Two-Stage Learning to Predict Human Eye Fixations via SDAEs

IEEE'16

cited 40

Junwei Han, Dingwen Zhang, Shifeng Wen, Lei Guo, Tianming Liu

Senior Member, IEEE

Xuelong Li

Fellow, IEEE

Background

- Visual Attention (Human Eye Fixations) - select information from visual input, where redundant information is filtered out

- Saliency model
  - Eye fixation prediction
  - Salient object detection

Motivation

- Previous studies of saliency detection
  - use hand-crafted features
  - contrast inference mechanisms
  - contrast integration

- To design powerful hand-crafted features and contrast inference mechanisms
  - domain-specific knowledge required
  - lack of understanding of the biological knowledge of human visual attention

  → Learn optimal features and contrast inference mechanism from image data by itself

Problem formulation

- Input : Image
- Output: Eye fixation maps

Outline

- **Related work**
- Eye fixation prediction Framework
  - SDAE
  - Learning stage 1 - Learning Feature Representation
  - Learning stage 2 - Learning mechanism for Contrast Inference and Integration
- Experiments
- Conclusion

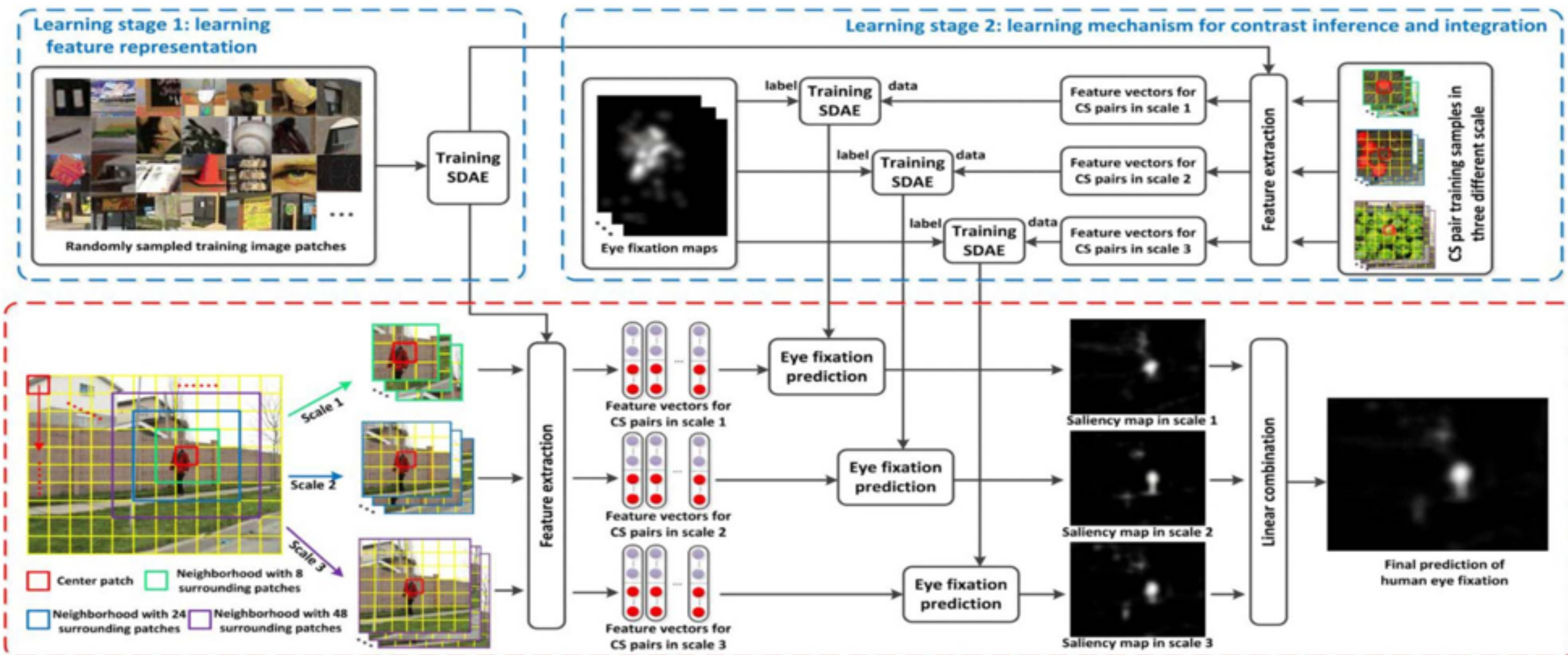
Related work

- Local contrast-based method - computing the contrast of an image location against its local and small neighbourhood
- Global contrast-based method - rarity of locations over the entire image for saliency prediction
- Combined local and global contrasts

Outline

- Related work
- **Eye fixation prediction Framework**
  - SDAE
  - Learning stage 1 - Learning Feature Representation
  - Learning stage 2 - Learning mechanism for Contrast Inference and Integration
- Experiments
- Conclusion

Eye fixation prediction Framework



SDAE - Stacked denoising autoencoders

- Autoencoder - one type of neural network
    - capture the informative hidden patterns and obtain powerful representation
- Goal
    - retain a significant amount of information from the original input
    - learned feature is sparse enough for powerful representation

SDAE - Stacked denoising autoencoders (cont.)

