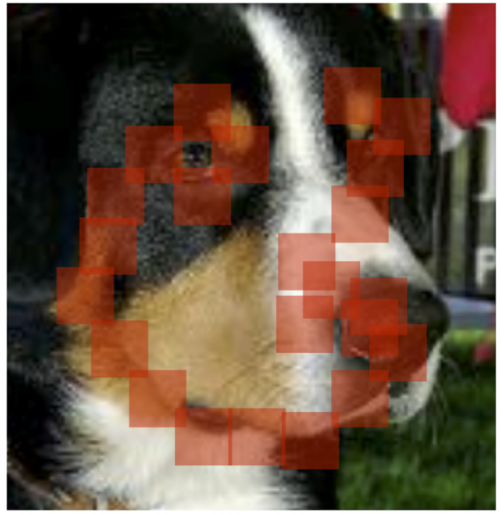
Learning what and where to attend

Drew Linsley, Dan Shiebler, Sven Eberhardt, Thomas Serre

ICLR 2019

Human clicks



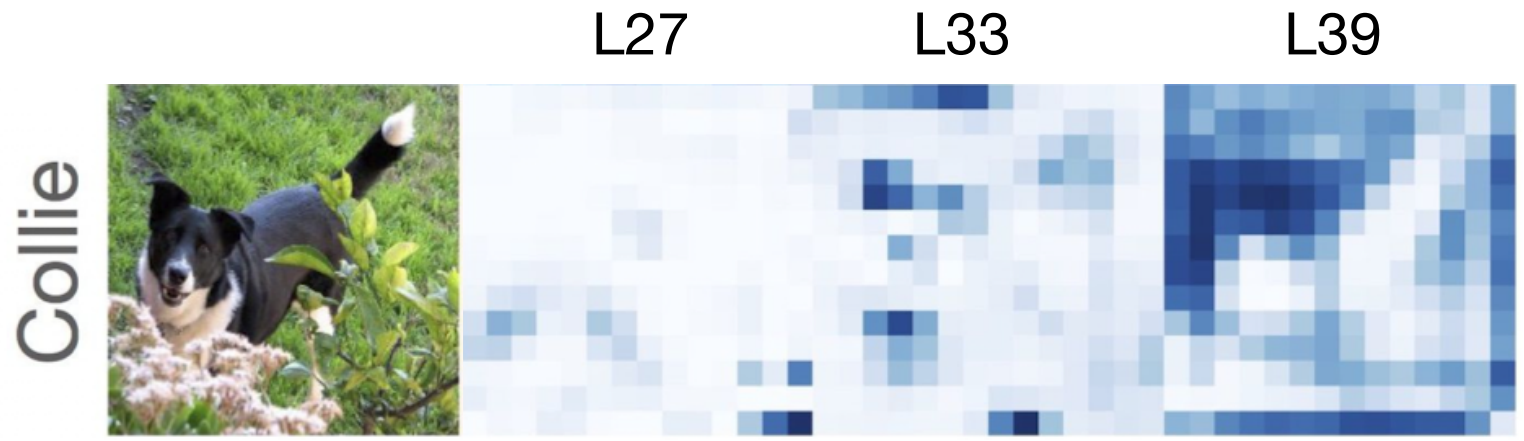
Presented by
Fiona Muntwyler

Visible to DCN

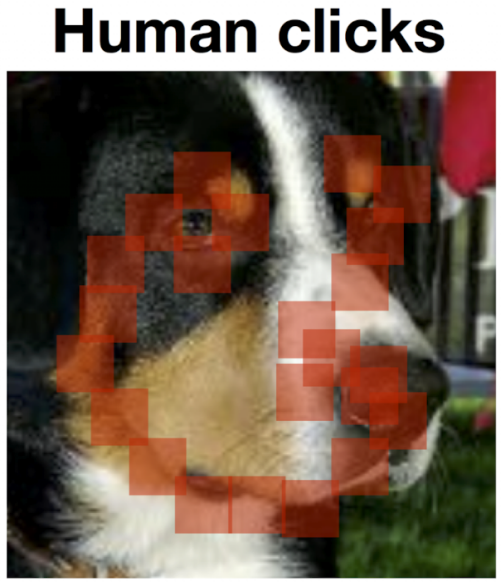
7 seconds to recognize

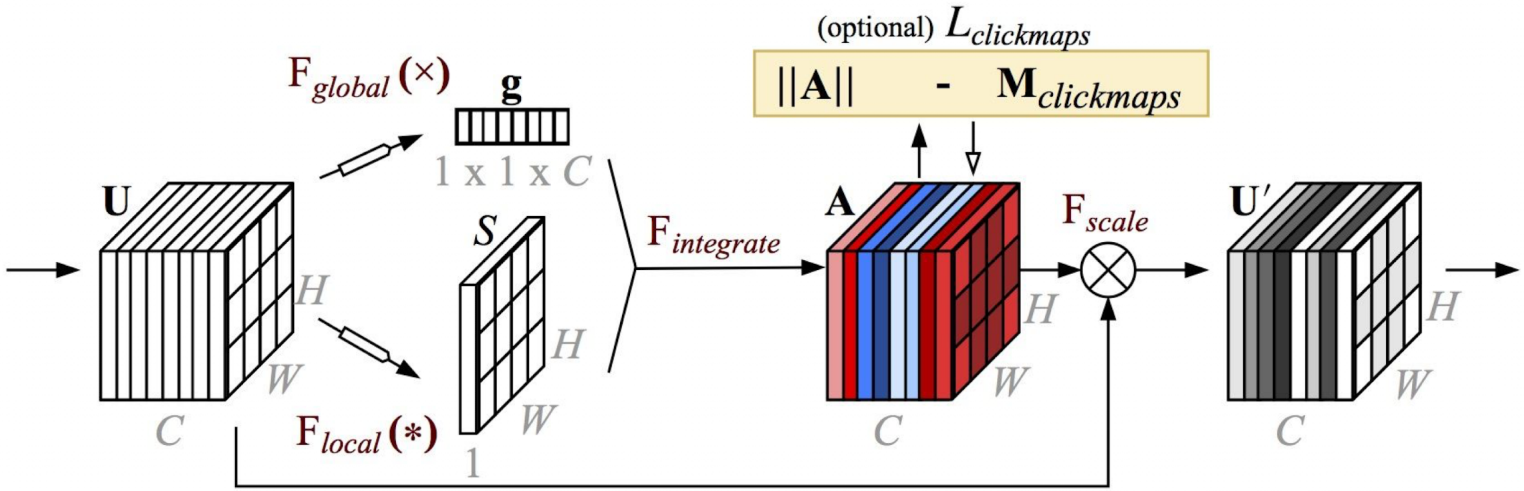
1. Problem description and motivation
2. Related work
3. ClickMe.ai
4. Network architecture
5. Evaluation
6. Discussion
7. Conclusion

1. **Problem description and motivation**
2. Related work
3. ClickMe.ai
