Believe It or Not, We Know What You Are Looking at!

Dongze Lian, Zehao, Shenghua Gao
School of Information Science and Technology, ShanghaiTech University

In ACCV 2018

2019.10.08

Hanyang univ. AILAB 정지은

Introduction

Introduction

- Gaze-following task

Where are they looking?(NIPS 2015)

- 어떤 인물이 어느 지점을 보고있는지 맞추는 task
- 그 지점에 어떠한 물체가 있는지 인식하는 것과는 별개임
- 기본적으로 한 이미지 내부에 있는 오브젝트를 보고 있다고 가정(이미지 외부의 오브젝트는 논외)

Introduction

- Gaze-following task

(a)

(b)

(c)

(d)

- 인간의 시선에 근거하여 사람의 의도를 유추할 수 있음
- Retailing 관점에서 소비자가 어떠한 제품에 관심을 두는지 시선을 바탕으로 유추할 수 있음

Introduction

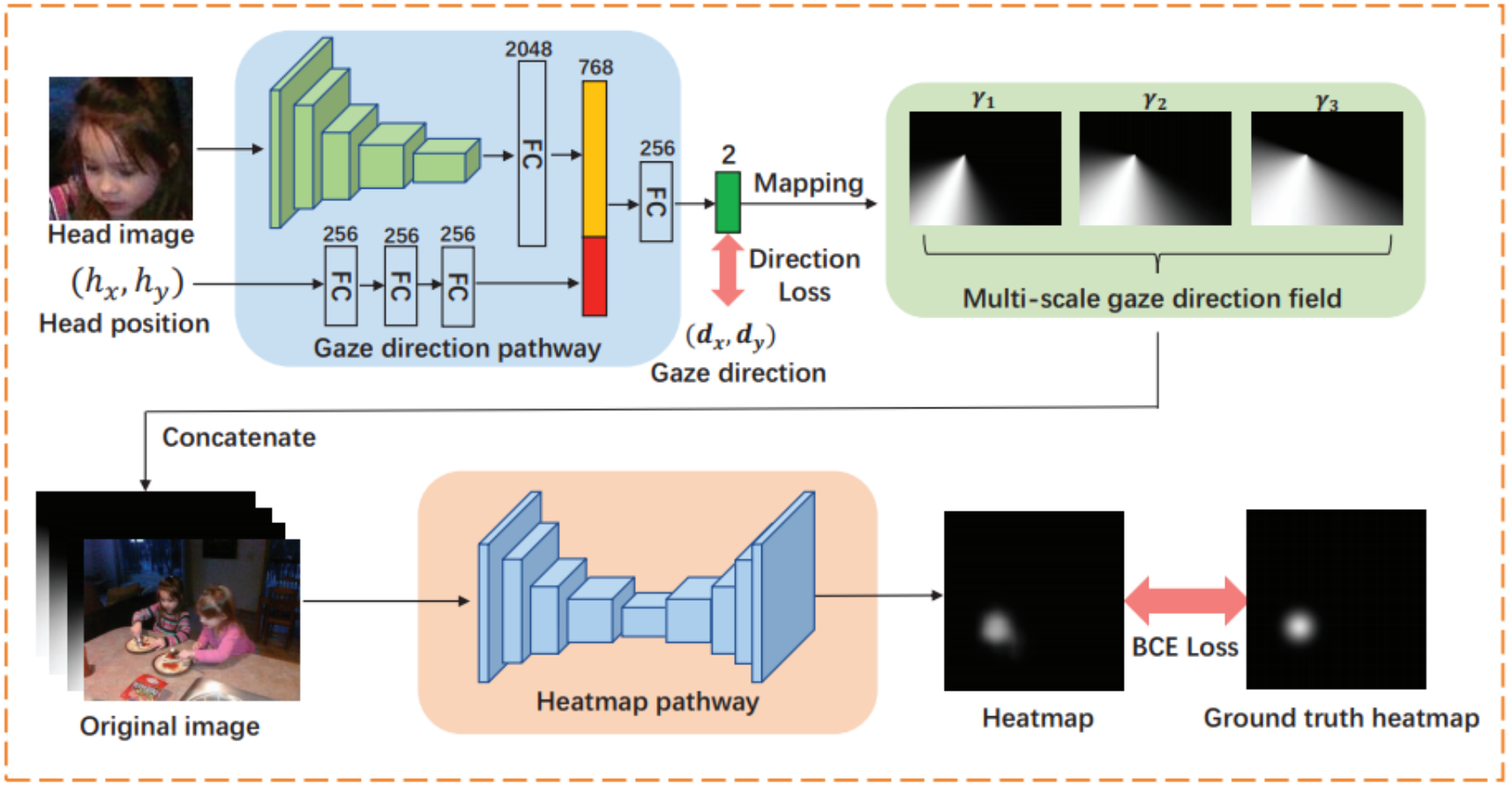
- Contributions

- **사람이 타인의 시선을 인식하는 방식을 모방하여 ‘심리적으로 직관적인’ 모델을 구현**
- **기저모델들보다 더 견고하게 학습이 가능하도록 모델을 변형하여 학습, Gaze-follow task에서 SOTA 찍음**
- **비디오 데이터셋 만듬 (95,000 frames)**