- Framework of auto encoder
  - stochastic mapping

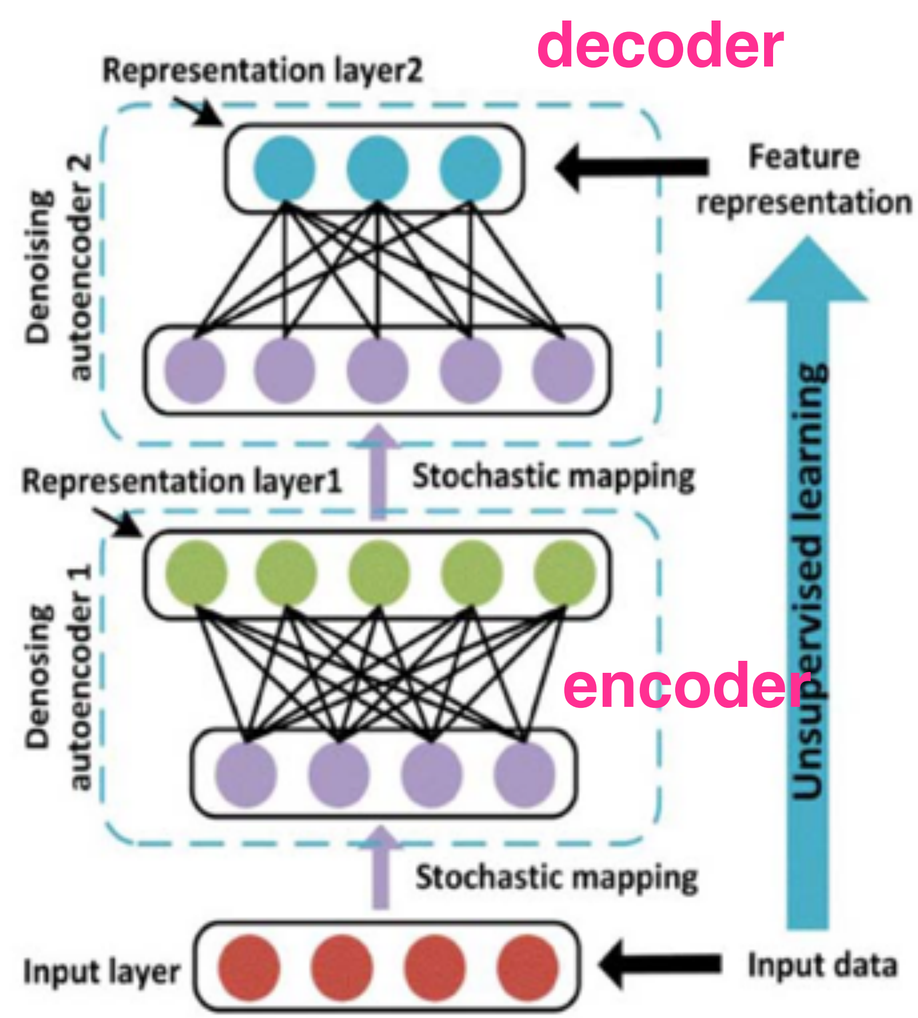
$$\tilde{\boldsymbol{x}}_i = qD(\tilde{\boldsymbol{x}}_i|\boldsymbol{x}_i)$$

  - encoder procedure - nonlinear mapping function

$$\boldsymbol{y}_i = f(\tilde{\boldsymbol{x}}_i, \theta_f) = sigm\left(\mathbf{W}^{(1)}\tilde{\boldsymbol{x}}_i + \boldsymbol{b}^{(1)}\right)$$

  - decoder procedure - nonlinear mapping function

$$z_i = g(\boldsymbol{y}_i, \theta_g) = sigm\left(\mathbf{W}^{(2)}\boldsymbol{y}_i + \boldsymbol{b}^{(2)}\right).$$



(a)

SDAE - Stacked denoising autoencoders (cont.)

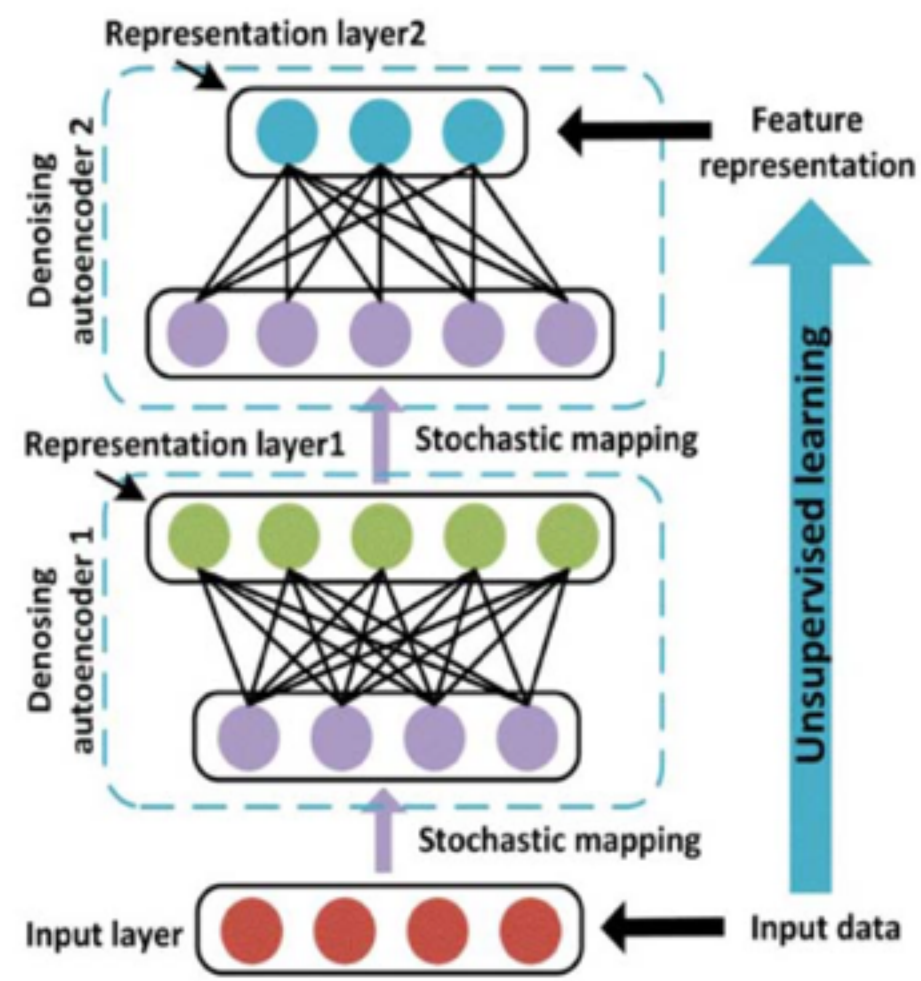
- Loss function

$$L = \frac{1}{2}\sum_{i=1}^{m} ||\boldsymbol{x}_i - \boldsymbol{z}_i||_2^2$$

enhance the probability of linear separability
→ add sparsity constraint

$$L_s = \frac{1}{2}\sum_{i=1}^{m} ||\boldsymbol{x}_i - \boldsymbol{z}_i||_2^2 + \beta \sum_{j=1}^{N} \mathrm{KL}(\rho||\hat{\rho}_j) + \omega \sum_{i=1}^{T}\sum_{j=1}^{N}\left(W_{ij}^{(1)}\right)^2$$

