Weakly Supervised Posture Mining for Fine-grained Classification

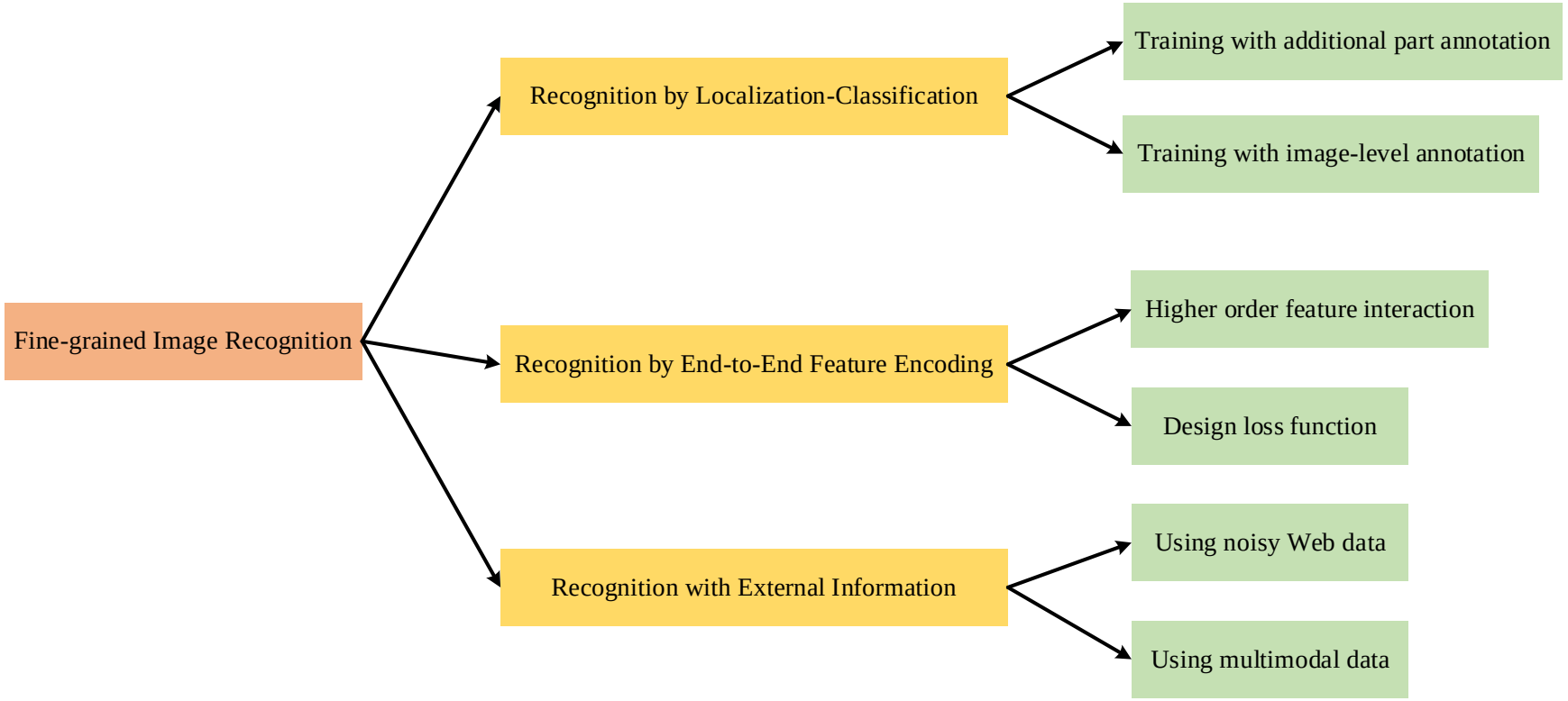
Zhenchao Tang[1,†], Hualin Yang[1,†], and Calvin Yu-Chian Chen[1,2,3,*]
[1]Sun Yat-sen University, [2]China Medical University Hospital, [3]Asia University
[†]Equal contribution, [*]Corresponding author