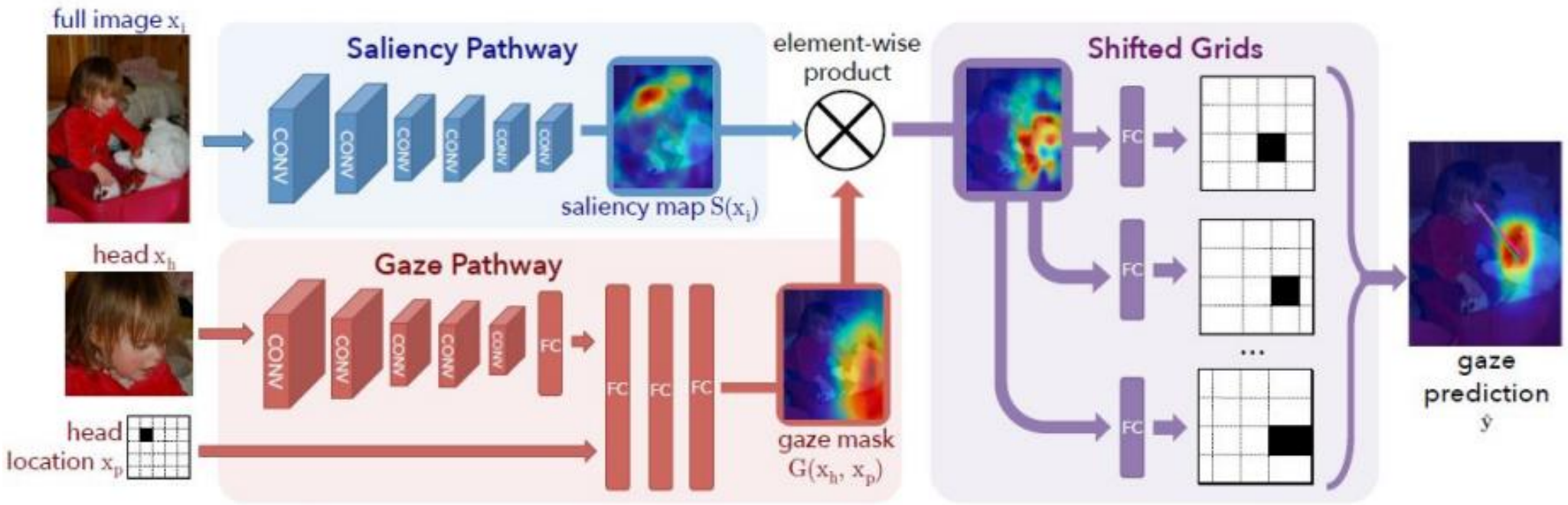
- (개인적으로) 모델구조의 변형을 위해 실험을 많이 했으며 이를 잘 정리한 논문임

