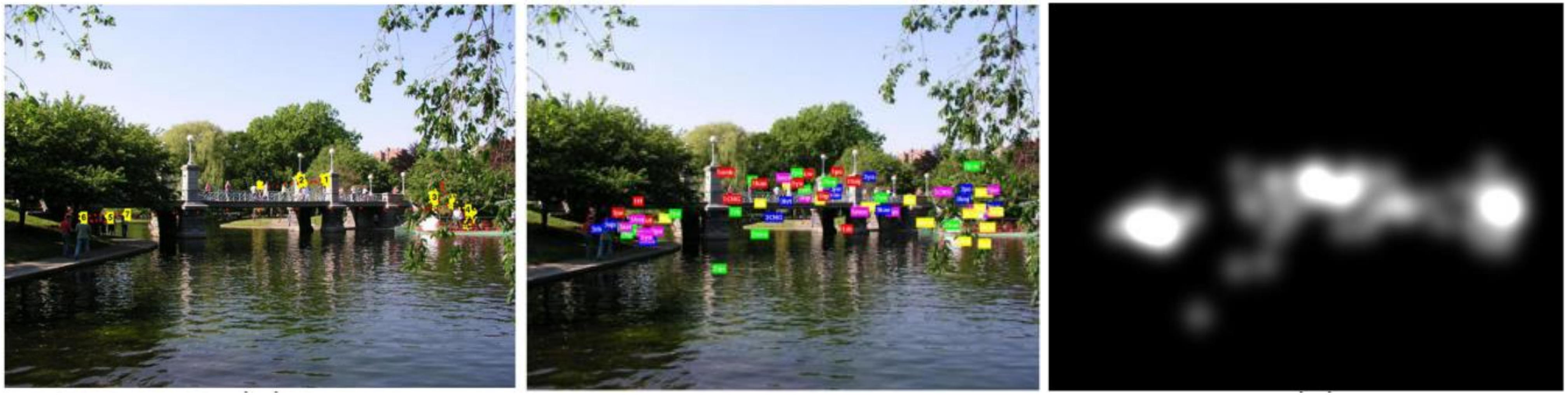
Scanpath

2018.11.10

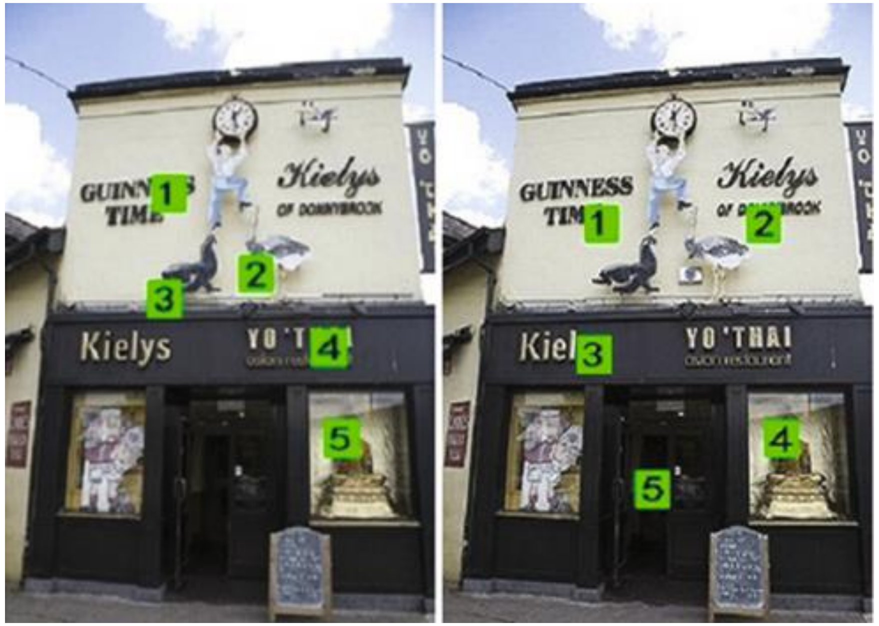
王前前

- 当人类观看一幅图片时，会花费大部分时间观看特定的区域。会将目光投向特定的点并开始探索图像从而得到一系列覆盖图像显著性区域的注视点。
- 视觉显著性预测是致力于估计吸引人类注意力的图像区域的计算机视觉领域，对这一过程的理解可以为人类图像理解提供线索，并在图像和视频压缩、传输、渲染等领域有应用。
- 注视点的一大特性：随机性。不同的观测者会产生非常不同的注视点。因此，显着性预测领域的研究者传统上整合多个观测者的注视点以生成一致性表示，显著性图。

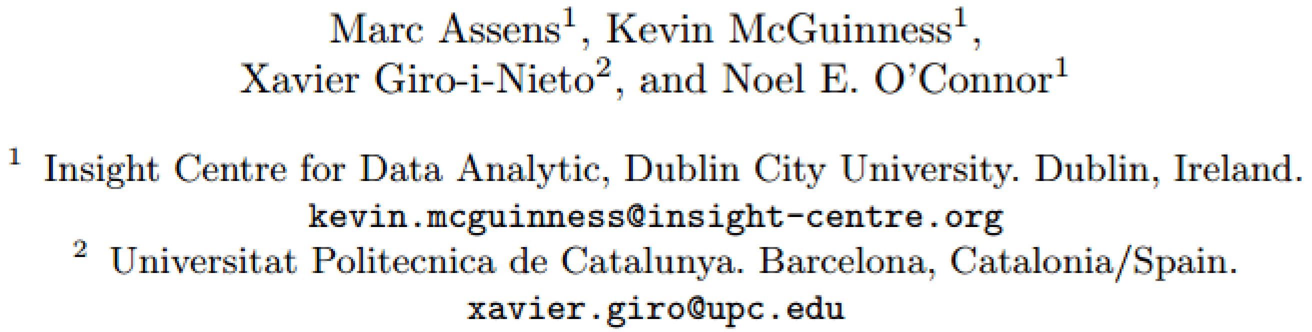


MIT dataset: 1003 images, 15 people.
http://people.csail.mit.edu/tjudd/WherePeopleLook/interactiveWebsite/seeFixations.html#

- 丢失了时域信息。在某些情况下，显着性图无法表示图像不同部分的相对重要性，人最初注视的区域可能更有意义(relevant)。
  - saliency map：为图像中每个像素分配注视概率值
  - scanpath prediction：预测注视点序列(gaze fixations sequence)
  - 后者所含信息更丰富，扫视路径的聚合将生成显著性图。



PathGAN: Visual Scanpath Prediction with Generative Adversarial Networks



Marc Assens[1], Kevin McGuinness[1],
Xavier Giro-i-Nieto[2], and Noel E. O'Connor[1]

[1] Insight Centre for Data Analytic, Dublin City University. Dublin, Ireland.
kevin.mcguinness@insight-centre.org
[2] Universitat Politecnica de Catalunya. Barcelona, Catalonia/Spain.
xavier.giro@upc.edu