4. Network architecture
5. Evaluation
6. Discussion
7. Conclusion


L27
L33
L39
Collie

Contributions





1. Problem description and motivation
2. **Related work**
3. ClickMe.ai
4. Network architecture
5. Evaluation
6. Discussion
7. Conclusion

LEARNING WHAT AND WHERE TO ATTEND

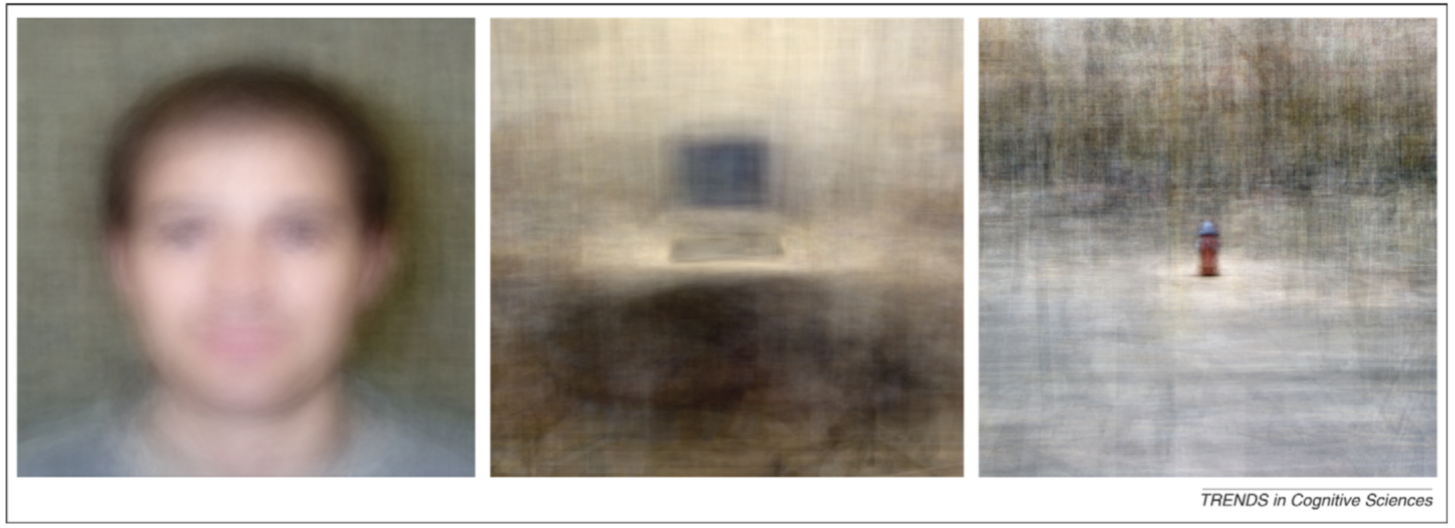
Drew Linsley, Dan Shiebler, Sven Eberhardt and Thomas Serre
Department of Cognitive Linguistic & Psychological Sciences
Carney Institute for Brain Science
Brown University
Providence, RI 02912
{drew_linsley,thomas_serre}@brown.edu

