The Robustness of Deep Saliency Models

Zhaohui Che

MMLab, Institute of Image Communication and Network Engineering,

Shanghai Jiao Tong University, China (PhD candidate supervised by Prof. Guangtao Zhai);

LS2N lab, Polytech Nantes, France (Visiting student, supervised by Prof. Patrick Le Callet);

chezhaohui@sjtu.edu.cn

https://github.com/CZHQuality

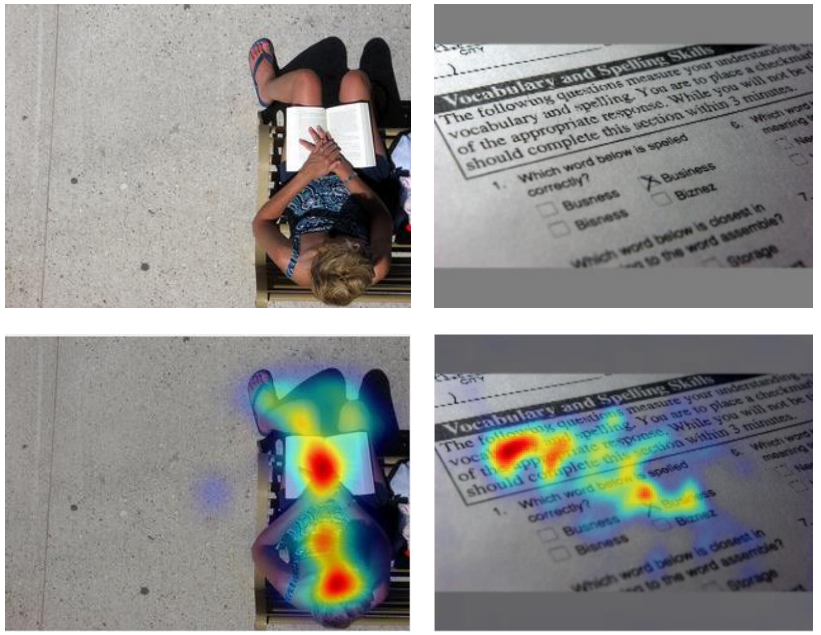
https://www.researchgate.net/profile/Zhaohui_Che

上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

1st Challenge: Robustness against Ubiquitous Transformations

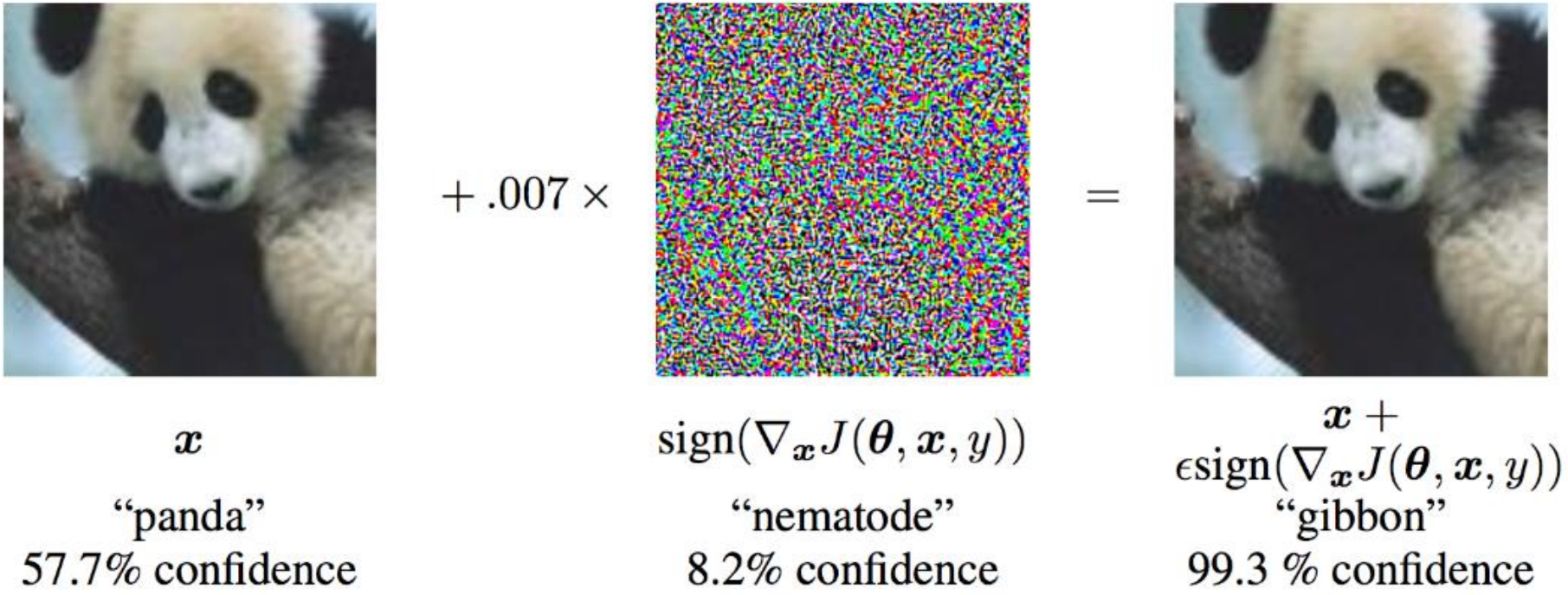
(a) Canonical high-quality stimulus



(b) Non-canonical stimulus with transformations

Most of current studies on human attention and saliency modeling have used canonical high-quality stimuli, as shown in (a). However, in real world, the captured images undergo **various transformations**, such as **viewpoint variation**, **poor photographing angles**, and **ubiquitous distortions and artifacts,** as shown in (b). The heat maps in the bottom row highlight the human gaze regions when viewing different stimuli.

2nd Challenge: Robustness against Malicious Adversarial Attack



Most of current deep neural networks are vulnerable to adversarial attacks: Adversarial Attacks are able to produce some deliberate and imperceptible adversarial perturbations. These adversarial perturbations are so subtle that a human observer cannot notice their presences, but the model output will be dramatically changed.

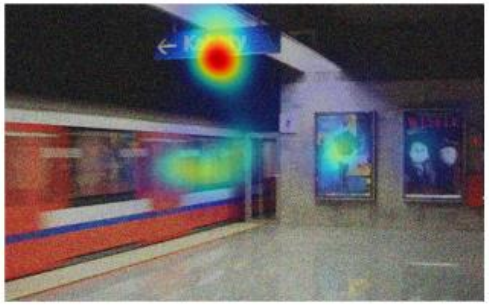
1st Challenge: *Ubiquitous Transformations*

1st Challenge: Eye-movement Data Analysis

(a) Original

(b) Ground Truth

Human Attention



(c) Noise

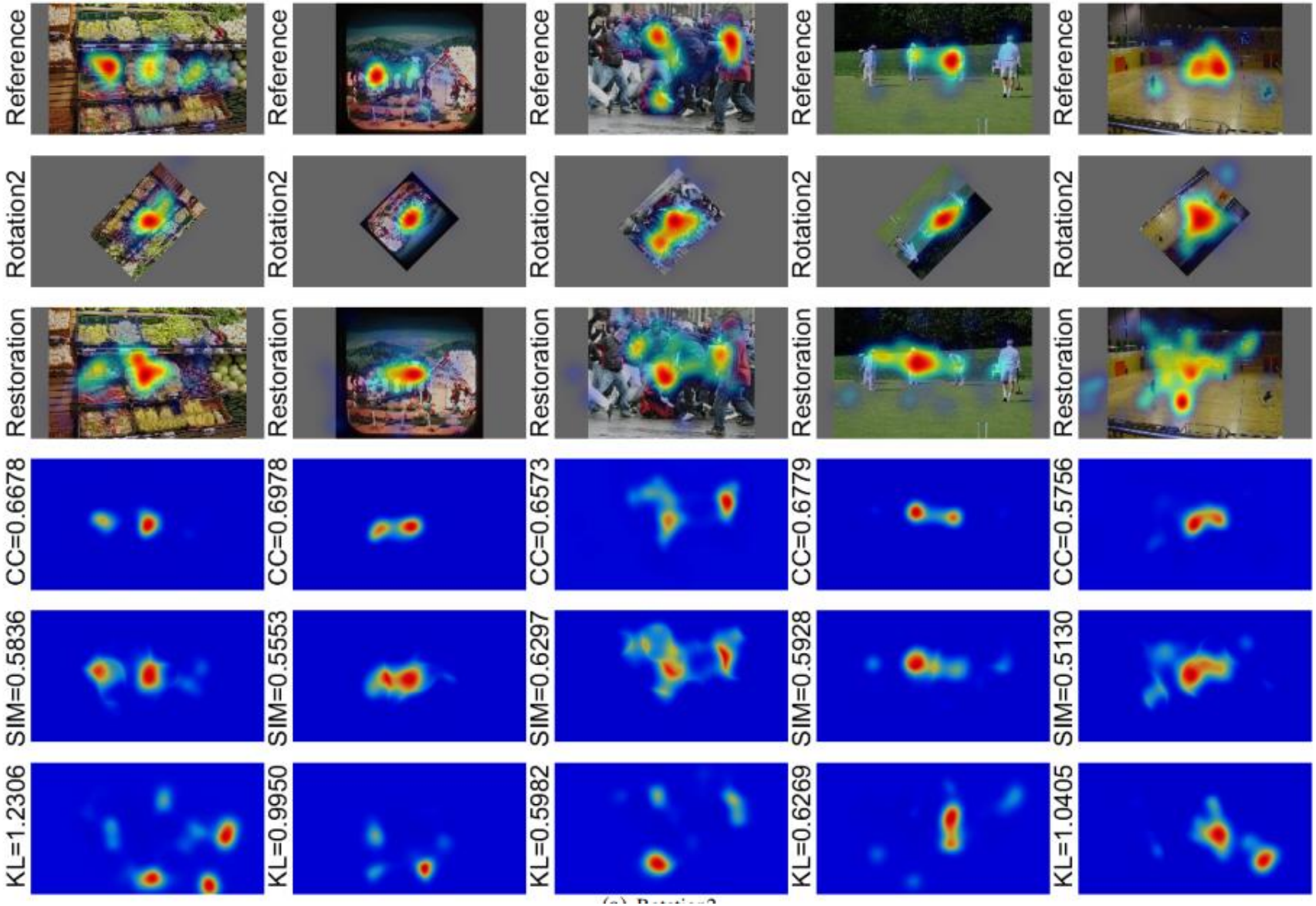
Model Attention

(d) SALICON Model

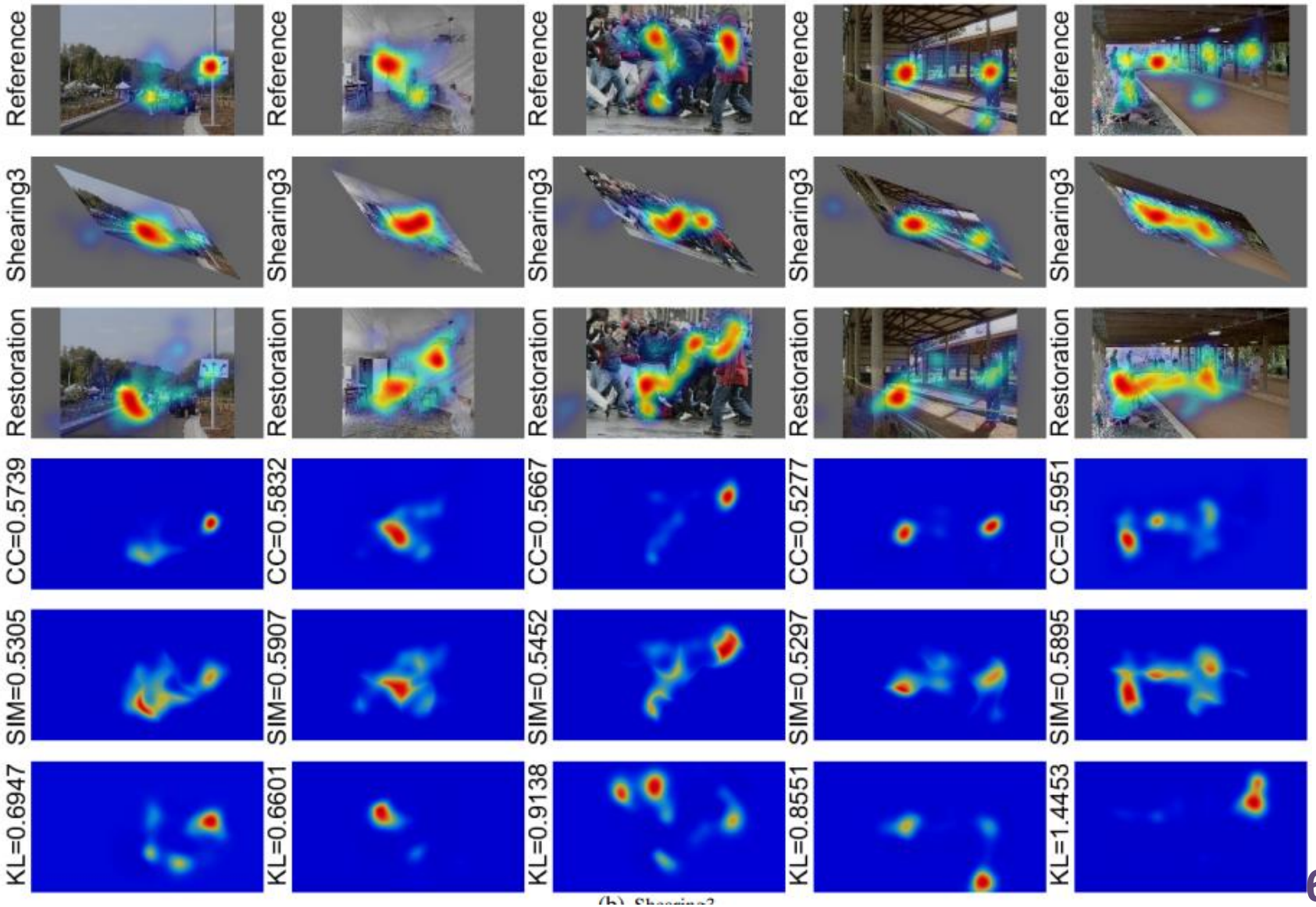
(e) Cropping

(f) Our Model

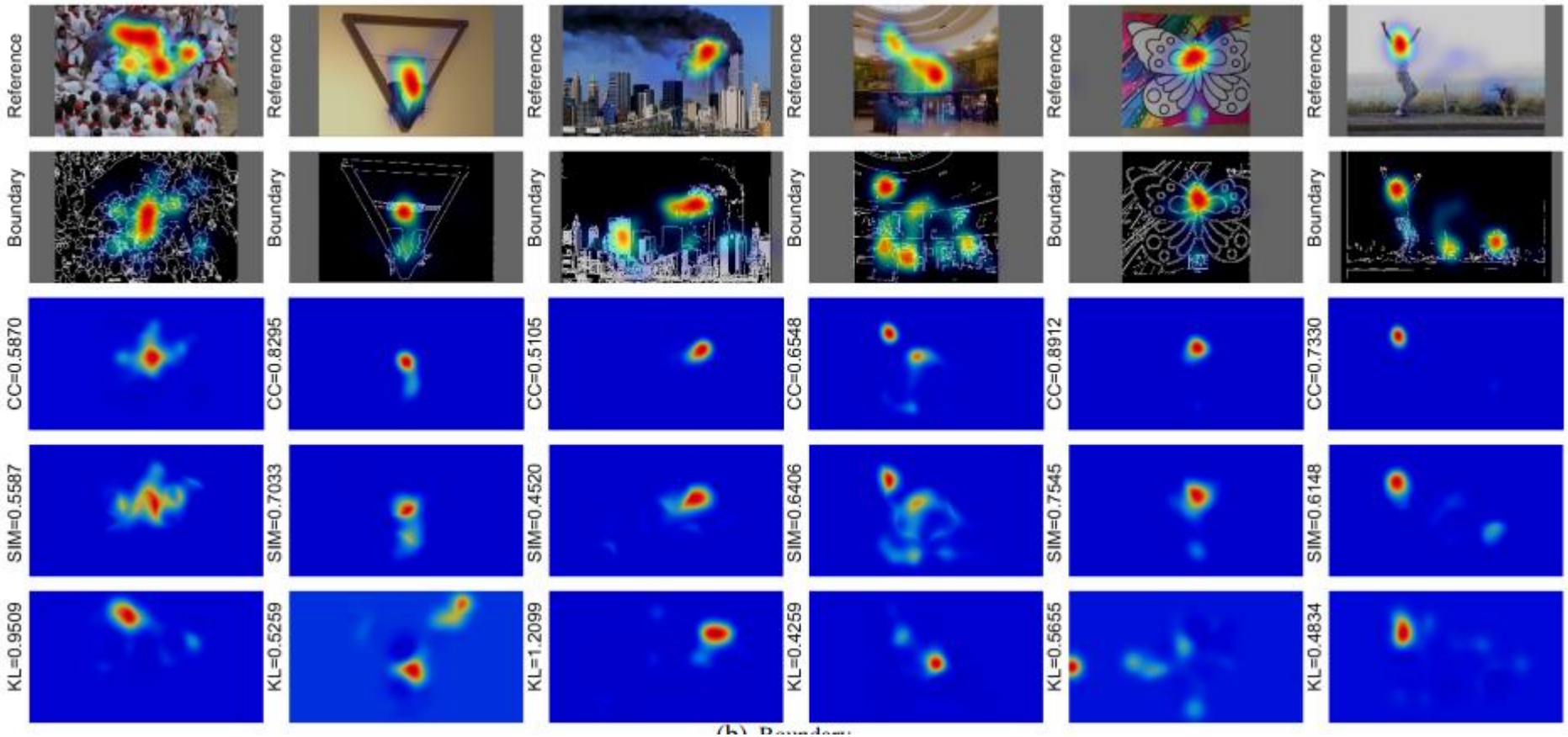
1st Challenge: Eye-movement Data Analysis



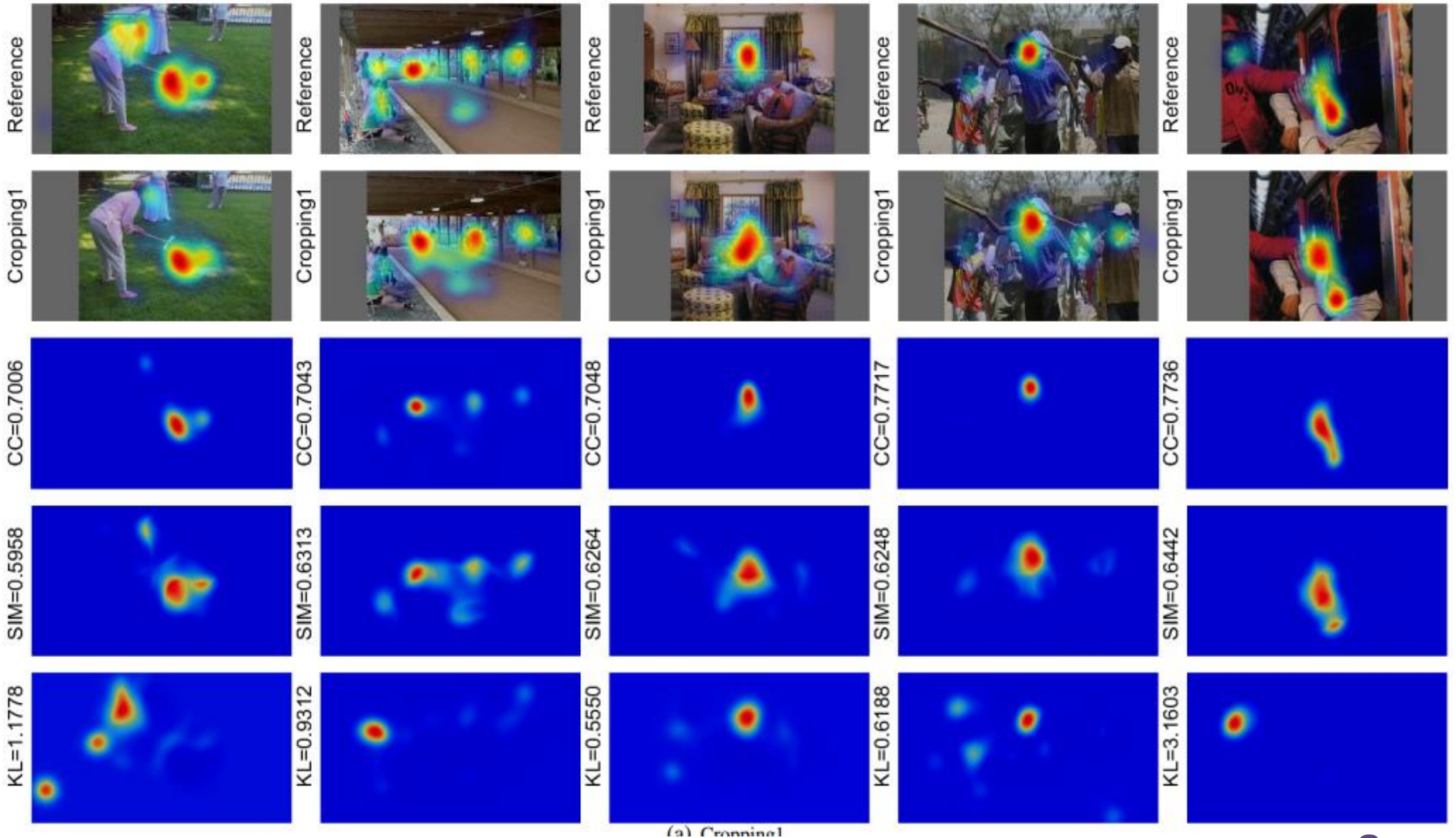
1st Challenge: Eye-movement Data Analysis



