REAPS: Towards Better Recognition of Fine-grained Images by Region Attending and Part Sequencing

Peng Zhang; Zhanzhan Cheng; Yi Niu
Hikvision Research Institute

Xinyu Zhu; Shuigeng Zhou
Fudan University

Contact Mail: zhangpeng23@hikvision.com

MOTIVATION

- The-state-of-the-art is the part/region-based approaches.
- The discriminative feature representation of an object is prone to be disturbed by complicated background.
- It is unreasonable to fix the number of salient parts.
- The spatial correlation among different salient parts has not been thoroughly exploited.

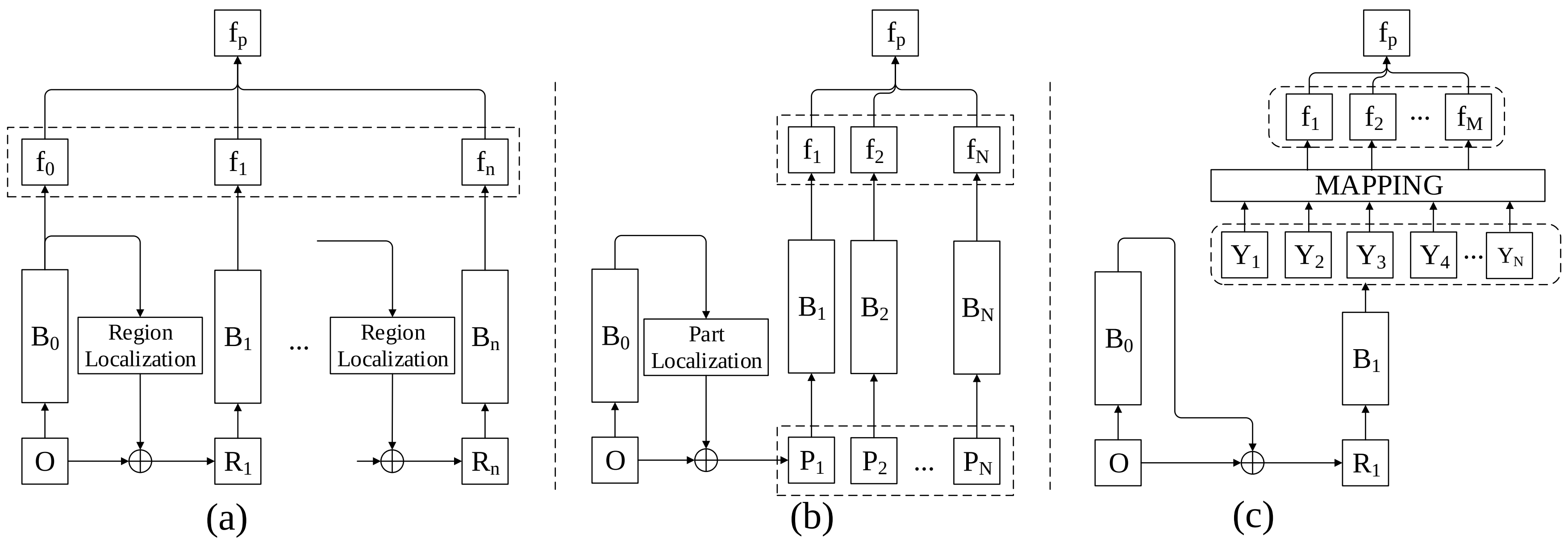


Fig.1. An illustrative comparison between our framework and two popular existing fine-grained recognition frameworks.

