Repository

Our related OptiCAM

CA-Stream: Attention-based pooling for interpretable image recognition

Felipe Torres[1], Hanwei Zhang[2],
Ronan Sicre[1], Stéphane Ayache[1],
Yannis Avrithis[3]

[1]Centrale Marseille, Aix Marseille Univ, CNRS, LIS, France
[2]Institute of Intelligent Software, China
[3]Institute of Advanced Research on Artificial Intelligence (IARAI), Austria

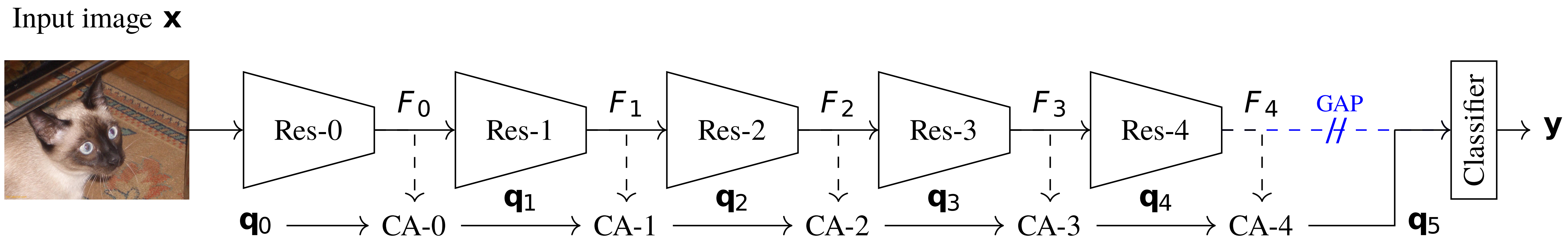

Figure 1: *Cross-Attention Stream* applied to ResNet-based architectures. During inference, we replace global average pooling (GAP) with our attention-based pooling mechanism, learned with a stream running in parallel to f

Abstract

Raw attention in transformer architectures acts as a class-agnostic saliency map and feature space mask. We designed the Cross-Attention Stream (CA-Stream) to replace Global Average Pooling (GAP) during inference. Our approach, incorporates cross-attention blocks at various network depths, improving interpretability metrics while maintaining recognition performance.

Class Activation Maps Self-Attention and Cross Attention

Class Activation Mapping (CAM)[5] shares similarities to self-attention[1], weighting feature maps. In a sense, CAM can be expressed as cross attention.

CAM computes the class-specific saliency map S^c, via linear combination of feature maps A^k_ℓ and a weighting coefficient w^c_k at layer ℓ:

$$S^c_\ell := h\left(\sum_k \alpha^c_k F^k_\ell\right), \qquad (1)$$

Self-Attention first computes dot-product similarities of the projections (Q) and (K) of embedding X_ℓ. Expressed in the attention matrix (A)

$$A = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_\ell}}\right). \qquad (2)$$

Self-attention is the average of all values (V) weighted by attention:

$$\mathrm{SA}(X_\ell) := AV \in \mathbb{R}^{t_\ell \times d_\ell}. \qquad (3)$$

Cross-Attention Considering the attention matrix in (2), the feature map matrix F, and a CLS token $q_\ell \in \mathbb{R}^d_\ell$; attention can be rewritten as:

$$\mathbf{a} = A^\top = \mathrm{softmax}\left(\frac{F_\ell \mathbf{q}_\ell}{\sqrt{d_\ell}}\right). \qquad (4)$$

Replacing q_ℓ with an arbitrary vector $\alpha \in \mathbb{R}^d_\ell$:

$$\mathbf{a} = h_\ell(F_\ell \boldsymbol{\alpha}) = h_\ell\left(\sum_k \alpha_k \mathbf{f}^k_\ell\right). \qquad (5)$$

Cross Attention Stream

Designed to run in parallel to convolutional neural networks, our *Cross-Attention Stream* takes input features at key depths within the network, utilizing cross-attention to build a global image representation, replacing GAP before the classifier, see Figure 1.

Set-Up We train the CA-Stream using a pretrained network f which remains frozen. The stream parameters are learned computing the cross entropy of the logits obtained by forwarding the CLS token through the frozen classifier.
To evaluate our approach, we compare the results obtained using the baseline architecture with GAP, and the outputs generated with the class-token to classify.

Results

We display class-specific saliency maps obtained through CAM for images contained in the validation set of ImageNet, as well as raw attention maps for images with categories that are not seen during training, from the MIT 67 Scenes dataset.