1st Challenge: Eye-movement Data Analysis

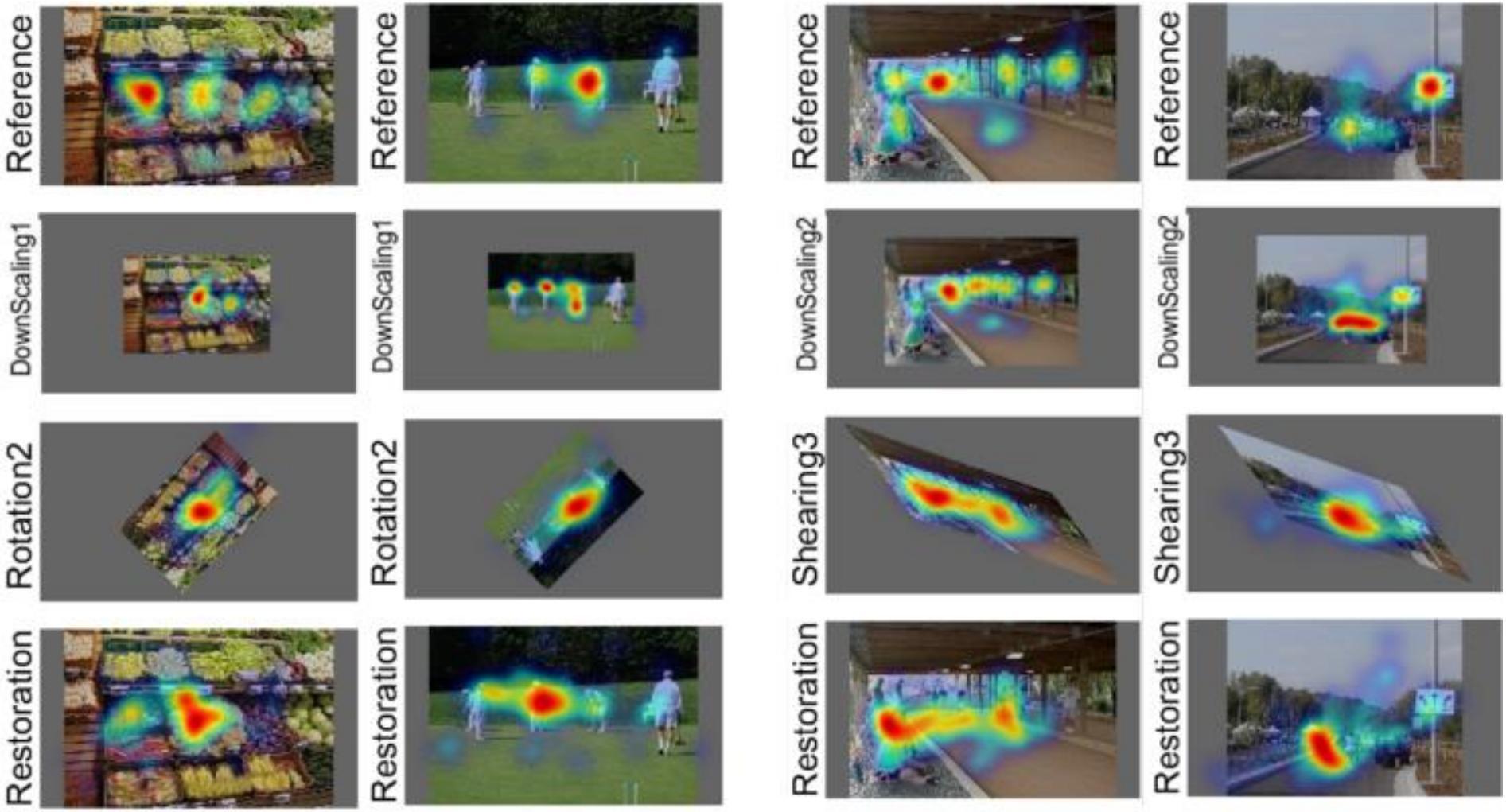


1st Challenge: Eye-movement Data Analysis



(a) Cropping1

1st Challenge: Eye-movement Data Analysis

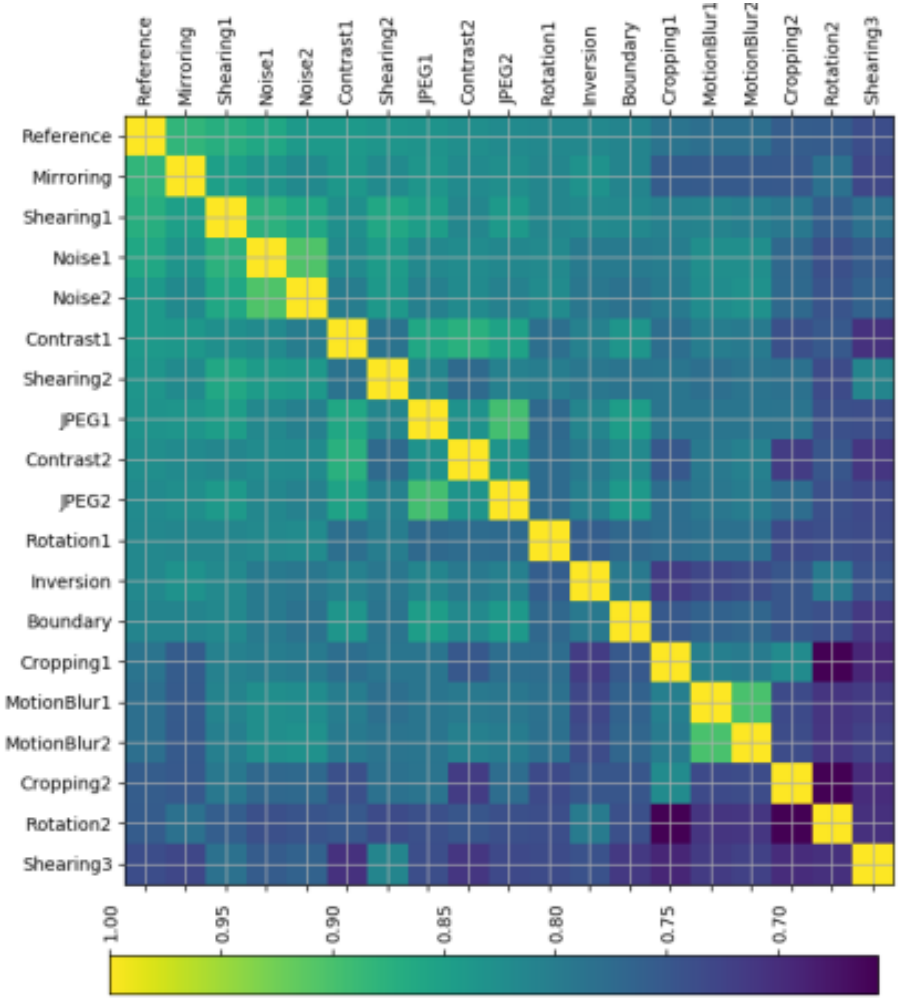


1st Challenge: Transformation Types

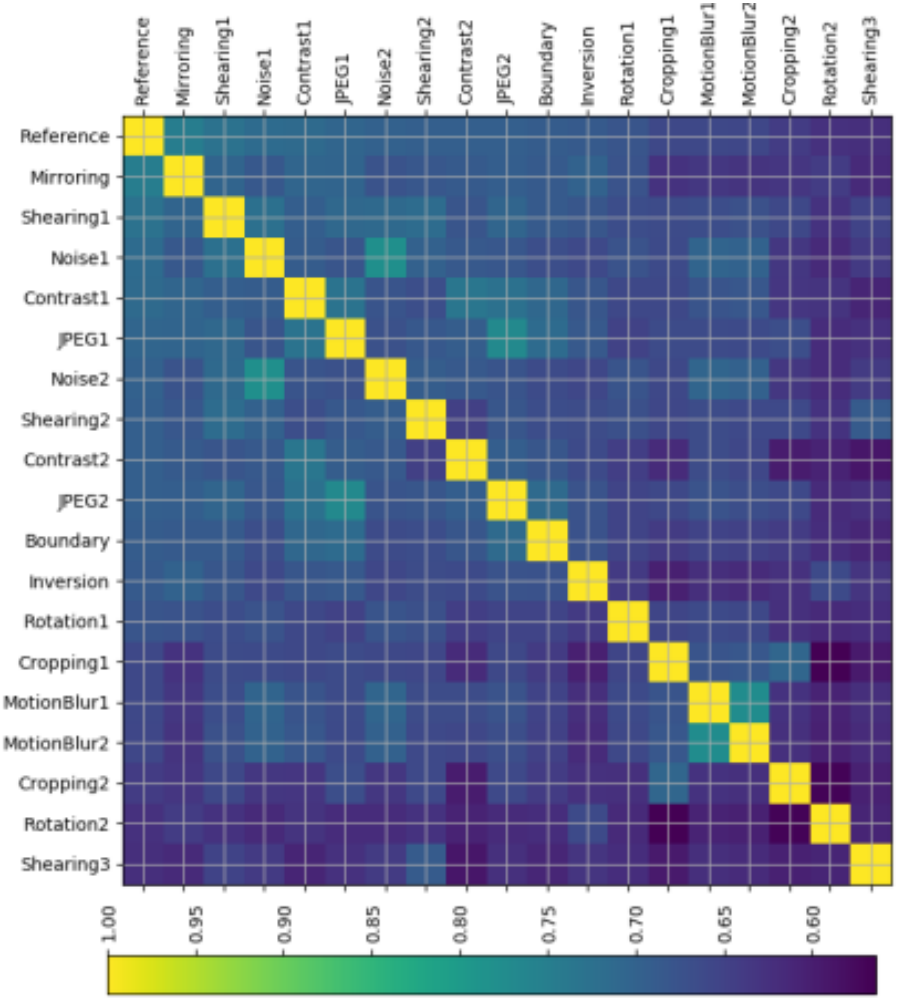
- **Image Acquisition Stage:**
- **Image Transmission Stage:**
- **Image Displaying Stage:**
- **Other Special Cases:**

Distortion	Generation code (using Matlab)	sAUC, CC, NSS
Reference	*100 distortion-free images* (img) from CAT2000	0.733, 0.954, 3.435
MotionBlur1	imfilter(img, fspecial('motion', **15, 0**))	0.664, 0.923, 2.572
MotionBlur2	imfilter(img, fspecial('motion', **35, 90**))	0.651, 0.920, 2.588
Noise1	imnoise(img, 'gaussian', 0, **0.1**)	0.706, 0.940, 3.032
Noise2	imnoise(img, 'gaussian', 0, **0.2**)	0.696, 0.939, 3.026
JPEG1	imwrite(img, saveroutine, 'Quality', **5**)	0.703, 0.902, 2.919
JPEG2	imwrite(img, saveroutine, 'Quality', **0**)	0.705, 0.903, 2.863
Contrast1	imadjust(img, [ ], **[0.3,0.7]**)	0.722, 0.931, 3.008
Contrast2	imadjust(img, [ ], **[0.4,0.6]**)	0.702, 0.931, 3.430
Rotation1	imrotate(img, **-45, 'bilinear', 'loose'**)	0.680, 0.893, 2.287
Rotation2	imrotate(img, **-135, 'bilinear', 'loose'**)	0.654, 0.892, 2.098
Shearing1	imwarp(img, **affine2d([1 0 0; 0.5 1 0; 0 0 1]**)	0.711, 0.943, 3.011
Shearing2	imwarp(img, **affine2d([1 0.5 0; 0 1 0; 0 0 1]**)	0.687, 0.927, 2.576
Shearing3	imwarp(img, **affine2d([1 0.5 0; 0.5 1 0; 0 0 1]**)	0.665, 0.888, 2.118
Inversion	imrotate(img, **-180, 'bilinear', 'loose'**)	0.695, 0.934, 3.062
Mirroring	mirror symmetry version of reference images	0.726, 0.930, 3.360
Boundary	edge(img, 'canny', 0.3, sqrt(2))	0.667, 0.888, 2.312
Cropping1	a 1080 × 200 band from the **left** of img	0.697, 0.934, 2.630
Cropping2	a 200 × 1920 band from the **top side** of img	0.692, 0.934, 2.641

1st Challenge: The similarity of human gaze when viewing different Transformed Stimuli versus Original Stimuli



(a) CC↑ similarity matrix



(b) SIM↑ similarity matrix

1st Challenge: Data Augmentation Application

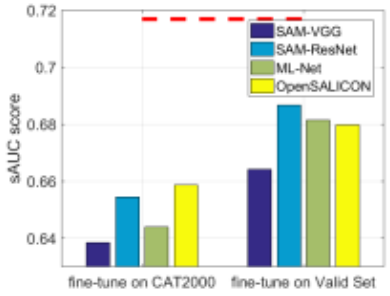
Label-preserving Valid Data Augmentation Transformations:

Reference, Mirroring, Inversion, Contrast1, Shearing1, JPEG1, Noise1

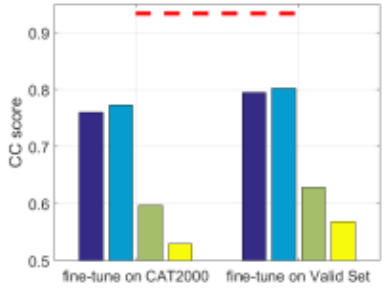
Label-damaging Invalid Data Augmentation Transformations:

Rotation1, Rotation2, Shearing2, Shearing3, Cropping1, Cropping2, MotionBlur2

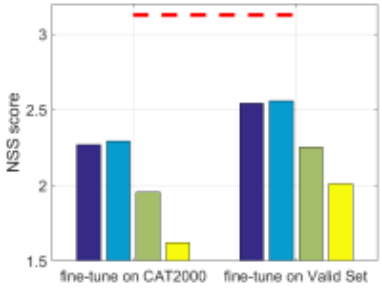
1st Challenge: Data Augmentation Application



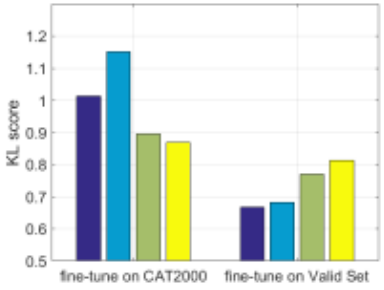
(a) sAUC↑



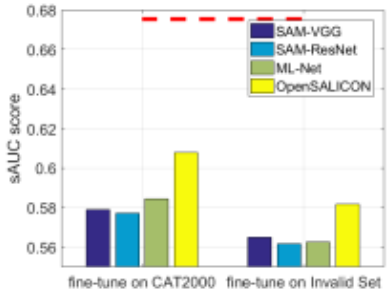
(b) CC↑



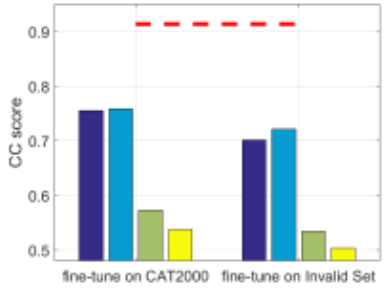
(c) NSS↑



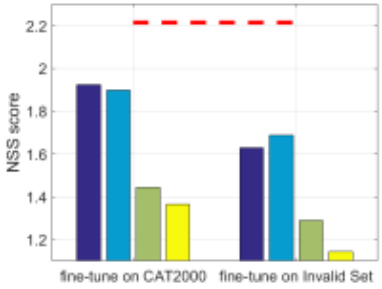
(d) KL↓



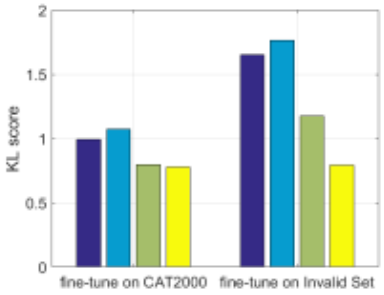
(e) sAUC↑



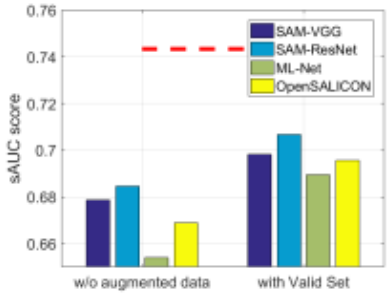
(f) CC↑



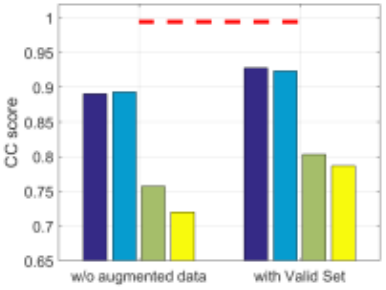
(g) NSS↑



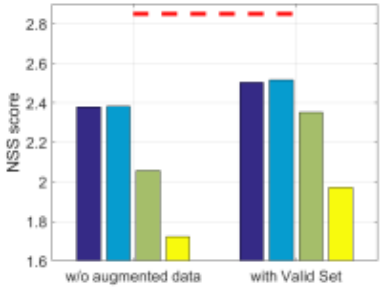
(h) KL↓



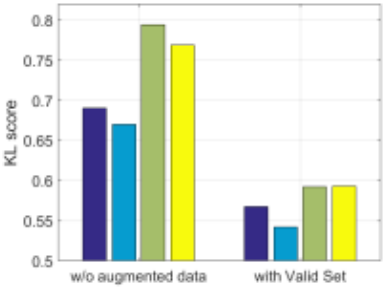
(i) sAUC↑



(j) CC↑

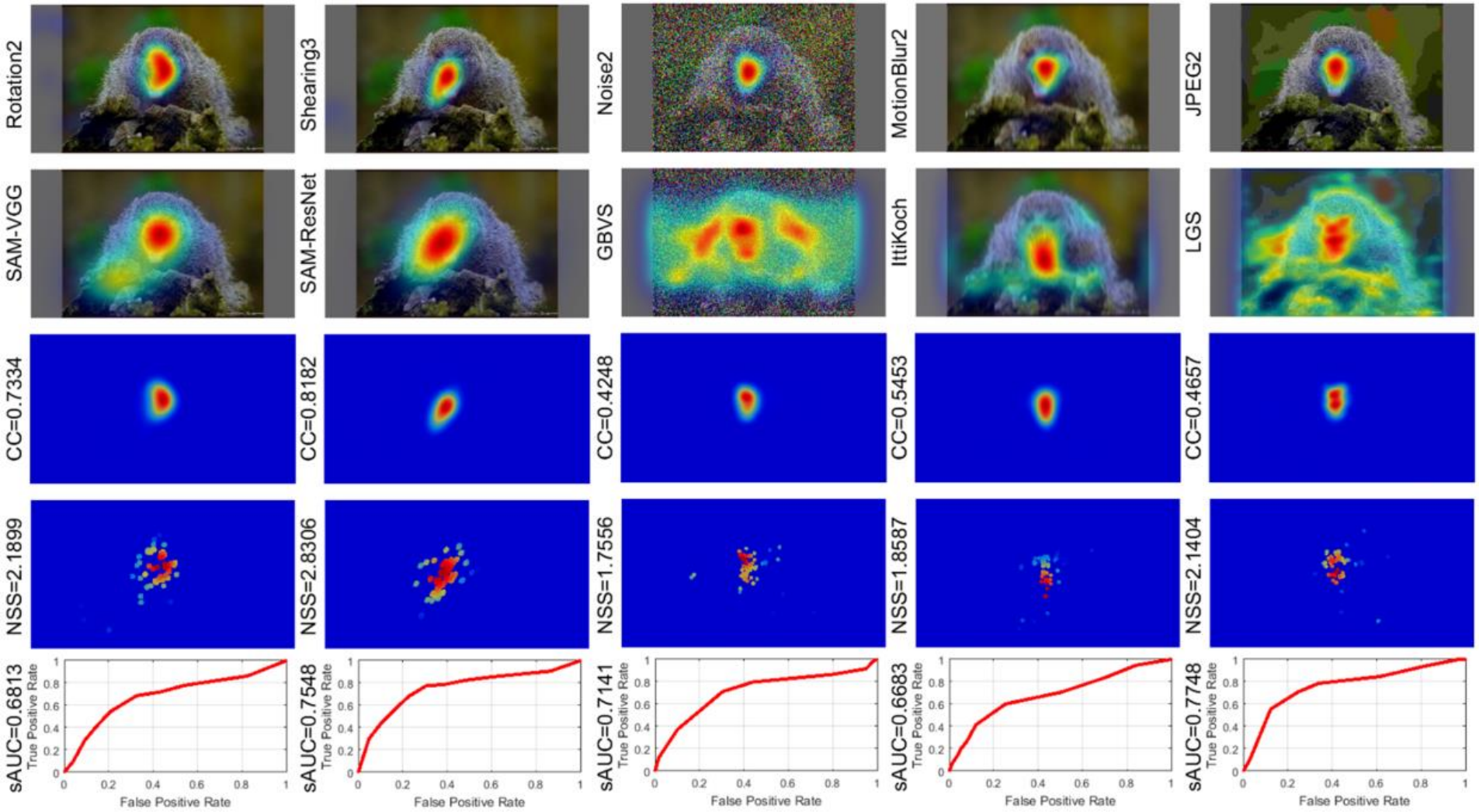


(k) NSS↑



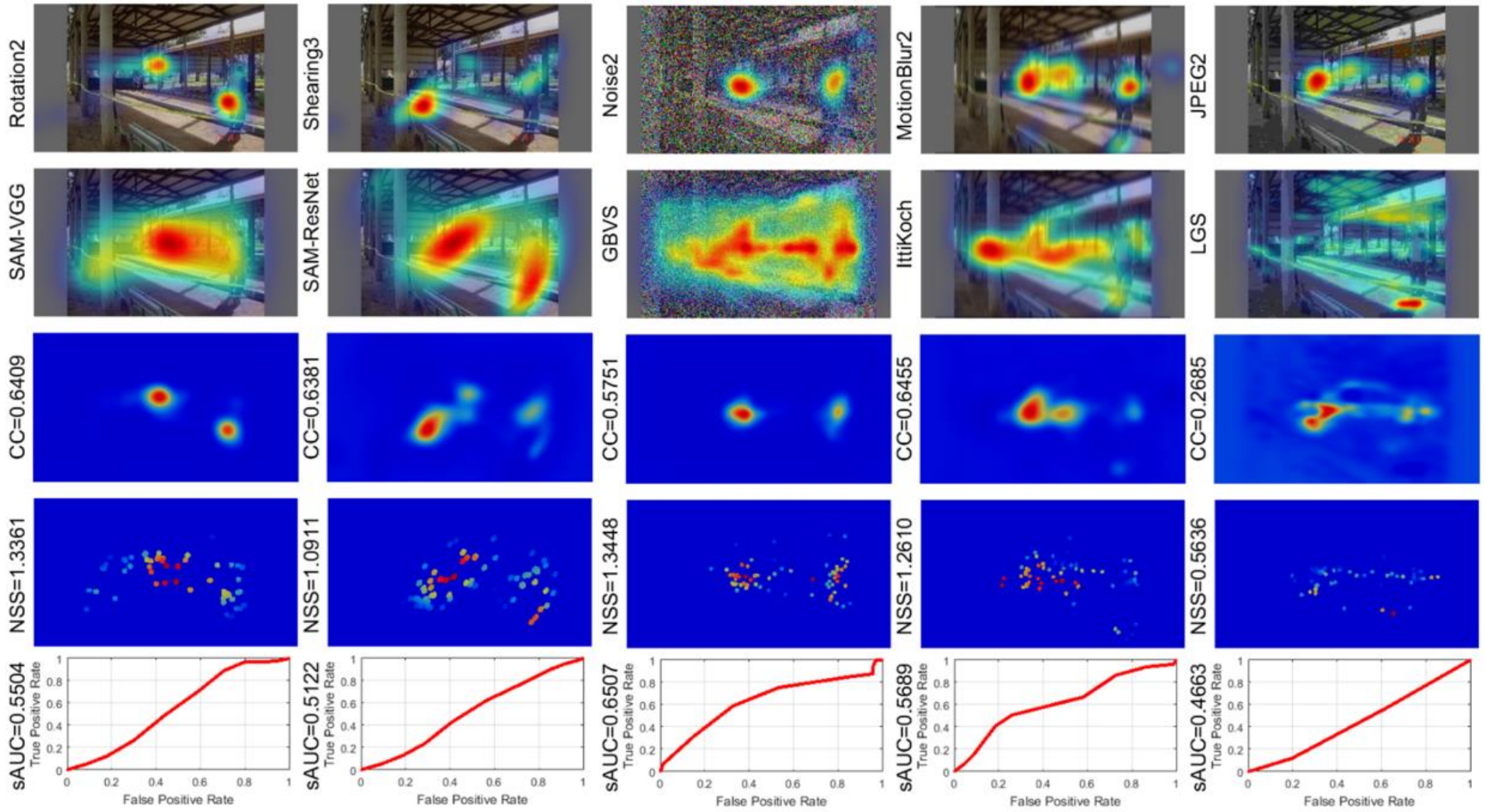
(l) KL↓

1st Challenge: Failure cases of other methods



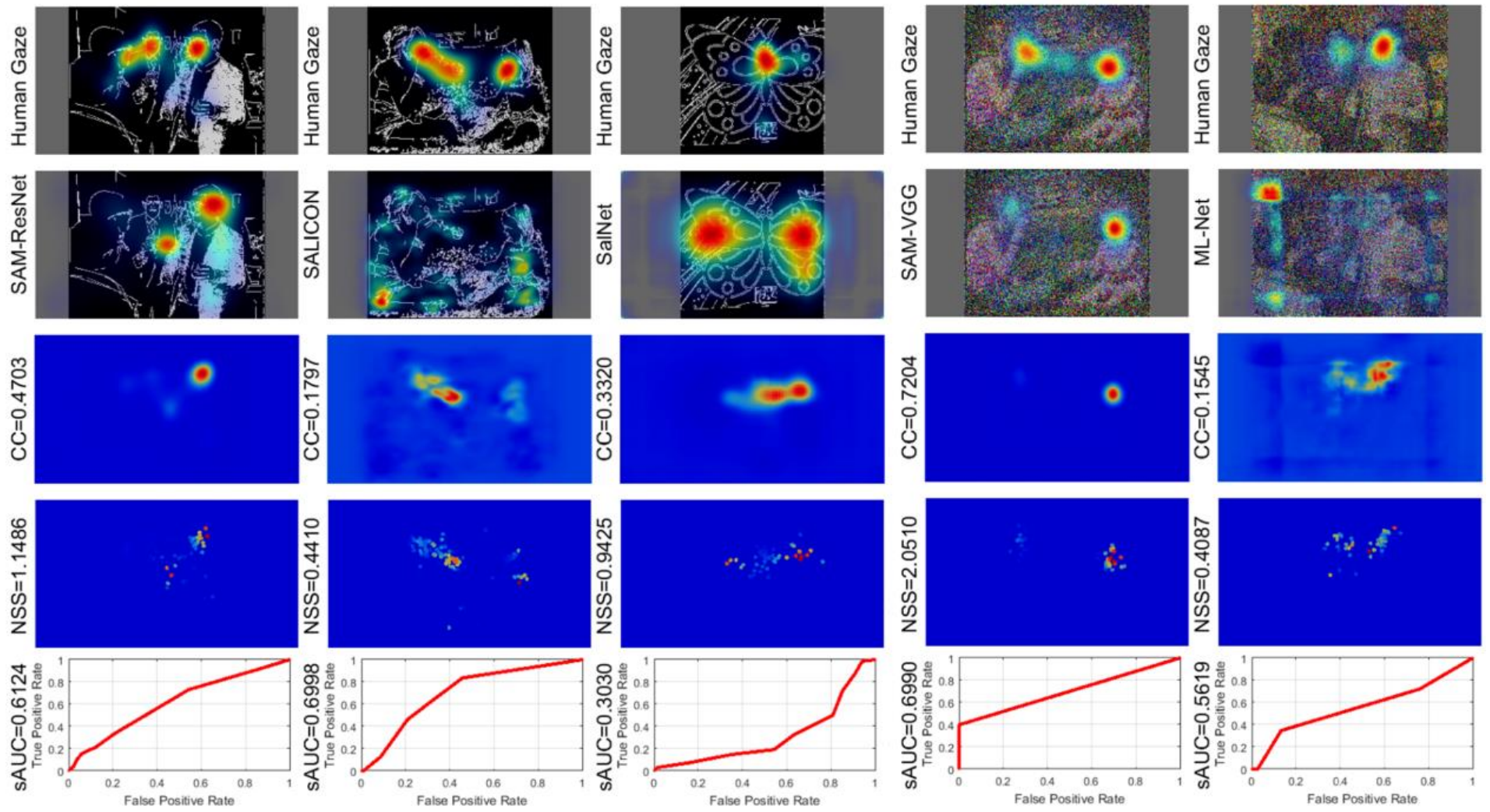
(a) Saliency maps produced by existing models on a simple scene with transformations