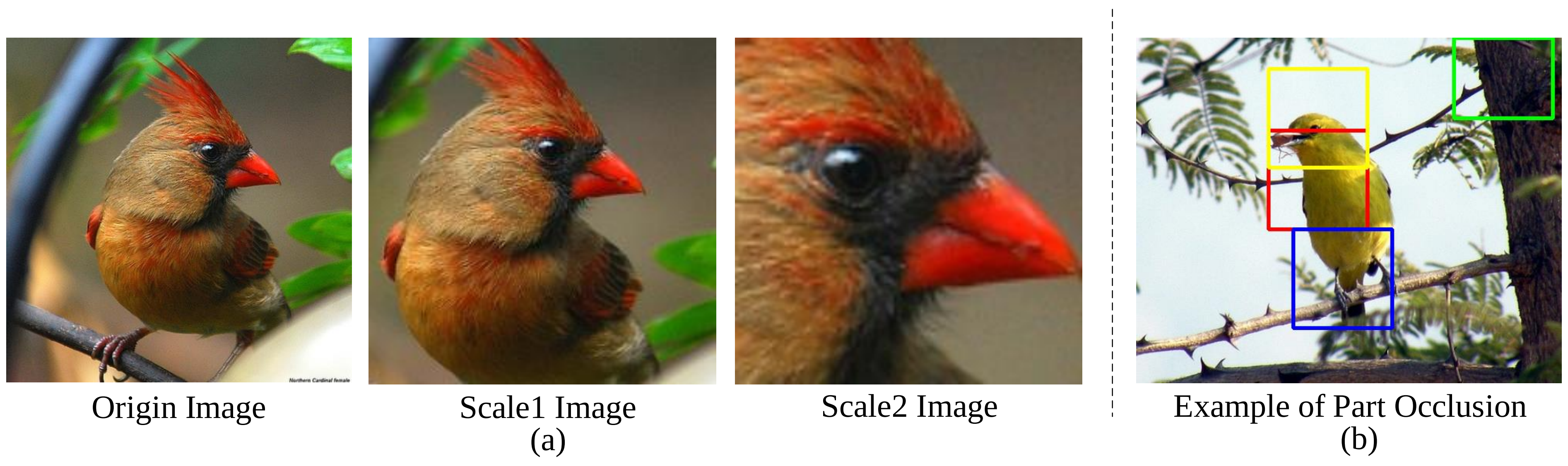
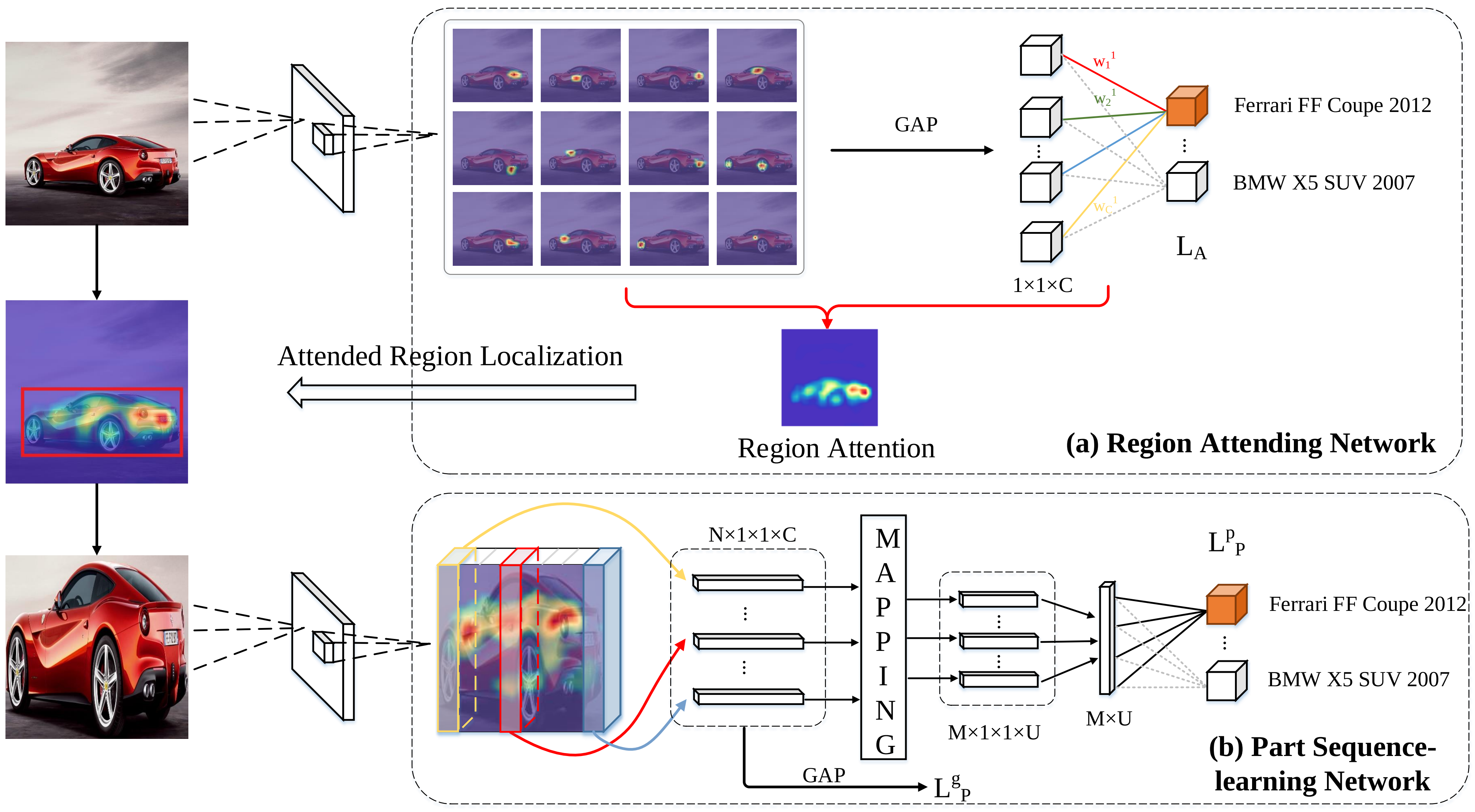