Approach

Approach : Model Architecture

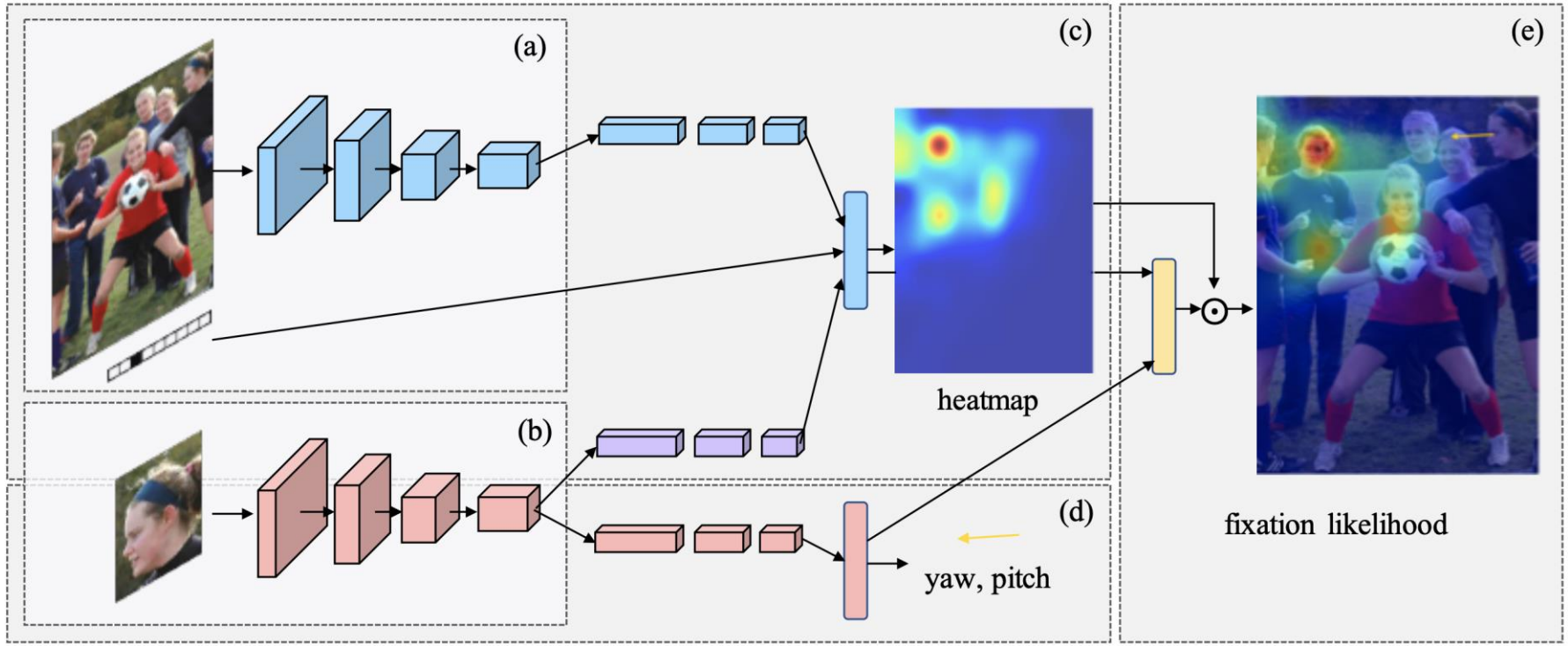


기저모델(NIPS 2015)



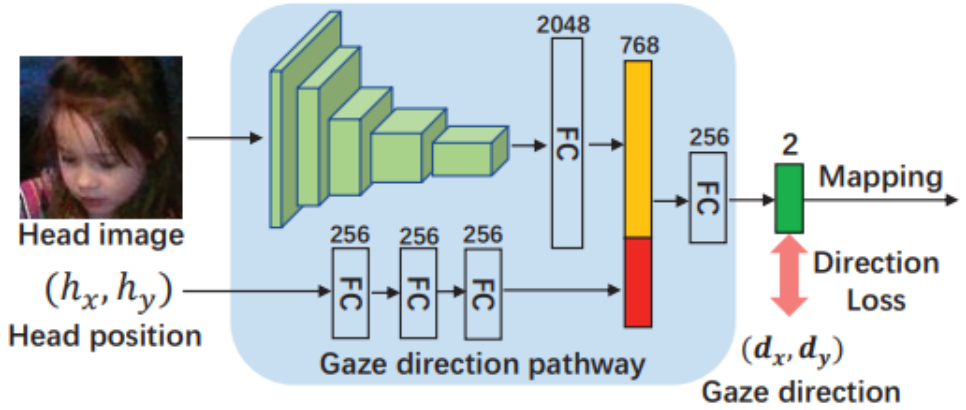
지난시간 발표했던 모델(CVPR 2018)

- Path (a), (b) : fully-conv layer (Resnet50)
- Path (d) : training for the gaze angle
- Path (c) : 2 conv pathways learn the heatmap
- Path (e) : learn "strength" of visual attention