$$\mathrm{KL}(\rho||\hat{\rho}_j) = \rho\log\frac{\rho}{\hat{\rho}_j} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}_j}$$



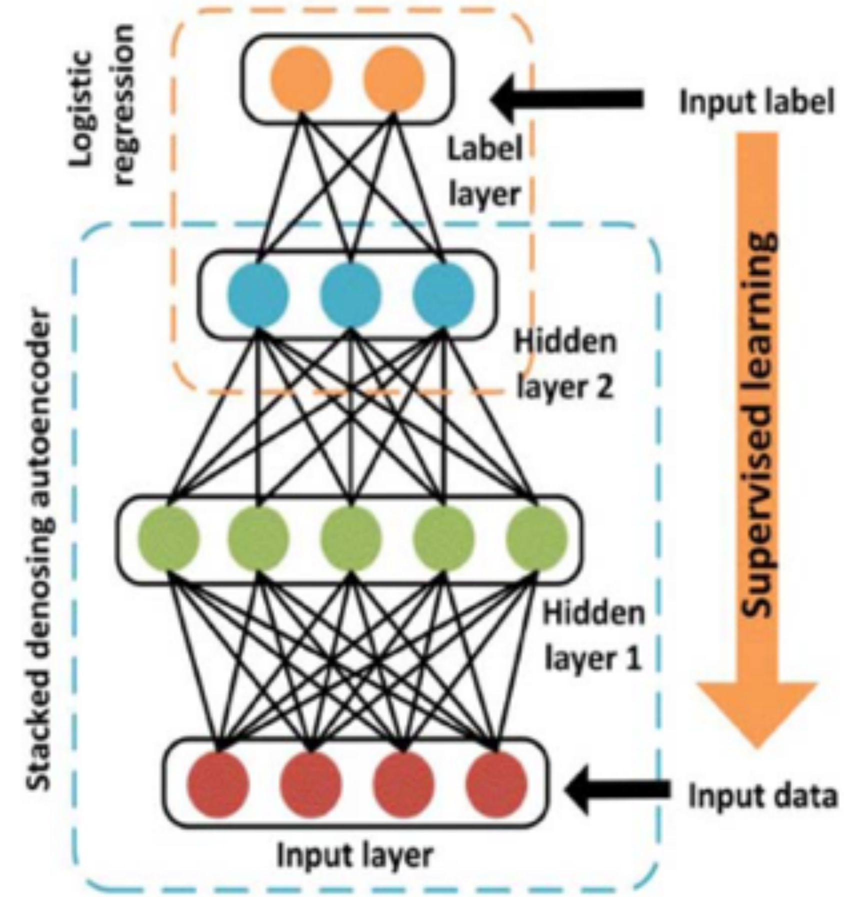
(a)

SDAE - Stacked denoising autoencoders (cont.)

- Framework of SDAE

$$h_{\Theta}\big(H_{V,d}(\boldsymbol{x}_i)\big) = \frac{1}{1+\exp\big(-\Theta^T H_{V,d}(\boldsymbol{x}_i)\big)}$$

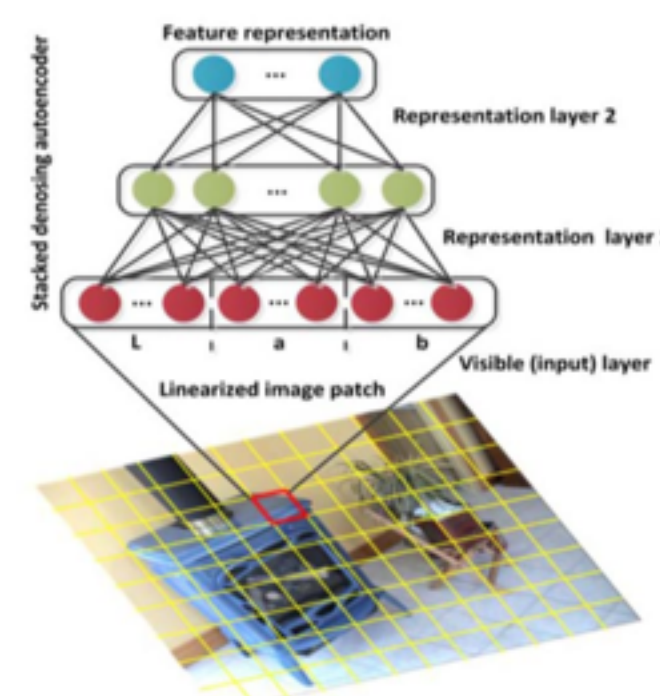
$$J = -\frac{1}{m}\Bigg[\sum_{i=1}^{m} \ell_i \log h_{\Theta}\big(H_{V,d}(\boldsymbol{x}_i)\big) + (1-\ell_i)\log\Big(1-h_{\Theta}\big(H_{V,d}(\boldsymbol{x}_i)\big)\Big)\Bigg] + \omega \sum_{k=1}^{R-1}\sum_{i=1}^{S_k}\sum_{j=1}^{S_{k+1}}\Big(Q_{ij}^{(k)}\Big)^2$$



(b)

Learning stage 1 - *Learning Feature Representation*

- Train SDAE
  - Randomly select 300 square image patches with the size of 8 × 8 pixels from each training image
  - Concatenate all the pixel values in each color channel