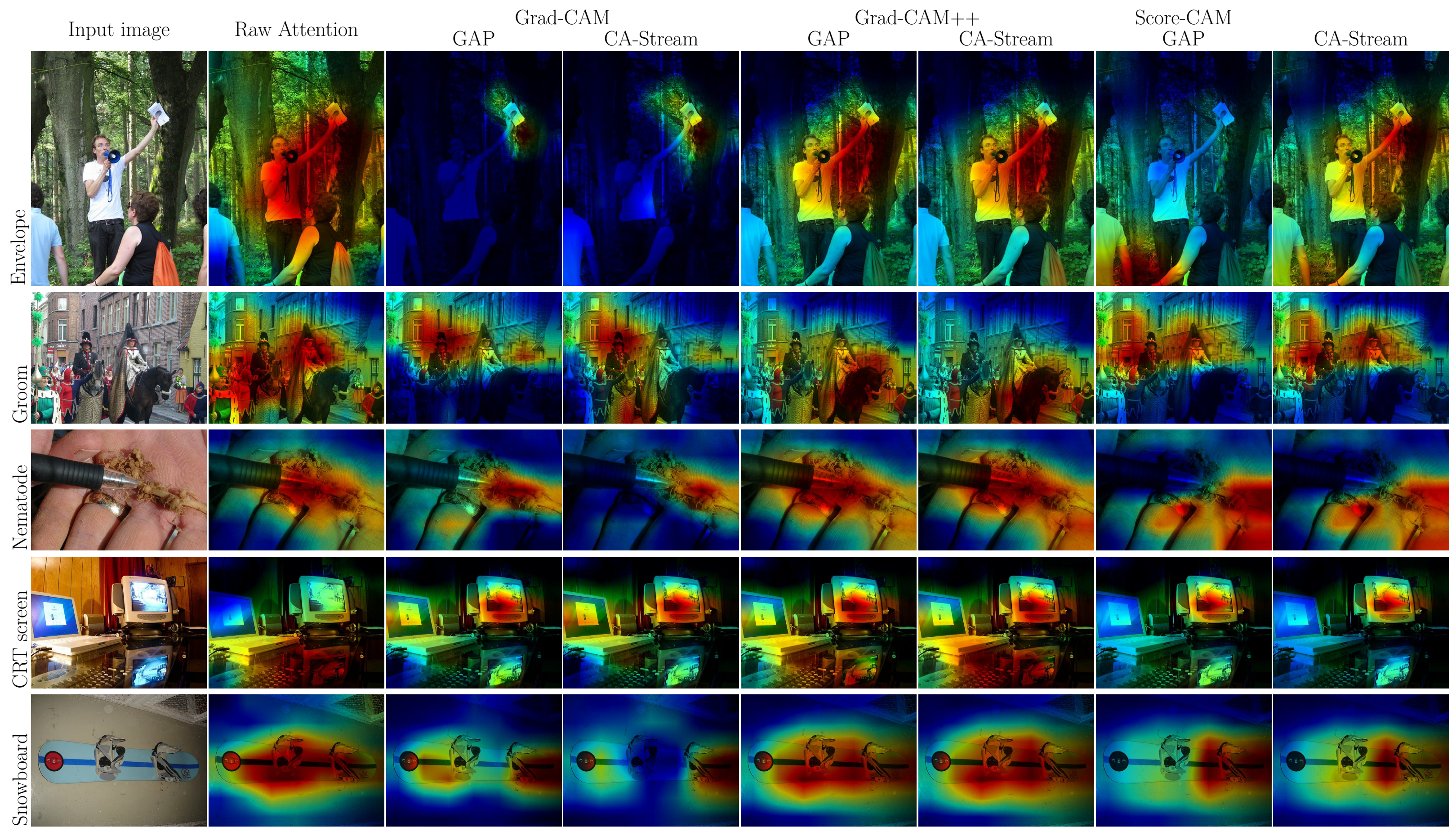

Figure 2: **Comparison of saliency maps** generated by different CAM-based methods, using GAP and our CA-Stream, on ImageNet images. The raw attention is the one used for pooling by CA-Stream.