1st Challenge: Failure cases of current saliency models



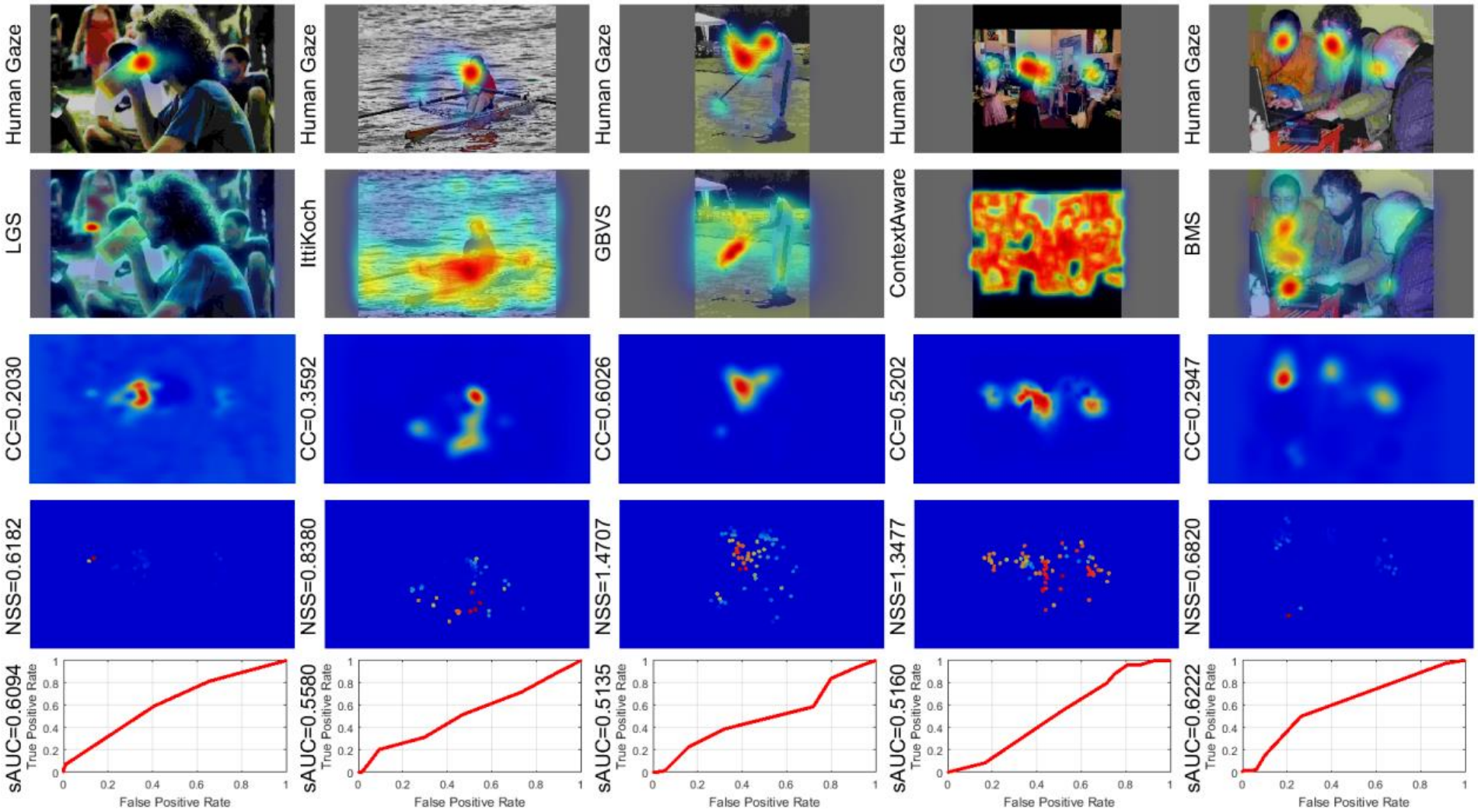
(b) Saliency maps produced by existing models on a complex scene with transformations

1st Challenge: Failure cases of current saliency models



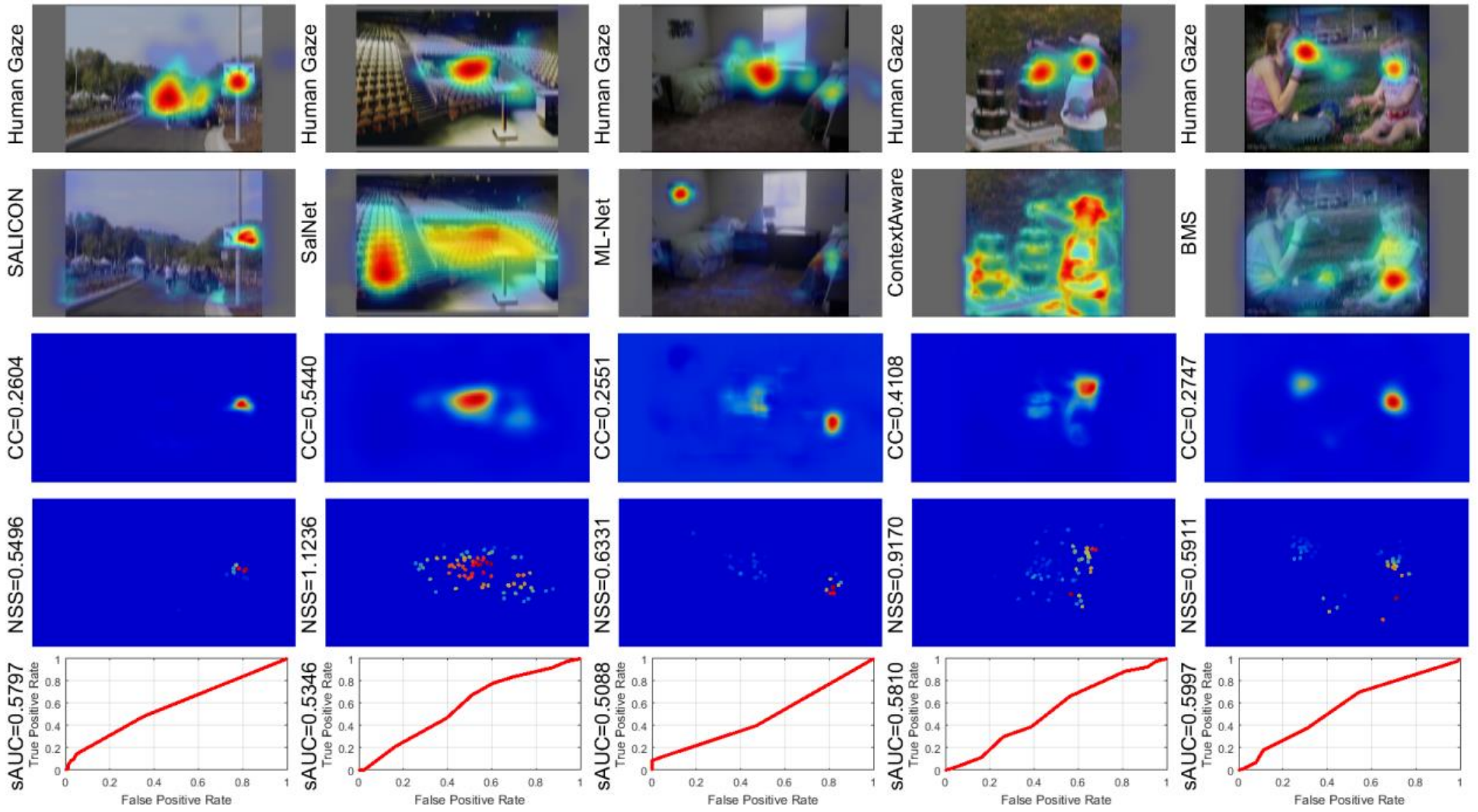
(a) Failure cases of existing models on scenes with Boundary and Noise

1st Challenge: Failure cases of current saliency models



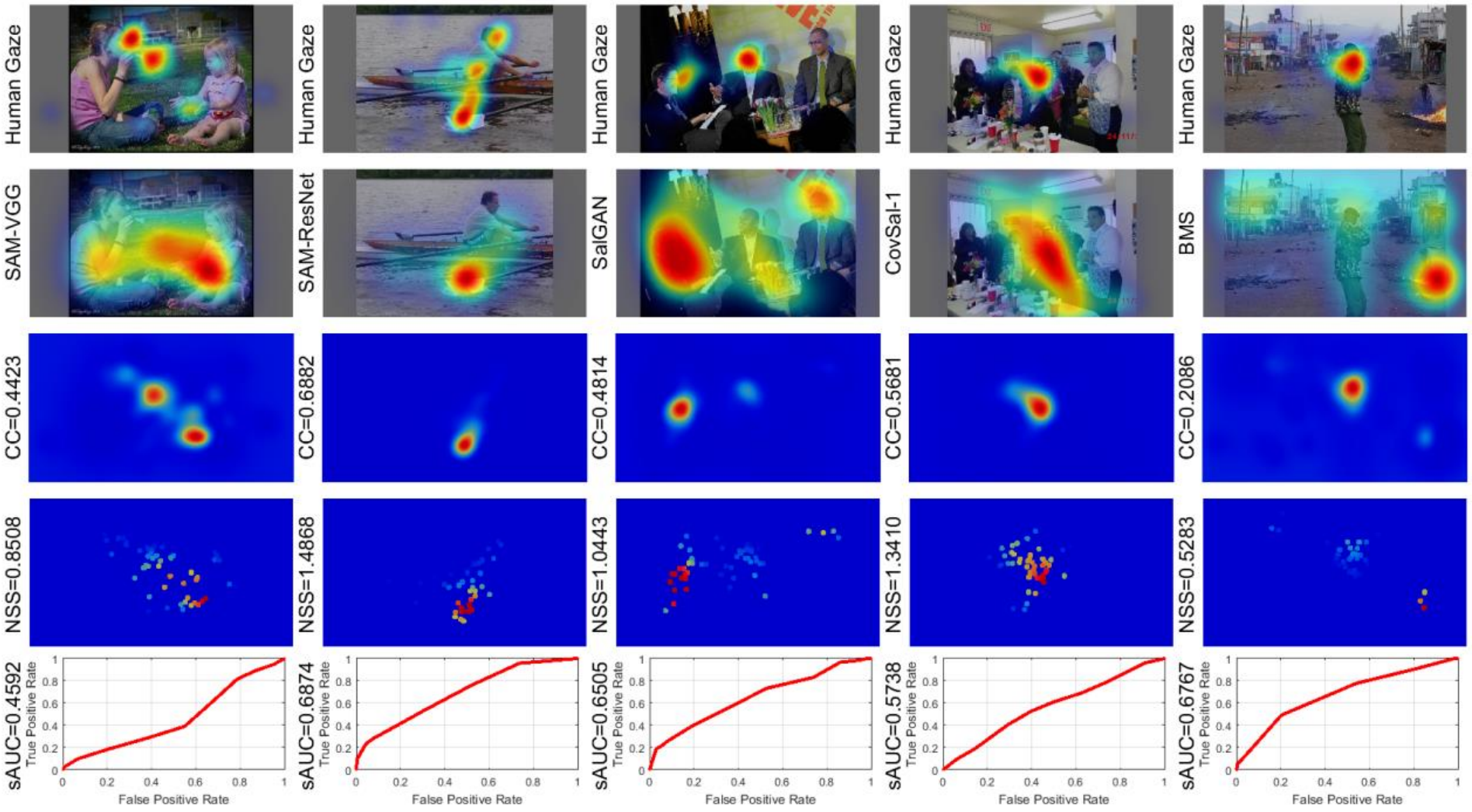
(b) Failure cases of existing models on scenes with Compression

1st Challenge: Failure cases of current saliency models



(a) Failure cases of existing models on scenes with MotionBlur

1st Challenge: Failure cases of current saliency models



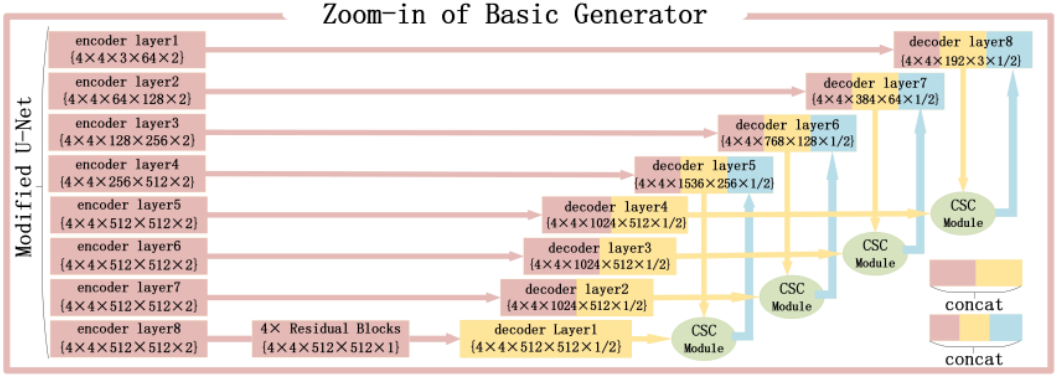
(b) Failure cases of existing models on scenes with Rotation2

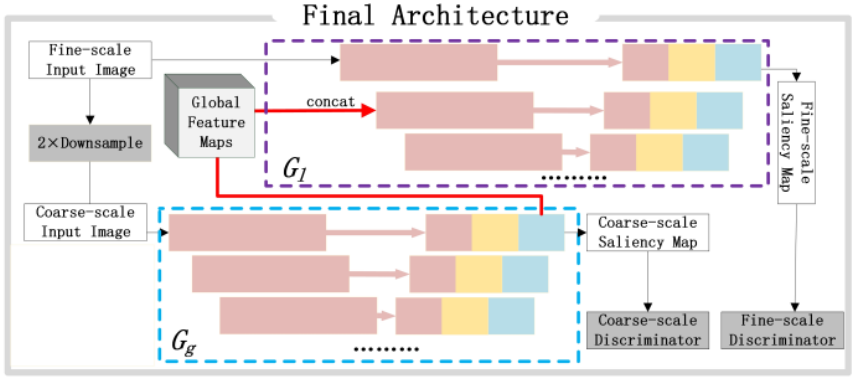
1st Challenge: Lessons from the analysis

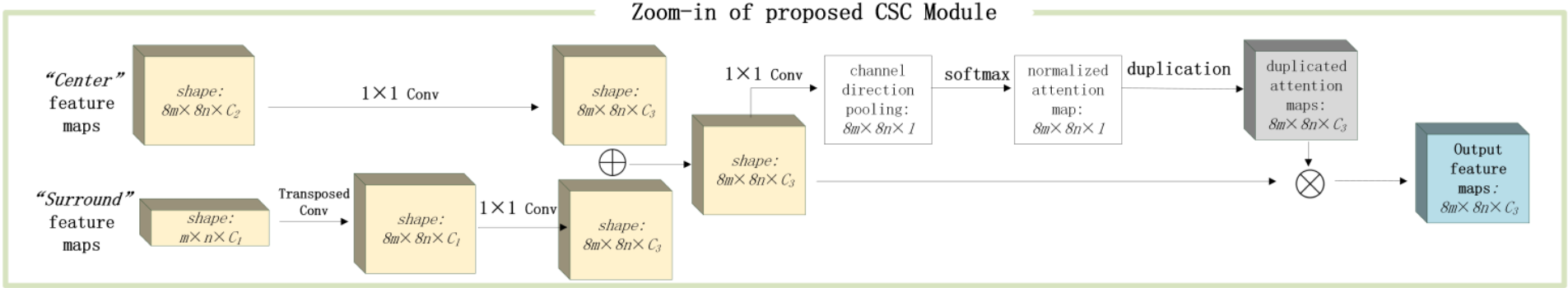
Any lessons?

- **Discriminative semantic objects**: When a scene is cropped, human attention tends to focus on the salient regions with more semantic information that help to understand the cropped scene and to recover from the information loss.
- **Highlighting semantic salient information while ignoring trivial artifacts**: We verified that human gaze focuses on semantic objects over various transformations, besides, human gaze tolerates the trivial artifacts caused by transformations such as JPEG and Noise distortions. In order to reach human level accuracy on transformed scenes, the robust saliency models should emphasize semantic salient regions while mitigating trivial artifacts.
- **Leveraging structural and texture information**: For upright and rigid scenes, low-level structural and texture information helps to detect the salient regions.
- **Combining multiple metrics**: There is no "perfect" metric that can accurately quantify human gaze on various transformations. However, they can complement each other. [2]

1st Challenge: Solution 1, Model Architecture

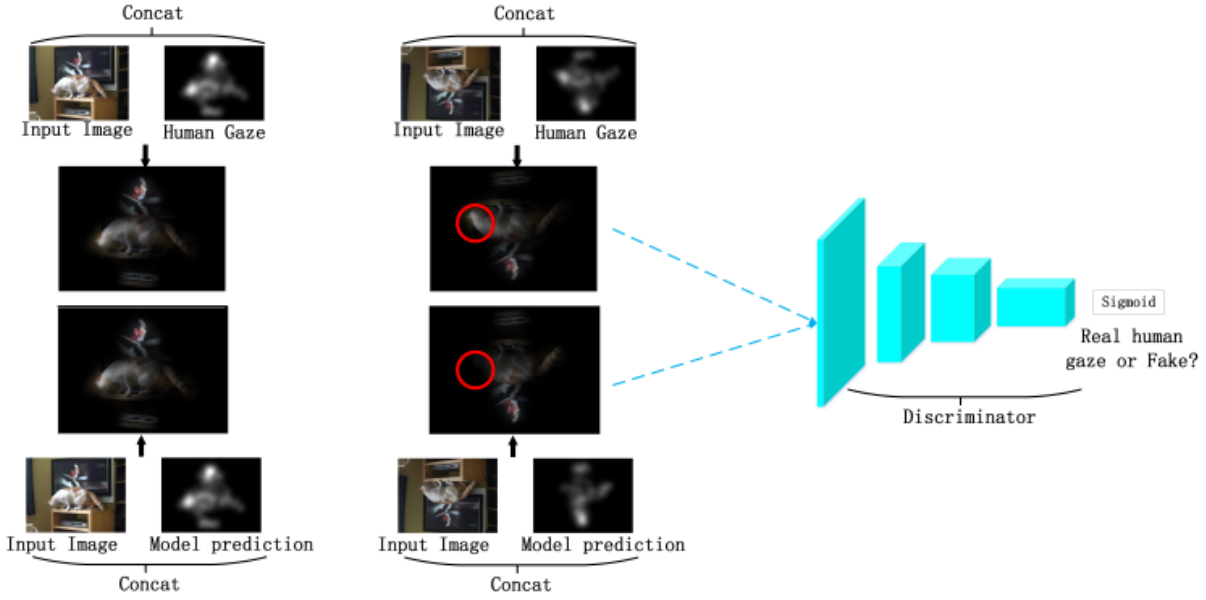




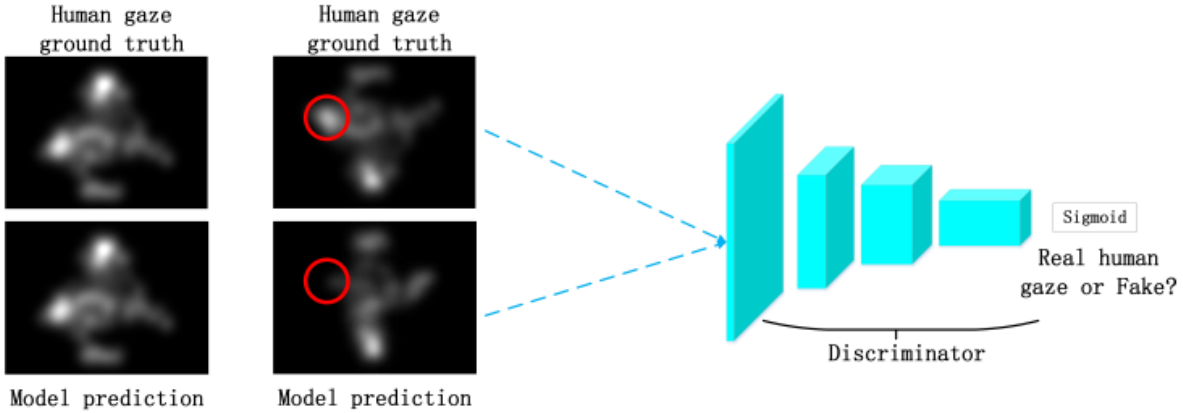


- Local-Global GAN architecture (more robust to scale transform)
- Skip-Connection of U-Net (for leveraging structural and texture information)
- Proposed Cross-Scale Center-Surround-Connection (for highlighting semantic information, while mitigating trivial artifacts)
- Conditional GAN (for discriminating semantic object)
- Proposed Differentiable Histogram Loss using the Alternative Chi-Square Distance
- Proposed Valid Data Augmentation Transformations

1st Challenge: Solution 1, Model Architecture



(a) Conditional discriminator



(b) Normal discriminator

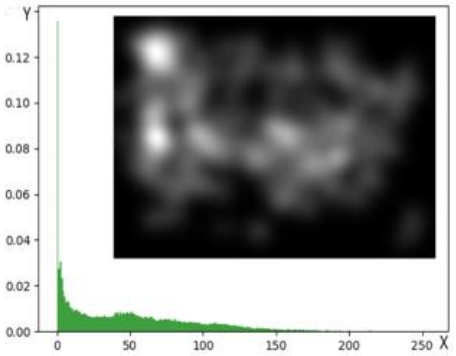
1st Challenge: Solution 1, Loss Function

$$p_k = \frac{1}{S} \times \Big(\sum_{l_i \in [b_{k-1}, b_k)} \frac{l_i - b_{k-1}}{\Delta} + \sum_{l_i \in [b_k, b_{k+1}]} \frac{b_{k+1} - l_i}{\Delta} \Big),$$

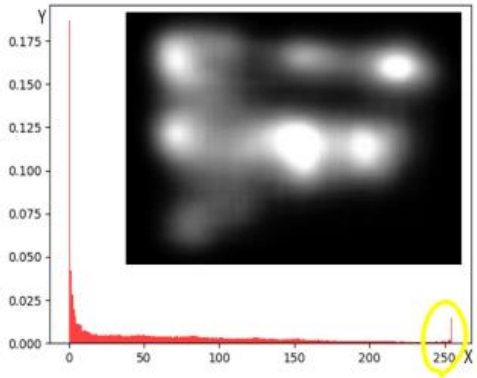
$$\bar{p_k} = \frac{p_k - \min(p_j, j \in [0, N])}{\max(p_j, j \in [0, N]) - \min(p_j, j \in [0, N])},$$

$$L_{ACS} = 2 \times \sum_{k=0}^{N} \frac{(\bar{p_k} - \bar{q_k})^2}{\bar{p_k} + \bar{q_k} + \epsilon},$$