□ Drawback:从随机数据中学习很困难，使用MSE loss的监督学习性能不好，因为最终的预测结果是所有数据的平均，通常处于图片的中心。

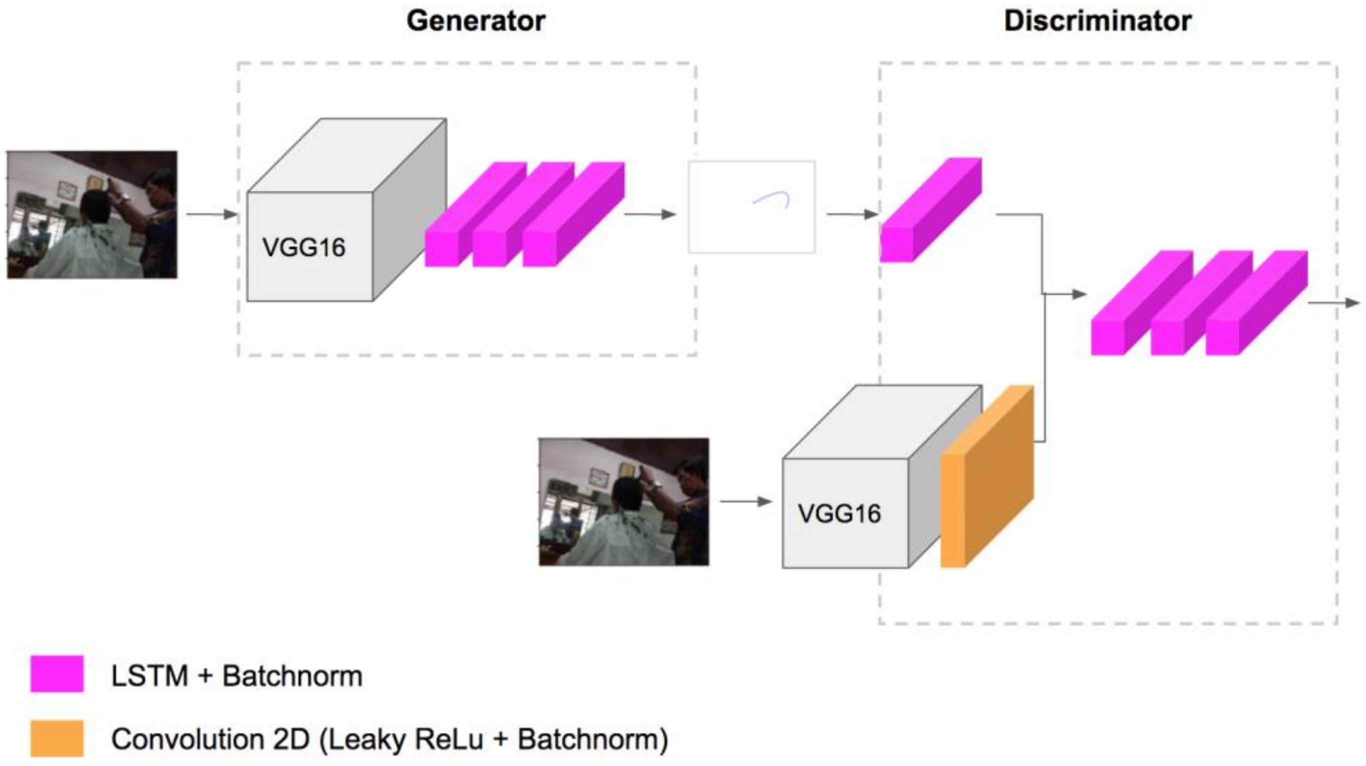
□ Architecture: 旨在从给定的图像预测真实的scanpath

Assens, Marc, et al. "PathGAN: Visual Scanpath Prediction with Generative Adversarial Networks." ECCV workshop EPIC (2018).

PathGAN: Visual Scanpath Prediction with Generative Adversarial Networks

- objective function $$L_{\text{cGAN}}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))]$$
- the task of the discriminator remains unchanged, but the generator is forced to output samples that are close to the ground truth $$L_{L^2}(G) = \mathbb{E}_{x,y,z}[||y - G(x, z)||^2]$$
- final loss function

$$L = L_{\text{cGAN}}(G, D) + \alpha L_{L^2}(G)$$

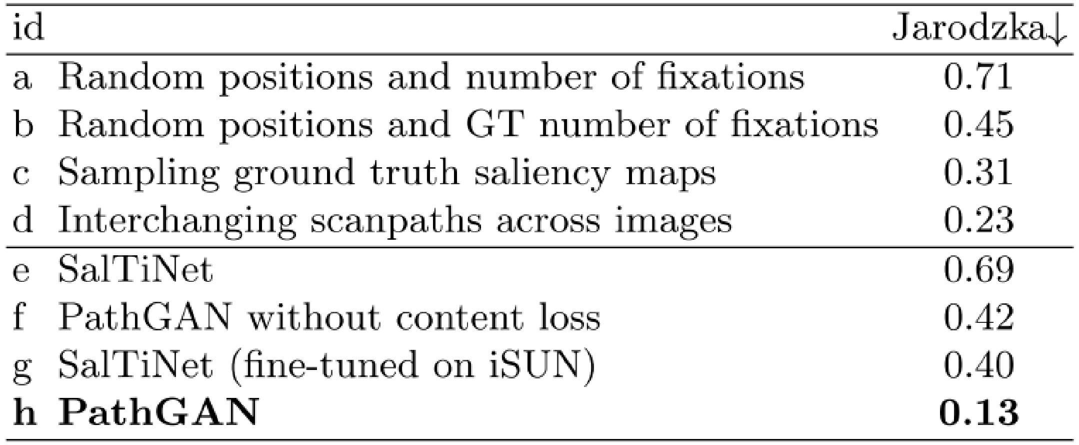


PathGAN: Visual Scanpath Prediction with Generative Adversarial Networks

- Generator
  - VGG+LSTM
  - 输入：image，输出：不定长注视点序列
  - 预测值包括注视点坐标、时间、结束概率 [x, y, t, EOS]
- Discriminator
  - 判断一条路径是否是合成的
  - 输入：scanpath + image
- Dataset：首先在iSUN上训练, contains 6000 training images，在全景数据集Salient360上fine-tune, which has 40 training images
- Metric: Jarodzka algorithm