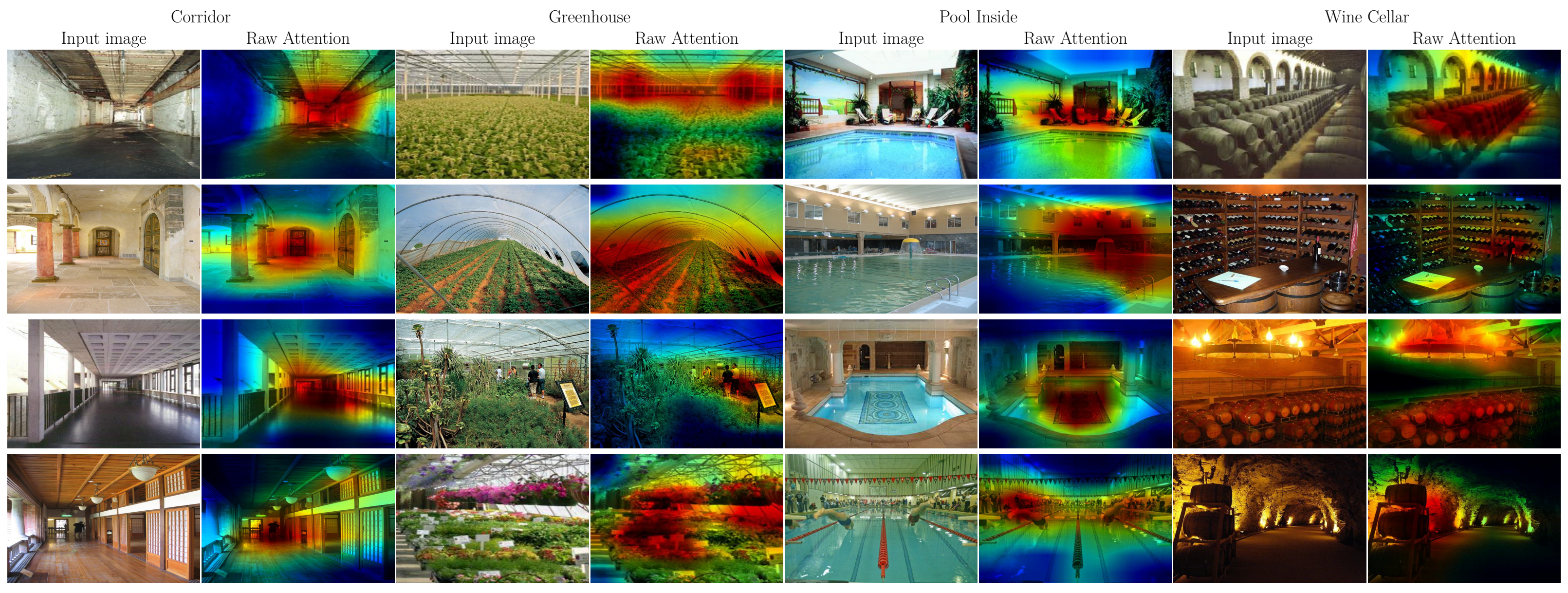

Figure 3: **Raw attention maps** obtained from our CA-Stream on images of the MIT 67 Scenes dataset [3] on classes not seen in ImageNet. The network sees these categories at inference for the first time.

To compare the effect of our CA *vs.* GAP, we measure interpretable image recognition metrics such as AD, AI, AG [4] and causal perturbations with Insertion and Deletion [2]. CA-Stream provides consistent improvements over GAP over most metric, while performing lower on Deletion.

Network	Pooling	Acc↑
ResNet-50	GAP	74.55
	CA	74.70
ConvNeXt-B	GAP	83.72
	CA	83.51

Network	Attribution	Pooling	AD↓	AG↑	AI↑	I↑	D↓
ResNet-50	Grad-CAM	GAP	13.04	17.56	44.47	72.57	**13.24**
		CA	**12.54**	**22.67**	**48.56**	**75.53**	13.50
	Grad-CAM++	GAP	**13.79**	15.87	42.08	72.32	**13.33**
		CA	13.99	**19.29**	**44.60**	**75.21**	13.78
	Score-CAM	GAP	8.83	17.97	48.46	71.99	**14.31**
		CA	**7.09**	**23.65**	**54.20**	**74.91**	14.68
ConvNeXt-B	Grad-CAM	GAP	33.72	2.43	15.25	52.85	**29.57**
		CA	**19.45**	**13.96**	**32.89**	**86.38**	45.29
	Grad-CAM++	GAP	**34.01**	2.37	15.60	52.83	**29.17**
		CA	36.69	**8.00**	**21.95**	**85.39**	53.42
	Score-CAM	GAP	43.55	2.23	15.67	50.96	**39.49**
		CA	**23.51**	**11.04**	**27.35**	**83.41**	60.53

Table 1: **Interpretability metrics** of CA-Stream *vs.* baseline GAP for different networks and interpretability methods on ImageNet.

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[2] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *BMVC*, 2018.

[3] Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE conference on computer vision and pattern recognition*, 2009.

[4] Hanwei Zhang, Felipe Torres, Ronan Sicre, Yannis Avrithis, and Stephane Ayache. Opti-cam: Optimizing saliency maps for interpretability. *arXiv preprint arXiv:2301.07002*, 2023.

[5] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.