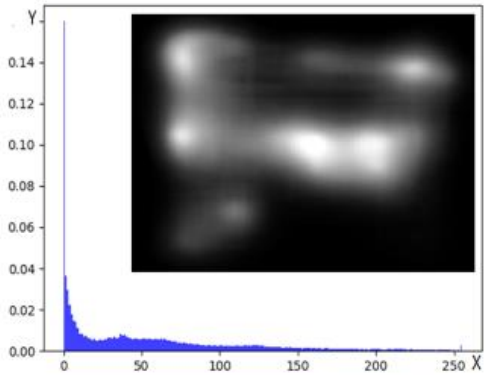
(a) Input Image



(b) Histogram of Human Gaze

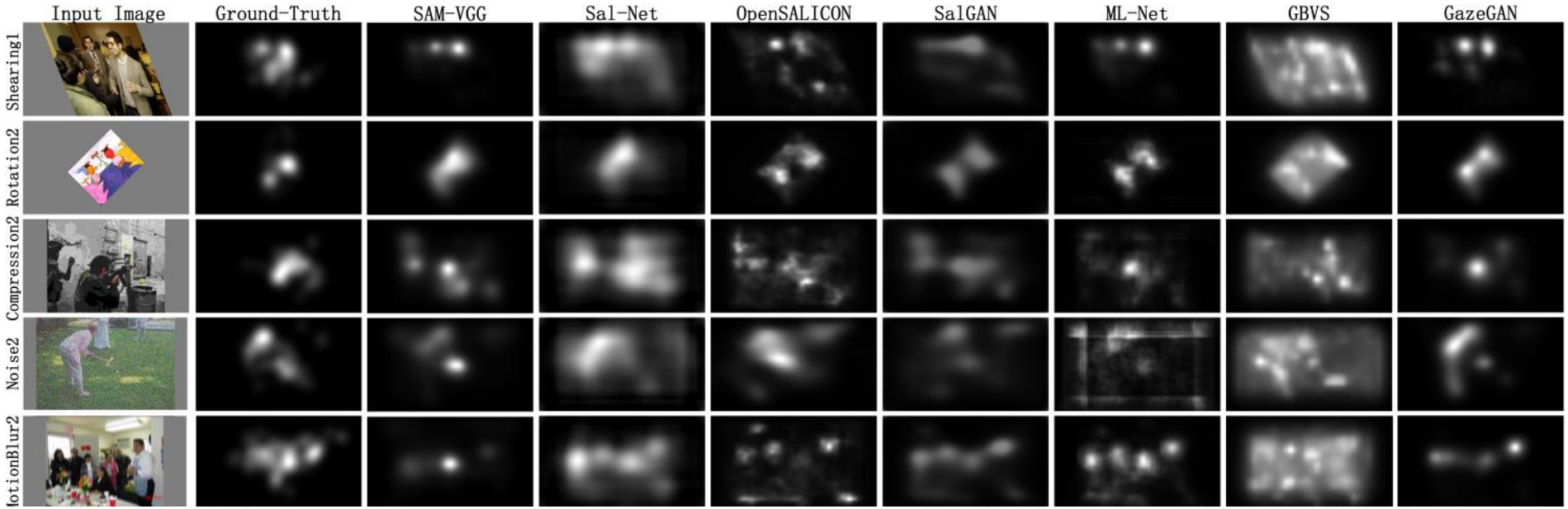


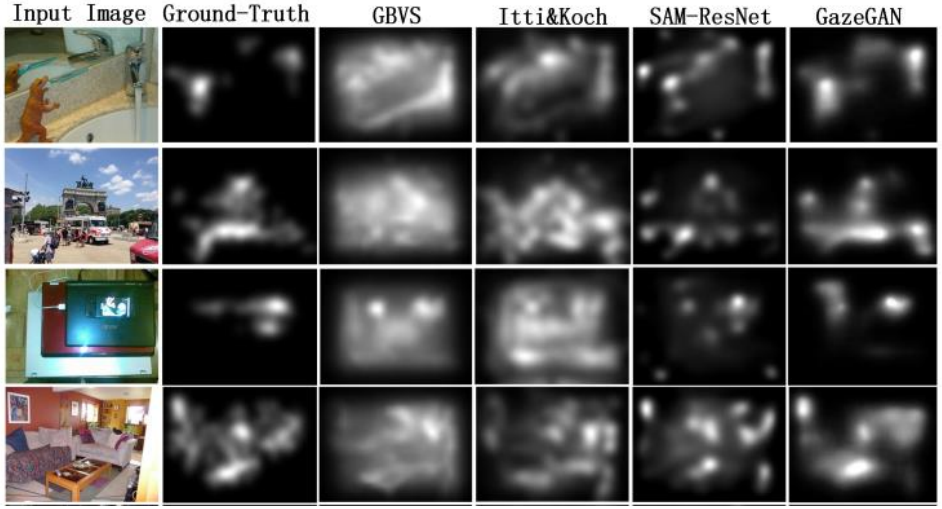
(c) Histogram of Saliency Map trained without HistLoss

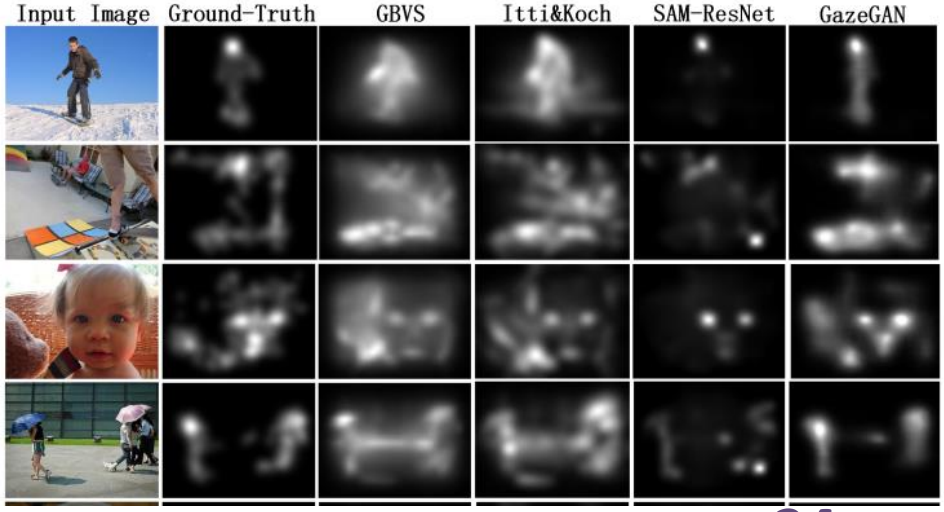


(d) Histogram of Saliency Map trained with HistLoss

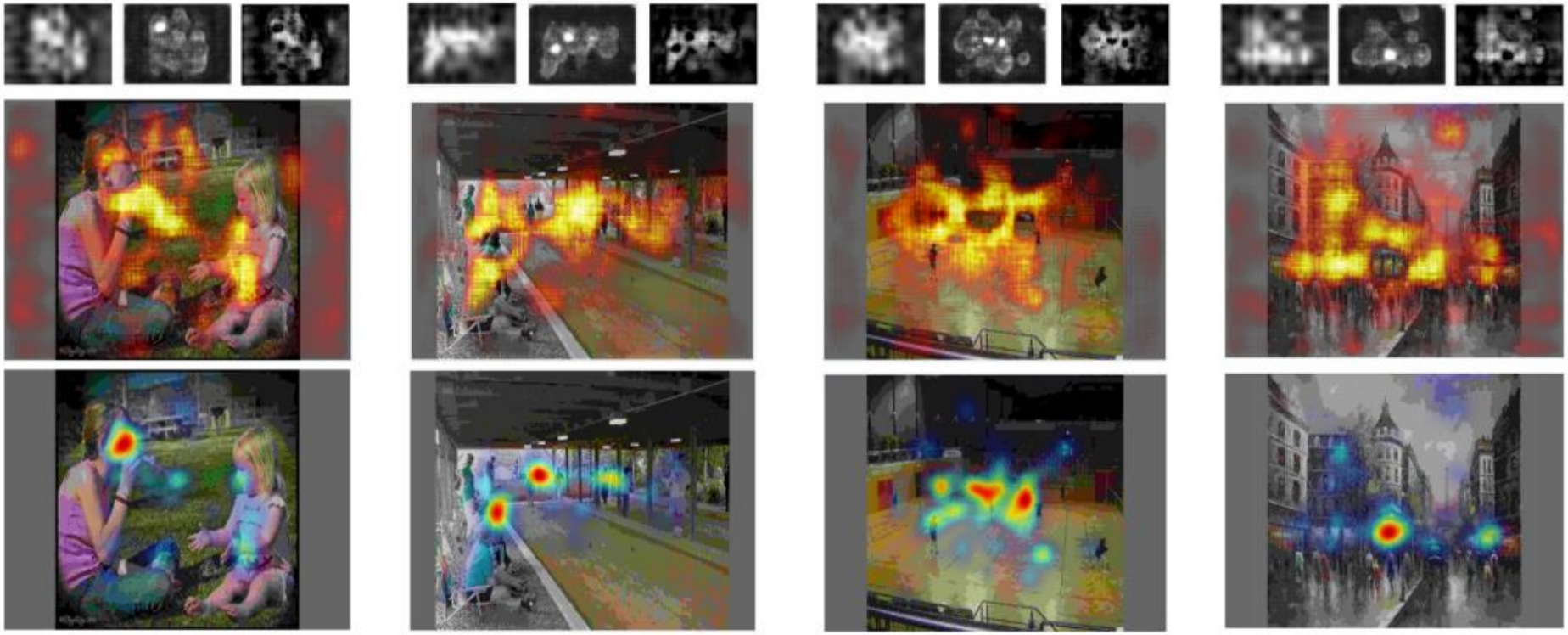
1st Challenge: Results of Solution 1







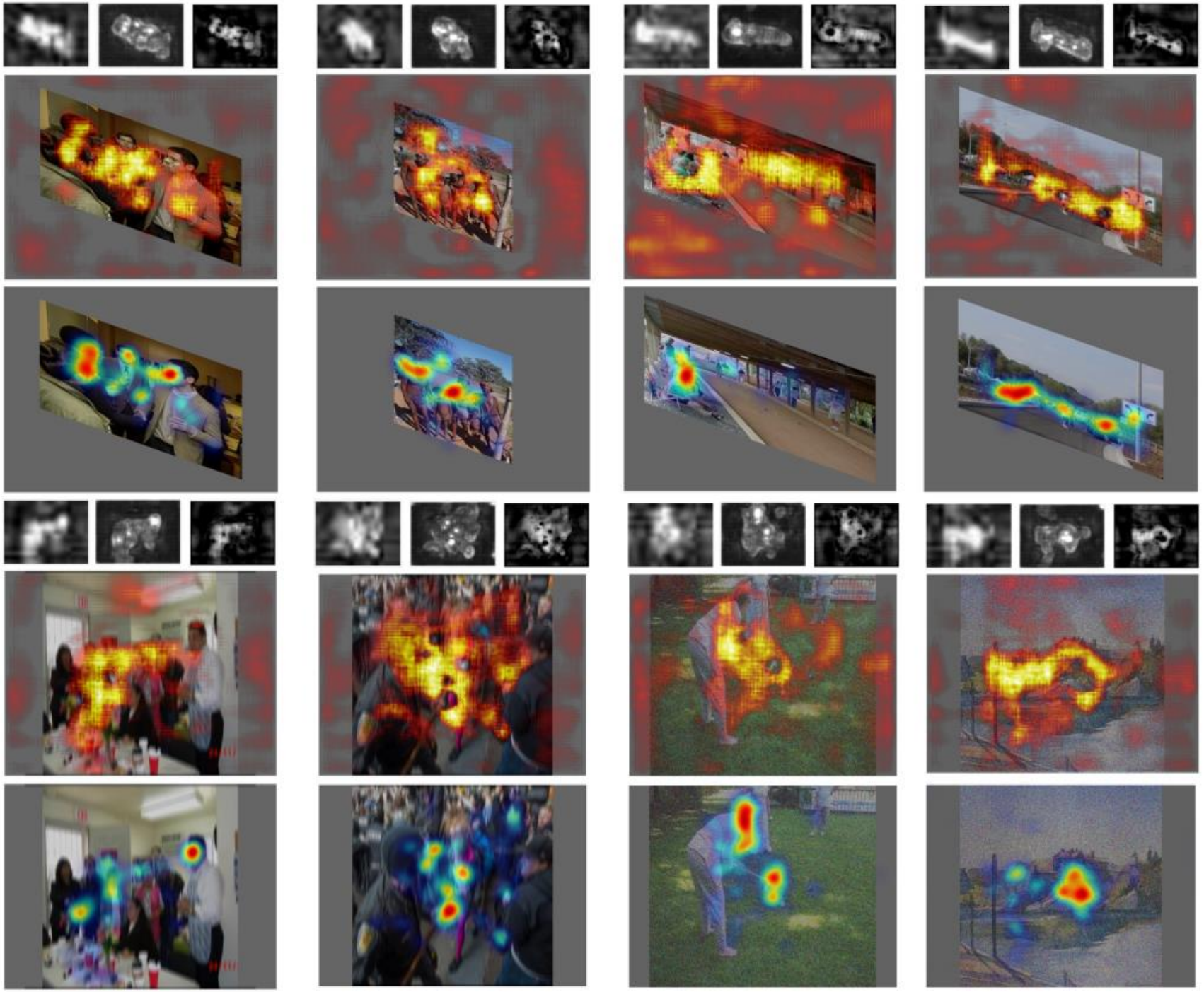
1st Challenge: Results of Solution 1



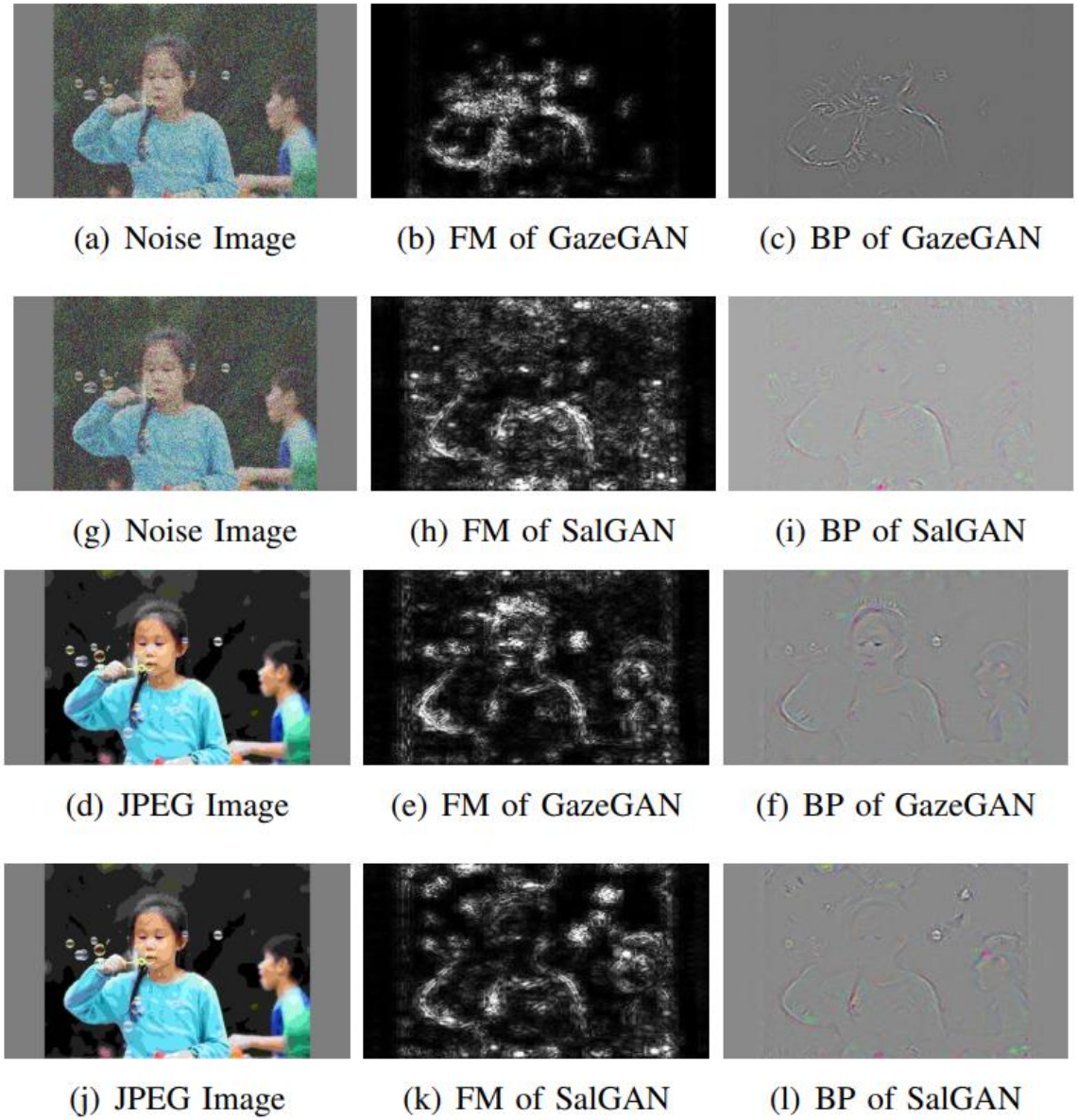
The 1*st* row represents the *surround* feature maps, *center* feature, and the *difference* maps of *surround* and *center*, respectively.

The 2*nd* row reflects the wrong predictions of *surround* feature maps caused by trivial artifacts.

The 3*rd* row reflects the final predictions processed by proposed CSC module that focus on semantic salient regions.



1st Challenge: Visualizations in Feature-Space



(a) Noise Image (b) FM of GazeGAN (c) BP of GazeGAN

(g) Noise Image (h) FM of SalGAN (i) BP of SalGAN

(d) JPEG Image (e) FM of GazeGAN (f) BP of GazeGAN

(j) JPEG Image (k) FM of SalGAN (l) BP of SalGAN

1st Challenge: Visualizations in Feature-Space


input image

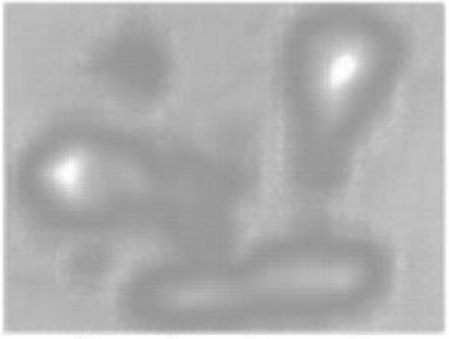

encoder layer1

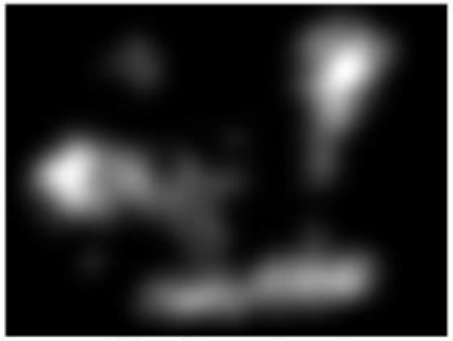

encoder layer2


encoder layer3


decoder layer1


decoder layer3


decoder layer5

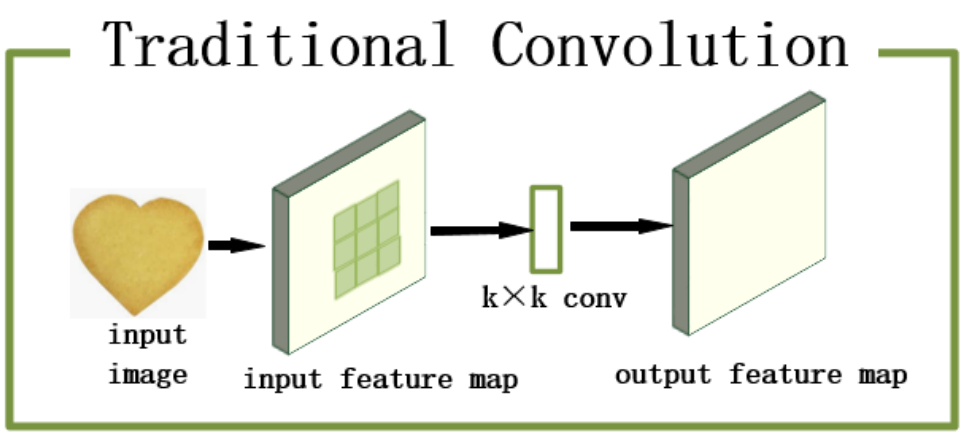

final output

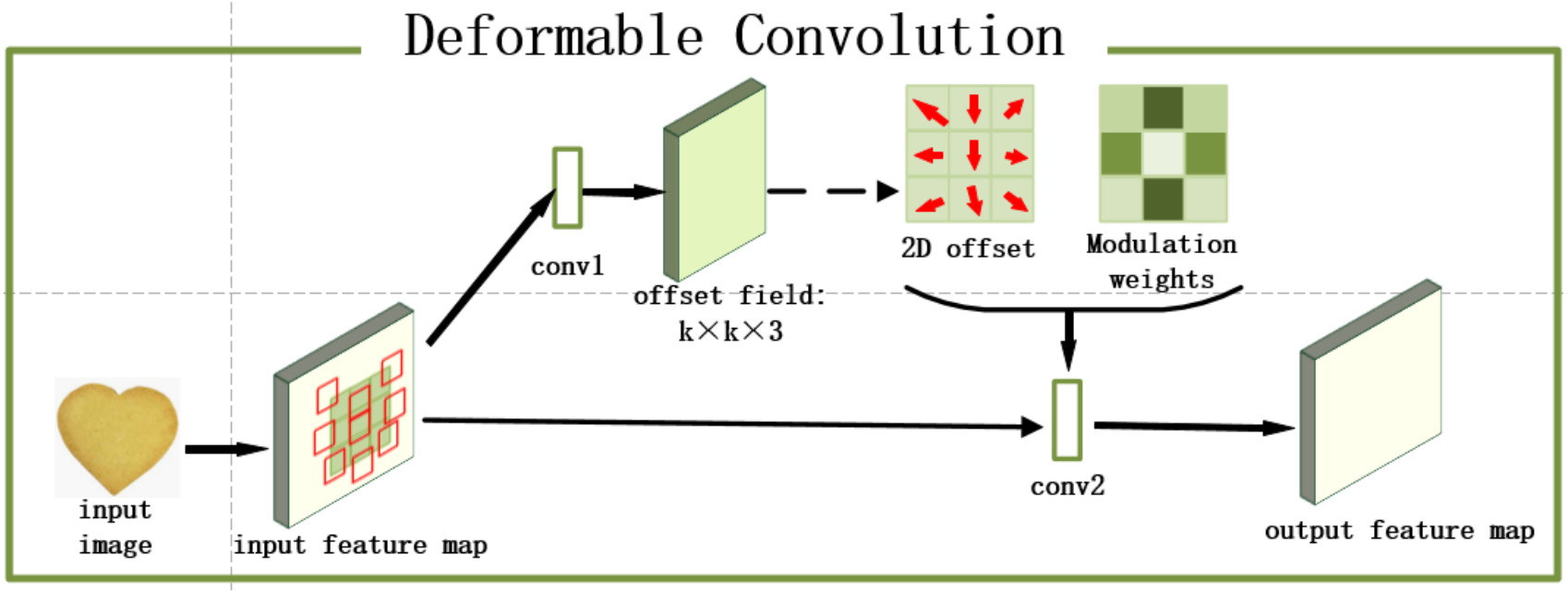
1st Challenge: Solution 2

Question:

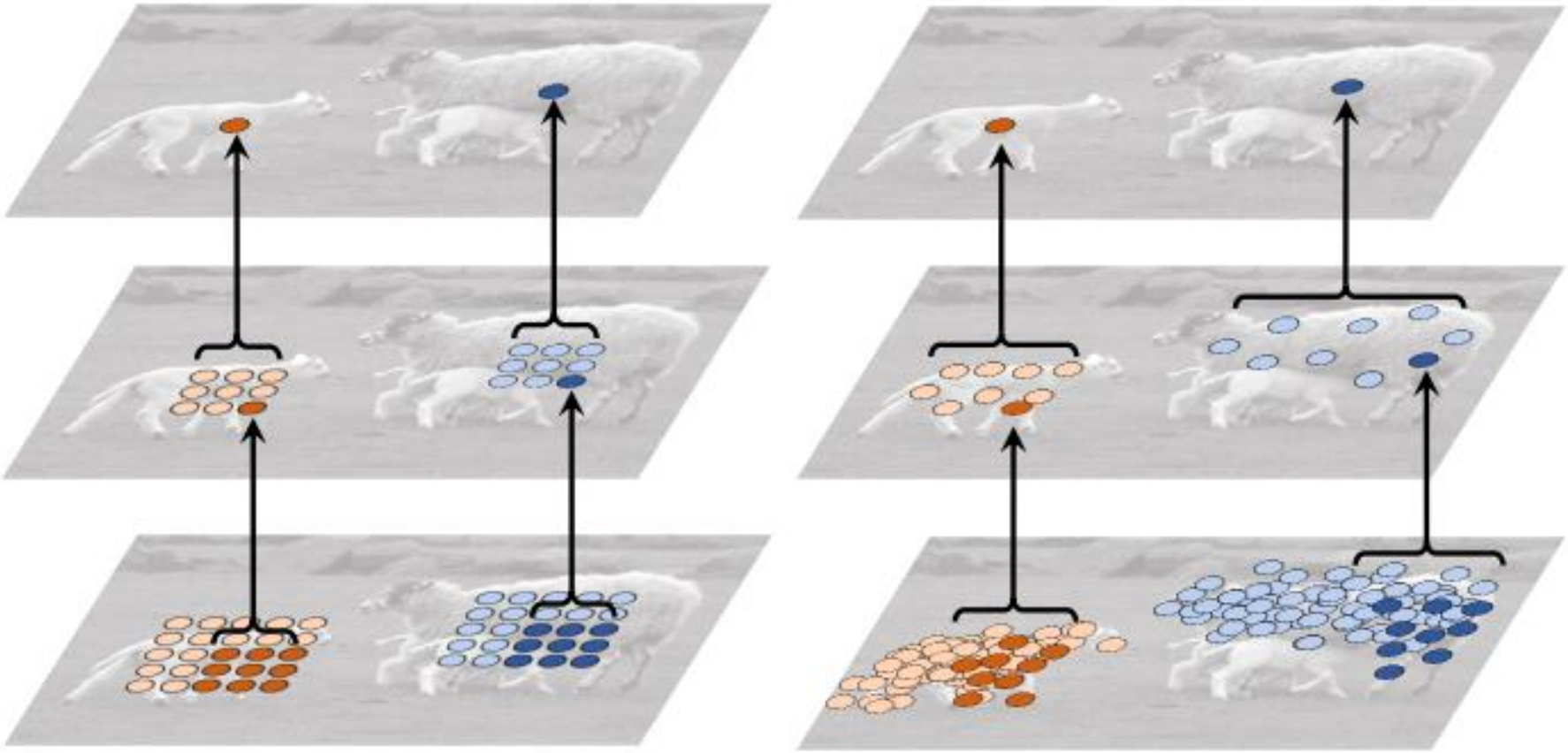
How many offsets should we learn?