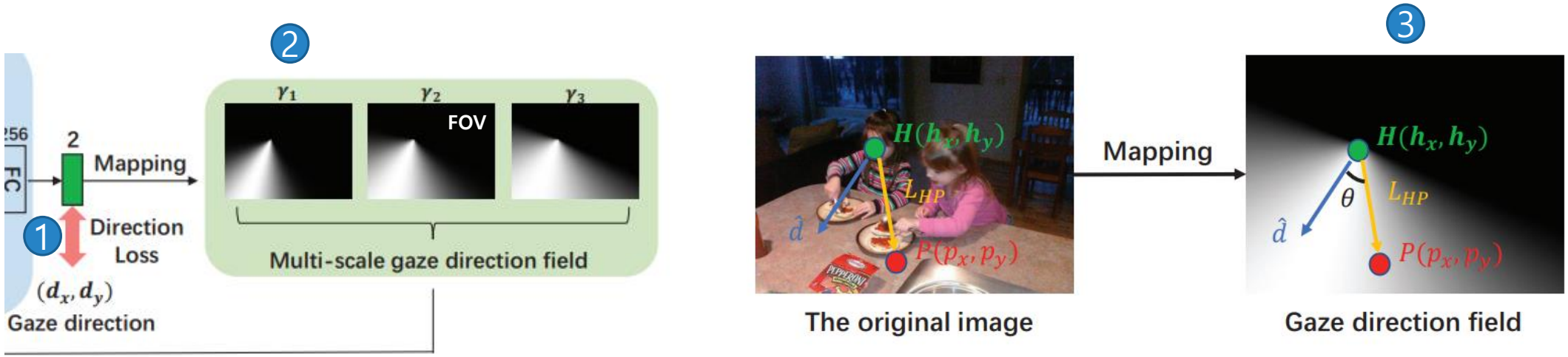
Yaw, pitch L1 loss

Approach : 3.1 Gaze direction pathway



- **Feature extractor : ResNet-50**
- **기저모델의 Gaze Mask가 아닌, Gaze direction 자체를 estimate 하는것에 집중함**
  - Gaze direction GT를 만들 수 있으므로 Direction Loss를 구할 수 있음
  - 전체 모델 중간에 지도학습하는 부분을 추가하여 더욱 Robust한 방식으로 학습이 가능
- **Head position도 기저모델에서는 5x5 matrix로 one-hot encoding 했지만, FC 3개를 거쳐 보다 세밀하게 위치정보를 encoding함**

Approach : 3.2 Gaze direction field



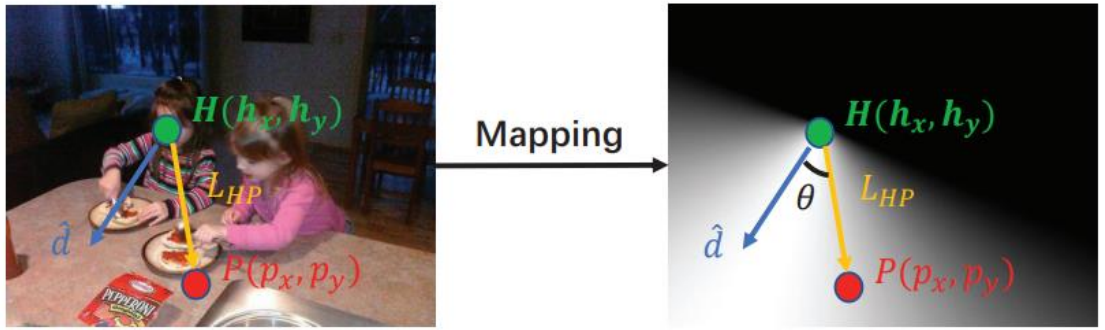
1. **Gaze direction 을 지도 학습함**
2. **Gaze direction 정보를 알면 머리위치로부터 cone 형태로 projection을 해서 FOV(field of view)를 만들 수 있음**
3. **Gaze direction field 의미 : 점 P(x,y)가 예측되었을때, 그 점이 정답일 확률들을 모아놓은 MAP**
   - L(HP) : 정답 시선벡터
   - d : 예측된 시선벡터
   - θ : 사이각
   - **각도로부터 확률값으로 매핑하기 위해 '코사인함수'를 사용**