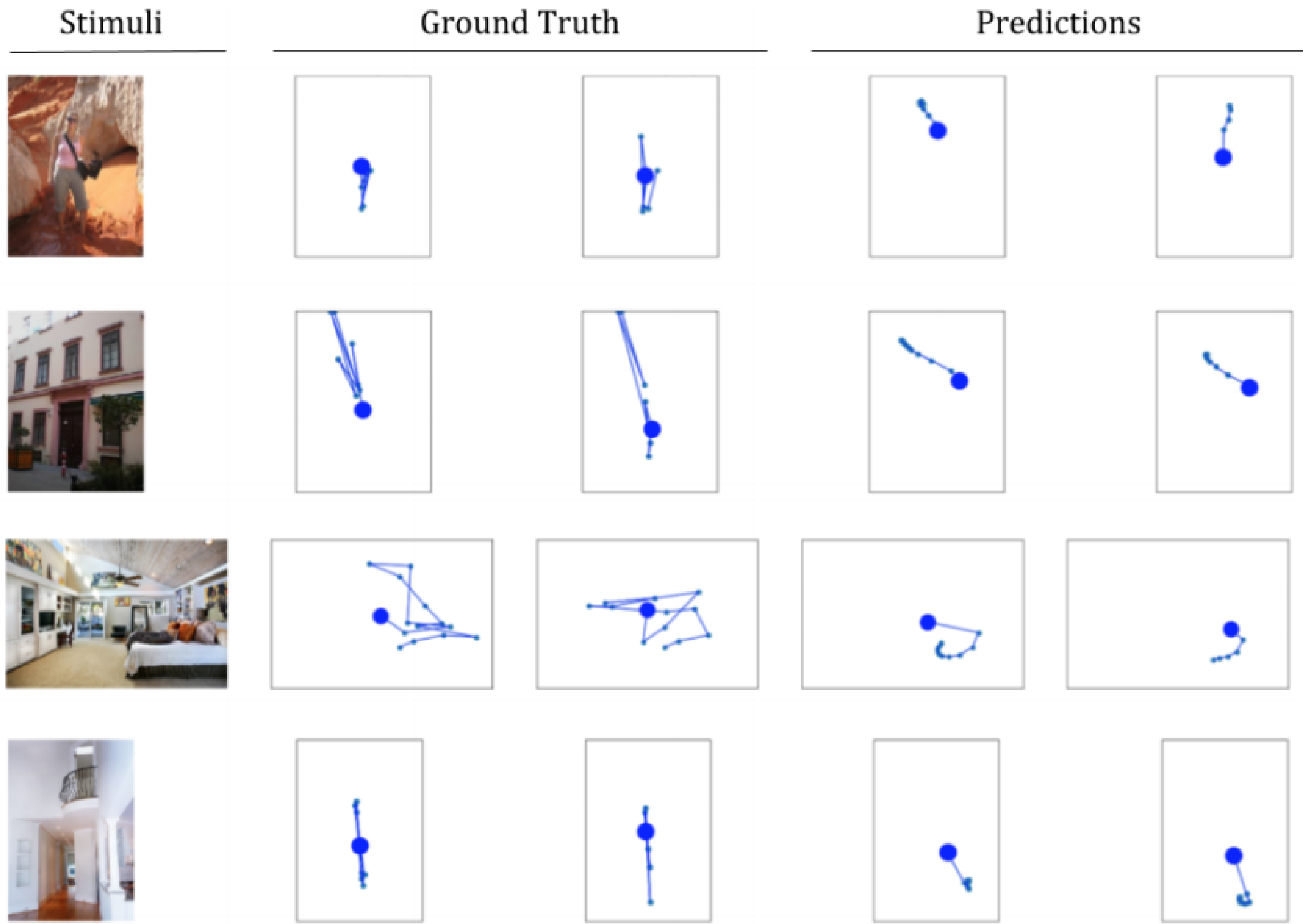
PathGAN: Visual Scanpath Prediction with Generative Adversarial Networks

- 结果
- Drawback: it neglects the characteristic of omni-directional images where points that are close to opposite corners are spatially close



id		Jarodzka↓
a	Random positions and number of fixations	0.71
b	Random positions and GT number of fixations	0.45
c	Sampling ground truth saliency maps	0.31
d	Interchanging scanpaths across images	0.23
e	SalTiNet	0.69
f	PathGAN without content loss	0.42
g	SalTiNet (fine-tuned on iSUN)	0.40
h	**PathGAN**	**0.13**

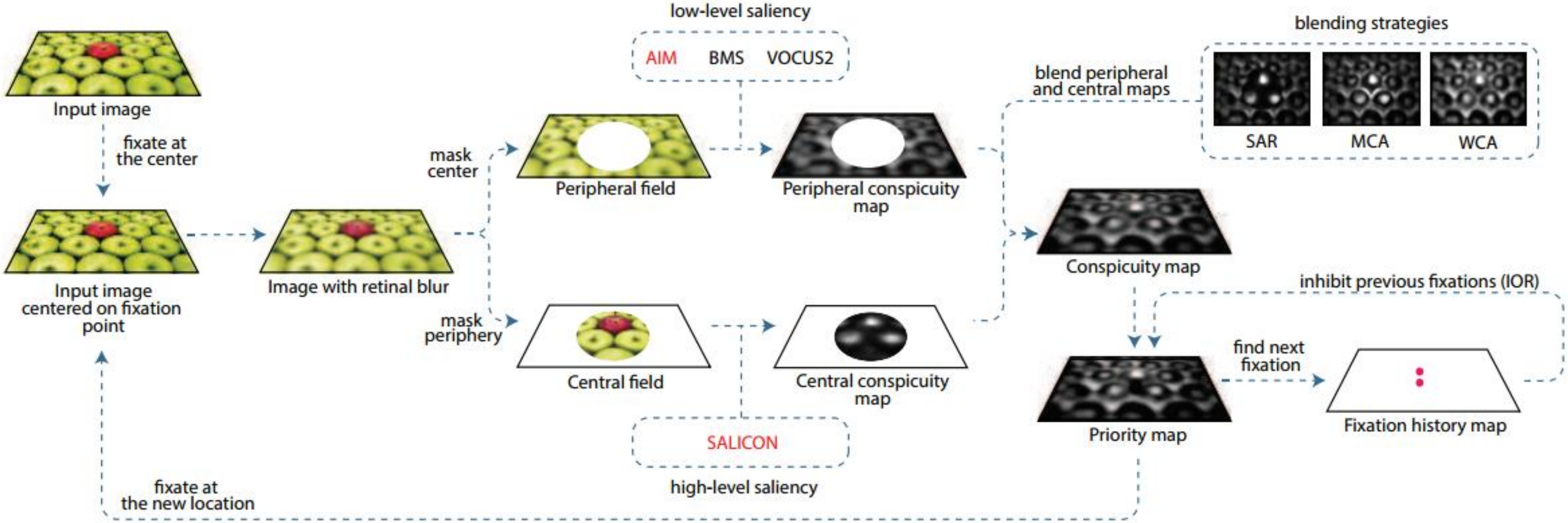
(a) Mean performance on iSUN with the Jarodzka metric



Active Fixation Control to Predict Saccade Sequences, CVPR 2018

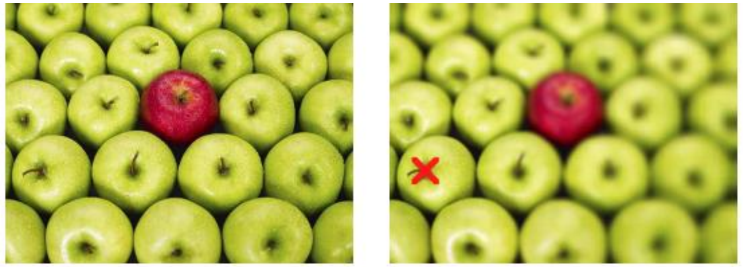
Calden Wloka　　Iuliia Kotseruba　　John K. Tsotsos
Department of Electrical Engineering and Computer Science
York University, Toronto, Canada
calden, yulia_k, tsotsos@cse.yorku.ca

- □ STAR-FC, a novel multi-saccade generator based on the integration of central high-level and object-based saliency and peripheral lower level feature-based saliency.