For the same group of feature maps, the channels are spatially consistent



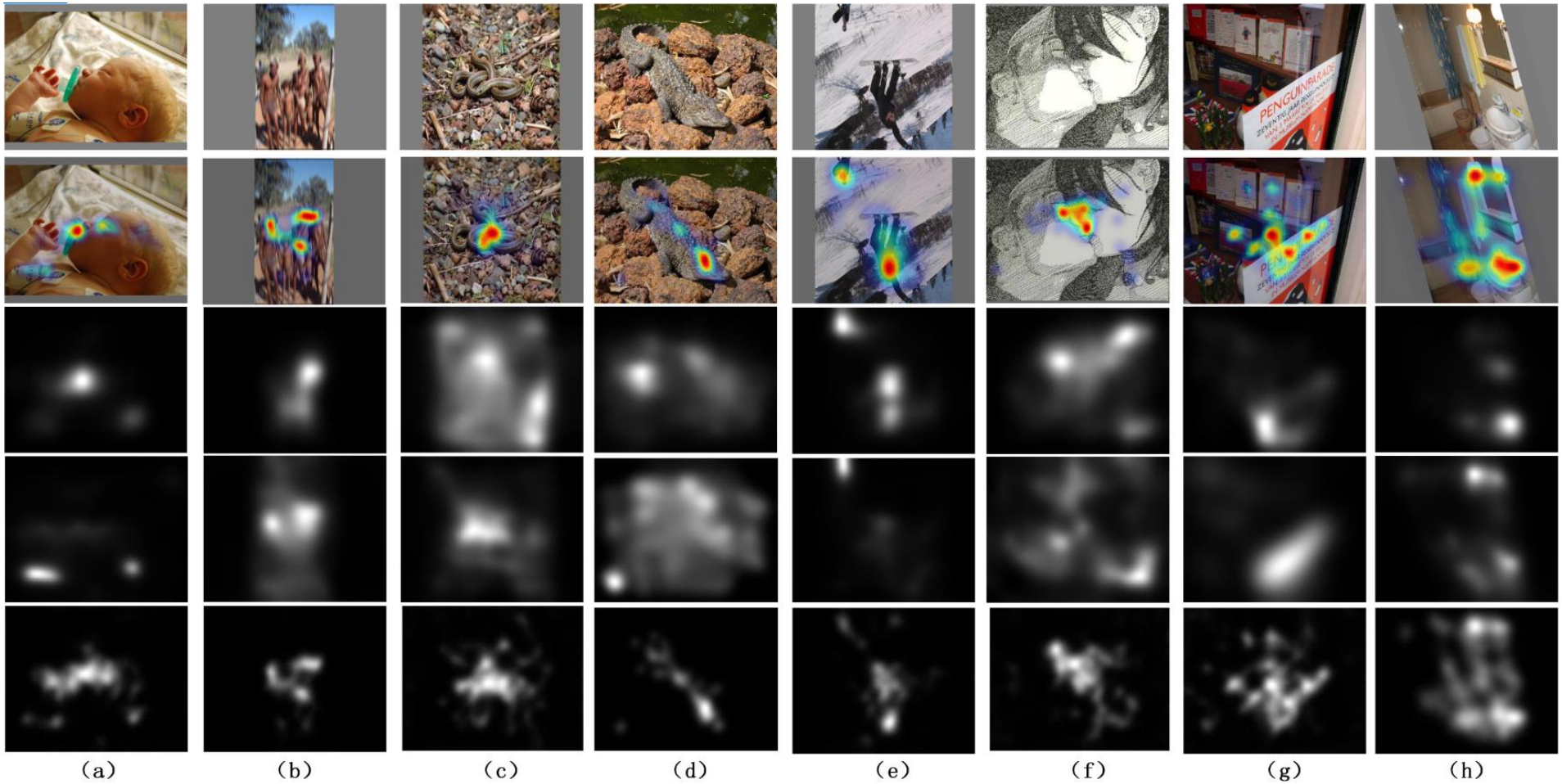


1st Challenge: Solution 2



(a) standard convolution (b) deformable convolution

1st Challenge: Results of Solution 2



The 1st -2nd rows represent the source stimuli and the human gaze ground-truth, while the 3rd -5th rows represent the prediction results of SALICON, DeepGaze2, and the proposed model, respectively.

Results

#	User	Entries	Date of Last Entry	SAUC ▲	IG ▲	NSS ▲	CC ▲	AUC ▲	SIM ▲	KL ▲
1	coolMatt1024	6	04/27/18	0.741 (4)	0.818 (1)	1.940 (6)	0.891 (3)	0.867 (1)	0.773 (5)	0.258 (1)
2	xubinwei	1	07/23/19	0.747 (1)	0.702 (5)	2.065 (1)	0.890 (4)	0.866 (2)	0.790 (3)	0.510 (9)
2	senjia	3	04/18/18	0.746 (2)	0.736 (3)	2.050 (2)	0.886 (5)	0.866 (2)	0.780 (4)	0.520 (10)
3	marcella.cornia	3	07/07/17	0.741 (4)	0.538 (13)	1.990 (4)	0.899 (2)	0.865 (3)	0.793 (2)	0.610 (14)
4	codacscgaze	2	08/04/19	0.736 (7)	0.720 (4)	1.899 (11)	0.879 (6)	0.864 (4)	0.773 (5)	0.376 (5)
5	ZhaohuiChe	8	08/02/19	0.733 (8)	0.696 (6)	1.884 (12)	0.872 (9)	0.862 (5)	0.766 (8)	0.394 (7)
5	Jack0521	4	06/12/18	0.745 (3)	0.646 (7)	1.928 (7)	0.874 (7)	0.862 (5)	0.771 (6)	0.489 (8)
5	ysyscool	2	01/17/18	0.739 (5)	0.195 (16)	1.959 (5)	0.902 (1)	0.862 (5)	0.795 (1)	0.864 (16)
5	fredqi	3	01/13/18	0.745 (3)	0.357 (14)	2.045 (3)	0.862 (10)	0.862 (5)	0.753 (10)	1.026 (17)
6	jackdong	5	06/01/18	0.745 (3)	0.611 (10)	1.923 (8)	0.872 (9)	0.861 (6)	0.769 (7)	0.539 (12)
6	codalabcdw	6	07/23/19	0.738 (6)	0.337 (15)	1.847 (13)	0.873 (8)	0.861 (6)	0.771 (6)	0.686 (15)
7	zhewuucas	3	07/13/17	0.726 (9)	0.738 (2)	1.841 (14)	0.860 (11)	0.859 (7)	0.756 (9)	0.318 (2)
8	sfdodge	5	07/19/17	0.720 (11)	0.646 (7)	1.911 (9)	0.821 (17)	0.856 (8)	0.722 (13)	0.527 (11)
9	hallazie	1	01/31/18	0.724 (10)	0.640 (8)	1.804 (17)	0.844 (15)	0.855 (9)	0.714 (14)	0.341 (3)
9	lvjincheng	6	06/12/18	0.726 (9)	0.613 (9)	1.829 (16)	0.856 (12)	0.855 (9)	0.705 (15)	0.352 (4)
10	hrtavakoli	6	02/17/18	0.717 (12)	0.541 (12)	1.773 (18)	0.848 (14)	0.854 (10)	0.684 (16)	0.384 (6)
11	lihan	2	07/21/19	0.746 (2)	0.554 (11)	1.907 (10)	0.833 (16)	0.851 (11)	0.731 (12)	0.564 (13)
12	RyanLui	8	06/11/18	0.724 (10)	-0.187 (17)	1.838 (15)	0.855 (13)	0.850 (12)	0.746 (11)	1.208 (18)
13	fwahhab	13	04/26/18	0.691 (13)	-4.777 (18)	1.611 (19)	0.677 (18)	0.794 (13)	0.535 (17)	4.942 (19)

2nd Challenge: *Malicious Adversarial Perturbations*

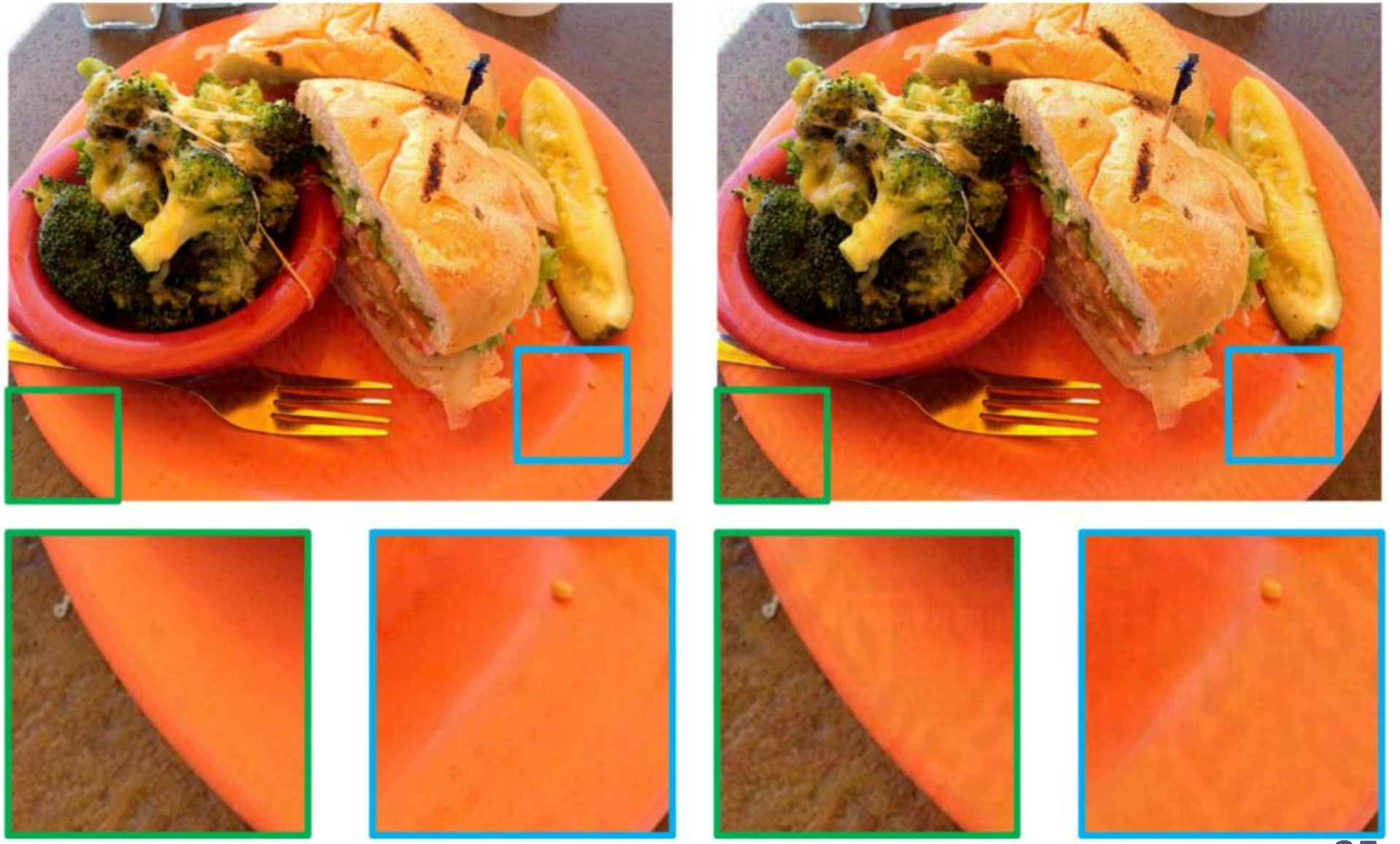
2nd Challenge: Targeted Adversarial Attack

(a) Guide Image (b) Original Image (c) Original Output

(d) Image-space Perturbation (e) Image-space Adversarial Example (f) Output of Image-space Adv. Example

(g) Feature-space Perturbation (h) Feature-space Adversarial Example (i) Output of Feature-space Adv. Example

2nd Challenge: Targeted Adversarial Attack



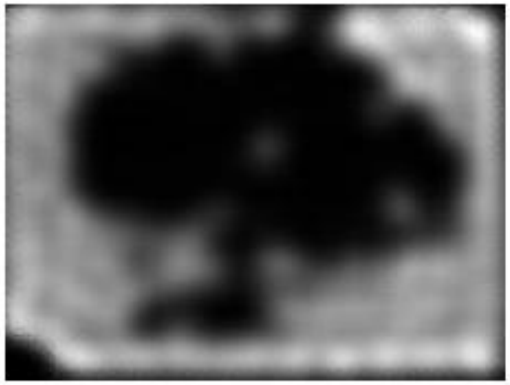
2nd Challenge: Non-targeted Adversarial Attack

(a) Nontargeted Adv. Example by KL loss

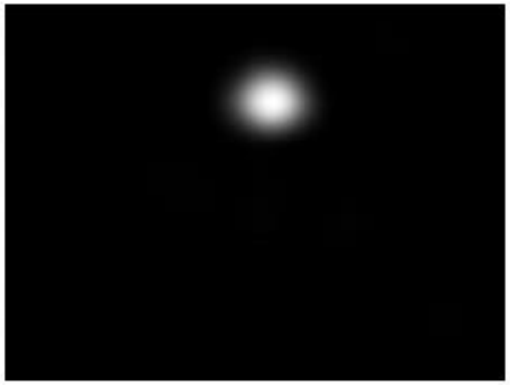
(b) Nontargeted Adv. Example by CC loss

(c) Nontargeted Adv. Example by L_1 loss

(d) Output of Adv. Example by KL loss

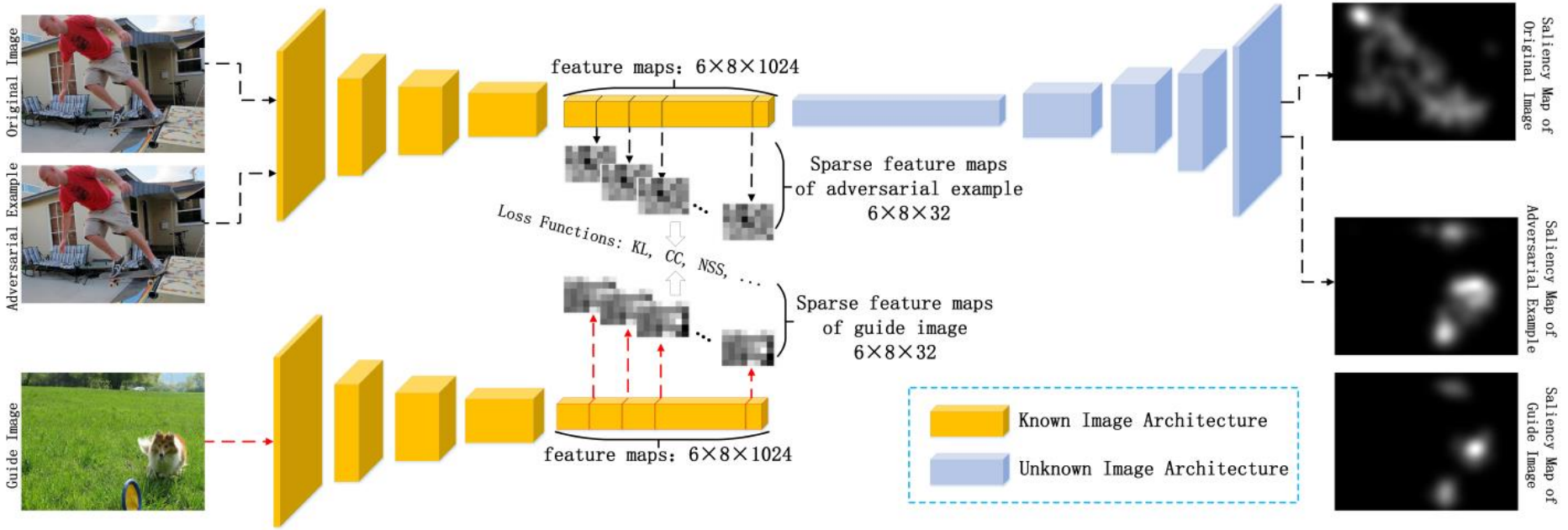


(e) Output of Adv. Example by CC loss



(f) Output of Adv. Example by L_1 loss

2nd Challenge: *White-box* and *Partially-White-box* Adversarial Attack



2nd Challenge: Objective functions

(1) Original objective function in output space (*white-box* case)

$$\begin{cases} \min \mathcal{L}_{img} = \mathcal{L}_1[\,\mathcal{F}(I^*),\ \mathcal{F}(G)] + \lambda_1 \cdot \mathcal{L}_2(I,\ I^*), \\ s.t. \quad \mathcal{L}_2(I,\ I^*) \leq \mathcal{T}. \end{cases} \tag{1}$$

(2) Objective function in subspace (*partially-white-box case*)

$$\begin{cases} \min \mathcal{L}_{sub} = \mathcal{L}_3[\,\mathcal{S}_i^{\downarrow}(I^*),\ \mathcal{S}_i^{\downarrow}(G)] + \lambda_1 \cdot \mathcal{L}_2(I,\ I^*), \\ s.t. \quad \mathcal{L}_2(I,\ I^*) \leq \mathcal{T}. \end{cases} \tag{2}$$

(3) Objective function in subspace for ensemble attack

$$\begin{cases} \min \mathcal{L}_{esb} = \mathcal{L}_3[\sum_{n=1}^{N} \omega_n \cdot \mathcal{S}_{n,i}^{\downarrow}(I^*),\ \sum_{n=1}^{N} \omega_n \cdot \mathcal{S}_{n,i}^{\downarrow}(G)] \\ \qquad\qquad + \lambda_1 \cdot \mathcal{L}_2(I,\ I^*), \\ s.t. \quad \mathcal{L}_2(I,\ I^*) \leq \mathcal{T}. \end{cases} \tag{3}$$

2nd Challenge: *Partially-White-Box* Subspace Attack Algorithm

Algorithm 1 : The *targeted partially-white-box* subspace attack using *Adam* optimizer under ℓ_∞ norm. $\mathbf{m}_t$ represents the 1^{st} gradient momentum vector, while $\mathbf{v}_t$ represents the 2^{nd} gradient momentum vector. $\odot$ is an element-wise product.

Input:

Original benign image I, guide image G (randomly selected); *Partially-white-box* substitute model $\mathcal{F}$;
Tunable hyper-parameters: 1. objective function $\mathcal{L}_3$ (in subspace); 2. index i of hidden layer of $\mathcal{F}$; 3. amount of channels p of sampled feature maps;
Decay factors of gradient momentums: μ_1, μ_2;
Maximum iterations X_1; Maximum perceptual constraint $\mathcal{T}_1$;
Step size of iterative gradient descent α_1; Smoothing term: $\epsilon = 1 \times 10^{-8}$.

Output:

The adversarial example $I^*_{X_1}$; The gradient prior $\mathbf{u}_{X_1}$;

Initialization: $I^*_0 \leftarrow I$, $\mathbf{m}_0 \leftarrow \mathbf{0}$, $\mathbf{v}_0 \leftarrow \mathbf{0}$, $\mathbf{u}_0 \leftarrow \mathbf{0}$, $t \leftarrow 0$
while $(0 \leq t < X_1$ and $\|I, \ I^*_t\|_\infty \leq \mathcal{T}_1)$ **do**
  $t \leftarrow t + 1$; (update the iteration epoch)
  Sampling p feature maps from the i_{th} hidden layer of $\mathcal{F}$ *w.r.t.* the adversarial example and the guide image respectively, *i.e.* $\mathcal{S}^{\downarrow}_i(I^*_{t-1})$ and $\mathcal{S}^{\downarrow}_i(G)$; (using the sampling strategies in Fig. 4)
  Normalizing the sampled feature maps; (following normalizations in Fig. 4)
  $\mathbf{g}_t \leftarrow \nabla_{I^*_{t-1}}(\mathcal{L}_3[\mathcal{S}^{\downarrow}_i(I^*_{t-1}), \mathcal{S}^{\downarrow}_i(G)] + \lambda_1 \cdot \|(I, \ I^*_{t-1})\|_\infty)$;
  $\hat{\mathbf{g}}_t \leftarrow \frac{\mathbf{g}_t}{\|\mathbf{g}_t\|_1}$; (gradient normalization)
  $\mathbf{m}_t \leftarrow \mu_1 \cdot \mathbf{m}_{t-1} + (1 - \mu_1) \cdot \hat{\mathbf{g}}_t$; (update the $\mathbf{m}_t$)
  $\mathbf{v}_t \leftarrow \mu_2 \cdot \mathbf{v}_{t-1} + (1 - \mu_2) \cdot \hat{\mathbf{g}}_t{}^2$; (update the $\mathbf{v}_t$)
  $\hat{\mathbf{m}}_t \leftarrow \mathbf{m}_t / (1 - \mu_1^t)$; (bias correction)
  $\hat{\mathbf{v}}_t \leftarrow \mathbf{v}_t / (1 - \mu_2^t)$; (bias correction)
  $I^*_t \leftarrow \mathrm{Clip}(I^*_{t-1} - \alpha_1 \cdot \frac{1}{\sqrt{\hat{\mathbf{v}}_t} + \epsilon} \odot \hat{\mathbf{m}}_t, [0, 1])$; (update the adversary)
  $\mathbf{u}_t \leftarrow \mathbf{u}_{t-1} + \alpha_1 \cdot \frac{1}{\sqrt{\hat{\mathbf{v}}_t} + \epsilon} \odot \hat{\mathbf{m}}_t$; (update the gradient prior)
end while
return $I^*_{X_1} \leftarrow I^*_t$; $\mathbf{u}_{X_1} \leftarrow \mathbf{u}_t$.

2nd Challenge: Boosting Black-box Attack by using *Partially-White-Box* Subspace Gradient Priors

Algorithm 2 : Using the gradient priors crafted by our subspace attack to boost the *query-based black-box attack* [24, 25].

Input:

Black-box model $\mathcal{Q}$; Benign image I, guide image G; Multiple *partially-white-box* substitute models $\mathcal{F}^1, \mathcal{F}^2, \ldots \mathcal{F}^N$; Objective function $\mathcal{L}_1$ (in output space); Maximum iterations X_2; Maximum perceptual constraint $\mathcal{T}_2$; Step size of gradient descent α_2; $\widetilde{g}$ is the estimation for the true gradient of $\mathcal{Q}$. Hyper-parameters: finite-difference step size τ and δ; learning rate $\eta_{\widetilde{g}}$; (these hyper-parameters follow the settings of [24])

Output:

The adversarial example $\widetilde{I}^*_{X_2}$ that fools the *black-box* target model $\mathcal{Q}$;

Initialization: $\widetilde{I}^*_0 \leftarrow I, t \leftarrow 0, \widetilde{g} \leftarrow \mathbf{0}$;

while $(0 \leq t < X_2$ and $\|I, \widetilde{I}^*_t\|_\infty \leq \mathcal{T}_2)$ **do**

  $t \leftarrow t + 1$; (update the iteration epoch)

  Choose a substitute model $\mathcal{F}^n$, where $n \in 1, 2, \ldots N$ is randomly selected;

  do Algorithm 1 to craft the gradient prior $\mathbf{u}^n_{X_1}$ from $\mathcal{F}^n$;

  $\widetilde{g}_+ \leftarrow \widetilde{g} + \tau \cdot \mathbf{u}^n_{X_1}, \widetilde{g}_- \leftarrow \widetilde{g} - \tau \cdot \mathbf{u}^n_{X_1}$; (coordinate descent estimation)

  $\widetilde{g}'_+ \leftarrow \frac{\widetilde{g}_+}{\|\widetilde{g}_+\|_1}$; $\widetilde{g}'_- \leftarrow \frac{\widetilde{g}_-}{\|\widetilde{g}_-\|_1}$; (gradient normalization)

  $\Delta_t \leftarrow \frac{\mathcal{L}_1[\mathcal{Q}(\widetilde{I}^*_{t-1} + \delta\widetilde{g}'_+), \mathcal{Q}(G)] - \mathcal{L}_1[\mathcal{Q}(\widetilde{I}^*_{t-1} + \delta\widetilde{g}'_-), \mathcal{Q}(G)]}{\tau\delta} \cdot \mathbf{u}^n_{X_1}$;
  (one update step for gradient estimation requires two queries to $\mathcal{Q}$)

  $\widetilde{g} \leftarrow \widetilde{g} + \eta_{\widetilde{g}}\Delta_t$; (update gradient estimation)

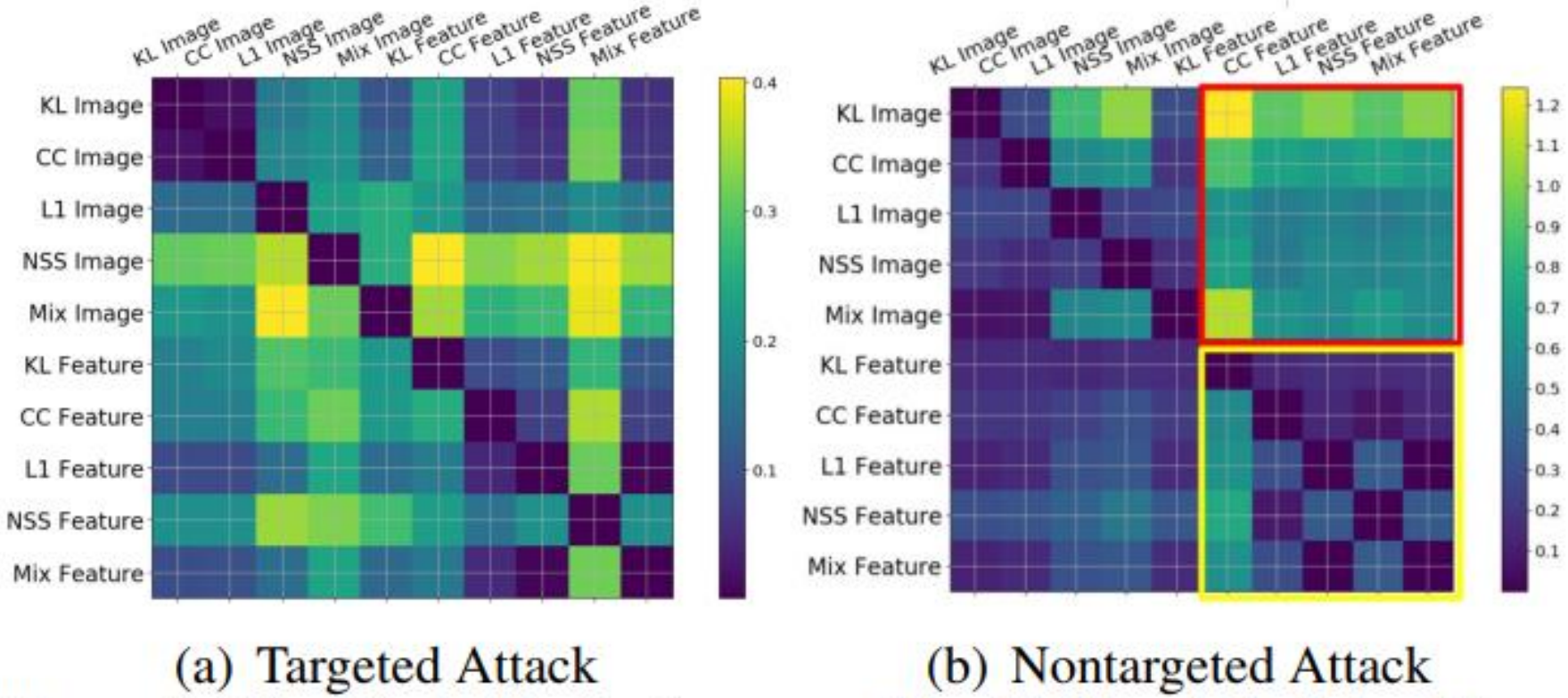
  $\widetilde{I}^*_t \leftarrow \text{Clip}(\widetilde{I}^*_{t-1} - \alpha_2 \cdot \widetilde{g}, [0, 1])$; (update the adversary)

  Update the hyper-parameters of Algorithm. 1, *i.e.* i, p and $\mathcal{L}_3$; (For crafting more diverse gradient priors from subspace)

end while

return $\widetilde{I}^*_{X_2} \leftarrow \widetilde{I}^*_t$.