Origin Image
Scale1 Image
Scale2 Image
(a)
Example of Part Occlusion
(b)

Fig.2. Drawbacks of two popular existing frameworks.

METHOD

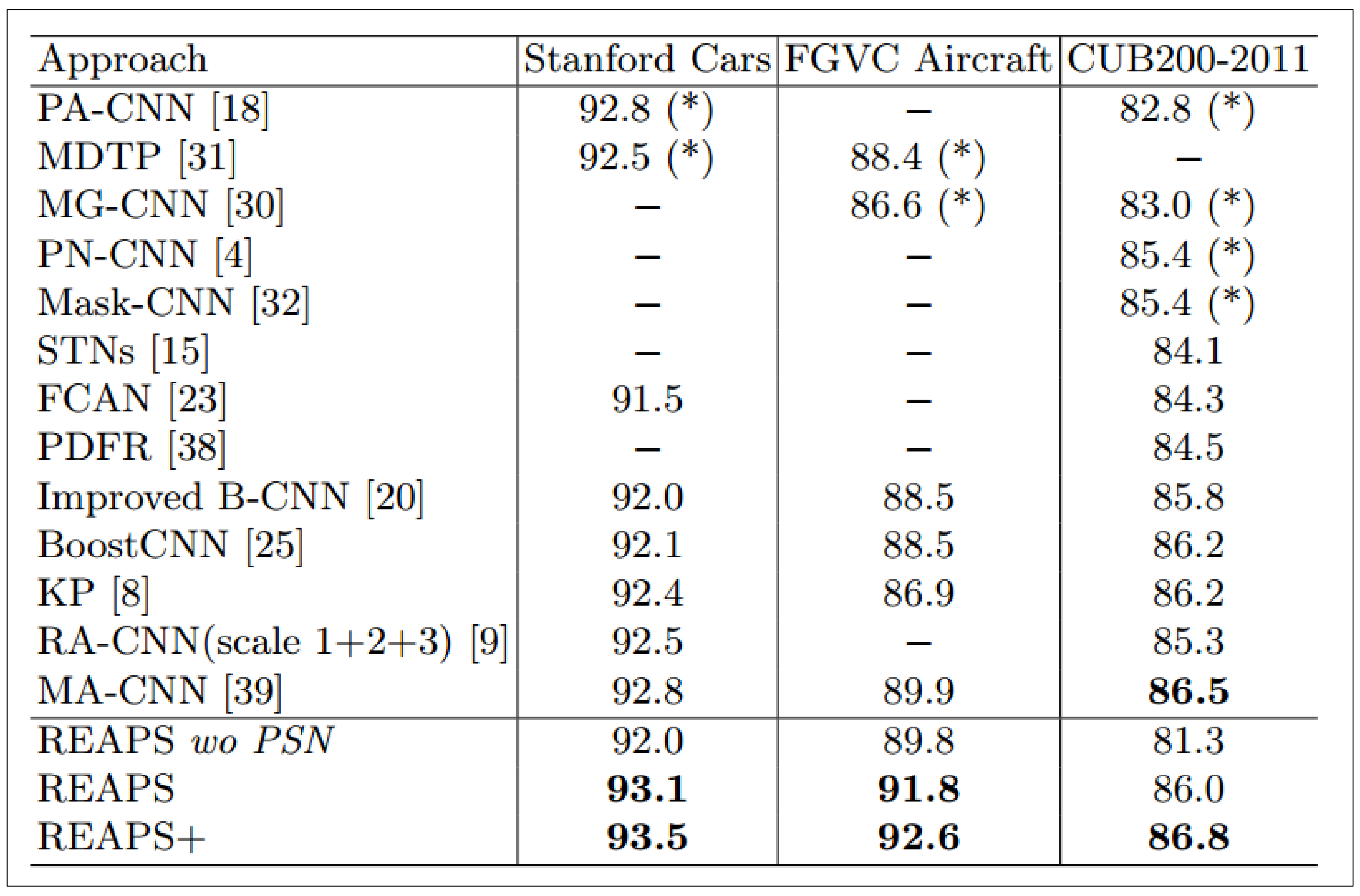


- **RAN**(Region Attending Network): Class Activation Mapping (**CAM**) mechanism is applied to generating the region attention. The attended region is cropped and amplified to depress the background noise.