Attention models in human vision

- Local features

Source: https://www.directindustry.de/prod/vag-group/product-26045-1884491.html

- Global features



Source: Oliva and Torralba, 2007

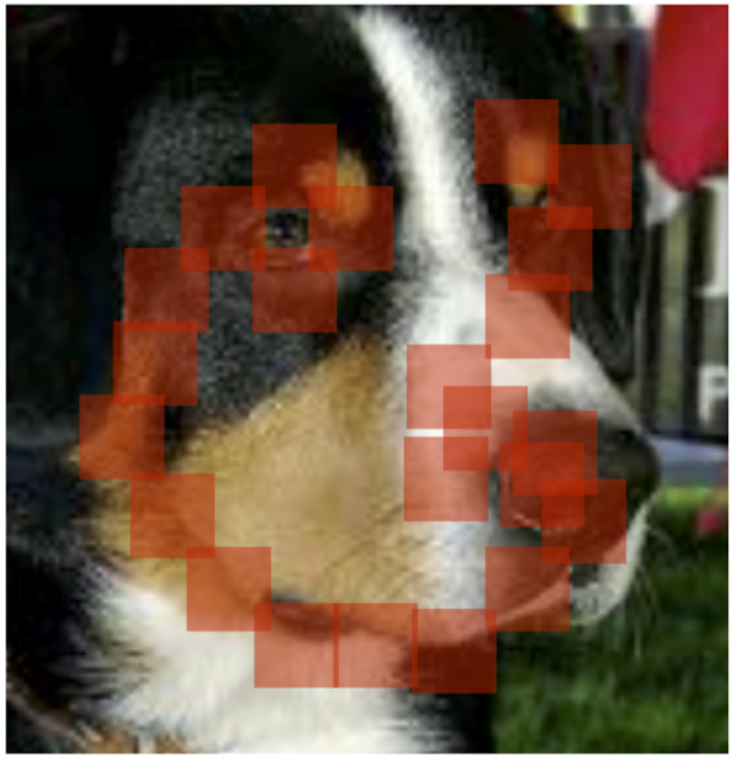
Humans-in-the-loop computer vision and attention datasets

- Online games
- Attention dataset: Salicon

1. Problem description and motivation
2. Related work
3. **ClickMe.ai**
4. Network architecture
5. Evaluation
6. Discussion
7. Conclusion