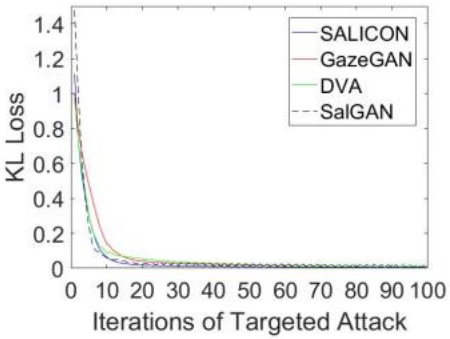
2nd Challenge: Diversity of Subspace Gradient Priors



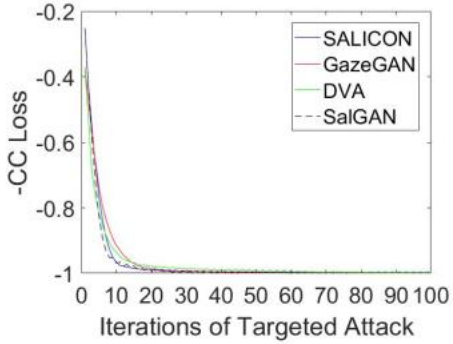
(a) Targeted Attack (b) Nontargeted Attack

Figure 5. The dissimilarity (measured by Kullback-Leibler Divergence [45]) of gradient priors crafted by our subspace attack using different losses and different hidden layers. The brighter the block, the bigger the dissimilarity.

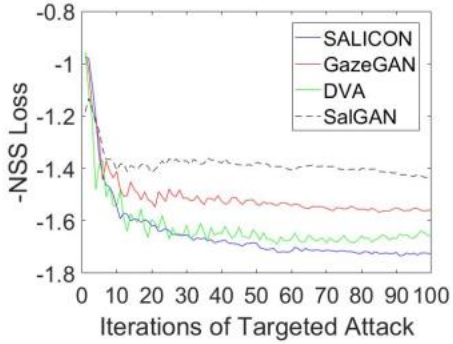
2nd Challenge: Convergence of Algorithm



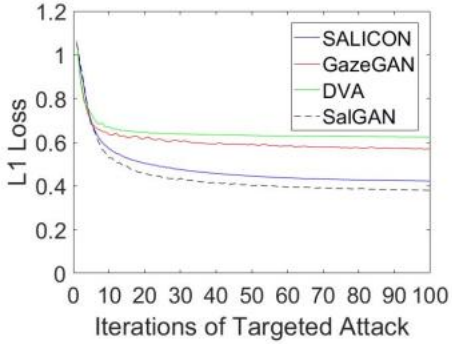
(a) Targeted Attack by KL loss



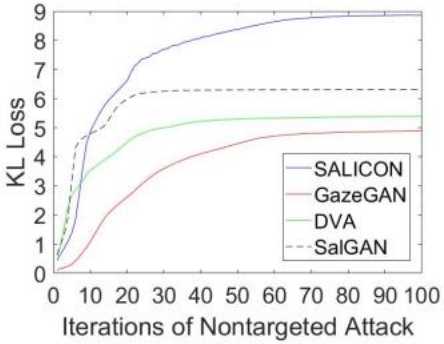
(b) Targeted Attack by CC loss



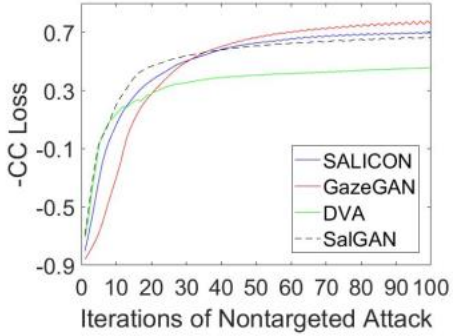
(c) Targeted Attack by NSS loss



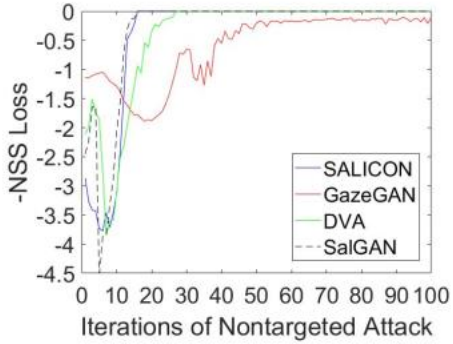
(d) Targeted Attack by L_1 loss



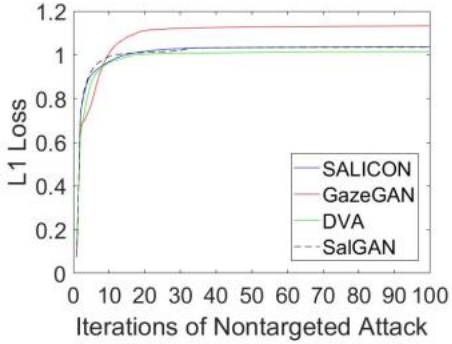
(e) Nontargeted Attack by KL loss



(f) Nontargeted Attack by CC loss

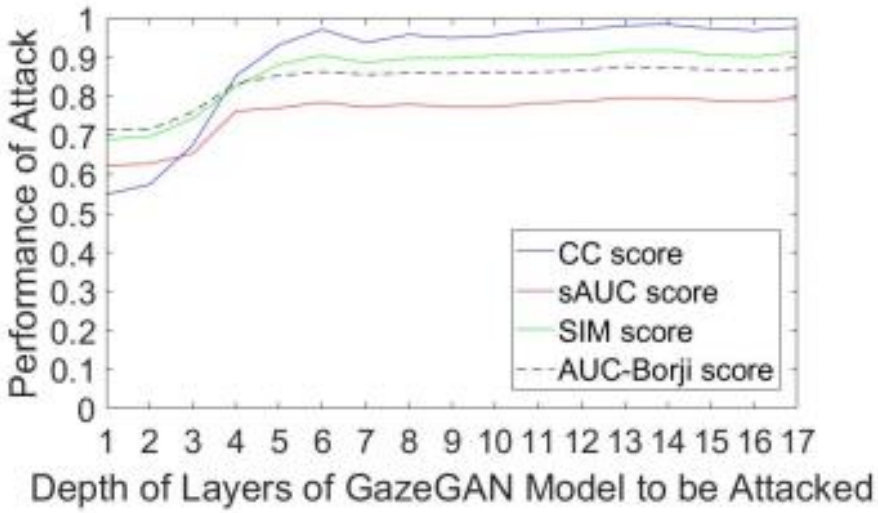


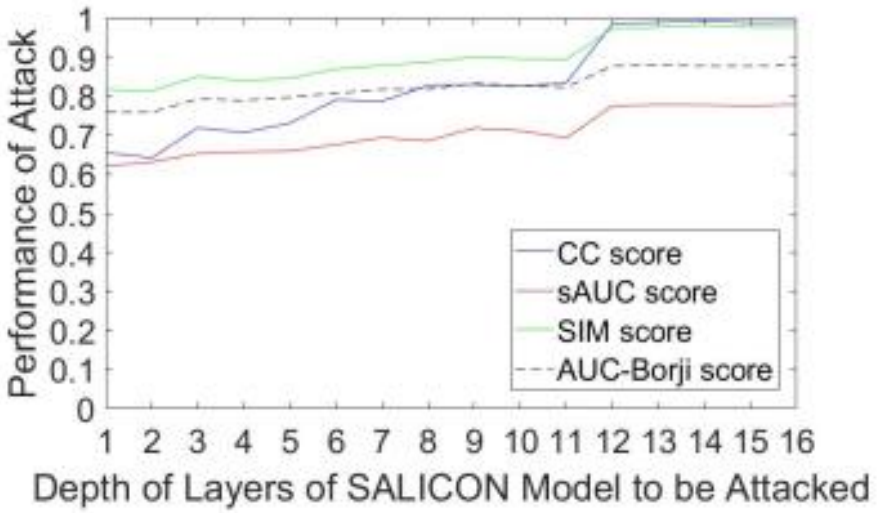
(g) Nontargeted Attack by NSS loss



(h) Nontargeted Attack by L_1 loss

2nd Challenge: Ablation studies of Algorithm

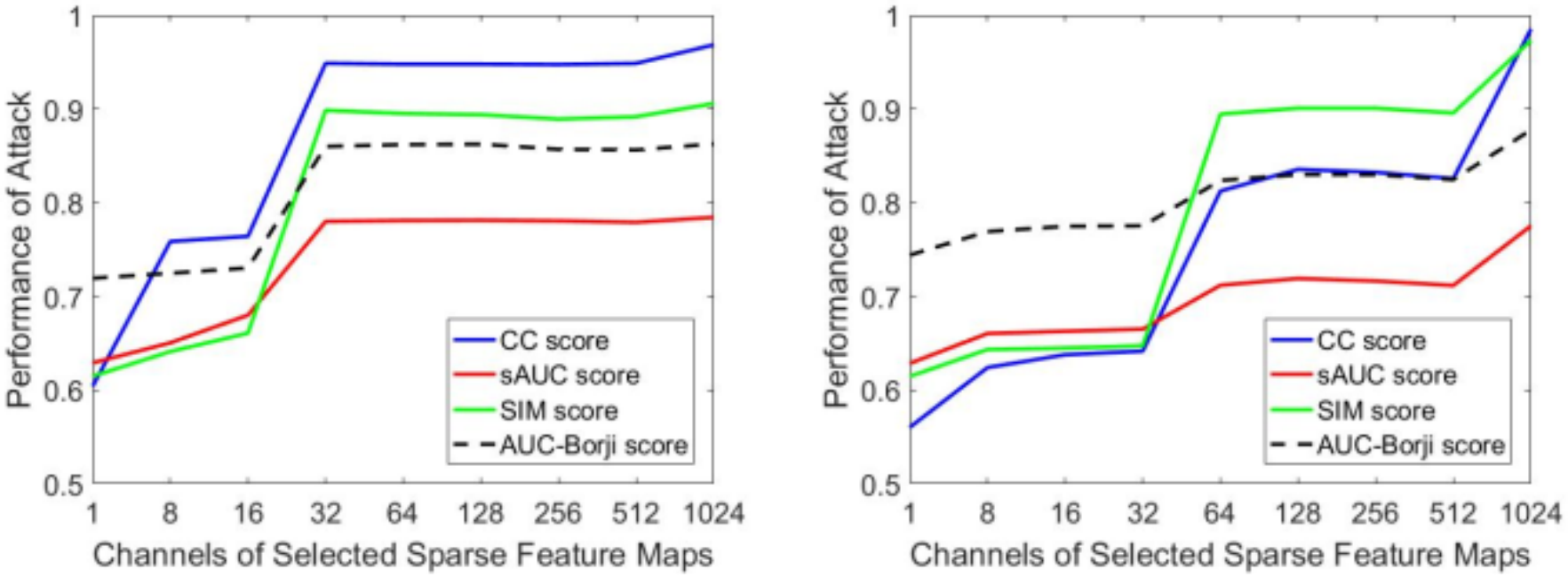




(a) Victim model: GazeGAN [41] (b) Victim model: SALICON [4]

Figure 7. The relationship between the *targeted white-box attack* performance and the depth of the hidden layer to be attacked.

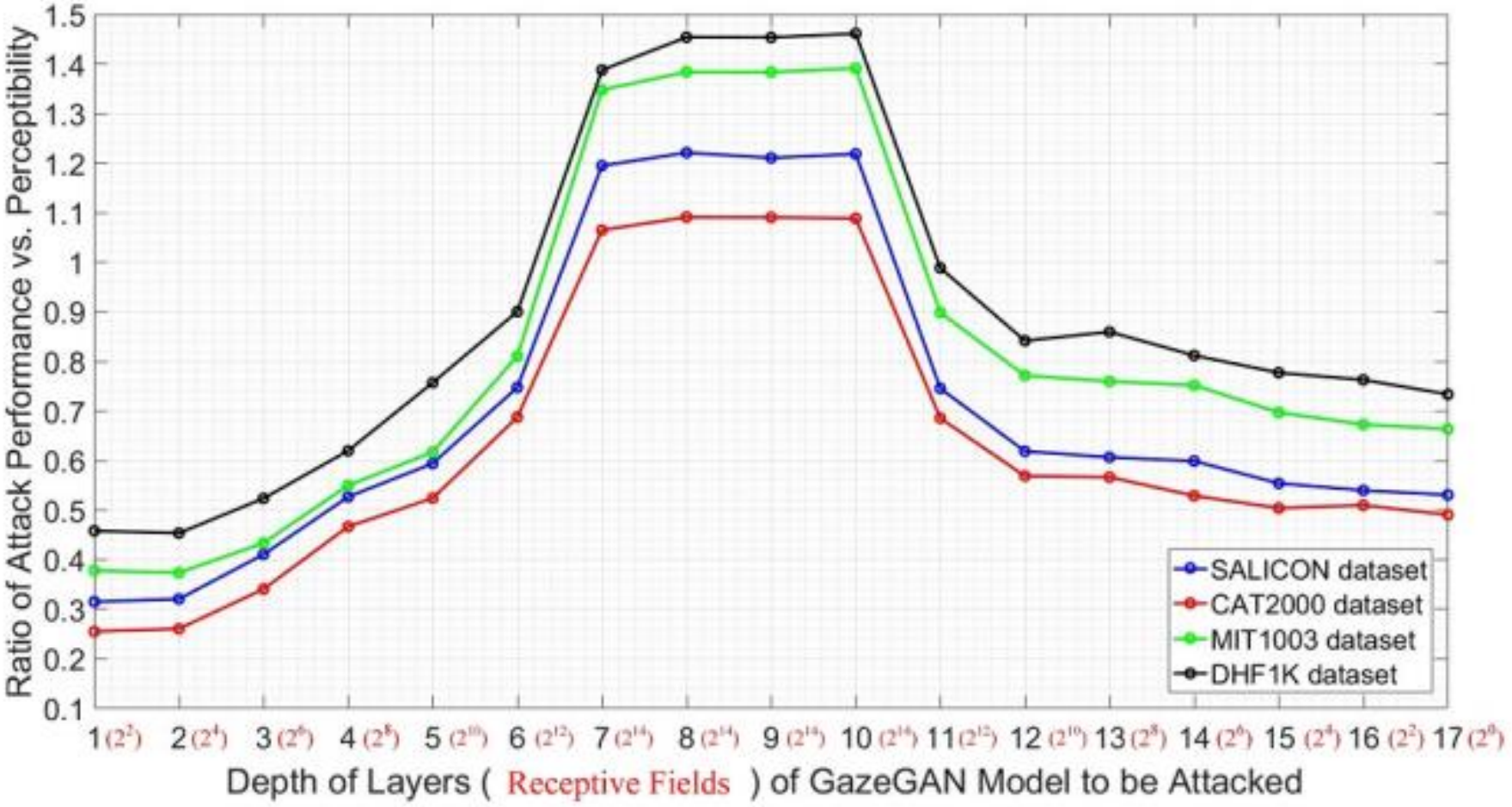
2nd Challenge: Ablation studies of Algorithm



(a) Victim model: GazeGAN [41] (b) Victim model: SALICON [4]

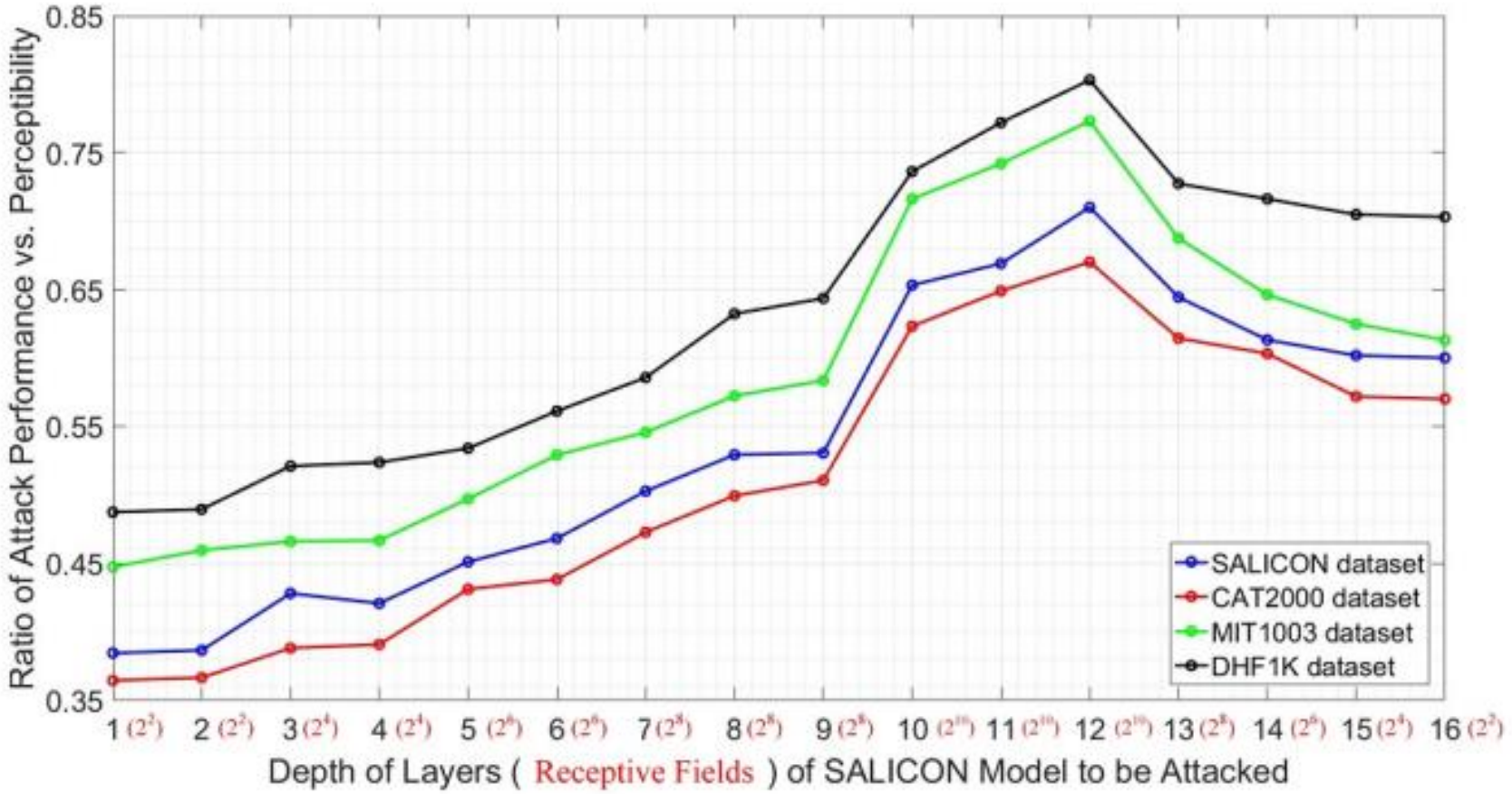
Figure 8. The relationship between the *targeted white-box attack* performance and the numbers of sampled feature map channels.

2nd Challenge: Ablation studies of Algorithm



(a) Victim model: GazeGAN [41]

2nd Challenge: Ablation studies of Algorithm



(b) Victim model: SALICON [4]

Figure 9. Tradeoff between *targeted white-box* attack ability (measured by $CC\uparrow$ metric) and perceptibility (measured by ℓ_2 distance).

2nd Challenge: Comparisons with State-of-the-arts

Table 2. Comparison of proposed attack with other *transfer-based black-box* attacks on different saliency datasets. Fooling ability is measured by performance drops (*i.e. sAUC*↑, *CC*↑, *SIM*↑ metrics). The bigger the drops, the better the attack. We exploit the gradient crafted by our subspace attack as prior information to boost *query-based black-box* attack. In our implementation, we normalize all images into the range [0, 1]. We set maximum ℓ_∞ norm as $\frac{17}{255}$.

Datasets	MIT1003 [51]			CAT2000 [52]			DHF1K [53]			LSUN'17 [49]		
Black-box victim model	GazeGAN [41]			SalGAN [43]			SAM-ResNet [6]			SALICON [4]		
Perceptual constraint (ℓ_∞ norm)	$\frac{17}{255}$			$\frac{17}{255}$			$\frac{17}{255}$			$\frac{17}{255}$		
Evaluation metrics	*CC*↑	*sAUC*↑	*SIM*↑	*CC*↑	*sAUC*↑	*SIM*↑	*CC*↑	*sAUC*↑	*SIM*↑	*CC*↑	*sAUC*↑	*SIM*↑
Random noise	0.0002	0.0002	0.0001	0.0003	0.0003	0.0002	0.0002	0.0003	0.0002	0.0003	0.0002	0.0002
Single-step FGSM [30]	0.0011	0.0006	0.0011	0.0014	0.0011	0.0013	0.0018	0.0012	0.0017	0.0017	0.0009	0.0014
Ensemble Attack using PGD [1]	0.2035	0.1429	0.1883	0.2136	0.1573	0.1939	0.2386	0.1669	0.2014	0.2557	0.1817	0.2125
Ensemble Attack using MIM [18]	0.2454	0.2044	0.2287	0.2600	0.1998	0.2105	0.2703	0.2199	0.2280	0.2821	0.2323	0.2424
Liu's Ensemble Attack [17]	0.2400	0.1853	0.2120	0.2593	0.1894	0.2052	0.2795	0.2080	0.2237	0.2936	0.2233	0.2418
Papernot's Substitute Attack [20]	0.2323	0.1955	0.2191	0.2500	0.1925	0.2099	0.2832	0.2138	0.2286	0.2955	**0.2338**	0.2499
Proposed Algorithm. 2	**0.2551**	**0.2084**	**0.2236**	**0.2890**	**0.2261**	**0.2466**	**0.3031**	**0.2227**	**0.2530**	**0.3115**	0.2307	**0.2580**

Table 3. Comparison of proposed attack with other *query-based black-box* attacks. We report the *average number of queries (AVG.Q*↓*)* together with *CC*↑ drop. The smaller the *AVG.Q*, the better the attack. We exploit the gradient crafted by our subspace attack as prior information to improve query efficiency.