Learning Stage 2: *Learning Mechanism for Contrast Inference and Integration*

- Contrast - the most significant factor to direct free-viewing visual attention

- Contrast inference - limited understanding of human attention mechanism

  → *abstract informative patterns hierarchically by SDAE*

  → *learn complex mapping relations between the designed CS pair input data and its eye fixation labels*

  CS pair - center surrounding pair

- Contrast inference and integration are addressed jointly in second learning stage

Learning Stage 2: *Learning Mechanism for Contrast Inference and Integration (cont.)*

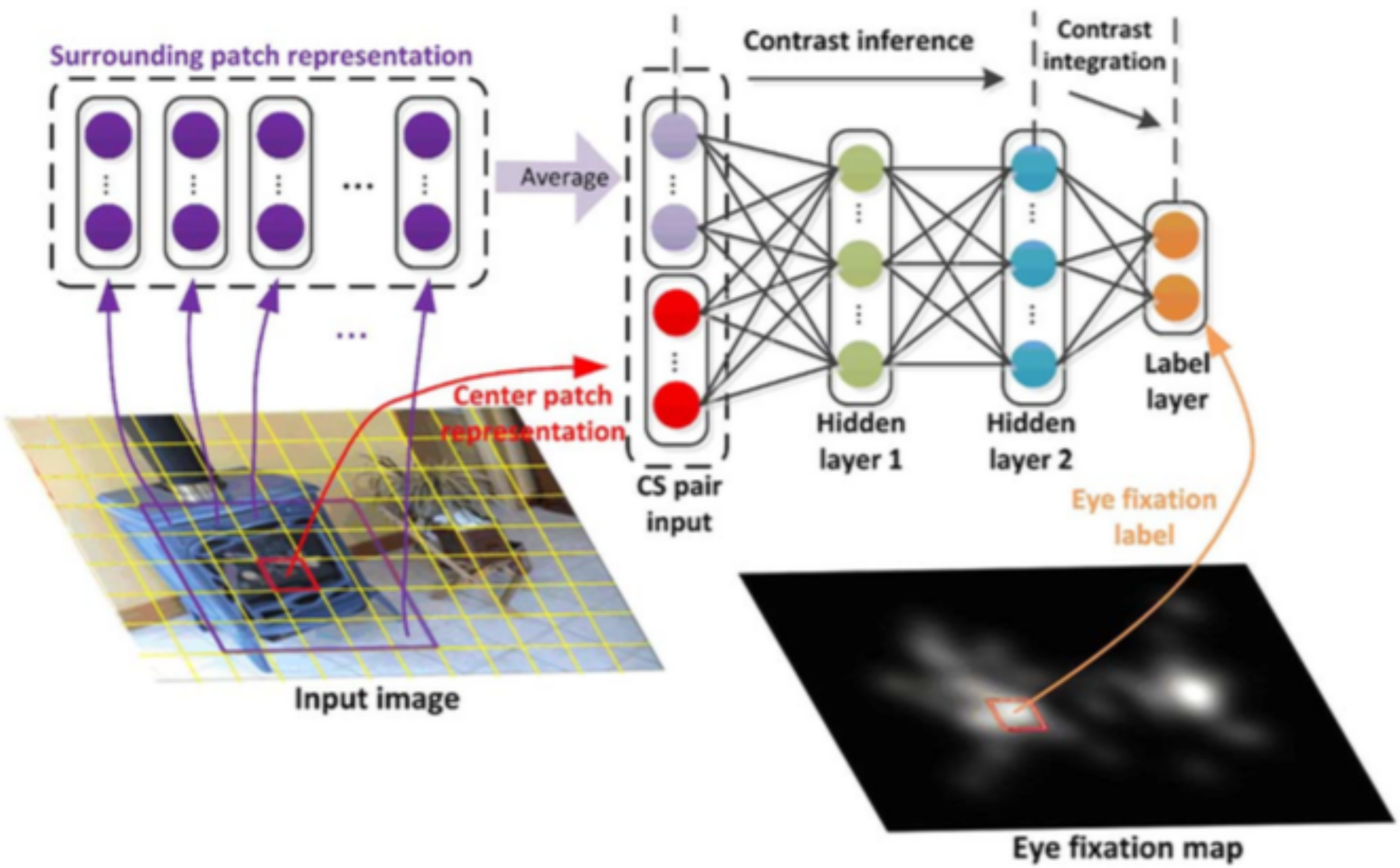
- Crop each square image patch with the size of 8 × 8 pixels centered at position of local maximum with its surrounding patches as one CS pair for generating positive examples (trained in different scale - **8,24,48**)

- Image patches in each CS pair are represented by the features learned in the first learning stage

- Train SDAE

Learning Stage 2: *Learning Mechanism for Contrast Inference and Integration (cont.)*

- Final saliency map is calculated by averaging each pixel from the saliency maps in three scales



Outline

- Related work
- Eye fixation prediction Framework
  - SDAE
  - Learning stage 1 - Learning Feature Representation
  - Learning stage 2 - Learning mechanism for Contrast Inference and Integration
- **Experiments**
- Conclusion

Experiments

- Publically available benchmark eye tracking datasets
  → (MIT) dataset,Toronto dataset, Cerf dataset

- Evaluation metrics - AUC
  → varying the quantization threshold within the range [0, 255]

$$\mathrm{TPR} = \frac{|\mathrm{SF} \cap \mathrm{PS}|}{|\mathrm{PS}|} \quad \mathrm{FPR} = \frac{|\mathrm{SF} \cap \mathrm{NS}|}{|\mathrm{NS}|}$$

Experiments(cont.)

TABLE I
HYPERPARAMETERS OF SDAE MODEL IN TWO LEARNING STAGES

	Learning stage 1		Learning stage 2	
	Representation layer 1	Representation layer 2	Hidden layer 1	Hidden layer 2
N	400	200	200	100
ε	.030	.040	.010	.010
ρ	.010	.010	.010	.010
β	.040	.005	.020	.020
ω	2e-4	2e-4	4e-4	2e-4

Experiments(cont.)