ClickMe.ai

Human clicks



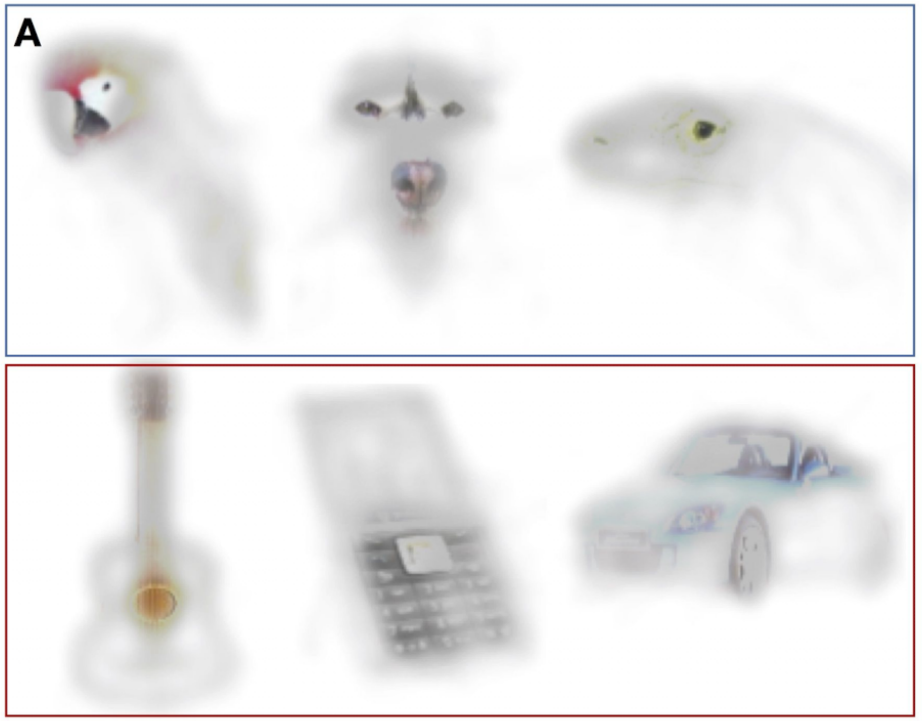
Visible to DCN

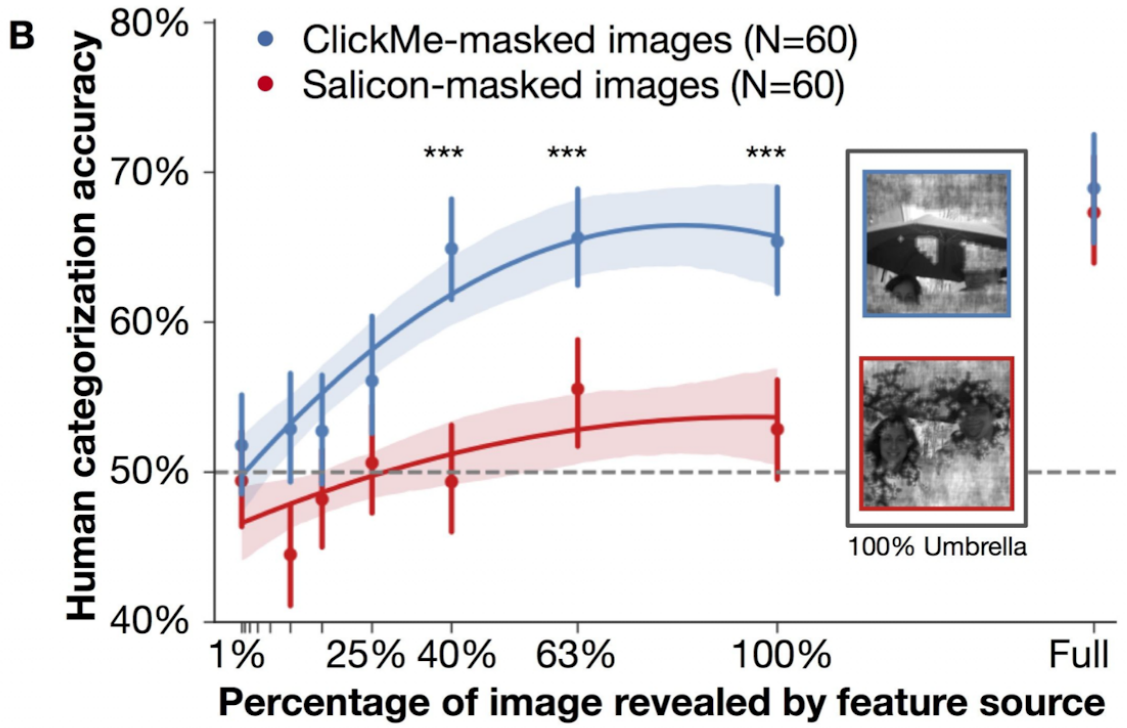
7 seconds to recognize

ClickMe.ai

- 1235 participants
- Each participant played on average 380 images
- Dataset: ILSVRC12 (10,000,000 labelled images)
- 472,946 ClickMe maps on 196,499 unique images
- Maps highly correlated with Clicktionary maps