Datasets	MIT1003 [51]		CAT2000 [52]		DHF1K [53]		LSUN'17 [49]	
Black-box victim model	GazeGAN [41]		SalGAN [43]		SAM-ResNet [6]		SALICON [4]	
Evaluation metrics	*CC*↑	*AVG.Q*↓	*CC*↑	*AVG.Q*↓	*CC*↑	*AVG.Q*↓	*CC*↑	*AVG.Q*↓
Random noise	0.0002	——	0.0003	——	0.0002	——	0.0003	——
Yan's zeroth-order attack [25]	**0.2750**	2848	0.2842	2715	0.3108	2654	**0.3303**	2532
Ilyas's bandit method [24]	0.2536	3353	0.2878	3054	**0.3122**	3126	0.3037	3059
Tu's AutoZoom [23]	0.2405	5389	0.2684	5322	0.2759	5543	0.2838	5561
Proposed Algorithm. 2	0.2551	**2320**	**0.2890**	**2173**	0.3031	**2132**	0.3115	**2054**

(a) Guide Image (b) Saliency Map of Guide Image

(c) Original Image (d) Saliency Map of Original Image (e) Adv. Exp. of Image-Space (f) Saliency Map of Adv. Exp. of Image-Space

(g) Original Image (h) Saliency Map of Original Image (i) Adv. Exp. of Feature-Space (j) Saliency Map of Adv. Exp. of Feature-Space

(k) Original Image (l) Saliency Map of Original Image (m) Adv. Exp. of Image-Space (n) Saliency Map of Adv. Exp. of Image-Space

(o) Original Image (p) Saliency Map of Original Image (q) Adv. Exp. of Feature-Space (r) Saliency Map of Adv. Exp. of Feature-Space

(a) Guide Image (b) Saliency Map of Guide Image

(c) Original Image (d) Saliency Map of Original Image (e) Adv. Exp. of Image-Space (f) Saliency Map of Adv. Exp. of Image-Space

(g) Original Image (h) Saliency Map of Original Image (i) Adv. Exp. of Feature-Space (j) Saliency Map of Adv. Exp. of Feature-Space

(k) Original Image (l) Saliency Map of Original Image (m) Adv. Exp. of Image-Space (n) Saliency Map of Adv. Exp. of Image-Space

(o) Original Image (p) Saliency Map of Original Image (q) Adv. Exp. of Feature-Space (r) Saliency Map of Adv. Exp. of Feature-Space

(a) Guide Image (b) Saliency Map of Guide Image

(c) Original Image (d) Saliency Map of Original Image (e) Adv. Exp. of Image-Space (f) Saliency Map of Adv. Exp. of Image-Space

(g) Original Image (h) Saliency Map of Original Image (i) Adv. Exp. of Feature-Space (j) Saliency Map of Adv. Exp. of Feature-Space

(k) Original Image (l) Saliency Map of Original Image (m) Adv. Exp. of Image-Space (n) Saliency Map of Adv. Exp. of Image-Space

(o) Original Image (p) Saliency Map of Original Image (q) Adv. Exp. of Feature-Space (r) Saliency Map of Adv. Exp. of Feature-Space

(a) Guide Image (b) Saliency Map of Guide Image

(c) Original Image (d) Saliency Map of Original Image (e) Adv. Exp. of Image-Space (f) Saliency Map of Adv. Exp. of Image-Space

(g) Original Image (h) Saliency Map of Original Image (i) Adv. Exp. of Feature-Space (j) Saliency Map of Adv. Exp. of Feature-Space

(k) Original Image (l) Saliency Map of Original Image (m) Adv. Exp. of Image-Space (n) Saliency Map of Adv. Exp. of Image-Space

(o) Original Image (p) Saliency Map of Original Image (q) Adv. Exp. of Feature-Space (r) Saliency Map of Adv. Exp. of Feature-Space

Rethinking the 2nd Challenge: *Stronger Attack?*

2nd Challenge: Optimization Literature

- **Zeroth-order optimizations:** *e.g.* bandit, coordinates descent …
- **First-order optimizations:** *e.g.* batch gradient descent, stochastic gradient descent, ,momentum-based stochastic gradient descent, AdaGrad, AdaDelta, RMSProp, Adam, RAdam …
- **Second-order optimizations:** *e.g.* newton method, BFGS, L-BFGS …

2nd Challenge: First-order Optimizations

(1) Original objective function in output space

$$\begin{cases} \min \mathcal{L}_{img} = \mathcal{L}_1[\mathcal{F}(I^*), \mathcal{F}(G)] + \lambda_1 \cdot \mathcal{L}_2(I, I^*), \\ s.t. \quad \mathcal{L}_2(I, I^*) \leq \mathcal{T}. \end{cases} \tag{1}$$

(2) General paradigm of the gradient descent based attacks

$$\begin{cases} g_t = \nabla_{I_t^*} \mathcal{L}_1(\mathbb{F}(G), \mathbb{F}(I_t^*)) + \lambda \cdot \nabla_{I_t^*} \mathcal{L}_2(I, I_t^*), \\ m_{t-1} = \phi(g_1, g_2, ..., g_{t-1}), \quad m_t = \mu_1 \cdot m_{t-1} + \mathbb{N}(g_t), \\ V_{t-1} = \psi(g_1^2, g_2^2, ..., g_{t-1}^2), \quad V_t = \mu_2 \cdot V_{t-1} + (1-\mu_2) \cdot g_t^2, \\ I_{t+1}^* = I_t^* - \alpha \cdot (\frac{1}{\sqrt{V_t + \epsilon}} \odot m_t). \end{cases}$$

- **raw gradient (denoted by *g*): the derivatives of *Eq.1* w.r.t the *I**
- **1st gradient momentum (denoted by *m*): the accumulation of *g*.**
  Goal: accelerate convergence, escape from poor local minima
- **2nd gradient momentum (denoted by *V*) : the accumulation of g^2**
  Goal: adapts the learning rates to different parameters

2nd Challenge: First-order Optimizations

$$\begin{cases} g_t = \nabla_{I_t^*}\mathcal{L}_1(\mathbb{F}(G),\ \mathbb{F}(I_t^*)) + \lambda \cdot \nabla_{I_t^*}\mathcal{L}_2(I,\ I_t^*), \\ m_{t-1} = \phi(g_1, g_2, ..., g_{t-1}), \quad m_t = \mu_1 \cdot m_{t-1} + \mathbb{N}(g_t), \\ V_{t-1} = \psi(g_1^2, g_2^2, ..., g_{t-1}^2), \quad V_t = \mu_2 \cdot V_{t-1} + (1-\mu_2) \cdot g_t^2, \\ I_{t+1}^* = I_t^* - \alpha \cdot (\frac{1}{\sqrt{V_t + \epsilon}} \odot m_t). \end{cases}$$

- **BGD and SGD:** only using raw gradient
- **MSGD:** only using the 1st gradient momentum
- **Adagrad:** only using the 2nd gradient momentum and raw gradient:

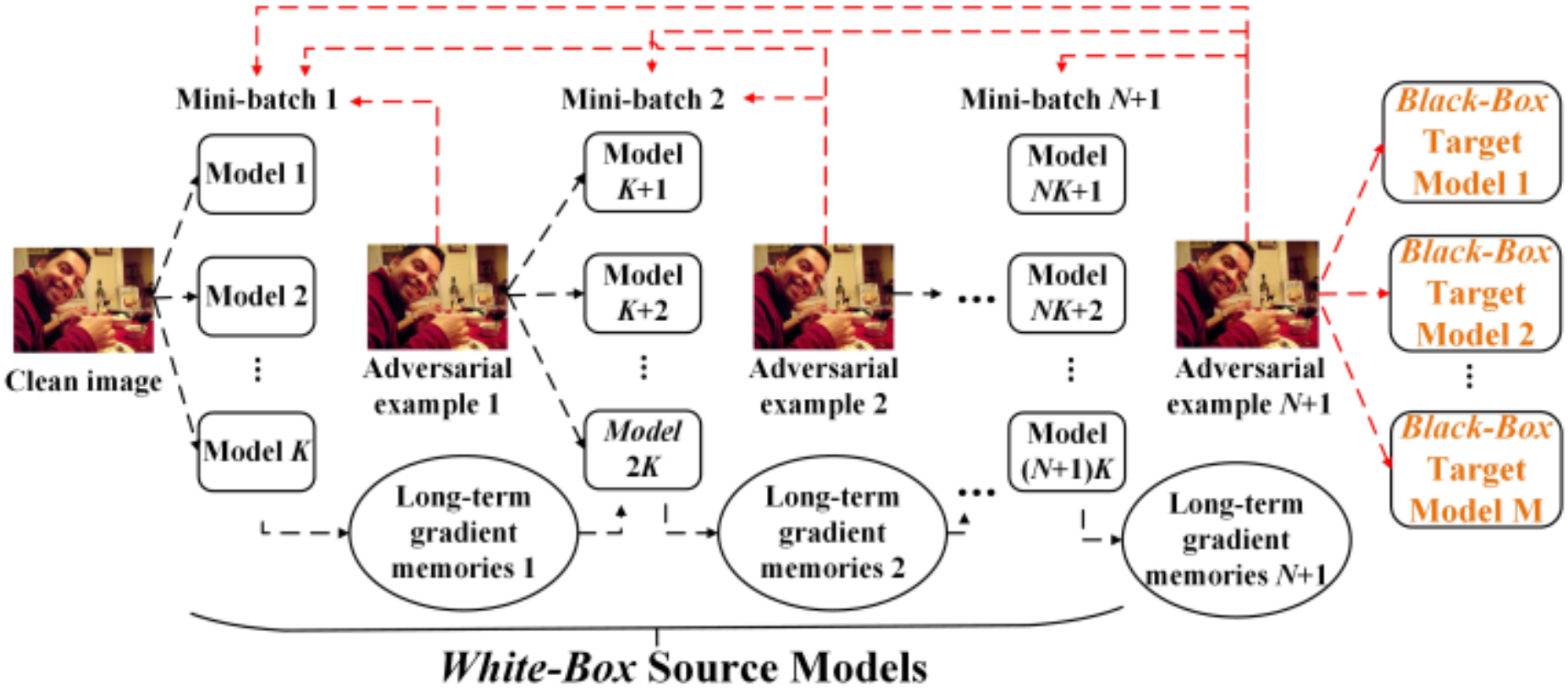
$$I_{t+1}^* = I_t^* - \alpha \cdot (\frac{1}{\sqrt{V_t + \epsilon}} \odot g_t)$$

- **Adadelta, RMSProp:** only using the 2nd gradient momentum and raw gradient, with the modification that the 2nd momentum is the exponential moving averages of the true expected gradient:

$$V_t = \mu_2 \cdot \mathbb{E}[g^2]_{t-1} + (1-\mu_2) \cdot g_t^2,$$

- **Adam:** using both 1st and 2nd momentums

2nd Challenge: Ensemble Attack



2nd Challenge: Intra-batch Ensemble Strategies

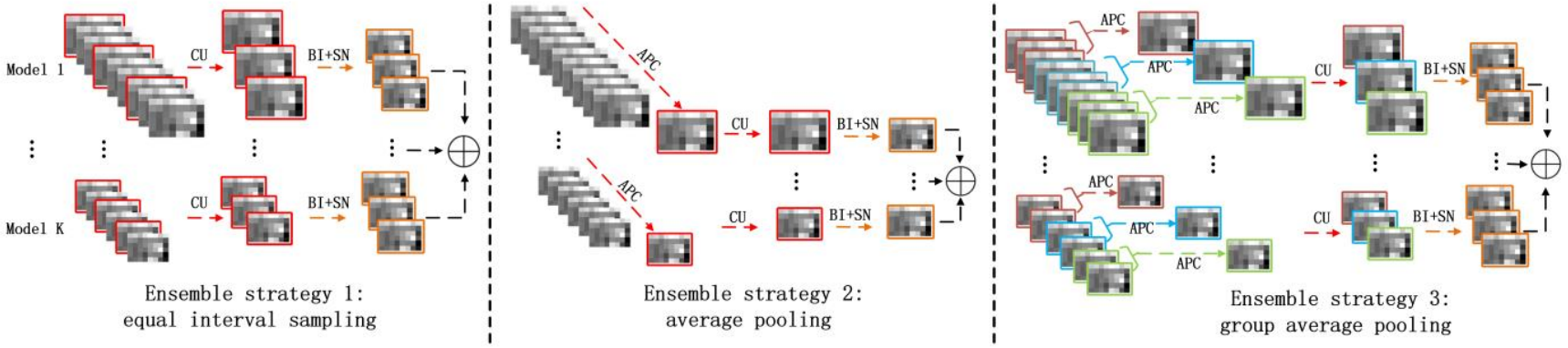


Figure 2: Visualizations of three ensemble strategies in *feature space*. The raw feature maps are processed by the batch normalization and ReLU activation function. **CU**: Channel amount Unification, **BI**: Bilinear Interpolation, **SN**: Softmax Normalization, **APC**: Average Pooling in Channel direction. $\oplus$ represents an elementwise weighted summation.

$$\begin{cases} \min \mathcal{L}_o = \mathcal{L}_1[\sum_{n=1}^{K} \sigma_n \cdot \mathbb{F}_n(I^*),\ \mathcal{F}(G)] + \lambda_1 \cdot \mathcal{L}_2(I,\ I^*), \\ s.t. \quad \mathcal{L}_2(I,\ I^*) \leq \mathcal{T}_1. \end{cases}$$

$$\begin{cases} \min \mathcal{L}_{o,f} = \mathcal{L}_o + \lambda_2 \cdot \mathcal{L}_3[\sum_{n=1}^{K} \omega_n \cdot \mathbb{D}_n(I^*), \sum_{n=1}^{K} \omega_n \cdot \mathbb{D}_n(G)], \\ s.t.\ \mathcal{L}_2(I,\ I^*) \leq \mathcal{T}_1. \end{cases}$$

Intra-batch update rules using short-term gradient momentums

Algorithm 1 : Intra-batch update rules of *SMBEA*. This algorithm is applicable to the first mini-batch. m_t represents the 1^{st} gradient momentum vector, while v_t represents the 2^{nd} gradient momentum vector. $\odot$ is an element-wise product.

Require:

Original clean image I, guide image G (randomly selected);
Intra-batch source models: $\mathbb{F}_1, \mathbb{F}_2, ..., \mathbb{F}_K$;
Decay factors of short-term gradient momentums: μ_1, μ_2;
Smoothing term: ϵ;
Maximum iterations X for single mini-batch;
Maximum perceptual constraint $\mathcal{T}_1$ for the first mini-batch;
Step size of iterative gradient descent α;

Ensure:

An adversarial example I_X^*; The ultimate 1^{st} momentum m_X, and the ultimate 2^{nd} momentum v_X.

Initialization: $I_0^* \leftarrow I, m_0 \leftarrow \mathbf{0}^d, v_0 \leftarrow \mathbf{0}^d, t \leftarrow 0$
while $(0 \leq t < X$ and $\|I, I_t^*\|_1 \leq \mathcal{T}_1)$ **do**
  $t \leftarrow t + 1$; (update the iteration epoch)
  $g_t \leftarrow \nabla_{I_{t-1}^*} \mathcal{L}_{o,f}$; ($\mathcal{L}_{o,f}$ is defined in *Eq.* 2)
  $\hat{g}_t \leftarrow \frac{g_t}{\|g_t\|_1}$; (gradient normalization)
  $m_t \leftarrow \mu_1 \cdot m_{t-1} + (1 - \mu_1) \cdot \hat{g}_t$; (update the m_t)
  $v_t \leftarrow \mu_2 \cdot v_{t-1} + (1 - \mu_2) \cdot \hat{g}_t{}^2$; (update the v_t)
  $\hat{m}_t \leftarrow m_t / (1 - \mu_1^t)$; (bias correction)
  $\hat{v}_t \leftarrow v_t / (1 - \mu_2^t)$; (bias correction)
  $I_t^* \leftarrow \text{Clip}(I_{t-1}^* - \alpha \cdot \frac{1}{\sqrt{\hat{v}_t} + \epsilon} \odot \hat{m}_t)$; (update the adversary)
end while
return $I_X^* \leftarrow I_t^*, m_X \leftarrow m_t, v_X \leftarrow v_t$.

Inter-batch update rules using long-term gradient memories

Algorithm 2 : Inter-batch update rules of *SMBEA*. This algorithm is applicable to all mini-batches, except for the first one, *i.e.* $i > 1$. Notice that the superscript in brackets denotes the batch number, *e.g.* $I_X^{*(i)}$ is the adversary of the i_{th} batch, while the superscript w/o brackets denotes the pow, *e.g.* β_3^i denotes the β_3 to the power i.

Require:

The adversarial example of the previous batch $I_X^{*(i-1)}$, the guide image G;
The 1^{st} gradient momentum of the previous batch $m_X^{(i-1)}$;
The 2^{nd} gradient momentum of the previous batch $v_X^{(i-1)}$;
Maximum perceptual constraint $\mathcal{T}_1^{(i-1)}$ of the previous batch;
Maximum perceptual constraint $\mathcal{T}_1^{(1)}$ of the first batch;
Maximum batch number N;
Intra-batch models of the current batch: $\mathbb{F}_1^{(i)}, \mathbb{F}_2^{(i)}, ..., \mathbb{F}_K^{(i)}$;
Decay factors of short-term gradient momentums: μ_1, μ_2;
Weights of long-term gradient momentums: $\beta_1, \beta_2 \in [0, 1]$;
Decay factor of perceptual constraint: $\beta_3 \in [0, 1]$;

Ensure:

An adversarial example of current batch $I_X^{*(i)}$; The ultimate 1^{st} momentum $m_X^{(i)}$, and ultimate 2^{nd} momentum $v_X^{(i)}$ of the current batch.

1: Initialization: $I_0^{*(i)} \leftarrow I_X^{*(i-1)}$, $m_0^{(i)} \leftarrow \beta_1 \cdot m_X^{(i-1)}$, $v_0^{(i)} \leftarrow \beta_2 \cdot v_X^{(i-1)}$, $\mathcal{T}_1^{(i)} \leftarrow \mathcal{T}_1^{(i-1)} + \beta_3^i \cdot \mathcal{T}_1^{(1)}$, $t \leftarrow 0$
2: **while** $(i \leq N$ and $0 \leq t < X$ and $\|I, I_t^*\|_1 \leq \mathcal{T}_1^{(i)})$ **do**
3: do Step.3 - Step.7 of the Algorithm. 1.
4: $\hat{m}_t^{(i)} \leftarrow \frac{m_t^{(i)}}{(1-\mu_1^t)} + \beta_1 \cdot m_X^{(i-1)}$; (bias correction)
5: $\hat{v}_t^{(i)} \leftarrow \frac{v_t^{(i)}}{(1-\mu_2^t)} + \beta_2 \cdot v_X^{(i-1)}$; (bias correction)
6: $I_t^{*(i)} \leftarrow \text{Clip}(I_{t-1}^{*(i)} - \alpha \cdot \frac{1}{\sqrt{\hat{v}_t^{(i)}} + \epsilon} \odot \hat{m}_t^{(i)})$;
7: **end while**
8: **return** $I_X^{*(i)} \leftarrow I_t^{*(i)}, m_X^{(i)} \leftarrow m_t^{(i)}, v_X^{(i)} \leftarrow v_t^{(i)}, i \leftarrow i + 1$.

2nd Challenge: four Inter-batch recursive variables

- $I_0^{*\,(i)} \leftarrow I_X^{*\,(i-1)}$: we adopt the adversarial example of the previous batch $I_X^{*\,(i-1)}$ as the initial state of the current batch, because $I_X^{*\,(i-1)}$ has learned some adversarial information against multiple models of previous batch.
- $m_0^{(i)} \leftarrow \beta_1 \cdot m_X^{(i-1)}$ *and* $v_0^{(i)} \leftarrow \beta_2 \cdot v_X^{(i-1)}$: we utilize the 1^{st} and 2^{nd} momentums of the previous batch to initialize the momentums of the current batch. These "long-term" gradient momentums preserve the learned adversarial information, and also serve as the regularization to prevent "over-fitting" on the following batch, thus boosting the inter-batch generalizability of the crafted adversary.
- $\mathcal{T}_1^{(i)} \leftarrow \mathcal{T}_1^{(i-1)} + \beta_3^i \cdot \mathcal{T}_1^{(1)}$: we recursively update the maximum perceptual constraint of the current batch (*i.e.* $\mathcal{T}_1^{(i)}$) by adding a loose factor $\beta_3^i \cdot \mathcal{T}_1^{(1)}$ to the perceptual constraint of the previous batch (*i.e.* $\mathcal{T}_1^{(i-1)}$), in order to prevent premature convergence that causes "under-fitting". Besides, by increasing the number of batches, the adversarial example tends to be converged, so we reduce the loose factor via a decay rate β_3^i, where β_3^i denotes the $\beta_3 \in [0, 1]$ to the power i (i is the batch number).

2nd Challenge: Mathematical properties

Property 1: *The effective step-size of inter-batch update rules is invariant to the scale transform of the gradient.*

Proof 1: As shown in Step.6 of Algorithm. 2, assuming $\epsilon = 0$, the effective step-size of the adversarial example at iteration t is $\Delta_t^{(i)} = \alpha \cdot \frac{1}{\sqrt{\hat{v}_t^{(i)}}} \odot \hat{m}_t^{(i)}$. The effective step-size $\Delta_t^{(i)}$ is invariant to the scale transform of gradient, because scaling raw gradient g_t with factor c will be normalized by $\mathbf{L_1}$ norm, *i.e.* $\frac{g_t}{\|g_t\|_1} = \frac{c \cdot g_t}{\|c \cdot g_t\|_1}$. Thus, $\hat{m}_t{}^{(i)}$, $\hat{v}_t{}^{(i)}$, $\Delta_t^{(i)}$ are invariant to the scale transform of the gradient.

2nd Challenge: Mathematical properties

Property 2: *The proposed inter-batch bias corrections can correct for the discrepancy between the expected value of the exponential moving averages* (i.e. $\mathbb{E}[m_t^{(i)}]$ *or* $\mathbb{E}[v_t^{(i)}]$) *and the true expected gradients* (i.e. $\mathbb{E}[\hat{g}_t]$ *or* $\mathbb{E}[\hat{g}_t{}^2]$).

$$\hat{m}_t{}^{(i)} \leftarrow \frac{m_t^{(i)}}{(1-\mu_1^t)} + \beta_1 \cdot m_X^{(i-1)}; \quad \text{(bias correction)}$$

$$\hat{v}_t{}^{(i)} \leftarrow \frac{v_t^{(i)}}{(1-\mu_2^t)} + \beta_2 \cdot v_X^{(i-1)}; \quad \text{(bias correction)}$$

Proof 2: The proposed inter-batch bias corrections are shown in Steps.4-5 of the Algorithm. 2. Here, we derive the bias correction for the 2^{nd} momentum estimate, and the derivation for the 1^{st} momentum is completely analogous.

$$v_t^{(i)} = \mu_2^t \cdot v_0^{(i)} + (1-\mu_2)\sum_{k=1}^{t} \mu_2^{t-k} \cdot \hat{g_k}^2,$$

$$\begin{cases} \mathbb{E}[v_t^{(i)}] = \mathbb{E}[\mu_2^t \cdot v_0^{(i)}] + \mathbb{E}[(1-\mu_2)\sum_{k=1}^{t} \mu_2^{t-k} \cdot \hat{g_k}^2] \\ \quad\quad = \mu_2^t \cdot v_0^{(i)} + \mathbb{E}[\hat{g_t}^2] \cdot (1-\mu_2)\sum_{k=1}^{t} \mu_2^{t-k} + \zeta \\ \quad\quad = \mu_2^t \cdot v_0^{(i)} + \mathbb{E}[\hat{g_t}^2] \cdot (1-\mu_2^t) + \zeta, \end{cases}$$

$$\frac{\mathbb{E}[v_t^{(i)}]}{1-\mu_2^t} = \frac{\mu_2^t \cdot v_0^{(i)}}{1-\mu_2^t} + \mathbb{E}[\hat{g_t}^2],$$

$$\frac{\mathbb{E}[v_t^{(i)}]}{(1-\mu_2^t)} - \frac{\mu_2^t \cdot v_0^{(i)}}{1-\mu_2^t} + \frac{1 \cdot v_0^{(i)}}{1-\mu_2^t} = \frac{1 \cdot v_0^{(i)}}{1-\mu_2^t} + \mathbb{E}[\hat{g_t}^2],$$

$$\frac{\mathbb{E}[v_t^{(i)}]}{(1-\mu_2^t)} + \beta_2 \cdot v_X^{(i-1)} = \frac{1}{1-\mu_2^t} \cdot \beta_2 \cdot v_X^{(i-1)} + \mathbb{E}[\hat{g_t}^2].$$

2nd Challenge: Performance comparison

Table 1: Comparison under *black-box* setting. Fooling ability is measured by performance drop: *CC*↑ for LSUN'17, *MSE*↓ for other datasets.