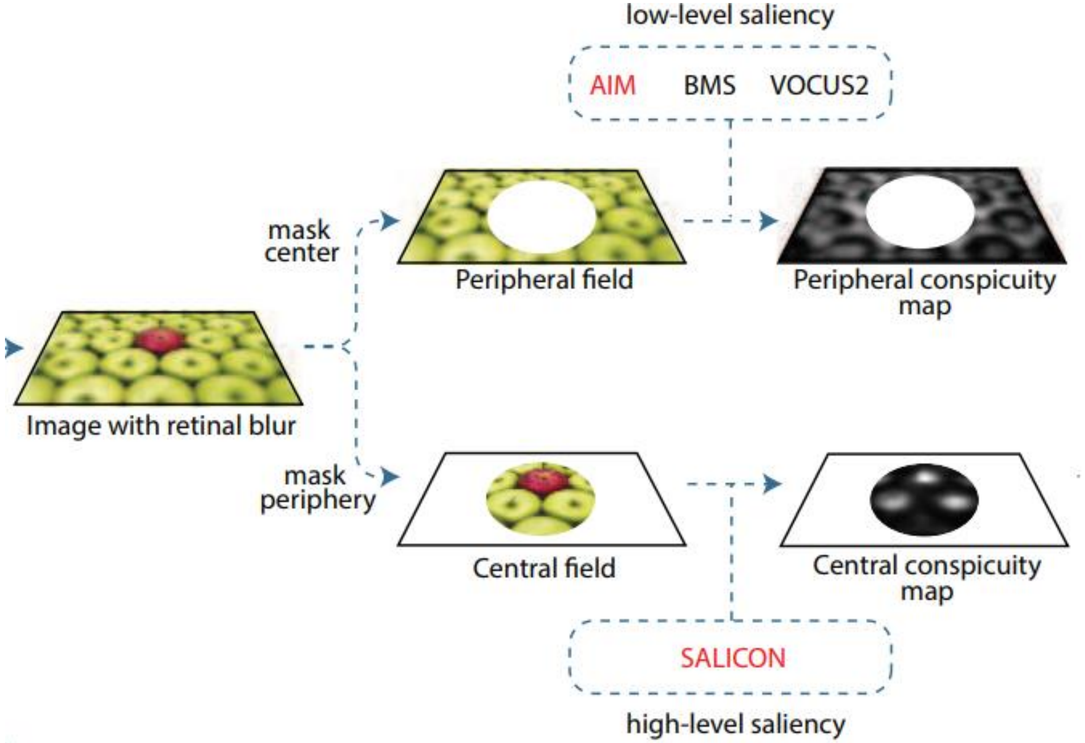
Active Fixation Control to Predict Saccade Sequences, CVPR 2018

- STARFC processes an input image iteratively through a chain of interacting modules
  - Retinal transform: 通过以当前注视点为中心的各向异性模糊重建人眼的视力场。 图像中的每个像素都是基于与注视点的距离从Gaussian pyramid采样，随着固定距离增加其模糊度。
  - Central-peripheral split：为了表示大脑皮层对于中央与边缘图片的不同处理机制。边缘注意力捕捉多依赖于low-level特征，而中央区则多依赖于high-level特征，倾向于object-based。



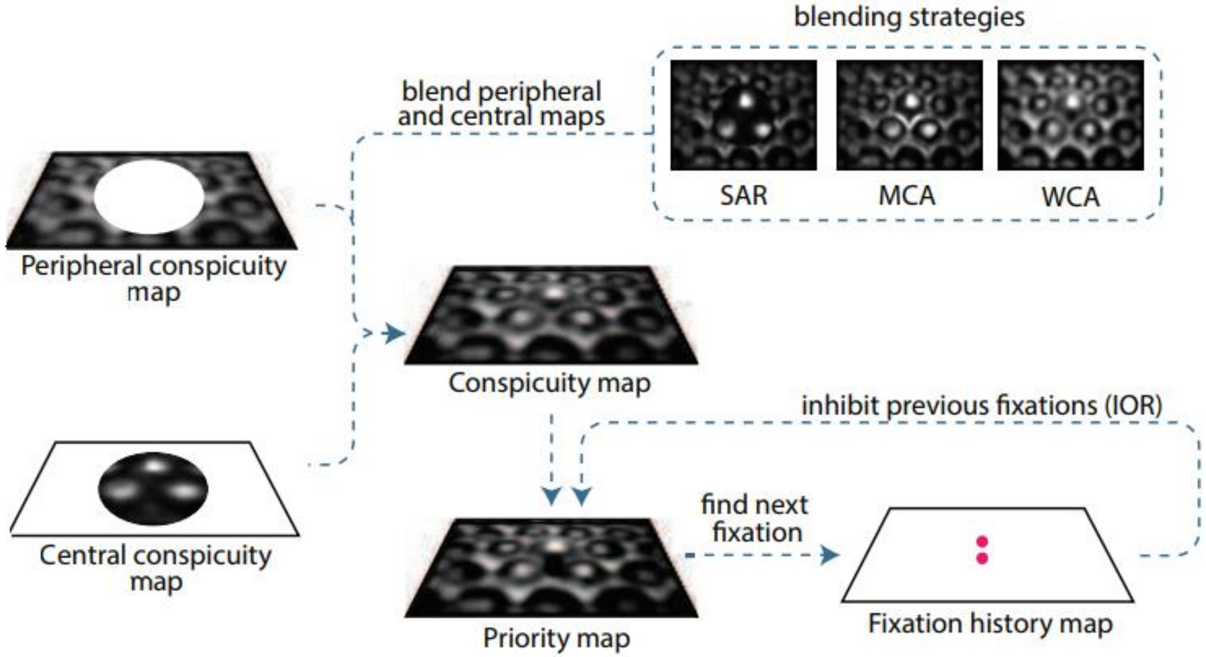
(a) Original image (b) With retinal transform

Figure 2: An example of the retinal transform: (a) original image; (b) fixated in the location marked with a red 'X'