Approach : 3.2 Gaze direction field

$$Sim(P) = \max\left(\frac{\langle G, \hat{d}\rangle}{|G||\hat{d}|}, 0\right)$$

the predicted gaze direction as $\hat{d} = (\hat{d}_x, \hat{d}_y)$,

L(HP) = $G = (p_x - h_x, p_y - h_y)$



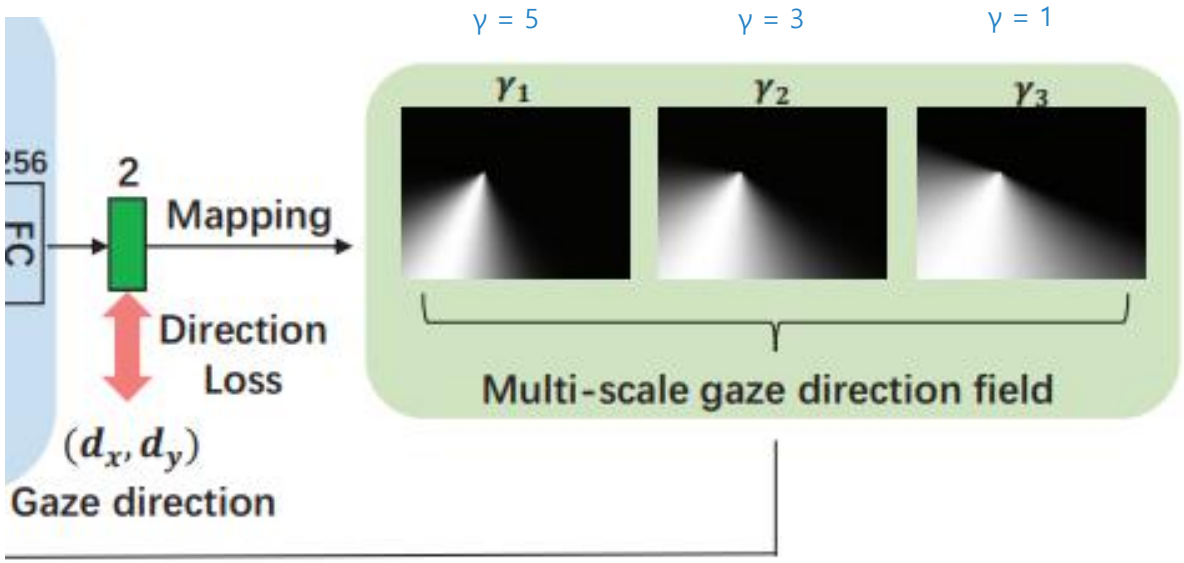
The original image

Gaze direction field

- θ 각이 작을수록, 코사인 유사도가 커지고 그 지점이 정답일 확률이 높음(90도가 넘어가면 음수이므로 0 처리)
- Gaze direction field = 점 P(x,y)가 예측되었을때, 그 점이 정답일 확률들을 모아놓은 MAP

.