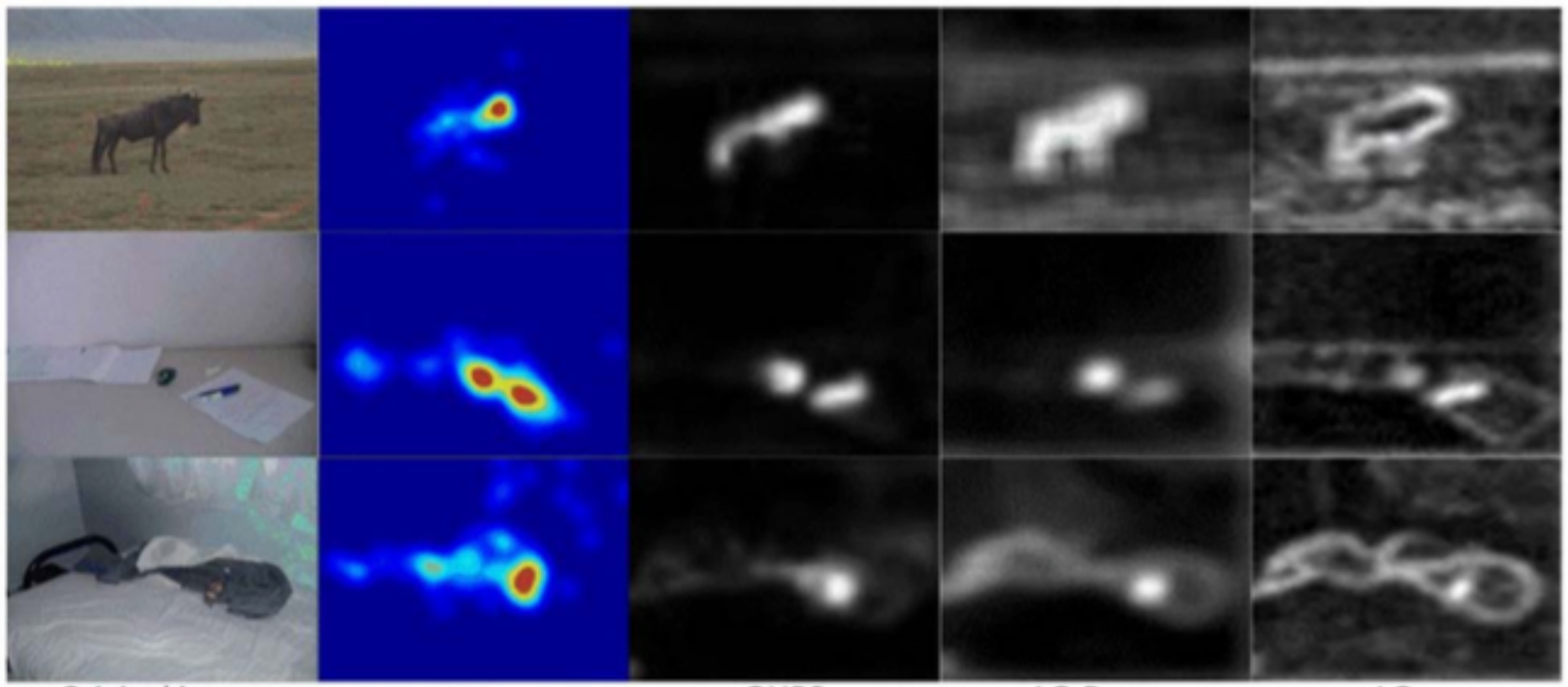


Fig. 5. Some experimental results of the LG method, the LG-deep method, and the proposed two-stage learning approach. GT denotes the ground-truth saliency map built by convolving the eye fixation locations with a Gaussian for smoothing, which is implemented by following [18], [54].

Experiments(cont.)