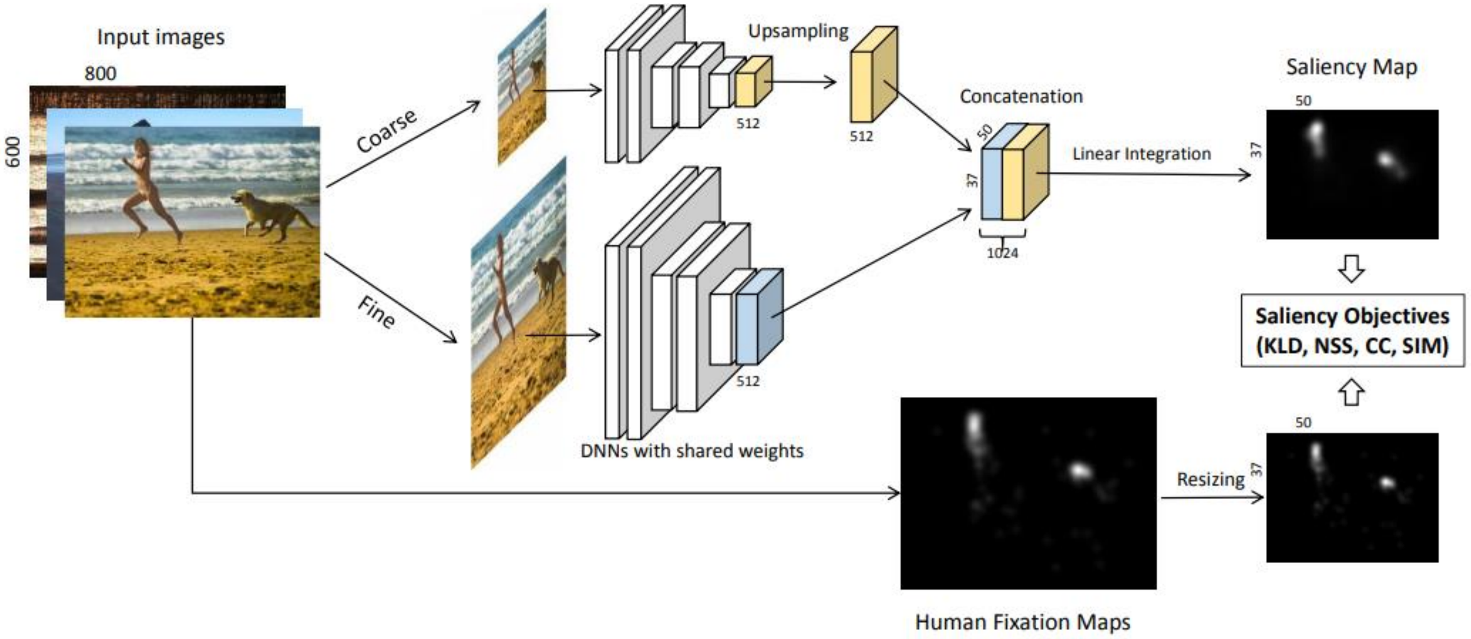
Active Fixation Control to Predict Saccade Sequences, CVPR 2018

- Conspicuity map: 中央和边缘流重新组合成单个map。
  - SAR/MCA/WCA
- Priority map: IOR(inhibition of return)
- Fixation history map: 存储历史注视点的坐标信息，这些位置受到圆形抑制区的抑制。
- Saccade control: 从priority map中找到下一注视点，并更新Fixation history map.



Active Fixation Control to Predict Saccade Sequences, CVPR 2018

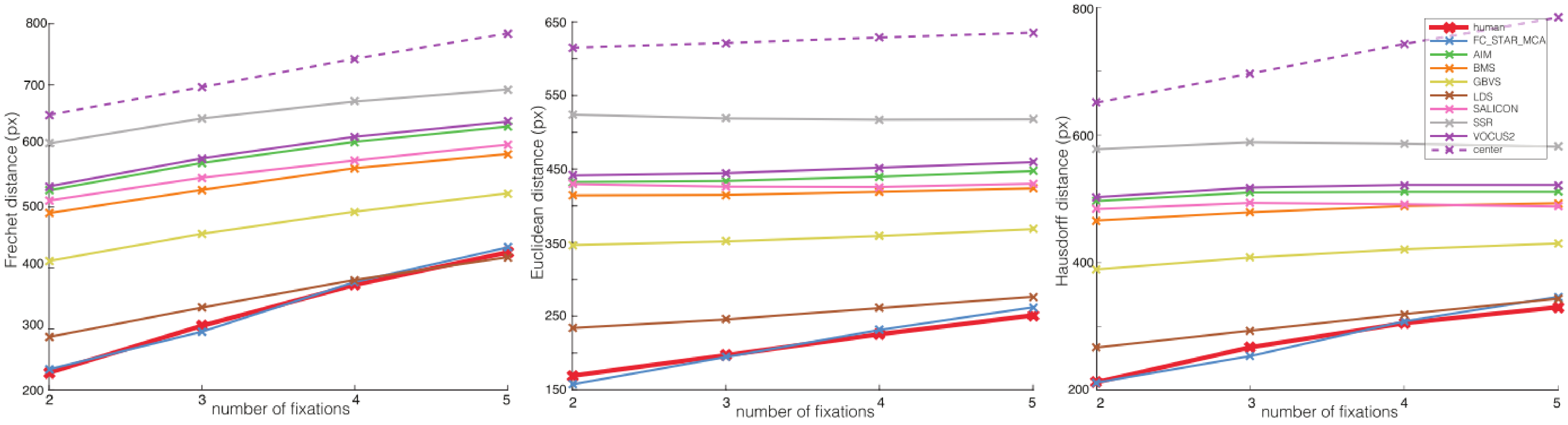
- SALICON



- Dataset: CAT2000, 20个类别x20.
- Evaluation Metrics:
  - saccade amplitude distributions
  - Euclidean Distance：计算注视点序列对应注视点的欧式距离的平均值
  - Frechet Distance：两条序列在给定点上的最大距离
  - Hausdorff Distance：一个序列中一个点到另一序列中最近点的最大距离

Active Fixation Control to Predict Saccade Sequences, CVPR 2018

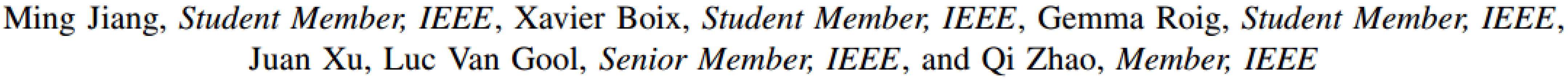
- 结果



(a) Mean FD (b) Mean ED (c) Mean HD

Figure 7: A comparison of fixation prediction scores for static saliency maps and STAR-FC. A sequence formed by always picking the center pixel is shown in a dashed line to provide a performance baseline.