Approach : 3.2 Gaze direction field

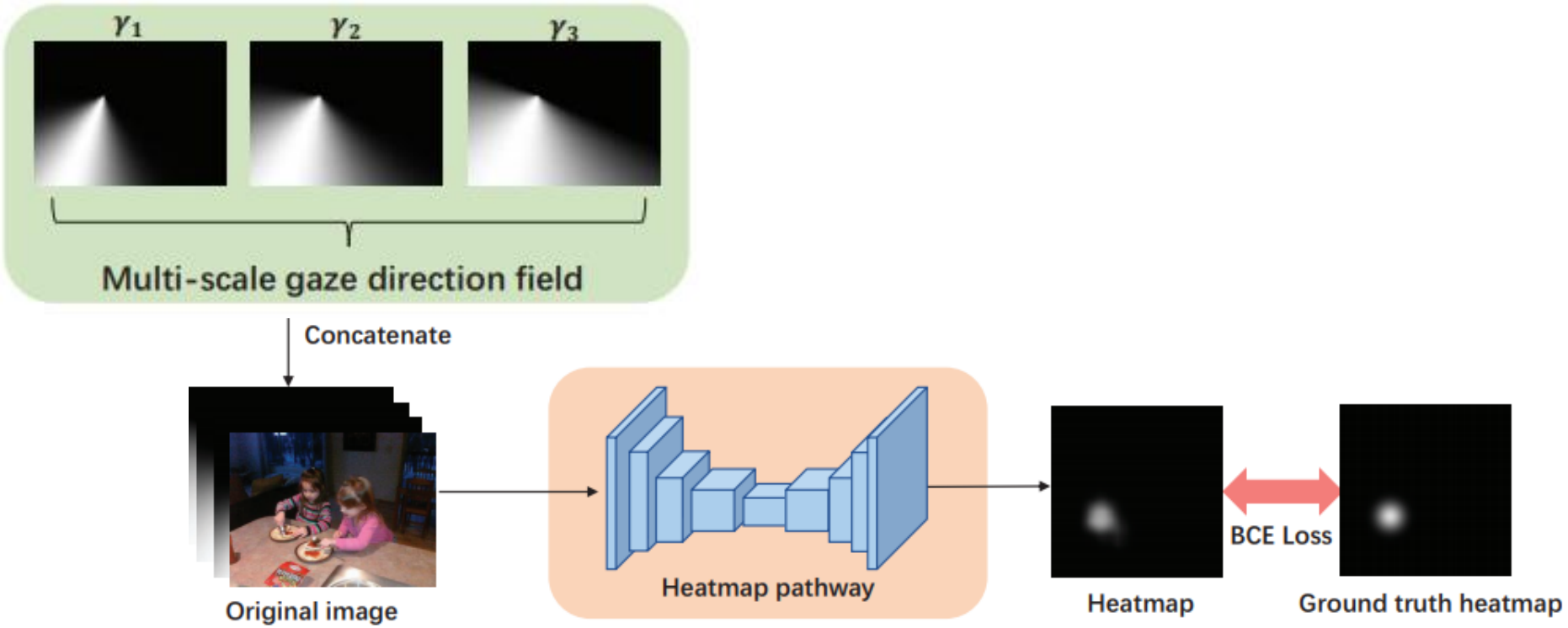


$$Sim(P, \gamma) = [Sim(P)]^{\gamma}$$

- 예측된 지점이 정답 시선지점일 확률이 Sim(P, γ)이며, 큰 γ 를 제곱해주면 확률의 분포의 절대값의 범위가 줄어드므로 더욱 좁은 영역에 집중해서 heatmap을 만들 수 있음
- 이때 하나의 Gaze direction field만 보지않고, 3가지 정도의 scale 을 가진 정보를 모두 주어서 학습하니 성능이 향상됨
- 만약 예측된 시선 방향이 정확하다면 시선 방향을 따라 cone의 형태를 더욱 좁혀 보게 될 것임

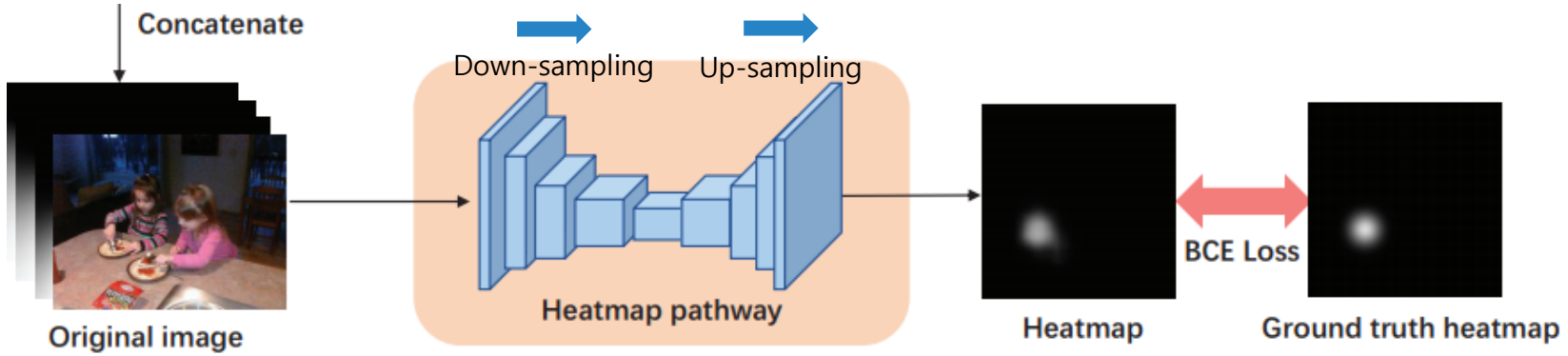
.

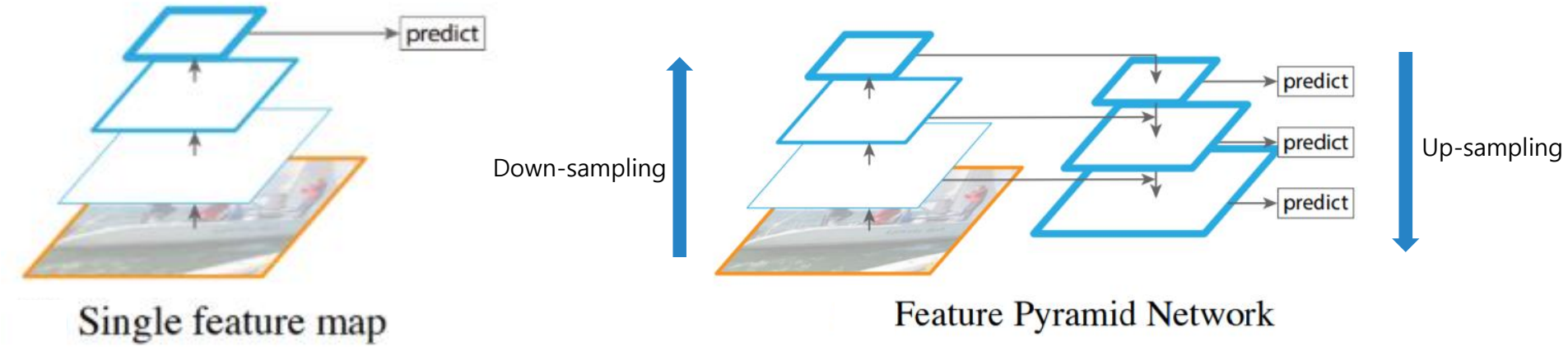
Approach : 3.3 Heatmap pathway



- Feature pyramid network(FPN) 사용
- 기저모델은 시선지점(x,y) 좌표를 예측하는 것을 objective로 학습했으나 본 모델은 Heatmap을 Objective로 함