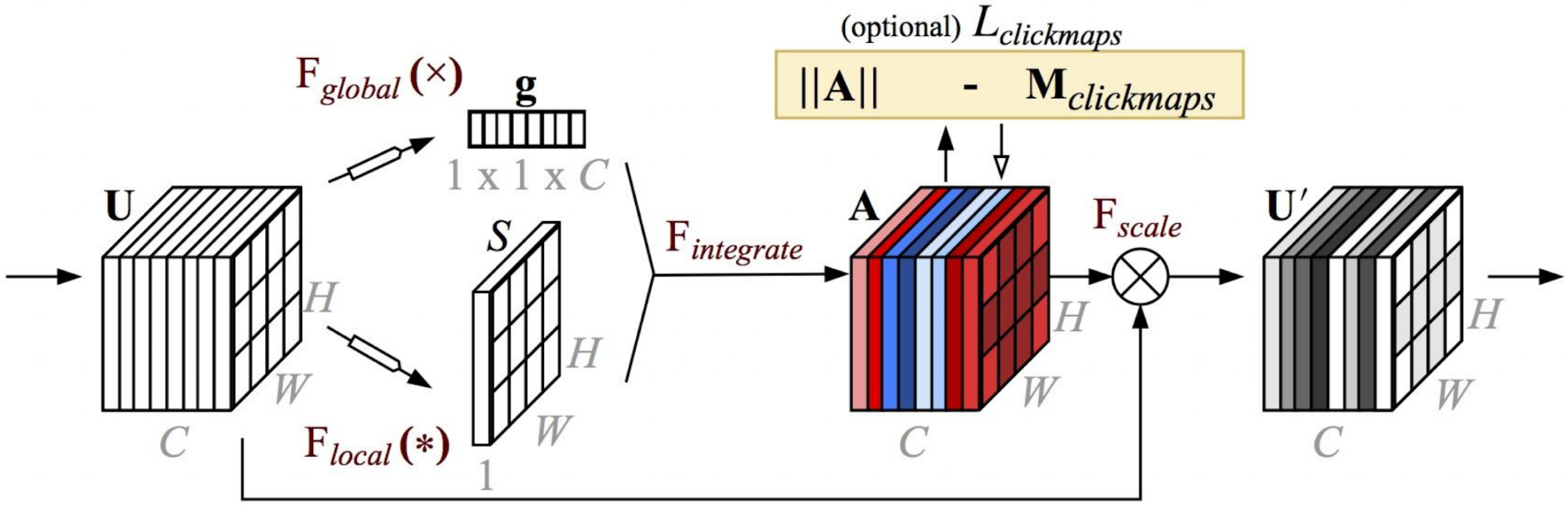
ClickMe.ai


A


B
ClickMe-masked images (N=60)
Salicon-masked images (N=60)
Human categorization accuracy
80%
70%
60%
50%
40%
***
***
***
100% Umbrella
1%
25%
40%
63%
100%
Full
Percentage of image revealed by feature source

1. Problem description and motivation
2. Related work
3. ClickMe.ai
4. **Network architecture**
5. Evaluation
6. Discussion
7. Conclusion

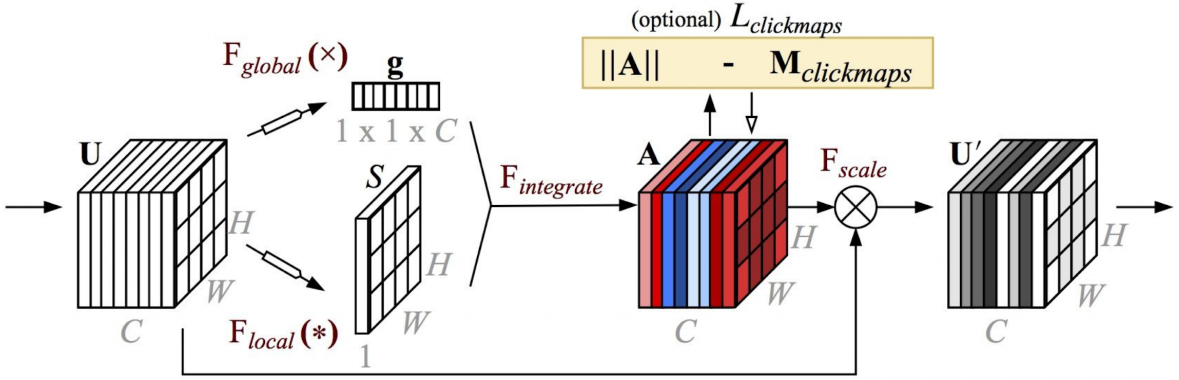
Global-and-local attention (GALA) block



$\mathbf{U}, \mathbf{A} \in \mathbb{R}^{H \times W \times C}$ $\quad \mathbf{g} \in \mathbb{R}^{1 \times 1 \times C}$

$S \in \mathbb{R}^{H \times W \times 1}$

Global attention g



$$\mathbf{p_k} = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} \mathbf{u}_{kxy}$$

$$\mathbf{g} = W_{\text{expand}} \left(\delta \left(W_{\text{shrink}} (\mathbf{p}) \right) \right)$$

$$\mathbf{U} = \left[\mathbf{u}_k \right]_{k=1 \ldots C} \in \mathbb{R}^{H \times W \times C}$$

$$\mathbf{p} = \left(p_k \right)_{k=1 \ldots C} \in \mathbb{R}^{1 \times 1 \times C}$$

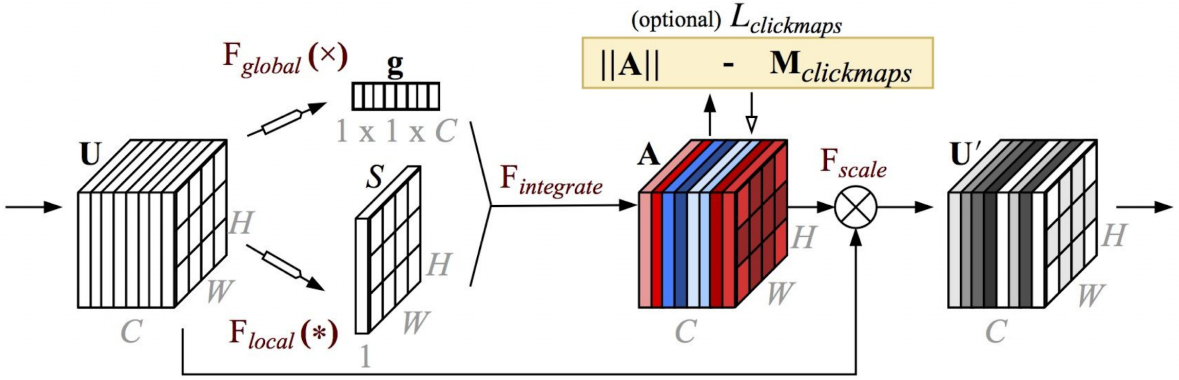
$$\mathbf{g} \in \mathbb{R}^{1 \times 1 \times C}$$