- **PSN**(Part sequence-learning Network): Intend to model the part representation in a soft-way, e.g. divide the object to N parts, learn discriminative part representation and capturing the spatial correlation among different salient parts simultaneously by using LSTM.

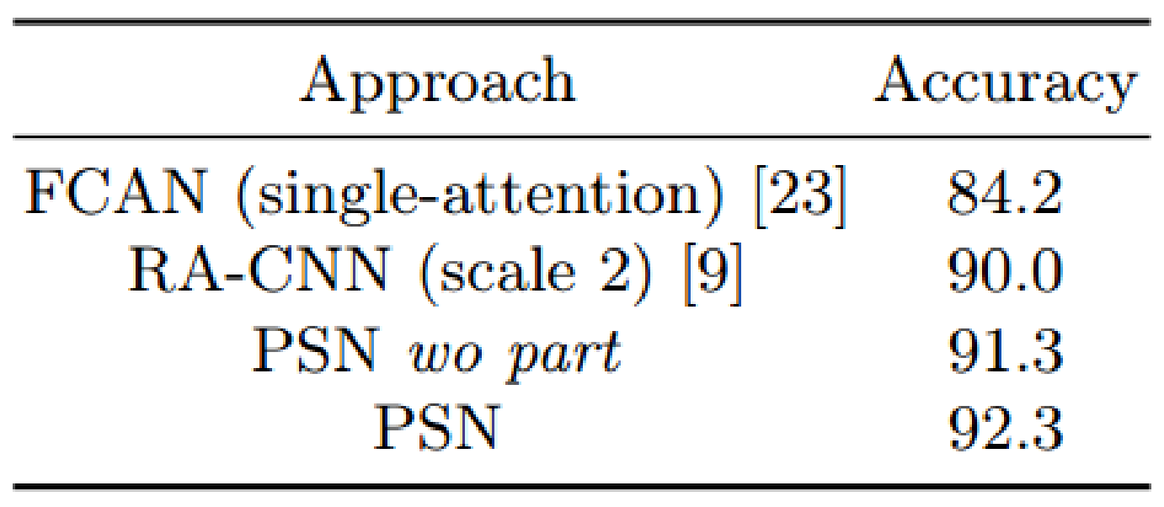
EXPERIMENT

Table 1. Performance comparison on the Stanford Cars, FGVC Aircraft and CUB-200-2011 datasets. (*) indicates whether bounding box or part annotation is used in training.



Approach	Stanford Cars	FGVC Aircraft	CUB200-2011
PA-CNN [18]	92.8 (*)	–	82.8 (*)
MDTP [31]	92.5 (*)	88.4 (*)	–
MG-CNN [30]	–	86.6 (*)	83.0 (*)
PN-CNN [4]	–	–	85.4 (*)
Mask-CNN [32]	–	–	85.4 (*)
STNs [15]	–	–	84.1
FCAN [23]	91.5	–	84.3
PDFR [38]	–	–	84.5
Improved B-CNN [20]	92.0	88.5	85.8
BoostCNN [25]	92.1	88.5	86.2
KP [8]	92.4	86.9	86.2
RA-CNN(scale 1+2+3) [9]	92.5	–	85.3
MA-CNN [39]	92.8	89.9	**86.5**
REAPS *wo PSN*	92.0	89.8	81.3
REAPS	**93.1**	**91.8**	86.0
REAPS+	**93.5**	**92.6**	**86.8**

Table 2. Performance comparison of attention localization.



Approach	Accuracy
FCAN (single-attention) [23]	84.2
RA-CNN (scale 2) [9]	90.0
PSN *wo part*	91.3
PSN	92.3

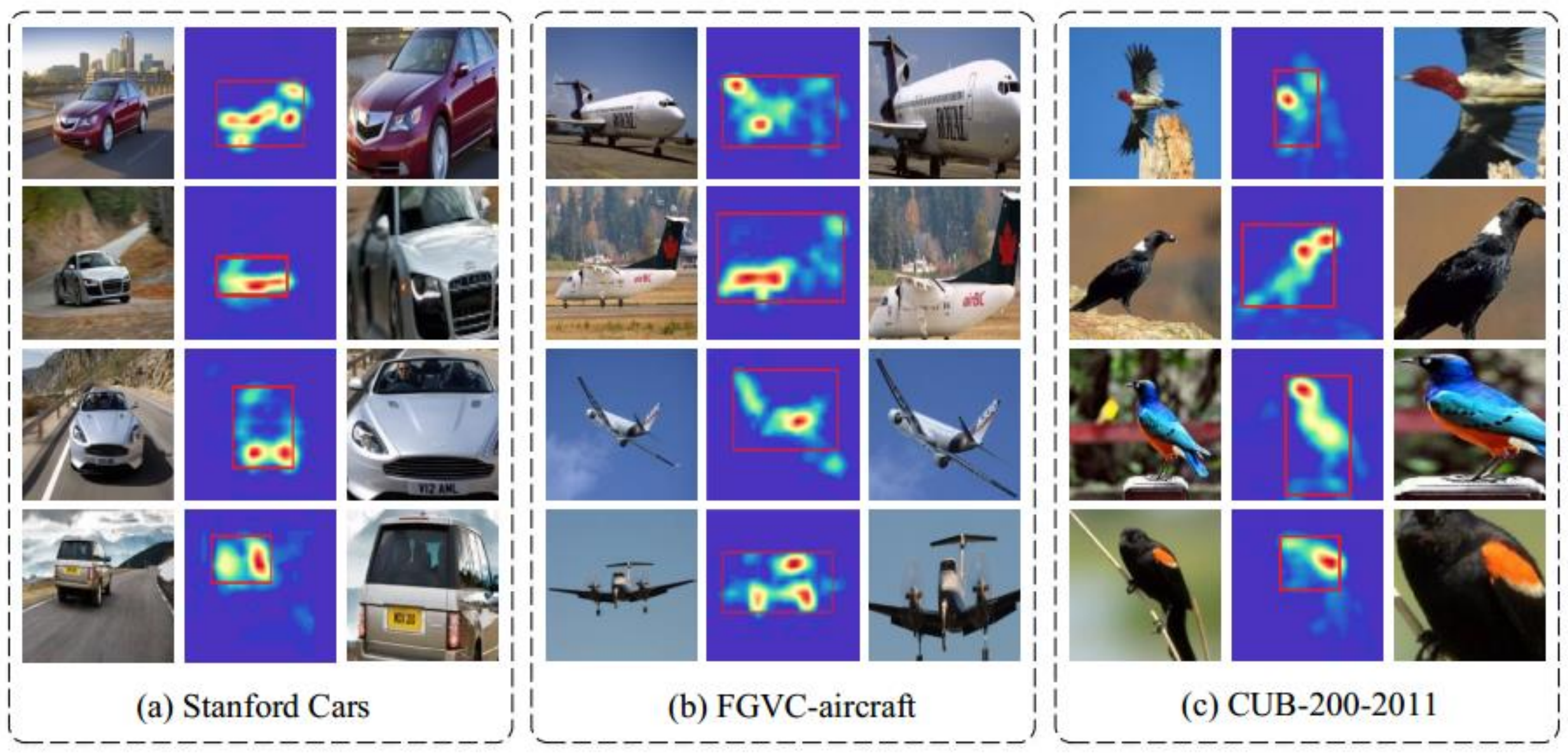

(a) Stanford Cars
(b) FGVC-aircraft
(c) CUB-200-2011

Fig. 3. Region attention localization results of RAN.

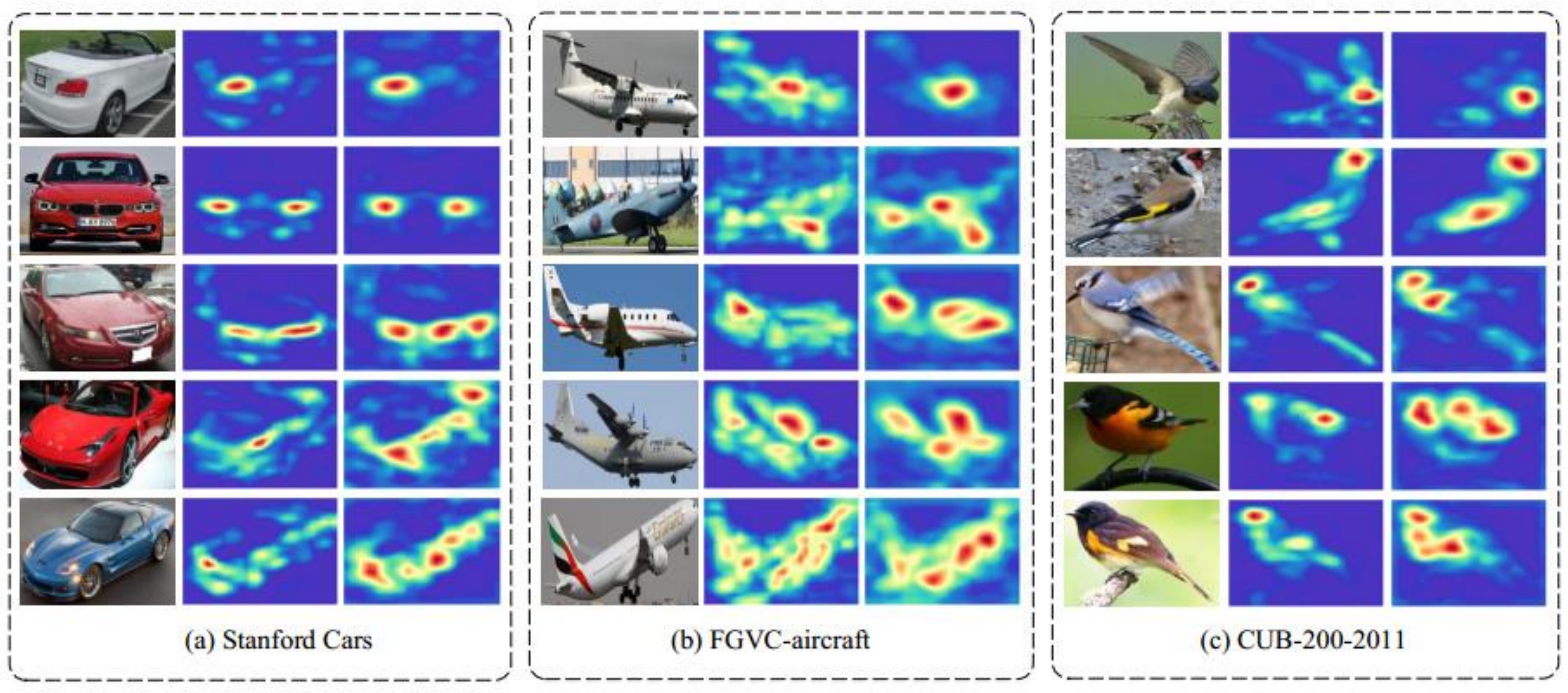

(a) Stanford Cars
(b) FGVC-aircraft
(c) CUB-200-2011

Fig. 4. Visualization of feature maps, pictures from left to right in (a-c) are the raw image, the feature map generated by PSN without part branch and the feature map generated by PSN with part branch.