Learning to Predict Sequences of Human Visual Fixations, Trans 2016



Ming Jiang, *Student Member, IEEE*, Xavier Boix, *Student Member, IEEE*, Gemma Roig, *Student Member, IEEE*,
Juan Xu, Luc Van Gool, *Senior Member, IEEE*, and Qi Zhao, *Member, IEEE*

- focus on learning how to dynamically combine different input cues for visual scanpath prediction, combining multiple cues at different stages of the scanpath
- MDP
  - State: all the information gathered through the visual exploration of the image, such as the locations of previous eye fixations, and features of the current visual exploration extracted from the image.
  - Actions: the location in the image where the gaze will be fixed next
    - 视觉注意力不会被单个像素所吸引，而是被一个区域所吸引，这个区域代表一个物体或物体的一部分。本文使用superpixels而非pixels，即将分割图像使其包含最多一个object。
    - superpixels extracted via energy-driven sampling superpixels
    - 将图像分为300 superpixels，质心作为注视点，动作数等于superpixels数

Learning to Predict Sequences of Human Visual Fixations, Trans 2016

- LSPI (least-squares policy iteration)

$$Q^{\pi}(s,a) \approx \hat{Q}^{\pi}(s,a) = \mathbf{w}^{T}\phi(s,a)$$

$$\pi(s) = \arg\max_{a\in\mathcal{A}} \hat{Q}^{\pi}(s,a) = \arg\max_{a\in\mathcal{A}} \mathbf{w}^{T}\phi(s,a).$$

- The generated sequences are composed by the state-action mapping before executing the action, the state-action mapping after executing the action, and the received reward $\phi(s^{n},a^{n})$, $\phi(s'^{n},\pi(s'^{n}))$, and $r(s^{n},a^{n})$

$$\mathbf{w}_{i+1} = \arg\min_{\mathbf{u}} \sum_{n}^{N} \| \mathbf{u}^{T}\phi(s^{n},a^{n}) - r(s^{n},a^{n}) \\ - \gamma\, \mathbf{w}_{i+1}^{T}\phi(s'^{n},\pi(s'^{n}))\|^{2}$$

Lagoudakis, Michail G., and Ronald Parr. "Least-squares policy iteration." *Journal of machine learning research* 4.Dec (2003): 1107-1149.

Learning to Predict Sequences of Human Visual Fixations, Trans 2016

- Evaluate
  - 计算人类所有注视点的mean-shift聚类
  - 为每一个聚类中心和相应的注视点分配字符，固每一个扫视路径都可以被表示为一串字符
  - 使用Needleman–Wunsch 字符串匹配算法来度量人类与模型预测scanpath之间的相似性
  - 与每个受试者之间路径度量的平均为最终值
- Dataset: OSIE dataset MIT dataset
  - OSIE dataset: 700 imgages, 15 participants, 12 semantic attributes, 800x600, . Each image was manually segmented into a collection of objects on which semantic attributes were manually labeled.
  - MIT dataset: 1003 images, 15 people.

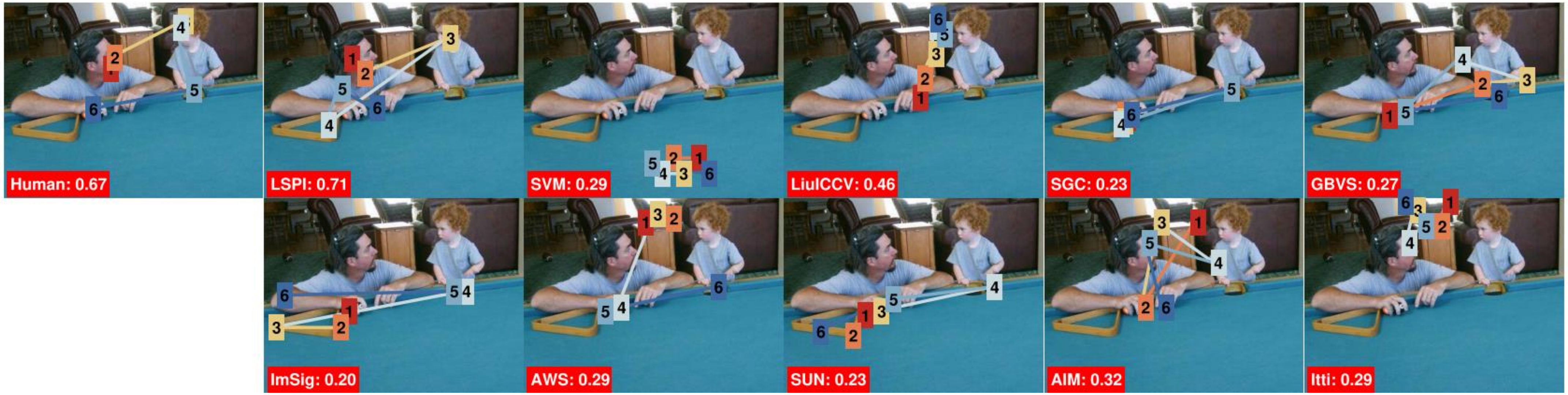
    http://people.csail.mit.edu/tjudd/WherePeopleLook/interactiveWebsite/seeFixations.html#

Learning to Predict Sequences of Human Visual Fixations, Trans 2016

- divide the visual scanpath into different temporal consecutive stages, use a total number of six stages.

$$\phi(s,a) = (\mathbf{I}[t=1]\phi'(s,a), \ldots, \mathbf{I}[t=6]\phi'(s,a))^T$$

- Feactures(19 for OSIE, 35 for MIT)
  - Low-level feacture: saliency feacture

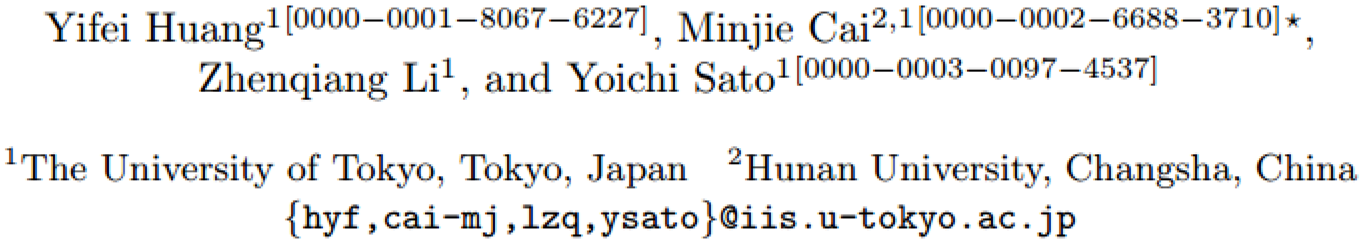


ner

ns

-

Predicting Gaze in Egocentric Video by Learning Task-dependent Attention Transition



Yifei Huang[1][0000–0001–8067–6227], Minjie Cai[2,1][0000–0002–6688–3710]★,
Zhenqiang Li[1], and Yoichi Sato[1][0000–0003–0097–4537]

[1]The University of Tokyo, Tokyo, Japan [2]Hunan University, Changsha, China
{hyf,cai-mj,lzq,ysato}@iis.u-tokyo.ac.jp

□ 基于第一人称视角拍摄的视频开始大量出现

□ 注视点及其周边区域包含了与相机穿戴者相交互的物体或该穿戴者的意图相关的重要信息

□ 对第一人称视频注视点的自动预测（gaze prediction）能够让计算机重点关注视频中与分析理解人的动作和意图最相关的重要区域，减少第一人称视觉的各种学习和推断任务所需的计算量，提高视觉模型的建模效率

□ 传统方法通常将这一问题构建成一个视觉显著性（visual saliency）的估计问题。但在包含复杂日常动作的视频中基于视觉显著性的方法并不能有效预测第一人称视频的注视点。

□ 任务相关的高层知识对于人的注视点转移有重要的影响。

Huang, Yifei, et al. "Predicting Gaze in Egocentric Video by Learning Task-dependent Attention Transition.“ ECCV (2018).