$$W_{\text{shrink}} \in \mathbb{R}^{\frac{c}{r} \times C}$$

$$W_{\text{expand}} \in \mathbb{R}^{C \times \frac{c}{r}}$$

$$r = 4$$

Local attention S



$$S = \mathbf{V}_{\text{collapse}} * \left(\delta \left(\mathbf{V}_{\text{shrink}} * \mathbf{U} \right) \right) \in \mathbb{R}^{H \times W \times 1}$$

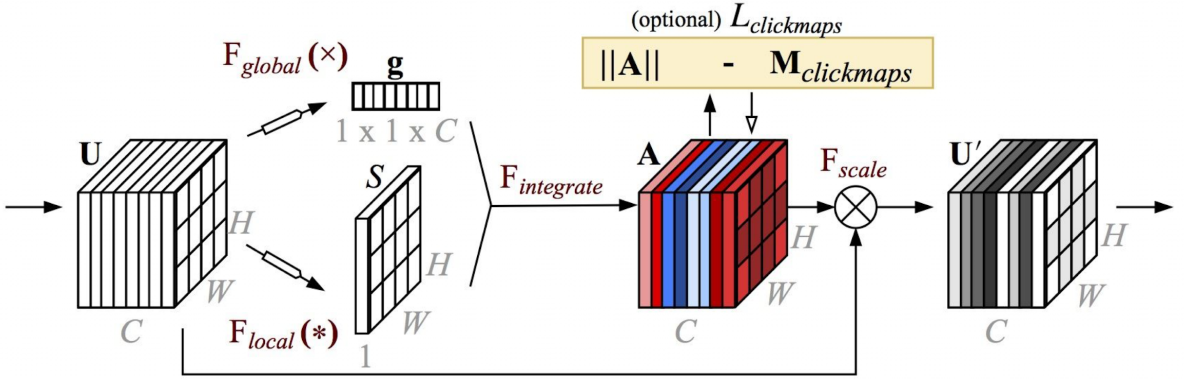
$$\mathbf{U} \in \mathbb{R}^{H \times W \times C}$$

$$S \in \mathbb{R}^{H \times W \times 1}$$

$$\mathbf{V}_{\text{shrink}} \in \mathbb{R}^{1 \times 1 \times C \times \frac{c}{r}}$$

$$\mathbf{V}_{\text{collapse}} \in \mathbb{R}^{1 \times 1 \times \frac{c}{r} \times 1}$$

Total attention A



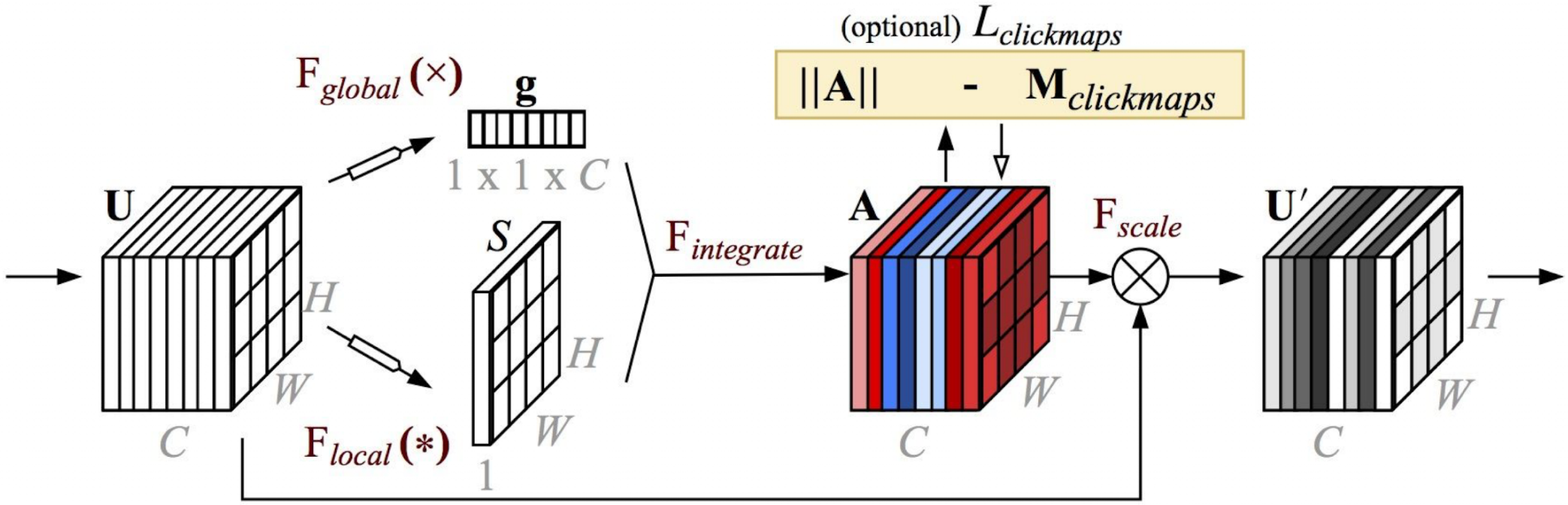
$$\mathbf{A}_{h,w,c} = \zeta\left(a_c\left(\boldsymbol{G}^*_{h,w,c} + \boldsymbol{S}^*_{h,w,c}\right) + m_c\left(\boldsymbol{G}^*_{h,w,c} \cdot \boldsymbol{S}^*_{h,w,c}\right)\right)$$

