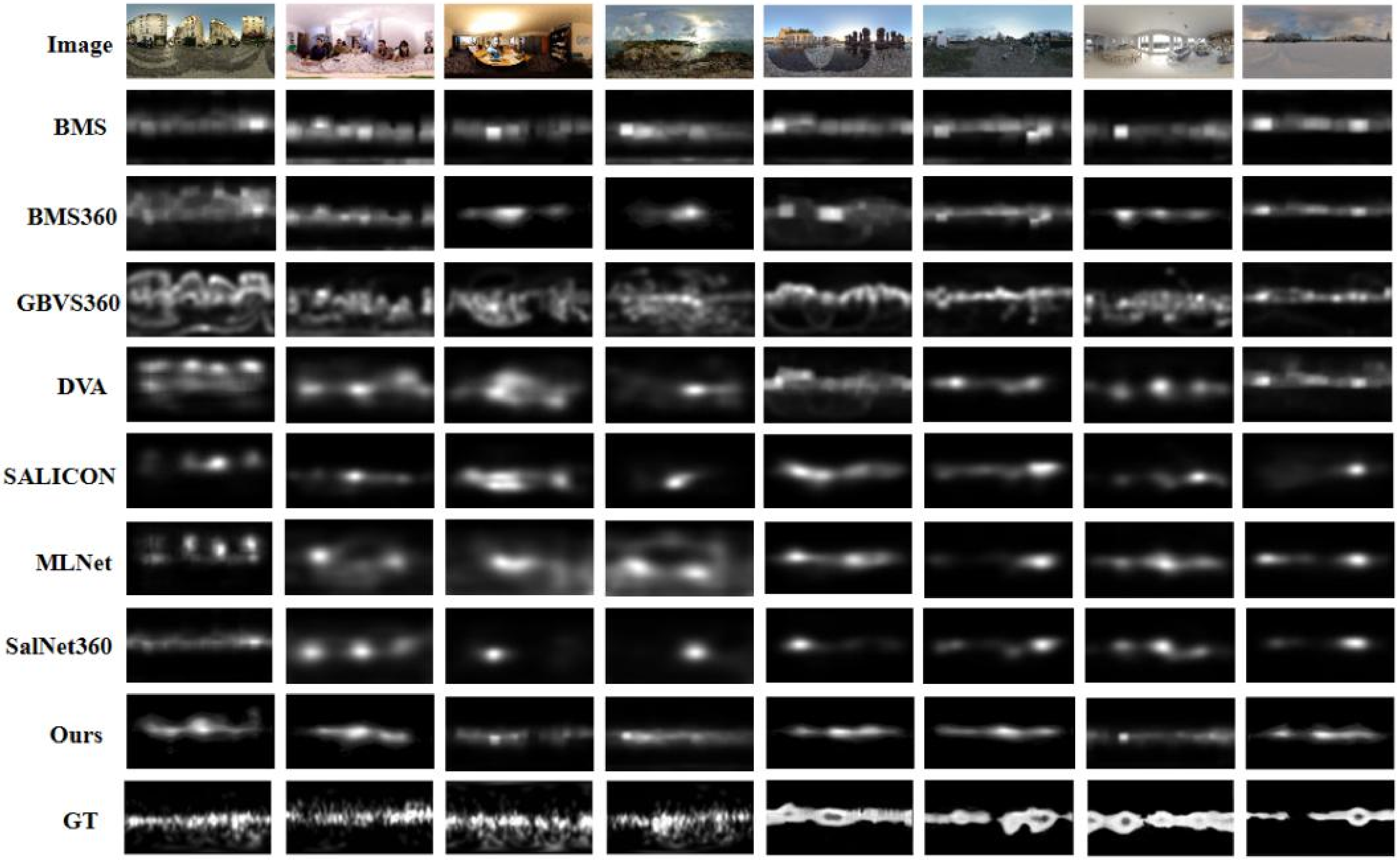
◆ Datasets

- Salient360!
  - ✓ 98 equirectangular images;
  - ✓ Ground-truth saliency maps and scan-path obtained from subjective experiments (free exploration + tracking of eye and head movements);
  - ✓ 40–42 subjects for each ODI， 5 seconds of black screen as break time.

- ODS
  - ✓ 22 high resolution equirectangular images;
  - ✓ Ground-truth Saliency maps and gaze trajectories are acquired under three viewing conditions (VR condition, VR seated condition, desktop condition);
  - ✓ 169 subjects for each ODI.

Experimental Results

◆ Qualitative comparison



Experimental Results

Quantitative comparison

Evaluation metrics: NSS, CC, AUC and KL divergence.

Methods	Salient360!				ODS			
	CC	AUC	NSS	KL divergence	CC	AUC	NSS	KL divergence
BMS [9]	0.562	0.721	0.963	0.589	0.545	0.687	0.942	0.634
BMS360 [4]	0.716	0.754	1.372	0.583	0.648	0.724	1.224	0.615
GBVS360 [4]	0.587	0.836	0.994	0.562	0.569	0.696	0.975	0.571
DVA [16]	0.728	0.772	1.394	0.594	0.612	0.765	1.327	0.541
SALICON [13]	0.745	0.781	0.998	0.554	0.724	0.769	0.987	0.538
MLNet [17]	0.764	0.812	1.012	0.713	0.745	0.797	1.081	0.686
SalNet360 [11]	0.795	0.843	1.581	0.514	0.776	0.821	1.565	0.534
Ours	**0.913**	**0.922**	**2.020**	**0.498**	**0.892**	**0.878**	**2.015**	**0.512**

Thank You!