Predicting Gaze in Egocentric Video by Learning Task-dependent Attention Transition

□ Architecture：基于视频的视觉显著性模型+基于任务的注视点转移模型

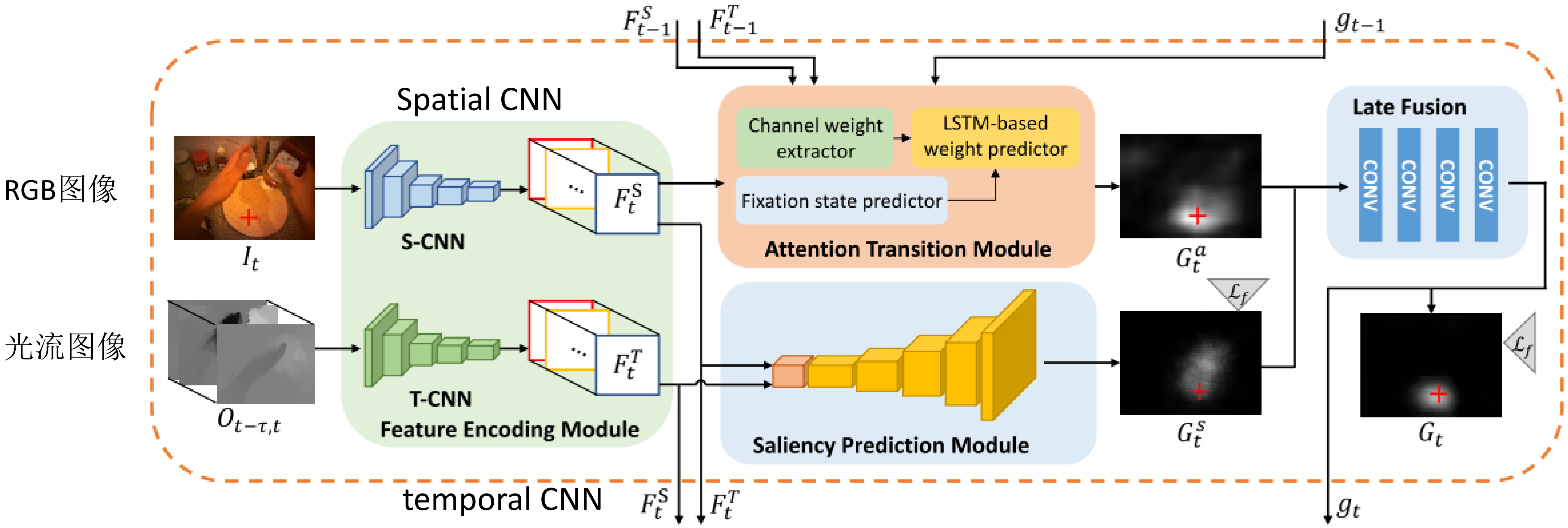
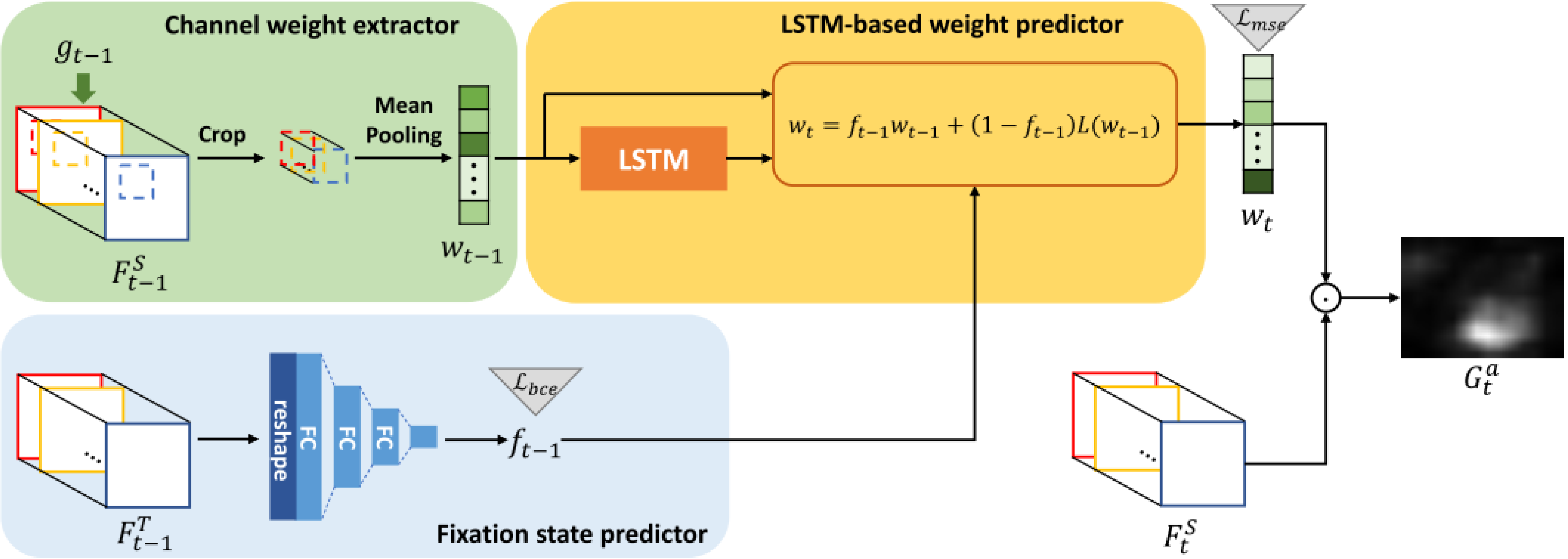


Fig. 1. The architecture of our proposed gaze prediction model. The red crosses in the figure indicate ground truth gaze positions.