$$\mathbf{g} \in \mathbb{R}^{1\times1\times C}$$
$$S \in \mathbb{R}^{H\times W\times 1}$$
$$\boldsymbol{G}^*, \boldsymbol{S}^* \in \mathbb{R}^{H\times W\times C}$$

$$\mathbf{U}' = \mathrm{F}_{\mathsf{scale}}(\mathbf{U}, \mathbf{A}) = \mathbf{U} \odot \mathbf{A}$$

$$\left(m_c\right)_{c\in1..C}$$
$$\left(a_c\right)_{c\in1\ldots C}$$

Global-and-local attention (GALA) block



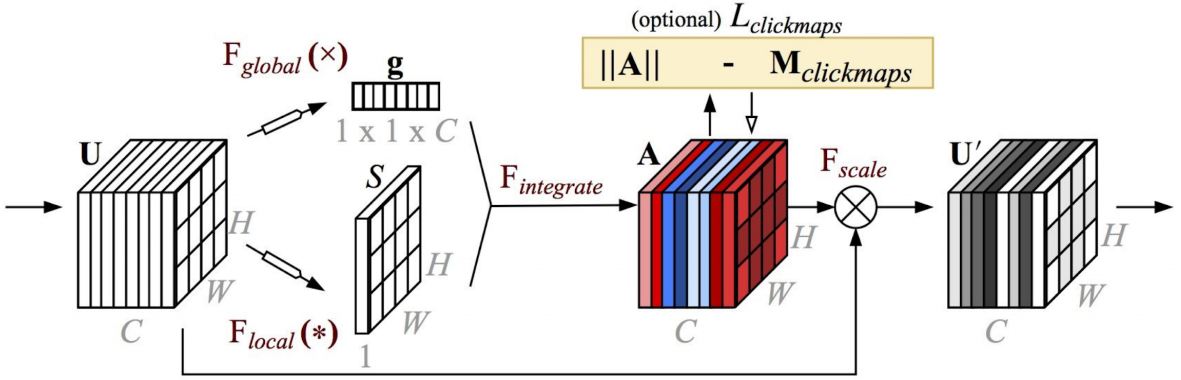
$\mathbf{U}, \mathbf{A} \in \mathbb{R}^{H \times W \times C}$ $\quad \mathbf{g} \in \mathbb{R}^{1 \times 1 \times C}$

$S \in \mathbb{R}^{H \times W \times 1}$

Network architecture

- ResNet-50
- Apply GALA module to final activity in a dense path of a residual module at 6 mid- to high-level feature layers (layers 24, 27, 30, 33, 36, 39)
  - 14x14

ClickMe Supervision



$$\mathscr{L}_T(\mathbf{X}, y) = \mathscr{L}_C(M(\mathbf{X}), y) + \mathscr{L}_{clickmaps}(\mathbf{X})$$

$$\mathscr{L}_{clickmaps}(\mathbf{X}) = \lambda \sum_{l \in \mathbf{L}} \left\| \frac{R^l(\mathbf{X})}{\| R^l(\mathbf{X}) \|_2} - \frac{\mathbf{A}^l(\mathbf{X})}{\| \mathbf{A}^l(\mathbf{X}) \|_2} \right\|_2$$

1. Problem description and motivation
2. Related work
3. ClickMe.ai
4. Network architecture
5. **Evaluation**
6. Discussion
7. Conclusion