Datasets (original performance)	Cityspaces (*MSE*=0.0139)			Facades (*MSE*=0.0521)			Google Satellite (*MSE*=0.0255)			LSUN'17 (*CC*=0.7748)		
Target *black-box* model	pix2pix U-Net			Global pix2pixHD			Local pix2pixHD			SALICON		
Number of mini-batches	1	3	5	1	3	5	1	3	5	1	3	5
Percep. cons. ($\mathbf{L_1}$ norm)	$1.0e^{-2}$	$2.0e^{-2}$	$2.4e^{-2}$	$1.2e^{-2}$	$2.4e^{-2}$	$2.8e^{-2}$	$1.0e^{-2}$	$2.0e^{-2}$	$2.4e^{-2}$	$3.4e^{-2}$	$6.6e^{-2}$	$7.8e^{-2}$
Random noise	+0.0002	+0.0008	+0.0011	+0.0003	+0.0010	+0.0013	+0.0002	+0.0006	+0.0011	-0.0002	-0.0002	-0.0003
Ensemble Attack using PGD	+0.0108	+0.0169	+0.0174	+0.0104	+0.0133	+0.0166	+0.0074	+0.0083	+0.0099	-0.0022	-0.0264	-0.0511
Ensemble Attack using C&W	+0.0113	+0.0170	+0.0173	+0.0097	+0.0131	+0.0168	+0.0068	+0.0087	+0.0096	-0.0763	-0.2117	-0.2452
Ensemble Attack using MIM	+0.0116	+0.0193	+0.0227	+0.0125	+0.0162	+0.0178	+0.0093	+0.0099	+0.0113	-0.0771	-0.2417	-0.2880
Liu's Ensemble Attack	+0.0118	+0.0194	+0.0230	+0.0129	+0.0165	+0.0184	+0.0098	+0.0105	+0.0116	-0.0780	-0.2533	-0.2941
Proposed *SMBEA*	**+0.0148**	**+0.0213**	**+0.0264**	**+0.0155**	**+0.0230**	**+0.0275**	**+0.0108**	**+0.0137**	**+0.0145**	**-0.0871**	**-0.3017**	**-0.4180**

Table 2: Evading defense: attack performance comparison against the adversarially trained *black-box* models. LSUN'17 is the test set.

Target model	Attack	No. of Batch: 3	No. of Batch: 5	No. of Batch: 7
GazeGAN	I-FGSM	-0.0125	-0.0177	-0.0206
GazeGAN	MIM	-0.1315	-0.1582	-0.2063
GazeGAN	*SMBEA*	**-0.2190**	**-0.2731**	**-0.3255**
SAM-ResNet	I-FGSM	-0.0164	-0.0238	-0.0295
SAM-ResNet	MIM	-0.1622	-0.1900	-0.2258
SAM-ResNet	*SMBEA*	**-0.2317**	**-0.2996**	**-0.3484**

2nd Challenge: Ablation studies

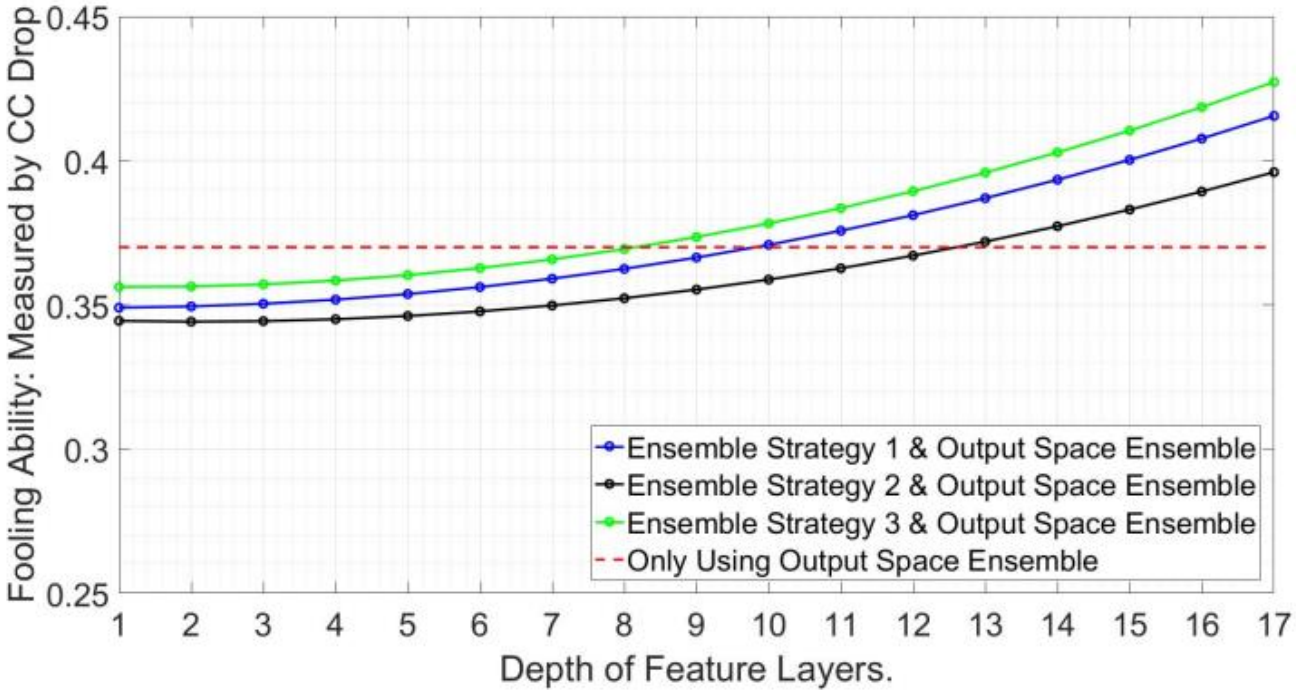


Figure 5: The relationship between fooling ability (against the source models) and the depth of feature layers. We compare different ensemble strategies when fusing 4 source models including GazeGAN (Che et al. 2019), Globalpix2pix (Wang et al. 2018), SAM-ResNet (Cornia et al. 2018), and SalGAN (Pan et al. 2017).

2nd Challenge: Ablation studies

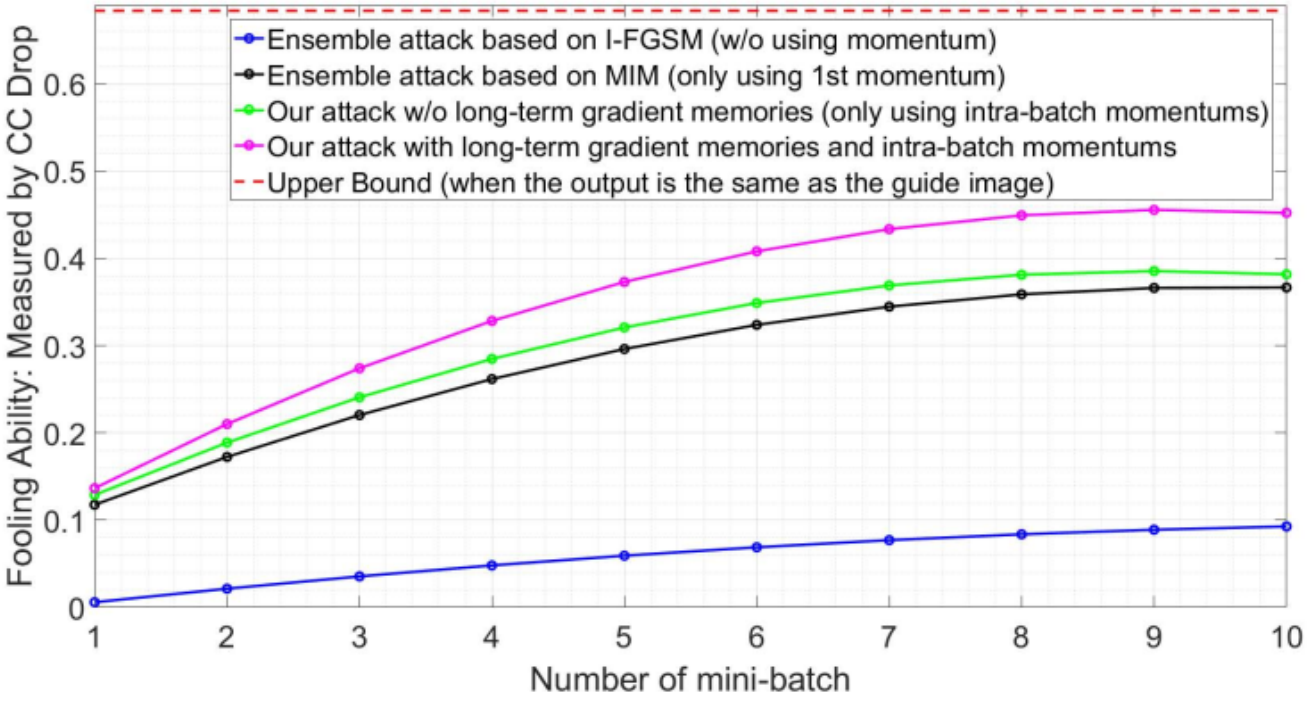
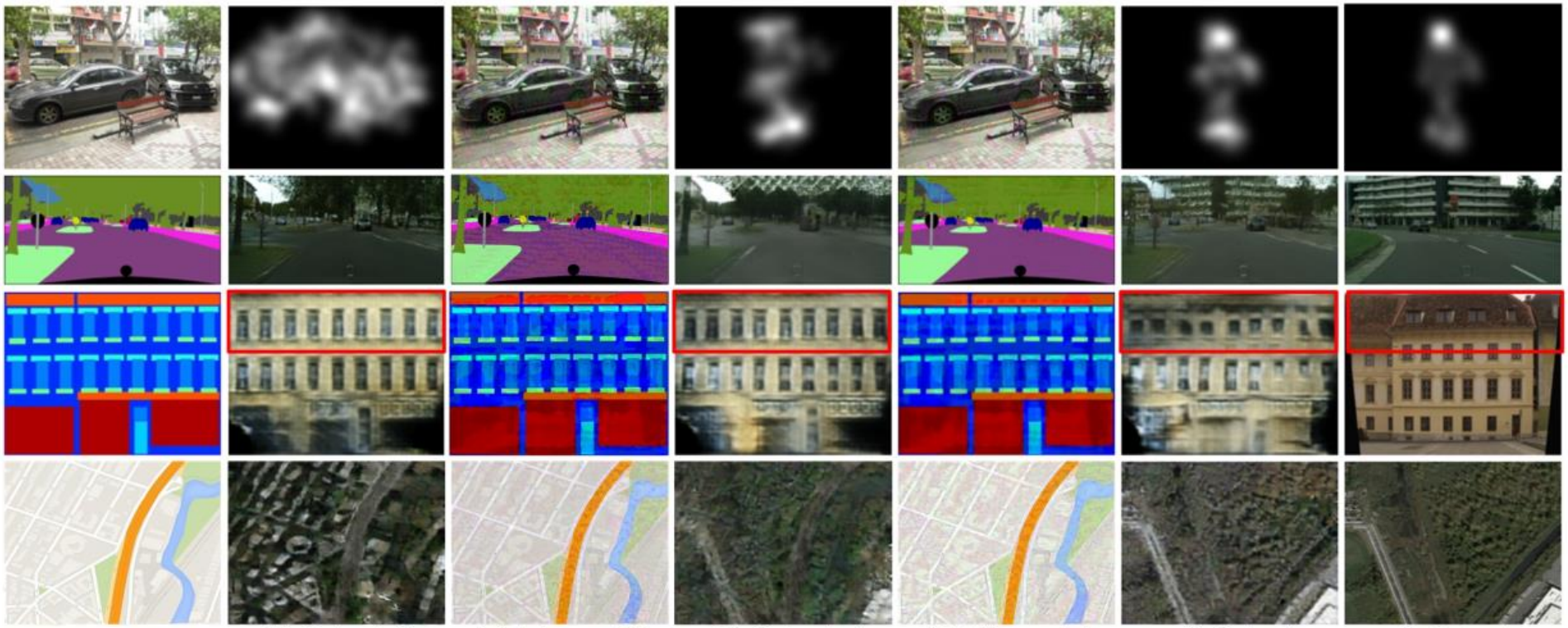


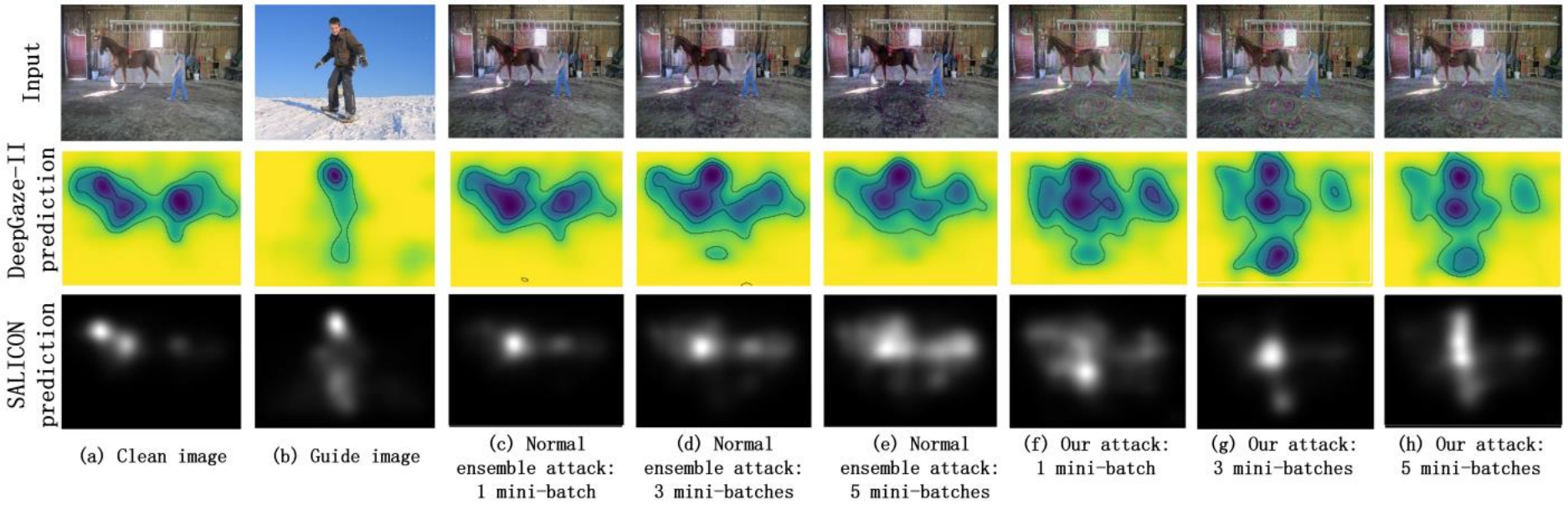
Figure 6: Ablation study of our attack. We use the online *black-box* SALICON model as the target model. With the increase of batch number, the perceptual constraint increases (as shown in Step.1 of Algorithm.2). For fair comparison, in the same batch, different competing attacks adopt the same perceptual constraint.

2nd Challenge: Qualitative comparisons on multiple pix2pix tasks, e.g. saliency region detection, image translations



(a) Clean image | (b) Output of Clean image | (c) Adversarial example of MIM | (d) Output of MIM | (e) Adversarial example of our attack | (f) Output of our attack | (g) Output of Guide image

2nd Challenge: Qualitative comparisons against two online *black-box* systems in real world



Online victim systems:

1. DeepGaze-II: https://deepgaze.bethgelab.org/
2. SALICON: http://salicon.net/demo/

2nd Challenge: Code repository of attacks

Table 1: The library of Gradient back-propagation based adversarial attack methods.

Number	Method	Gradient Descent Method	Frequency	Perceptual Constraint	Transferability	Goal	Dataset	Box-constraint
$attack_1$	*AdaEq2M*	1_{st}-randn momentum & 2_{nd}-plus momentum	Iterative	L_1, L_2, L_∞, *SSIM*, *MS-SSIM*	white-box & black-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_2$	Adam-decay-IM	Adam: vanilla 1_{st} momentum & 2_{nd}-decay momentum	Iterative	L_1, L_2, L_∞, *SSIM*, *MS-SSIM*	white-box & black-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_3$	Adam-vanilla-IM	Adam: vanilla 1_{st} momentum & vanilla 2_{nd} momentum	Iterative	L_1, L_2, L_∞, *SSIM*, *MS-SSIM*	white-box & black-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_4$	MSGD-MIM	MSGD: vanilla 1_{st} momentum	Iterative	L_1, L_2, L_∞, *SSIM*, *MS-SSIM*	white-box & black-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_5$	SGD-IFGSM	SGD: no momentum	Iterative	L_1, L_2, L_∞, *SSIM*, *MS-SSIM*	white-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_6$	SGD-IFGV	SGD: no momentum	Iterative	L_1, L_2, L_∞, *SSIM*, *MS-SSIM*	white-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_7$	AdaGrad-plus-IM	AdaGrad: 2_{nd}-plus momentum	Iterative	L_1, L_2, L_∞, *SSIM*, *MS-SSIM*	white-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped

2nd Challenge: Code repository of attacks

$attack_8$	AdaGrad-decay-IM	AdaGrad: 2_{nd}-decay momentum	Iterative	$L_1, L_2, L_\infty,$ *SSIM*, *MS-SSIM*	white-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_9$	*AdaEq2M-V2*	vanilla 1_{st} momentum & 2_{nd}-plus momentum	Iterative	$L_1, L_2, L_\infty,$ *SSIM*, *MS-SSIM*	white-box & black-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_{10}$	Hot&Cold	SGD: no momentum	Iterative	*SSIM*	white-box	Targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_{11}$	PGD	SGD: no momentum Start from a random state	Iterative	L_1, L_2, L_∞	white-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped
$attack_{12}$	Adam-decay-C&W	Adam: only using short-term momentums	Iterative	L_0, L_2, L_∞	white-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	automatic
$attack_{13}$	*AdaEq2M*-C&W	1_{st}-randn momentum & 2_{nd}-plus momentum	Iterative	L_0, L_2, L_∞	white-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	automatic
$attack_{14}$	**proposed *SMBEA***	short-term momentums long-term momentums	Iterative	L_1, L_2, L_∞	white-box & black-box	Targeted & Non-targeted	SALICON, MIT1003 CAT2000, DHF1K	clipped

2nd Challenge: Code repository of victim models

Table 2: The library of source models: state-of-the-art deep saliency models.

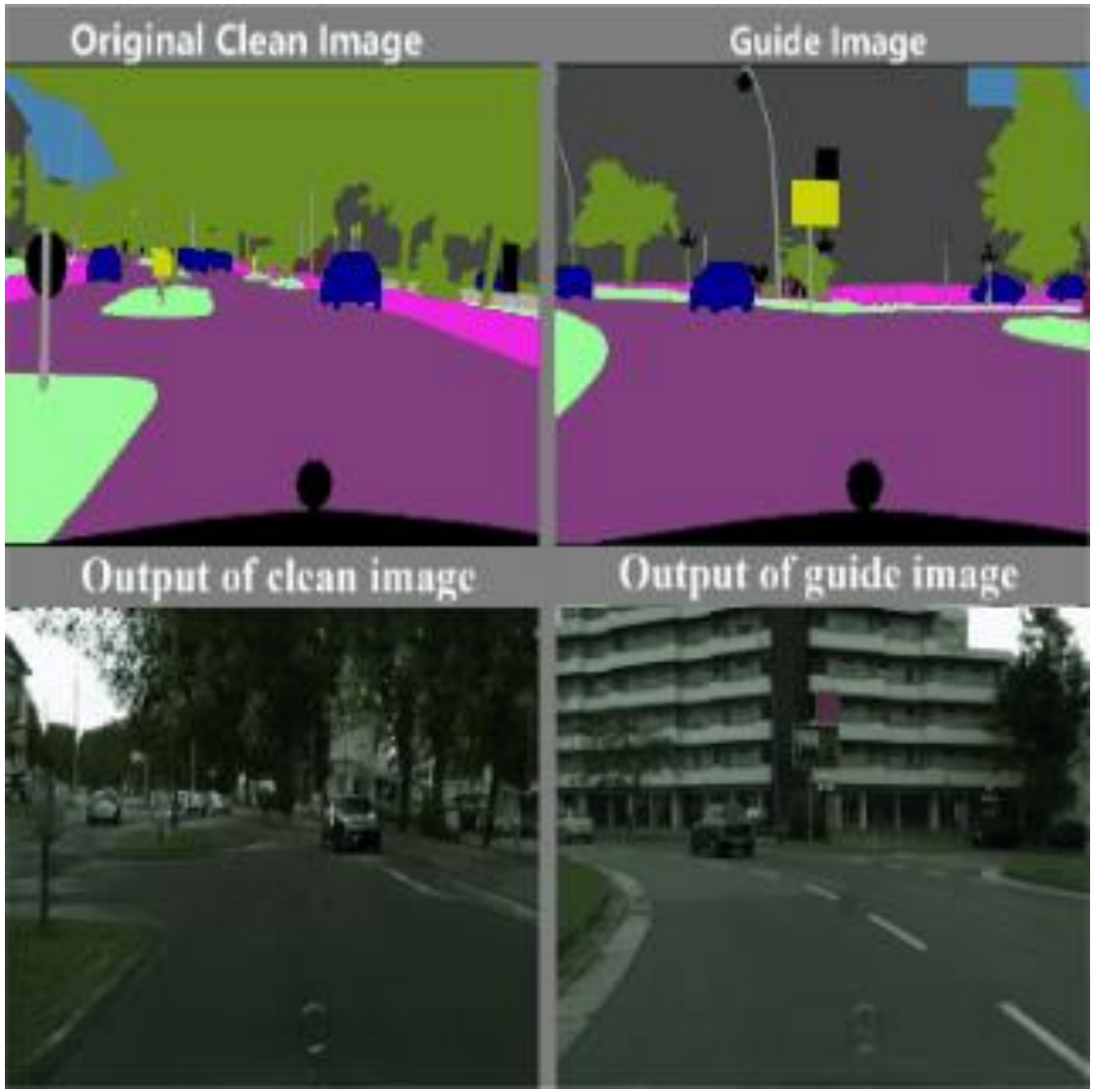
Number	Model	Backbone Network	Multi-scale	Convolution Form	Attention Module	Refinement	Adversarial Training
$model_1$	$GazeGAN_1$	U-Net	Single-stream	Standard Convolution	∅	∅	∅
$model_2$	$SALICON_2$	VGG16	Single-stream	Standard Convolution	∅	∅	∅
$model_3$	Global-pix2pix	ResNet	Single-stream	Standard Convolution	∅	∅	∅
$model_4$	$GazeGAN_2$	U-Net	Multi-stream	Standard Convolution	∅	∅	YES
$model_5$	DCN_LSTM_1	VGG16 & Inception-ResNet	Multi-stream	Standard Convolution & Modulated Deformable Convolution	Squeeze-and-Excitation & Pyramid Spatial Attention	ConvLSTM	∅
$model_6$	DCN_2	VGG16 & Inception-ResNet	Multi-stream	Standard Convolution & Modulated Deformable Convolution	Squeeze-and-Excitation & Pyramid Spatial Attention	∅	∅
$model_7$	SAM_VGG_1	VGG16	Single-stream	Standard Convolution & Dilated Convolution	∅	ConvLSTM	∅
$model_8$	SAM_VGG_2	VGG16	Single-stream	Standard Convolution & Dilated Convolution	Softmax Spatial Attention	ConvLSTM	∅
$model_9$	SAM_ResNet	ResNet	Single-stream	Standard Convolution & Dilated Convolution	Softmax Spatial Attention	ConvLSTM	YES

2nd Challenge: Code repository of victim models

$model_{10}$	CSC_Net	U-Net	Multi-stream	Standard Convolution	Softmax Spatial Attention	∅	∅
$model_{11}$	SalGAN_BCE	VGG16	Single-stream	Standard Convolution	∅	∅	∅
$model_{12}$	DCN_Inception	Inception-ResNet	Single-stream	Standard Convolution	∅	∅	∅
$model_{13}$	DeepGaze_only_VGG	VGG19	Single-stream	Standard Convolution	∅	∅	∅
$model_{14}$	DCN_SAM_VGG	VGG19	Single-stream	Standard Convolution & Modulated Deformable Convolution	Softmax Spatial Attention	ConvLSTM	∅
$model_{15}$	Local-pix2pix	ResNet	Multi-stream	Modulated Deformable Convolution	Squeeze-and-Excitation & Pyramid Spatial Attention	∅	∅
$model_{16}$	DenseSal	DenseNet	Single-stream	Standard Convolution	∅	∅	∅

2nd Challenge: Visualizations

2nd Challenge: Visualizations



2nd Challenge: Visualizations

How to defend these attacks?

1. *Adversarial training*
2. *Detection/Denoising*
3. *Robust architecture*
4. *Loss function design*
5. *Your solution? ……*

References

Publications:

- Z. Che, A. Borji, G. Zhai, X. Min, G. Guo, P.L. Callet. *How is Gaze Influenced by Image Transformations? Dataset and Model.* IEEE TIP 2019.
- Z. Che, A. Borji, G. Zhai, S. Ling, J. Li, P.L. Callet. *A New Ensemble Adversarial Attack Powered by Long-term Gradient Memories.* AAAI 2020.
- Z. Che, A. Borji, G. Zhai, S. Ling, Y. Tian, J. Li, G. Guo, P.L. Callet. *Boosting Black-Box Attack by using Partially-White-Box Subspace Gradient Priors against Deep Saliency Models.* Under review.

Code repositories:

- https://github.com/CZHQuality/Sal-CFS-GAN
- https://github.com/CZHQuality/AAA-Pix2pix

Dataset:

- https://drive.google.com/drive/folders/1qXVU6deYqdM2ZTyJQTxJyVWId9bTUceJ?usp=sharing

Thank you very much!

LS2N

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY