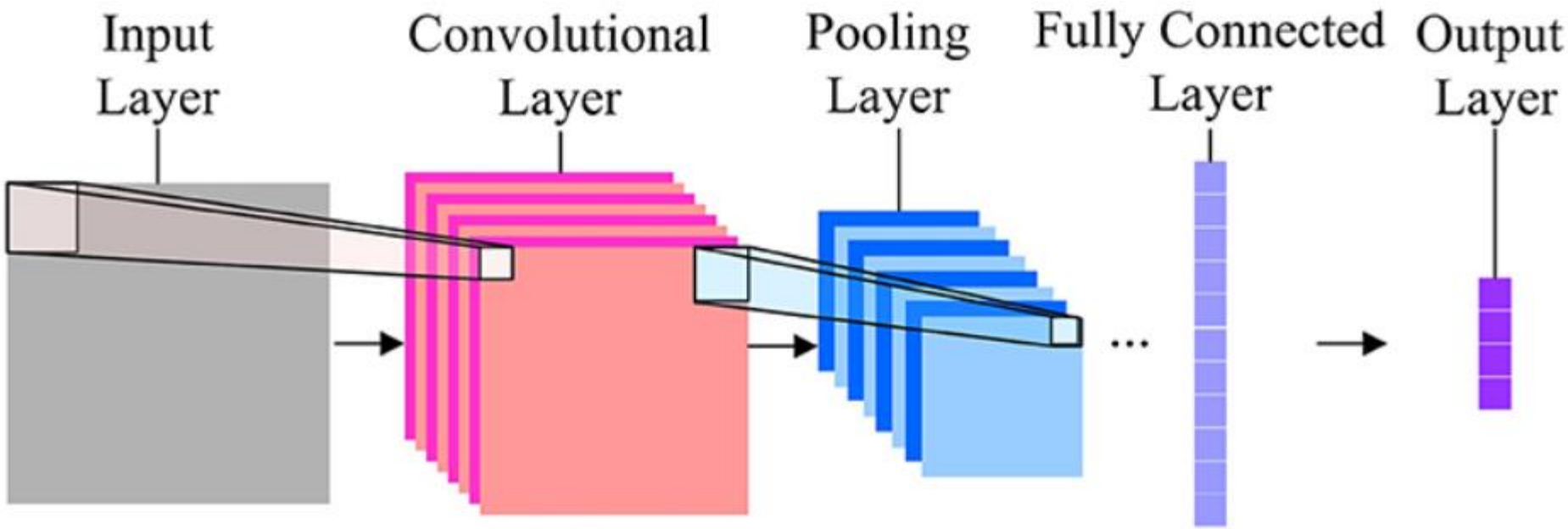
Multi-view Feature Augmentation with Adaptive Class Activation Mapping

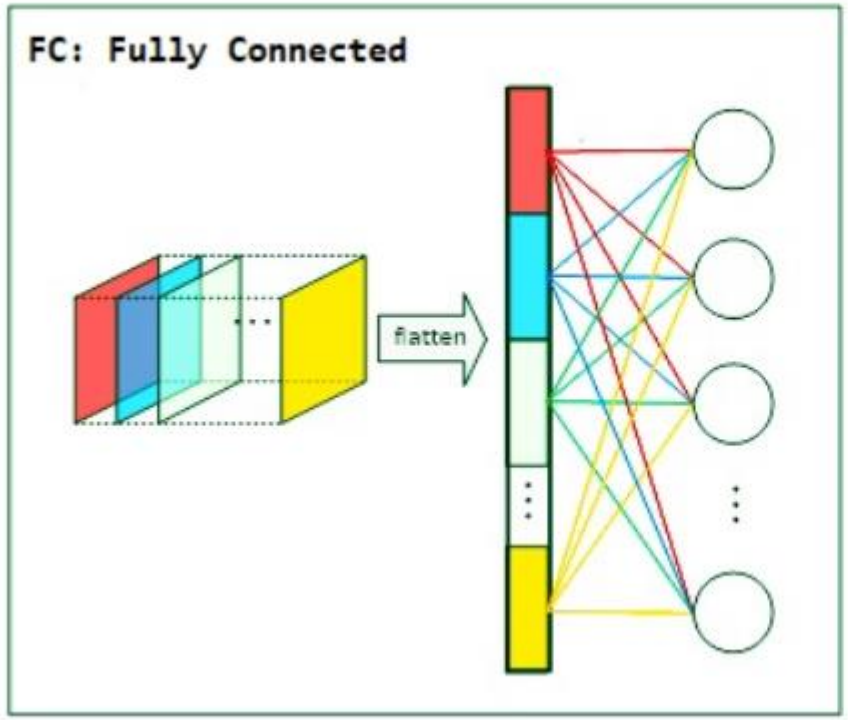
Xiang Gao Yingjie Tian Zhiquan Qi

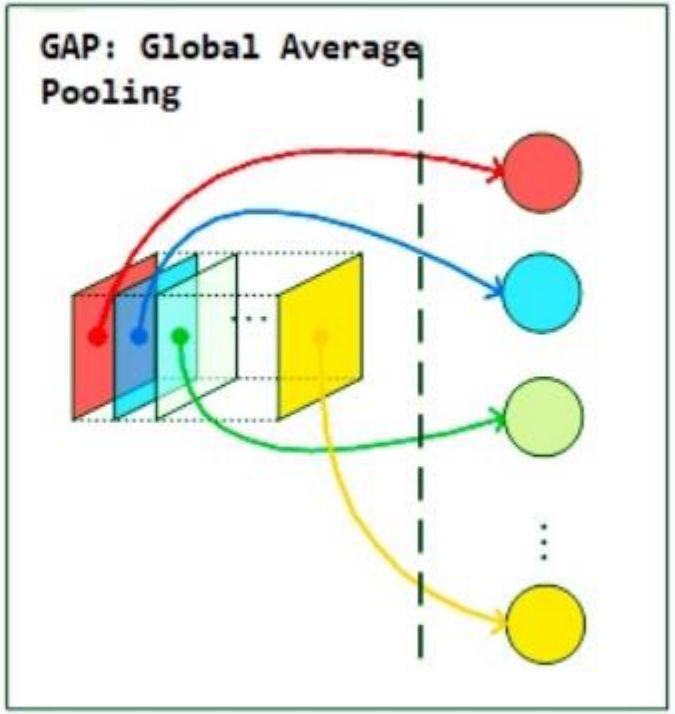
IJCAI 2021

CNN


Input Layer
Convolutional Layer
Pooling Layer
Fully Connected Layer
Output Layer

CNN → GAP(Global Average Pooling)





Motivation

Framework

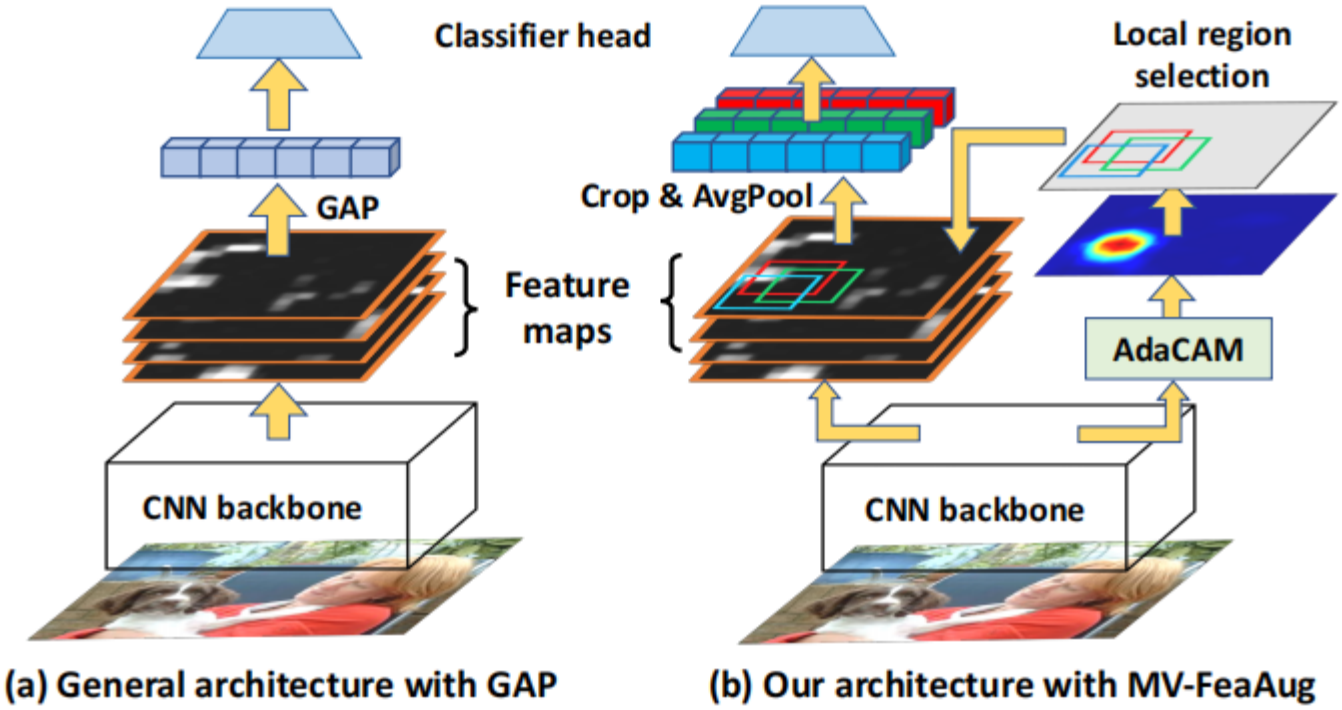


Figure 2: Comparison between the image classification architecture with the general GAP (left) and our MV-FeaAug module (right). We sample diverse local features around the class-discriminative region of the final convolutional feature maps as multi-view local image representations for ensembled classification, as compared with GAP that extracts only global-view image representation.

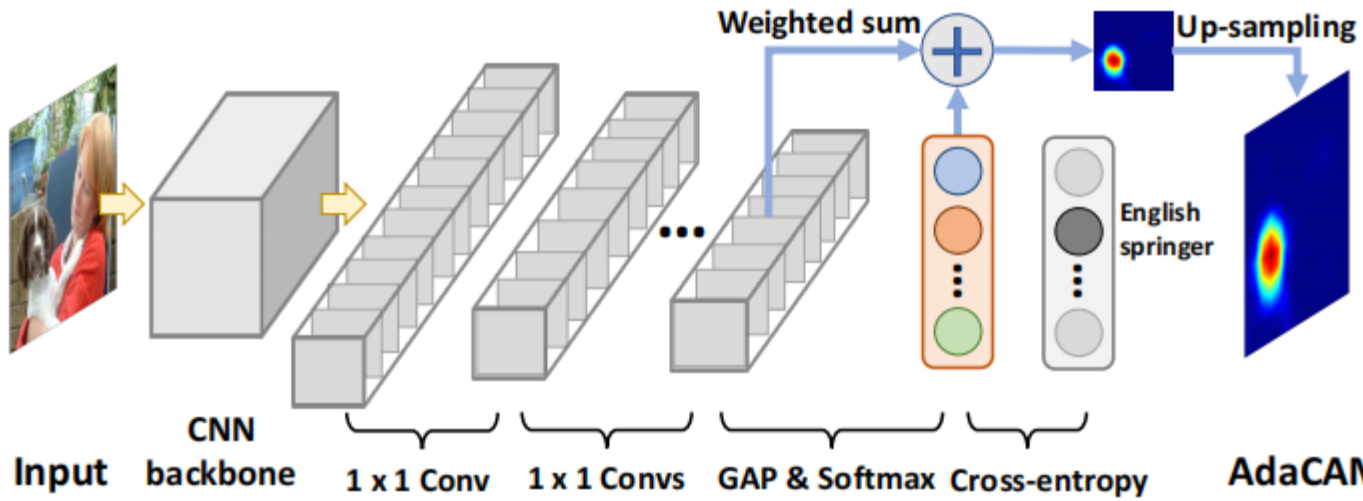
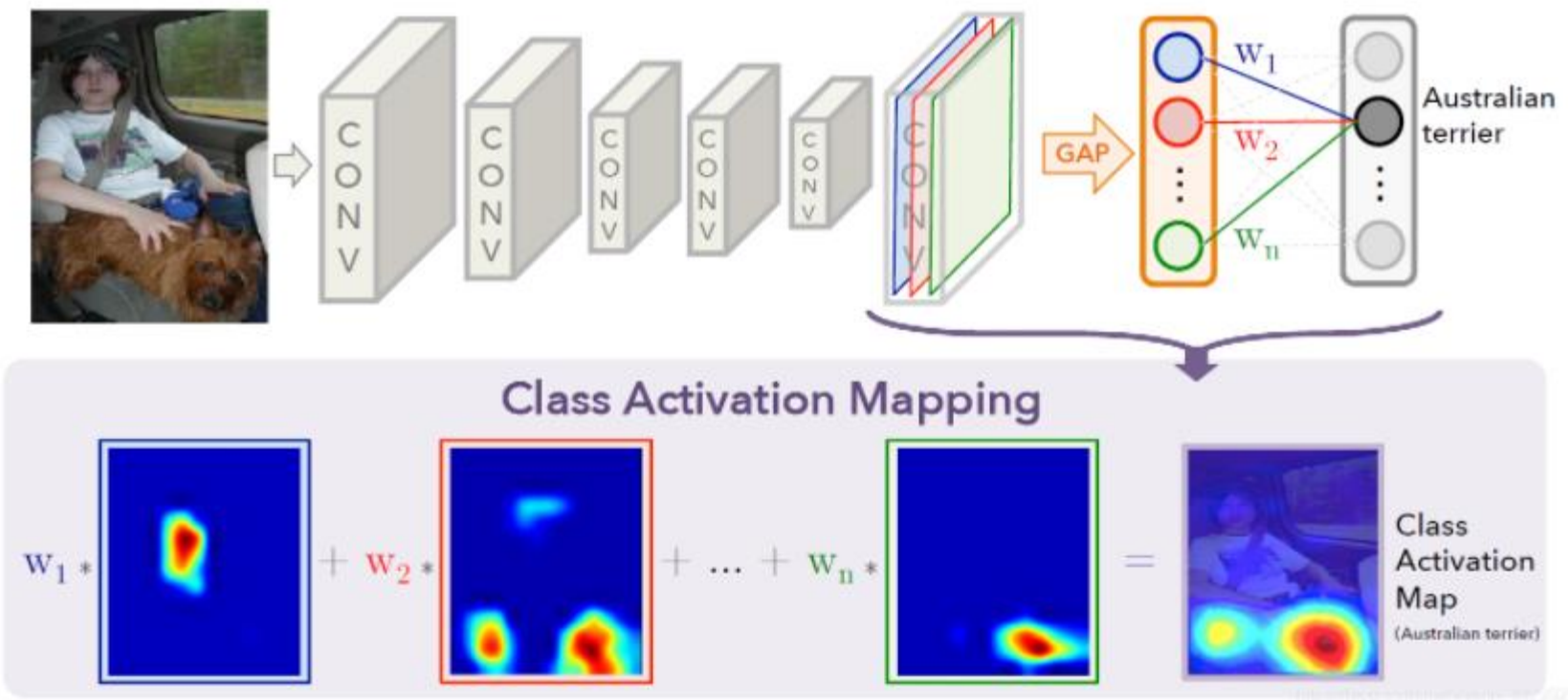


Figure 3: Adaptive class activation mapping (AdaCAM). We replace the traditional classifier head made up of [GAP→MLP→Softmax] with [MLPConv→GAP→Softmax] (MLPConv comprises consecutive $\text{Conv}_{1\times1}$ layers joined by non-linear activations) to maintain spatial resolution of feature maps. The AdaCAM is obtained by performing channel-wise weighted sum of the last convolutional feature maps with respect to the softmax logit vector.

CAM


CONV
CONV
CONV
CONV
CONV
CONV
GAP
W_1
W_2
W_n
Australian terrier
Class Activation Mapping
W_1 *
+ W_2 *
+ ... + W_n *
=
Class Activation Map
(Australian terrier)

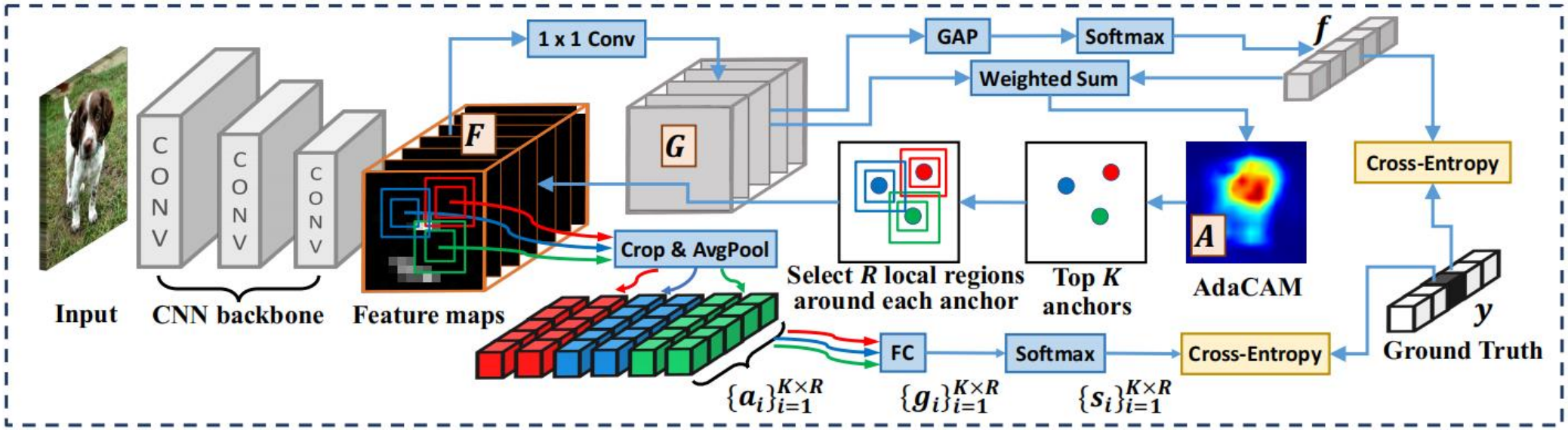


Figure 4: Overview of MV-FeaAug. We concurrently train an auxiliary classifier head comprised of only one 1×1 convolutional layer for dynamic generation of AdaCAM, based on which we sample multiple local representations on the final convolutional feature maps as multi-view inputs to the main classifier head. The main classifier head comprises (but not restricted to) a single fully-connected layer.

$$
\begin{aligned}
L_{total} = & -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} y_c^{(n)} log(f_c^{(n)}) \\
& -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{K\times R}\sum_{c=1}^{C} y_c^{(n)} log(s_{i,c}^{(n)}),
\end{aligned}
\tag{9}
$$

Experiments

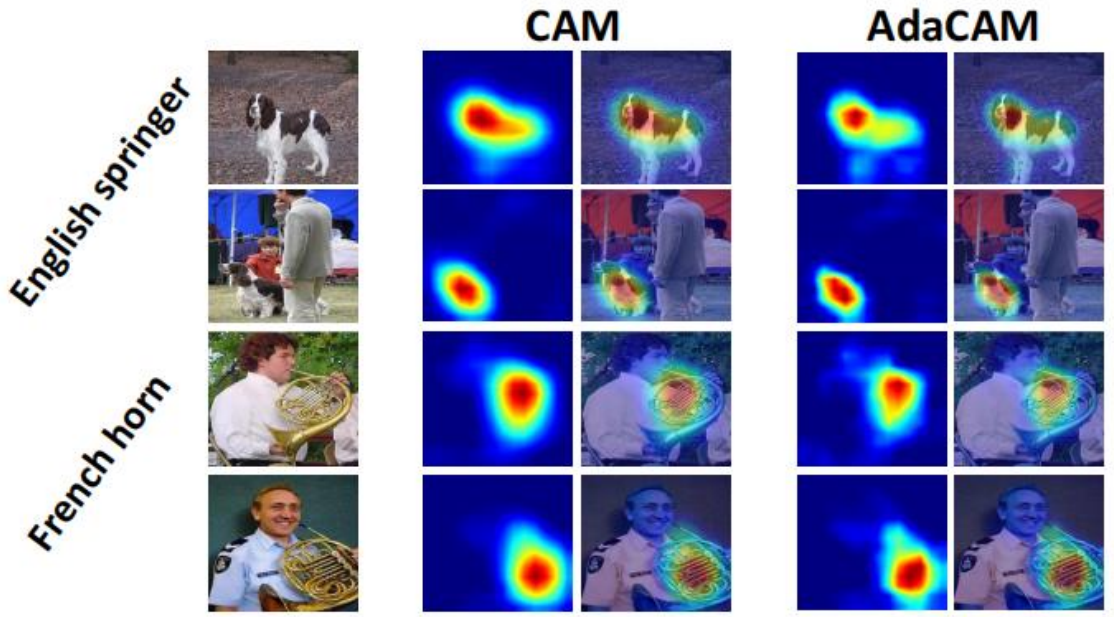
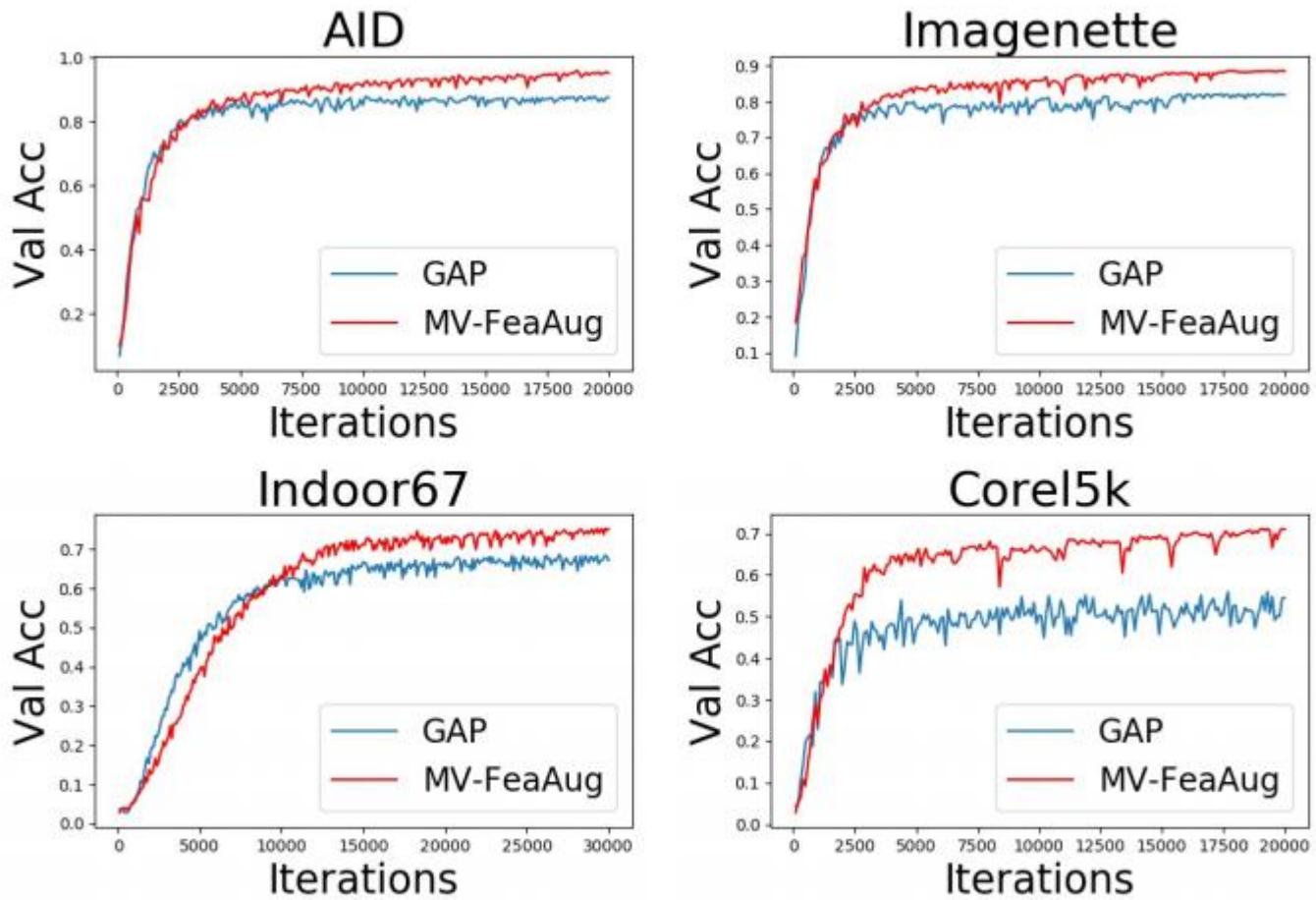


Figure 5: Visual comparison between CAM and our AdaCAM (evaluated on Imagenette validation set) in object localization. Refer to supplementary materials for more results.


AID
Val Acc
Iterations
GAP
MV-FeaAug
Imagenette
Val Acc
Iterations
GAP
MV-FeaAug
Indoor67
Val Acc
Iterations
GAP
MV-FeaAug
Corel5k
Val Acc
Iterations
GAP
MV-FeaAug

Model	Datasets									
	Imagenette	Caltech101	Corel5k	Scene15	Indoor67	Action40	Event8	UCMLU	RSSCN7	AID
VGG16_GAP	82.04%	70.42%	54.31%	80.86%	68.25%	52.19%	80.00%	90.95%	91.25%	87.61%
VGG16_MFA	**88.45%**	**76.05%**	**70.74%**	**85.53%**	**75.27%**	**60.36%**	**85.75%**	**96.67%**	**95.29%**	**94.55%**
ResNet50_GAP	83.66%	71.84%	54.55%	81.41%	70.14%	53.43%	81.08%	91.67%	91.43%	89.48%
ResNet41_GAP	82.47%	70.75%	52.72%	80.17%	68.90%	52.56%	79.88%	90.58%	90.54%	88.31%
ResNet41_MFA	**88.60%**	**76.19%**	**69.84%**	**85.00%**	**75.59%**	**60.61%**	**85.25%**	**96.43%**	**94.71%**	**94.89%**
ResNeXt50_GAP	84.10%	72.65%	56.20%	83.18%	71.26%	54.71%	82.95%	92.65%	92.32%	90.24%
ResNeXt41_GAP	82.79%	71.54%	55.66%	81.84%	69.95%	53.50%	81.89%	91.79%	91.18%	88.96%
ResNeXt41_MFA	**88.86%**	**76.78%**	**71.48%**	**86.29%**	**76.38%**	**61.77%**	**87.14%**	**97.31%**	**95.14%**	**95.29%**
MobileNet_GAP	81.26%	70.26%	53.86%	80.24%	67.79%	52.48%	80.50%	90.24%	90.54%	86.87%
MobileNet24_GAP	80.60%	69.78%	51.77%	79.34%	66.96%	51.73%	79.25%	89.37%	89.61%	86.04%
MobileNet24_MFA	**87.55%**	**75.74%**	**69.31%**	**84.97%**	**74.53%**	**59.95%**	**84.75%**	**95.48%**	**93.82%**	**93.52%**

Table 1: Comparison between GAP and our MV-FeaAug (MFA) module with $K = 50$ in validation accuracy based on different CNN backbones. Our MV-FeaAug impressively boosts model performance with simply one more $Conv_{1\times1}$ layer than GAP-based counterpart.

Model	Datasets			Backbone Parameters
	Caltech101	AID	Indoor67	
VGG16_GAP	70.42%	87.61%	68.25%	14.03M
VGG16_SE_GAP	70.96%	88.44%	68.75%	14.18M
VGG16_CBAM_GAP	71.53%	89.37%	69.24%	14.61M
VGG16_SK_GAP	71.25%	89.63%	69.08%	19.55M
VGG16_GC_GAP	71.87%	90.04%	69.51%	14.76M
VGG16_mixup_GAP	71.55%	89.50%	69.67%	14.03M
VGG16_cutout_GAP	70.90%	88.25%	69.14%	14.03M
VGG16_cutmix_GAP	72.10%	90.16%	70.03%	14.03M
VGG16_MFA (ours)	**76.05%**	**94.55%**	**75.27%**	14.03M

Table 2: Comparison of our MV-FeaAug (MFA) with related visual attention modules and data augmentation methods in validation accuracy. We use VGG16 backbone for all the related methods, and set $K = 50$, $L_R = [3, 5, 7, 9]$ in our MV-FeaAug module.

Datasets	GAP	MV-FeaAug				
		K=20	K=30	K=40	K=50	K=60
Imagenette	82.04%	86.27%	87.62%	88.13%	88.45%	88.72%
Indoor67	68.25%	73.51%	74.34%	74.89%	75.27%	75.55%
Action40	53.21%	58.24%	59.36%	59.98%	60.36%	60.68%
UCMLU	90.95%	94.52%	95.33%	96.02%	96.67%	97.05%

Table 3: Study of the impact of K to validation accuracy on different datasets. The backbone network is VGG16.