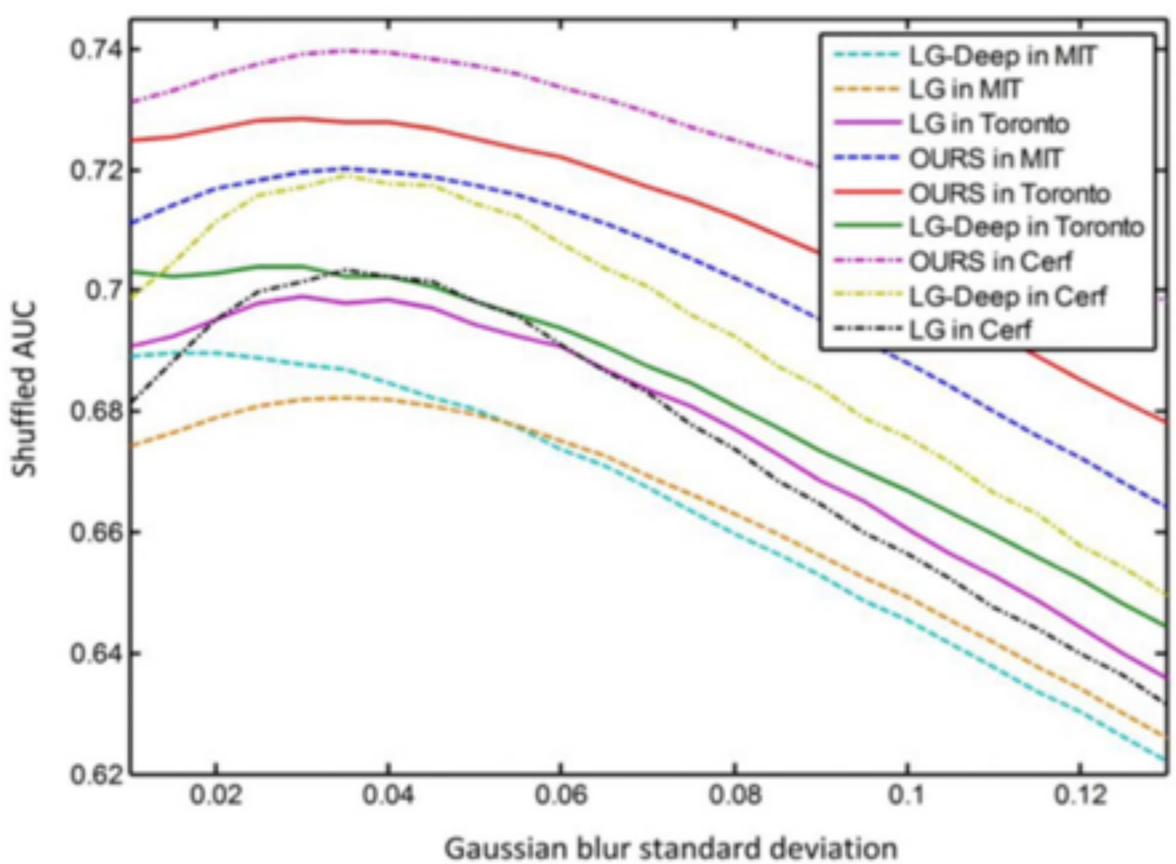


Fig. 6. Evaluation of the proposed feature representation over three datasets. x-axis represents the Gaussian blur standard deviation σ (in image width) by which maps are smoothed and y-axis represents the shuffled AUC score on one dataset.

TABLE II
MAXIMUM PERFORMANCE OF MODELS SHOWN IN FIG. 6. NUMBERS IN THE SECOND ROW OF EACH DATASET ARE THE OPTIMAL σ WHERE MODELS TAKE THE MAXIMUM PERFORMANCE

Dataset	LG	LG-Deep	OURS
MIT	.682	.690	.719
Opt. σ	.035	.015	-
Toronto	.699	.704	.728
Opt. σ	.030	.025	-
Cerf	.704	.719	.740
Opt. σ	.035	.035	-

Experiments(cont.)