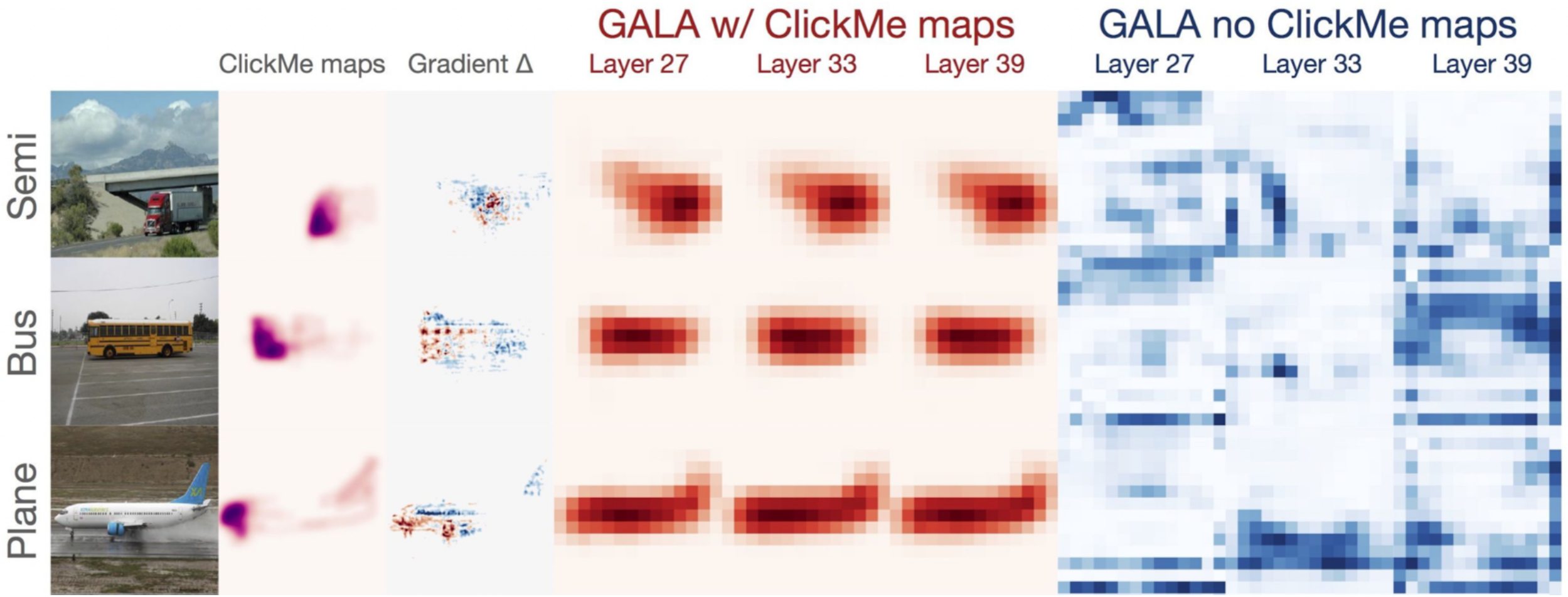
Evaluation

- Object classification accuracy
- Similarity between ClickMe maps and model attention maps
- Analysis over different values of λ -> $\lambda = 6$
- Model interpretability

Results

	top-1 err	top-5 err	maps
SE-ResNet-50	66.17	42.48	64.36**
ResNet-50	63.68	40.65	43.61
GALA-ResNet-50 no ClickMe	53.90	31.04	64.21**
GALA-ResNet-50 w/ ClickMe	**49.29**	**27.73**	**88.56****

	Reference		Ours	
	top-1 err.	top-5 err.	top-1 err.	top-5 err.
ResNet-50 (He et al., 2016)	24.70	7.80	23.88	6.86
SE-ResNet-50 (Hu et al., 2017)	23.29	6.62	23.26	6.55
GALA-ResNet-50 no ClickMe	-	-	**22.73**	**6.35**


GALA w/ ClickMe maps
GALA no ClickMe maps
ClickMe maps
Gradient Δ
Layer 27
Layer 33
Layer 39
Layer 27
Layer 33
Layer 39
Semi
Bus
Plane

1. Problem description and motivation
2. Related work
3. ClickMe.ai
4. Network architecture
5. Evaluation
6. **Discussion**
7. Conclusion

Discussion and Limitations

- ClickMe supervised GALA training
  - Improved performance when **data is limited**
  - Selects features similar to features that humans deem important -> possible application in areas where **interpretability** is important, e.g. in medicine
- For large datasets no improvement

1. Problem description and motivation
2. Related work
3. ClickMe.ai
4. Network architecture
5. Evaluation
6. Discussion
7. **Conclusion**

Conclusion

- **ClickMe dataset**
- **GALA module:** attention inspired by human attention
- **GALA + ClickMe:** Leverage human supervision to co-train an attention network

- Authors exaggerate how well their method performs
- A lot of results not presented in main part, but only in appendix
- Global attention in humans interesting insight

Source: http://www.home-designing.com/workstation-setup-ideas-photos

Source: https://whc.unesco.org/en/list/1505/

Thank you for your attention!

Source: http://www.home-designing.com/workstation-setup-ideas-photos

Source: https://whc.unesco.org/en/list/1505/