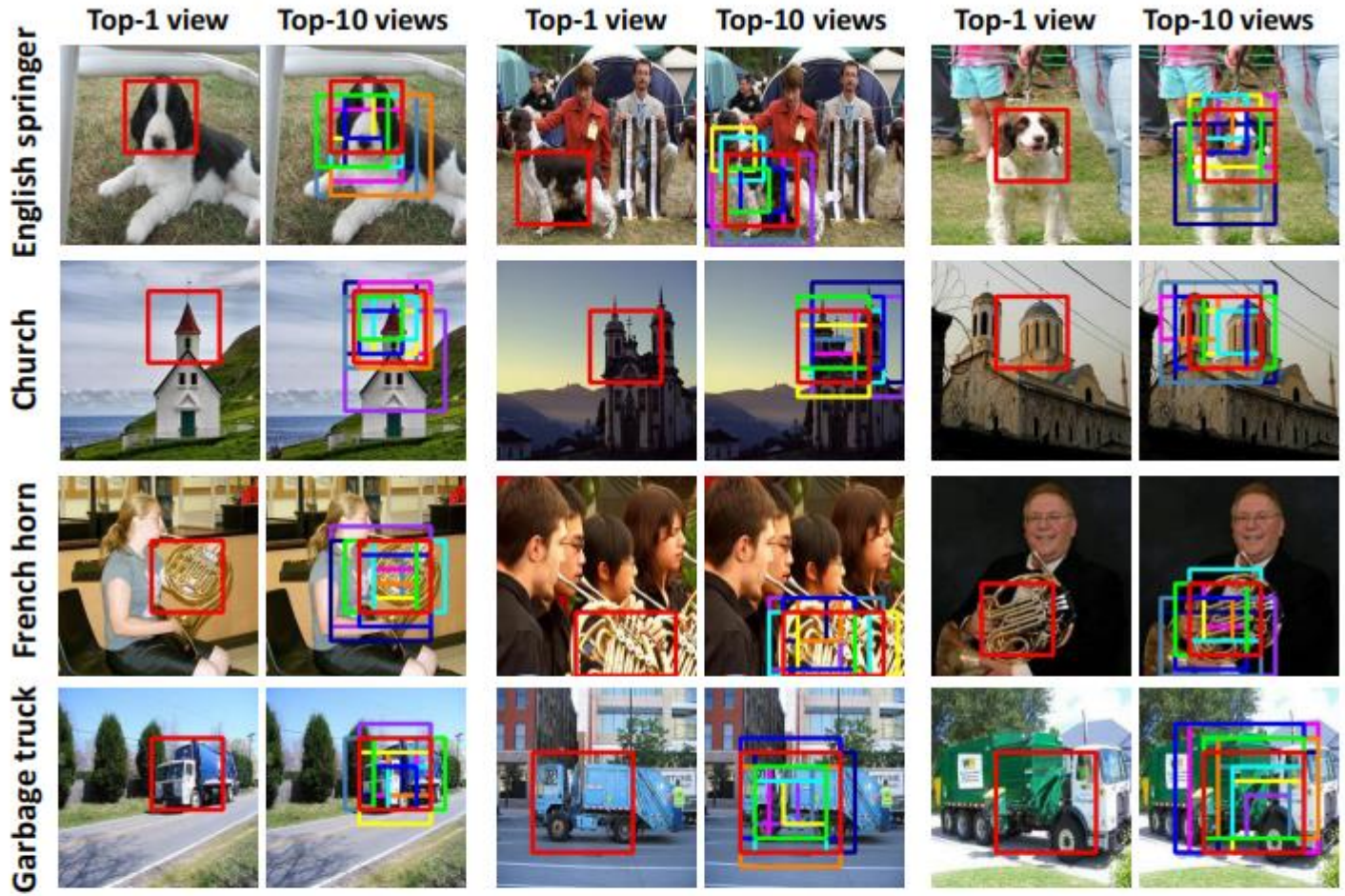


Figure 7: Visualization of the top-1 and top-10 local regions that correspond to local features with the highest prediction confidence of the corresponding image class. Please zoom in for better resolution and refer to supplementary materials for more results.