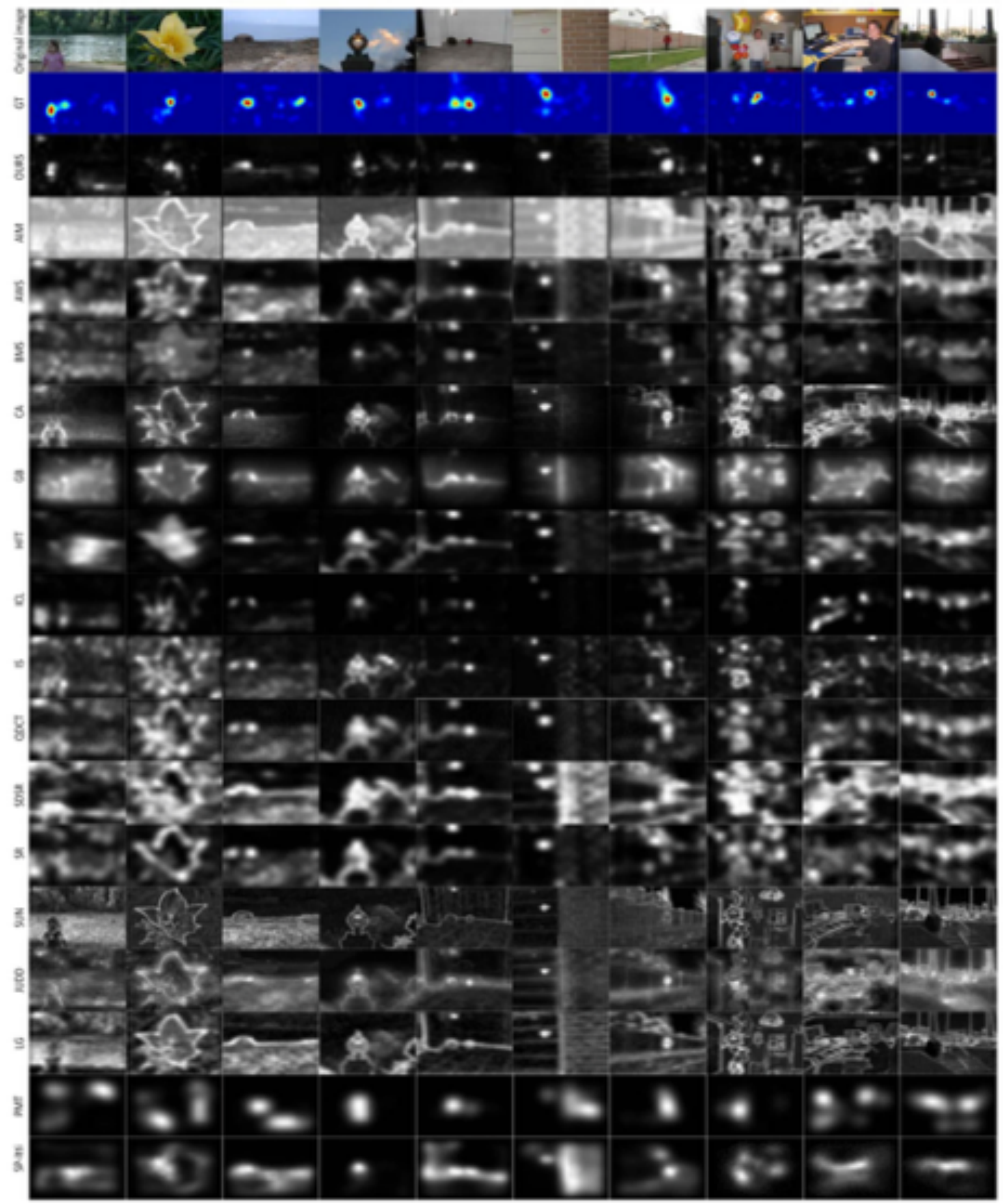
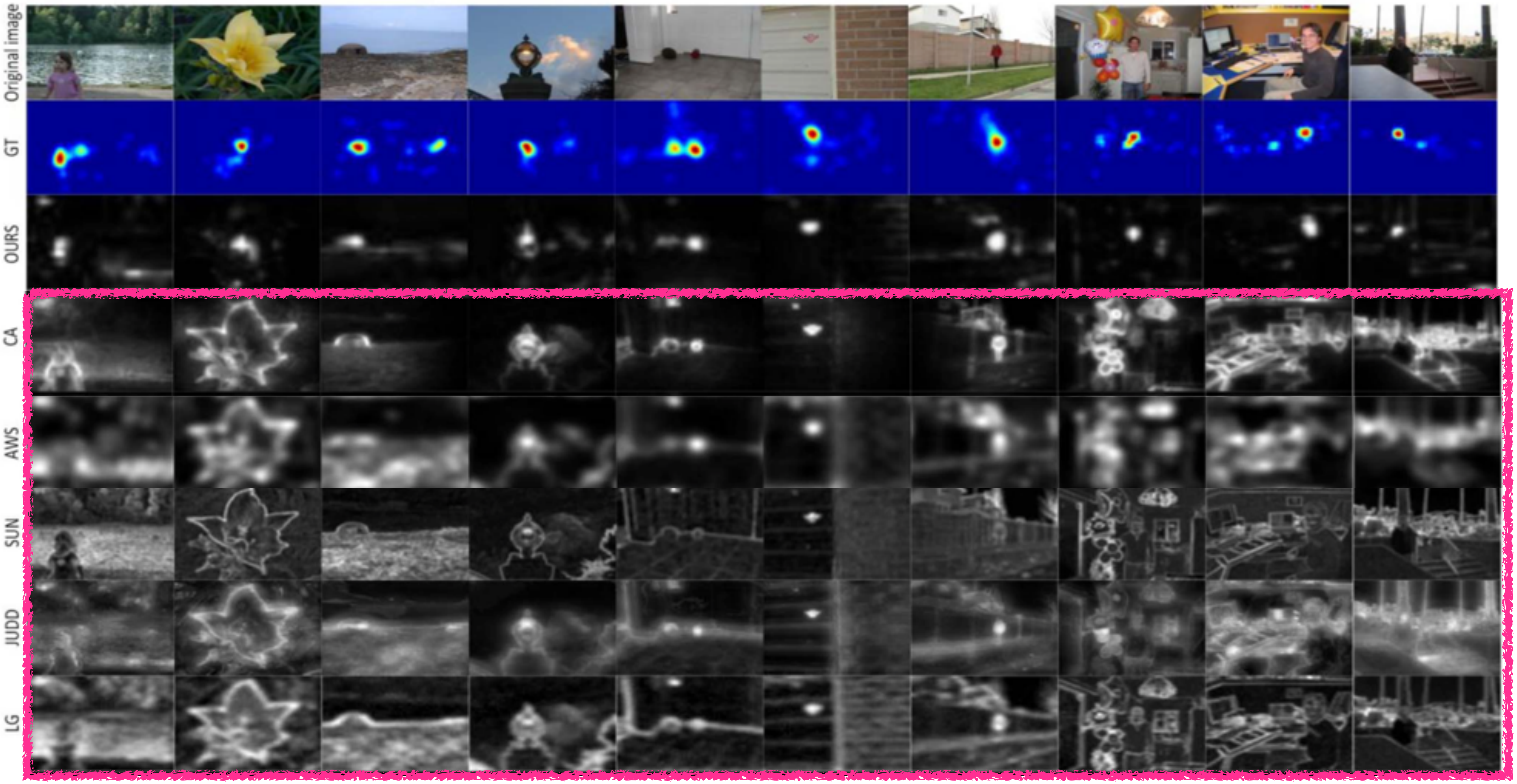
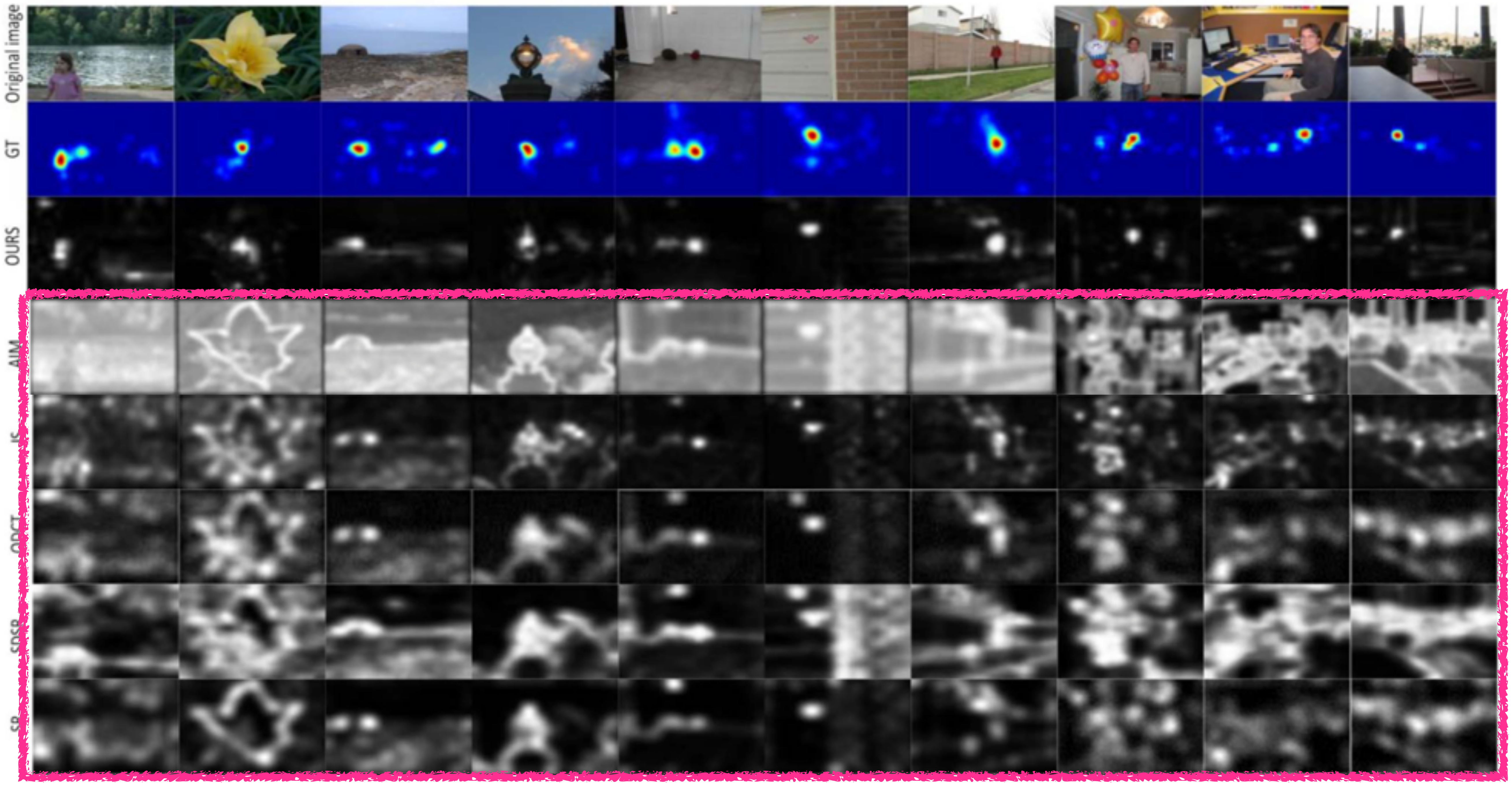