FPN : Feature Pyramid Network





Approach : 3.3 Heatmap pathway

- **Objective 를 왜 변경했는가?**

  1. 점(x,y)를 예측하는 것은 높은 non-linear function을 학습해야 하는 일이므로 학습하기 어려움

  2. 풀려고 하는 문제의 목적에 집중해서 생각해보면,

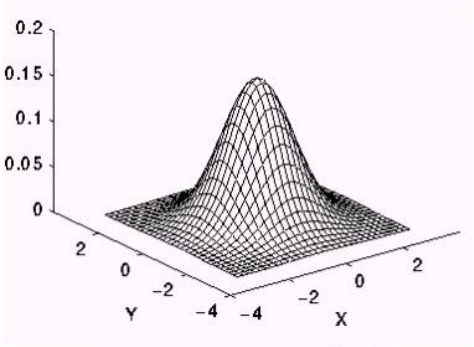
  - 인간도 타인의 시선을 정확히 산정해내기 어려움
  - 딥러닝 학습관점에서 보면 Output이 여러 개인 Multimodal task
  - 만약 heatmap을 objective로 학습한다면 heatmap의 사이즈에 따라 어느정도의 여유를 두고 학습할 수 있으므로 정답의 불명확성이라는 전제조건을 만족하면서도 더 안정적으로 학습할 수 있음

- **Heatmap 의 생성**

  주어진 데이터를 고차원 특징 공간으로 사상(매핑)해주는 '가우시안 커널'을 사용

$$H(i,j) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(i-g_x)^2 + (j-g_y)^2}{2\sigma^2}}$$

- H(i,j) : heatmap
- gx, gy : GT gaze point
- σ : variance of Gaussian kernel (set 3)



$$G(x,y) = \frac{1}{2\pi\sigma^2} \exp \frac{-(x^2+y^2)}{2\sigma^2}$$

Approach : 3.3 Heatmap pathway

- 뇌피셜이 아님
- 다양한 objective로 학습 후, **GT는 모두 gaze point를 기준**으로 평가함

Table 5. The evaluation of different objectives.

Methods	AUC	Dist	MDist	Ang	MAng
Point	0.892	0.173	0.103	21.9°	10.5°
Multi-task point	0.900	0.165	0.097	20.4°	10.1°
Shifted grid [22]	0.899	0.171	0.096	21.4°	10.3°
Heatmap (our)	**0.903**	**0.156**	**0.088**	**18.2°**	**9.2°**

* Multi-task regression: predicts both **gaze direction** and **gaze point simultaneously**

Approach : 3.4 Network training

- **Loss**

Heatmap loss : BCE_loss(predict_heatmap, gt_heatmap)

$$\ell_h = -\frac{1}{N}\sum_{i=1}^{N} H_i \log(\hat{H}_i) + (1 - H_i)\log(1 - \hat{H}_i)$$

N = heatmap size 56x56

Middle_angle_loss : 1 - cosine_similarity(direction, gt_direction)

gt_direction= gt_position – eye_position

GT → d, Predicted → $\hat{d}$

$$\ell_d = 1 - \frac{\langle d, \hat{d} \rangle}{|d||\hat{d}|}$$

최종 Loss : $\ell = \ell_d + \lambda \ell_h$

where λ is the weight to balance ℓ_d and ℓ_h. We set $\lambda = 0.5$ in our experiments.

Experiments

Experiments

One-scale and multi-scale correspond to the number of gaze direction fields in our model. For one-scale model, $\gamma = 1$.

Methods	AUC	Dist	MDist	Ang	MAng
Center [22]	0.633	0.313	0.230	49.0°	-
Random [22]	0.504	0.484	0.391	69.0°	-
Fixed bias [22]	0.674	0.306	0.219	48.0°	-
SVM + one grid [22]	0.758	0.276	0.193	43.0°	-
SVM + shift grid [22]	0.788	0.268	0.186	40.0°	-
Judd *et al.* [9]	0.711	0.337	0.250	54.0°	-
SalGAN [19]	0.848	0.238	0.192	36.7°	22.4°
SalGAN for heatmap	0.890	0.181	0.107	19.6°	9.9°
Recasens *et al.* [22]	0.878	0.190	0.113	24.0°	-
Recasens *et al.** [22]	0.881	0.175	0.101	22.5°	11.6°
One human [22]	0.924	0.096	0.040	11.0°	-
Ours (one-scale)	0.903	0.156	0.088	18.2°	9.2°
Ours (multi-scale)	**0.906**	**0.145**	**0.081**	**17.6°**	**8.8°**

Experiments : Ablation study

- 입력 데이터 or 구조의 요소를 변경하며 실험

Table 3. The results of ablation study.