Predicting Gaze in Egocentric Video by Learning Task-dependent Attention Transition

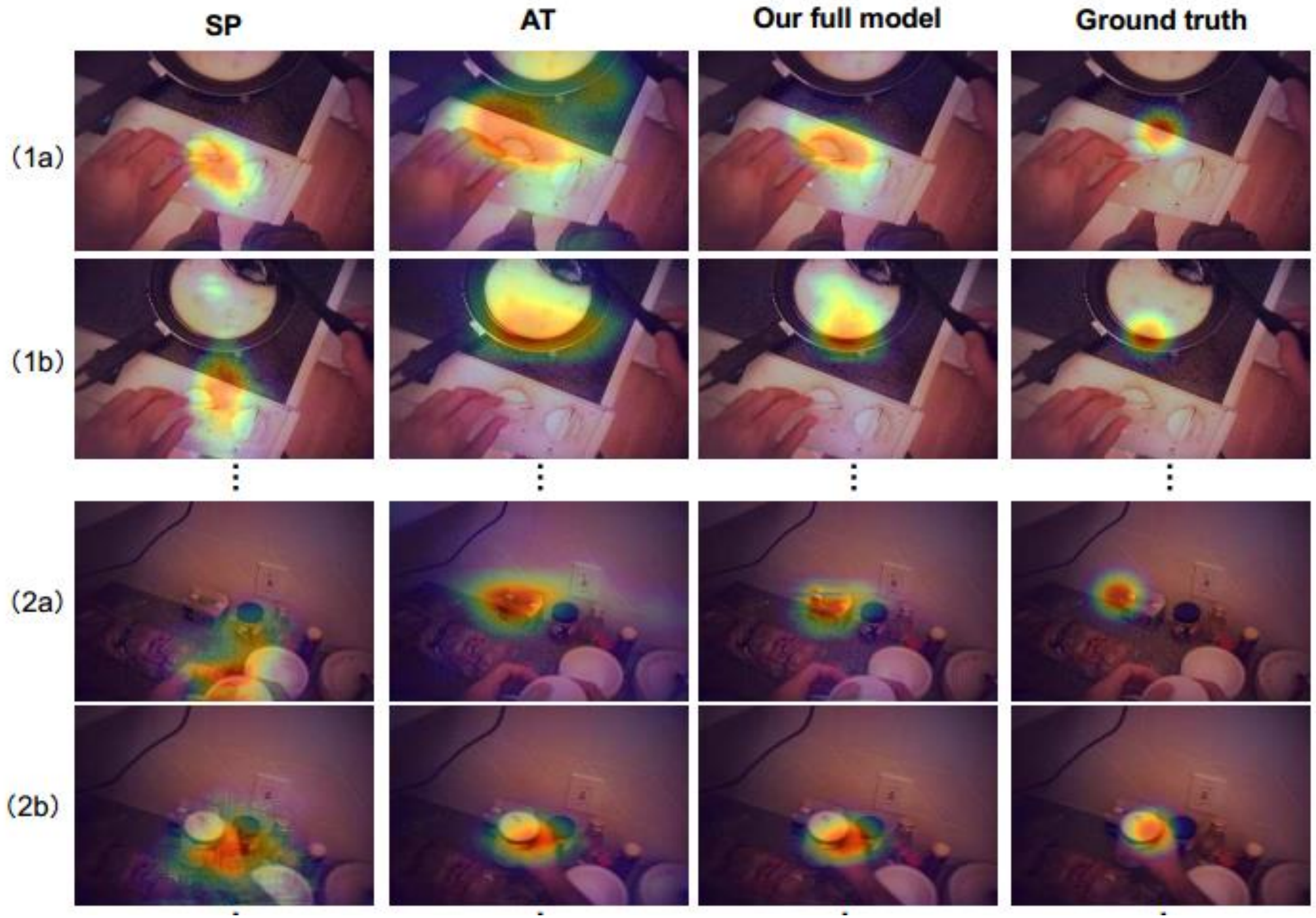
Attention transition module：对关注物体在时间上的转移过程建模

- Channel weight extractor: 在注视点附近对上一帧的feacture map进行裁剪，取平均得到一个表示卷积层不同 channel 权重的向量。ω_{t-1}是注视点周围关注区域的特征表示
- Fixation state predictor：输出注视点的score $\in [0,1]$，tells how likely fixation is occurring in the frame t-1.
- 使用LSTM通过学习权重变化来学习注意力的转移。



Predicting Gaze in Egocentric Video by Learning Task-dependent Attention Transition

- 实验：GTEA Gaze，kitchen tasks



Predicting Gaze in Egocentric Video by Learning Task-dependent Attention Transition

- 实验：GTEA Gaze，kitchen tasks

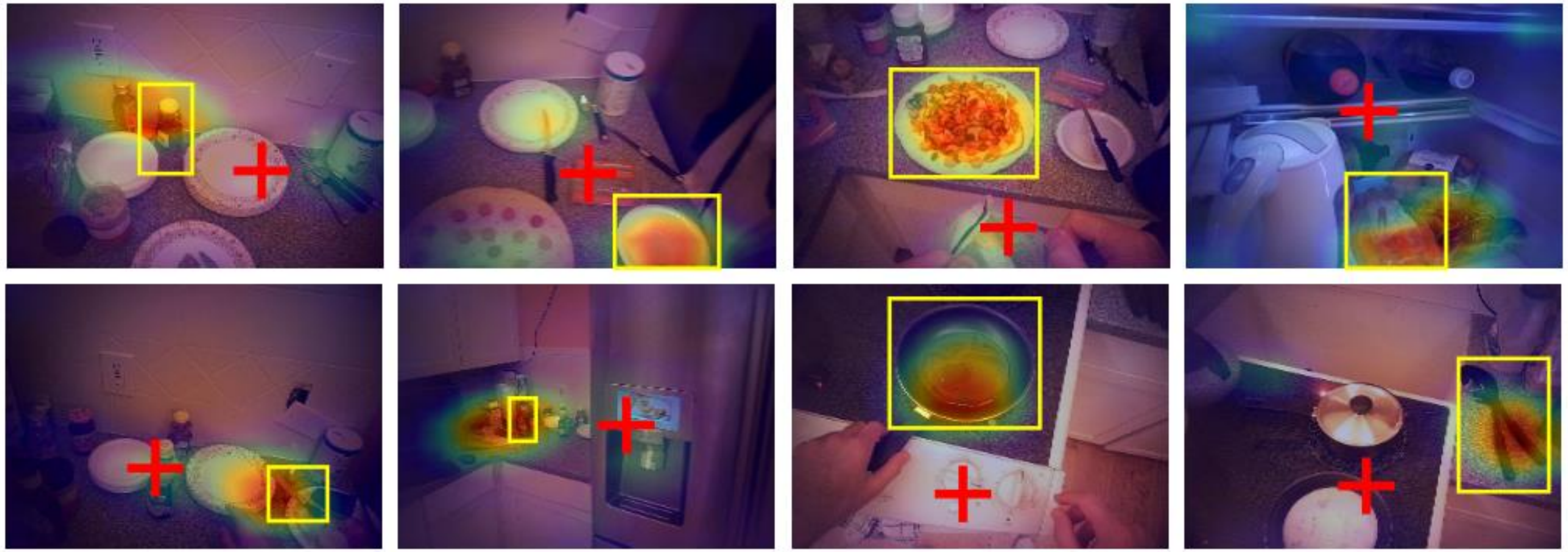


Fig. 5. Qualitative results of attention transition. We visualize the predicted heatmap on the current frame, together with the current gaze position (red cross) and ground truth bounding box of the object/region of the next fixation (yellow box).