Fig. 7. Comparison results of 16 state-of-the-art approaches, ours, and the GT saliency map built by convolving the eye fixation locations with a Gaussian for smoothing [18], [54].

Experiments(cont.)



Experiments(cont.)



Experiments(cont.)

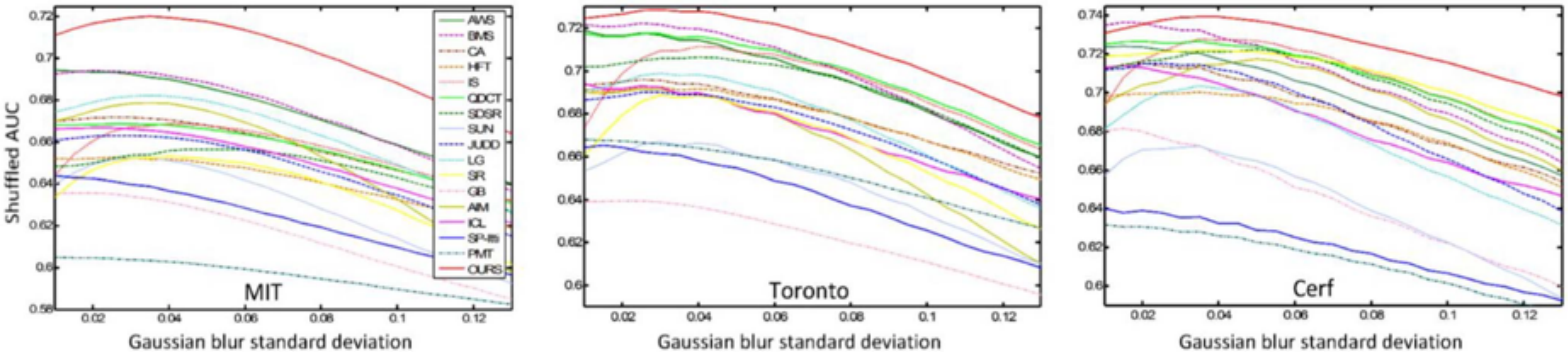


Fig. 8. Quantitative model comparisons. Fixation prediction accuracy of our saliency model along with 16 state-of-the-art models over three benchmark datasets. *x*-axis indicates the Gaussian blur standard deviation σ (in image width) by which maps are smoothed and *y*-axis indicates the shuffled-AUC score.

TABLE III
MAXIMUM PERFORMANCE OF MODELS SHOWN IN FIG. 8. NUMBERS IN THE SECOND ROW OF EACH DATASET ARE THE OPTIMAL σ WHERE MODELS TAKE THE MAXIMUM PERFORMANCE. ACCURACIES OF THE BEST MODELS OVER EACH DATASET ARE UNDERLINED AND SHOWN IN BOLD FACE FONT

Dataset	AIM	AWS	BMS	CA	GB	HFT	ICL	IS	JUDD	LG	*PMT*	QDCT	SDSR	*SP-Itti*	SR	SUN	OURS
MIT	.679	.695	.694	.672	.636	.653	.667	.669	.663	.682	*.605*	.669	.659	*.644*	.653	.652	**.719**
Opt. σ	.035	.010	.020	.025	.020	.025	.020	.040	.025	.035	*.010*	.025	.045	*.010*	.040	.030	-
Toronto	.692	.718	.722	.696	.640	.693	.694	.712	.690	.699	*.668*	.717	.707	*.665*	.689	.667	**.728.**
Opt. σ	.025	.010	.025	.025	.025	.030	.010	.040	.030	.030	*.010*	.025	.040	*.015*	.030	.030	-
Cerf	.716	.724	.736	.715	.681	.700	.714	.728	.715	.704	*.632*	.727	.726	*.640*	.722	.672	**.740**
Opt. σ	.050	.015	.015	.020	.015	.035	.015	.035	.025	.035	*.020*	.020	.035	*.010*	.040	.035	-
Average	.696	.712	.717	.694	.652	.682	.692	.703	.689	.695	*.635*	.704	.697	*.650*	.688	.664	**.729**

[1]In our experiments, we compared with the baseline model in SP approach [58], which is based on Itti's model.

Conclusion

- Suffer sufficient training data
- Used concepts from contrast inference mechanism