	Methods	AUC	Dist	MDist	Ang	MAng
1	Original image	0.839	0.212	0.146	32.6°	21.6°
2	Original image + ROI head	0.887	0.182	0.118	22.9°	10.7°
3	W/O mid-layer supervision	0.875	0.178	0.101	24.4°	12.5°
4	Ours (one-scale)	**0.903**	**0.156**	**0.088**	**18.2°**	**9.2°**

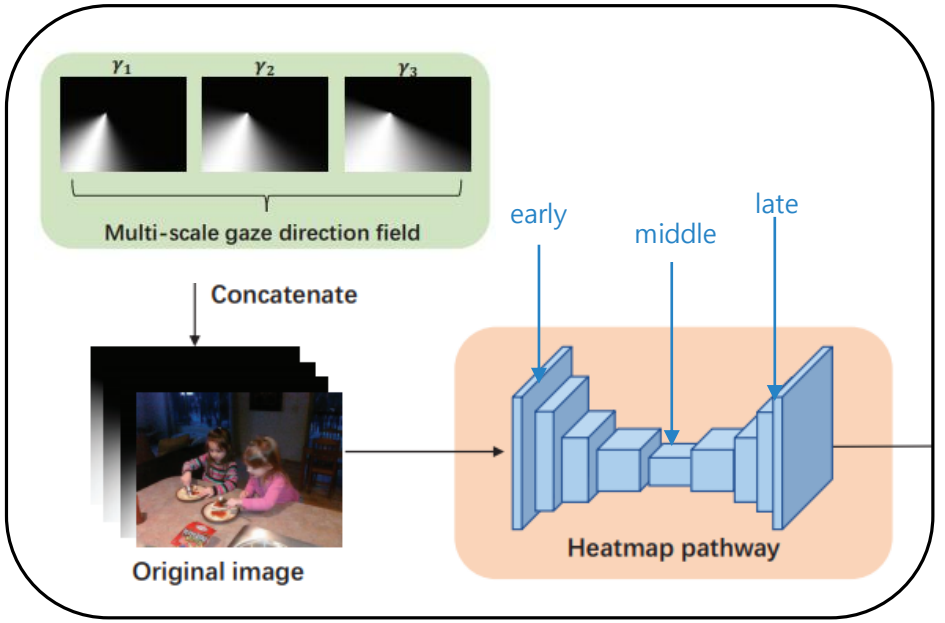
- Original image : 전체 이미지만 주고 feature extract 한다음 heatmap 생성
- Original image + ROI head : 전체 이미지를 입력으로 주고, 머리부분 영역의 피쳐를 뽑아서
  gaze direction regression하도록 multi-task learning
- W/O mid-layer supervision : 중간에 gaze direction 을 따로 지도학습 하지 않고, 하나의 direction field를 가지고 학습

Experiments : Ablation study

- Information fusion 방식을 변경하며 실험

Table 4. Different information fusion strategies.

Methods	AUC	Dist	MDist	Ang	MAng
Middle fusion (mul)	0.882	0.183	0.118	21.7°	10.7°
Middle fusion (concat)	0.884	0.177	0.105	21.0°	10.5°
Early fusion (mul)	0.898	0.160	0.098	18.7°	9.6°
Late fusion (mul)	0.888	0.176	0.102	20.1°	10.1°
Image fusion (mul)	0.895	0.163	0.096	19.3°	9.7°
Ours (concat)	**0.903**	**0.156**	**0.088**	**18.2°**	**9.2°**



1. How to choose the position : early, middle, late fusion

=> Heatmap encoding부분의 앞단에서 fusion하는 것이 쓸모없는 장면맥락을 최대한 억제하고 heatmap을 예측하므로 성능이 좋음

2. Way : multiplication or concatenation

⇒ Multiple은 예측된 gaze direction이 정확하지 않을때 이미지와 곱해지면 픽셀의 강도를 변화시키고 정보손실을 불러옴

⇒ Concat은 모든 정보가 그대로 보존되고, Gaze direction이 살짝 부정확하더라도, heatmap path 에서 올바르게 교정할 기회가 존재할 수 있으므로 더 성능이 좋은 것으로 추정됨

Experiments : 4.4 Visualization of predicted results

(a) Some accurate preditions.

Experiments : 4.4 Visualization of predicted results

(b) Some failures.

Thank You

Reference

- Paper : https://arxiv.org/pdf/1907.02364.pdf
- Paper2 (2015) : http://people.csail.mit.edu/khosla/papers/nips2015_recasens.pdf
- Paper3 (2018) : https://arxiv.org/pdf/1807.10437.pdf
- FPN 참고 : https://eehoeskrap.tistory.com/300
- FPN 참고2 : http://jeonseoungseon.blogspot.com/2017/06/fpn.html