THU-PM-299

Motivation

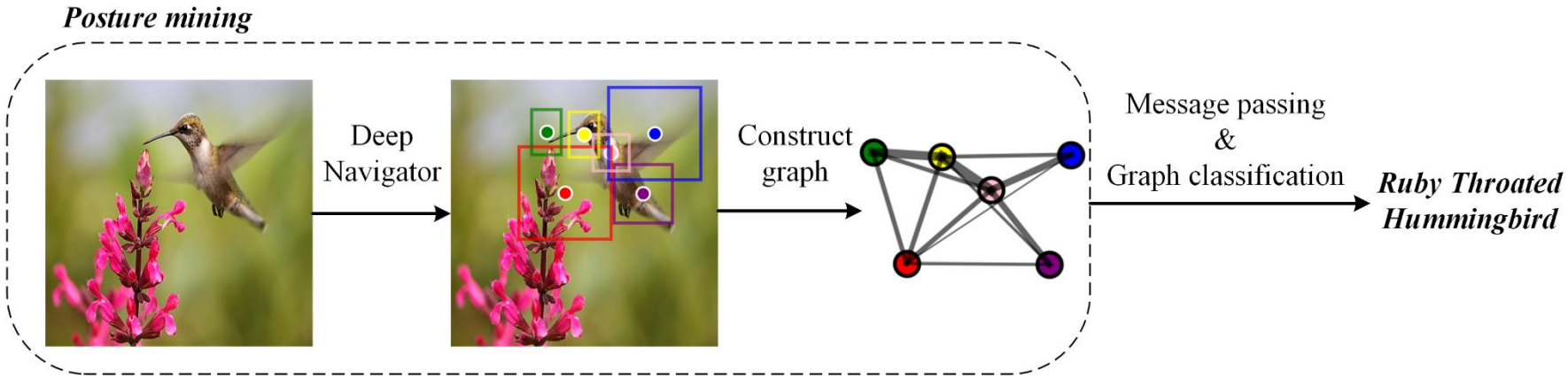
Basic-level category

Fine-grained

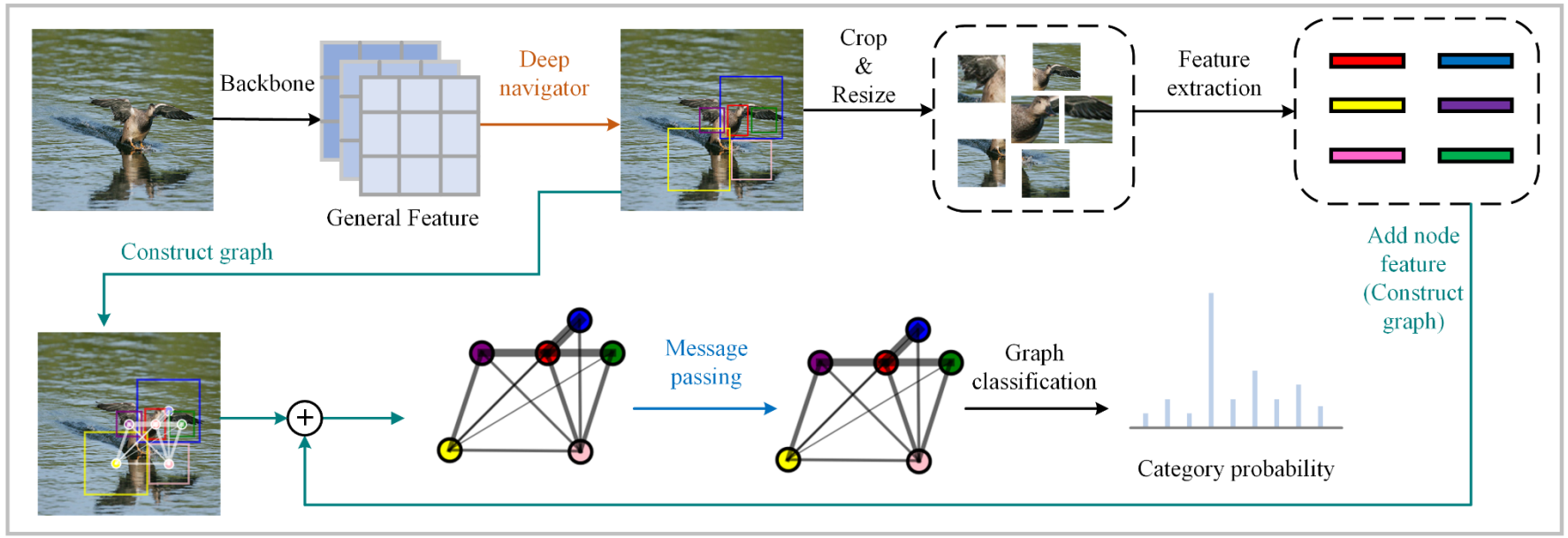


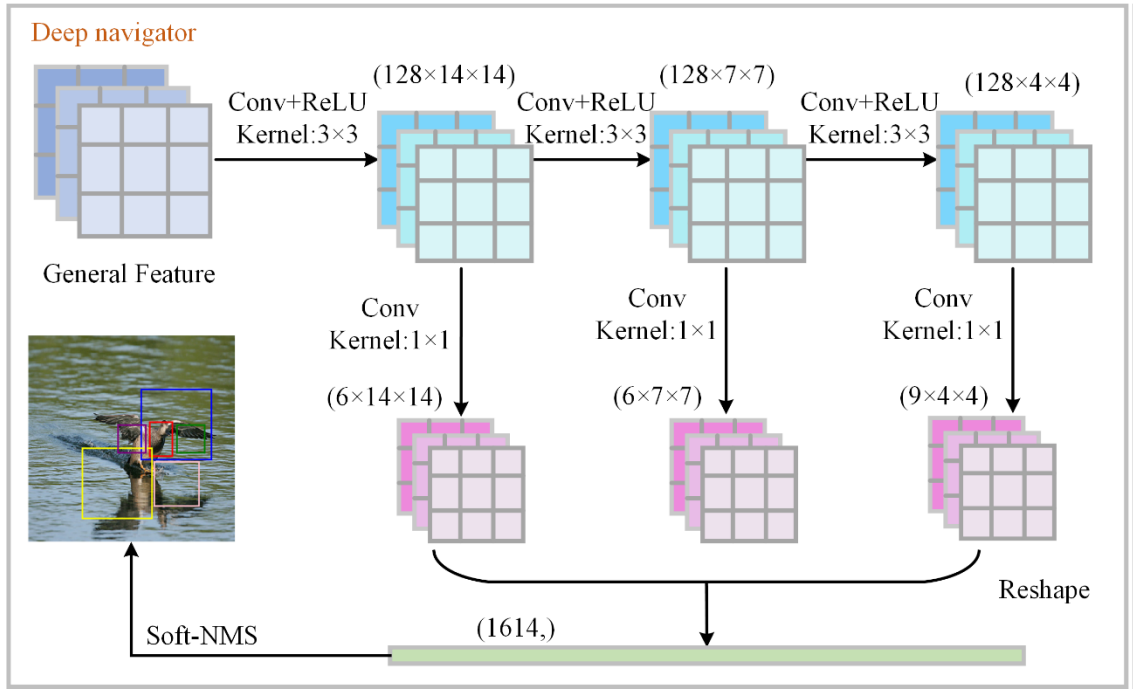
Training with image-level annotation.

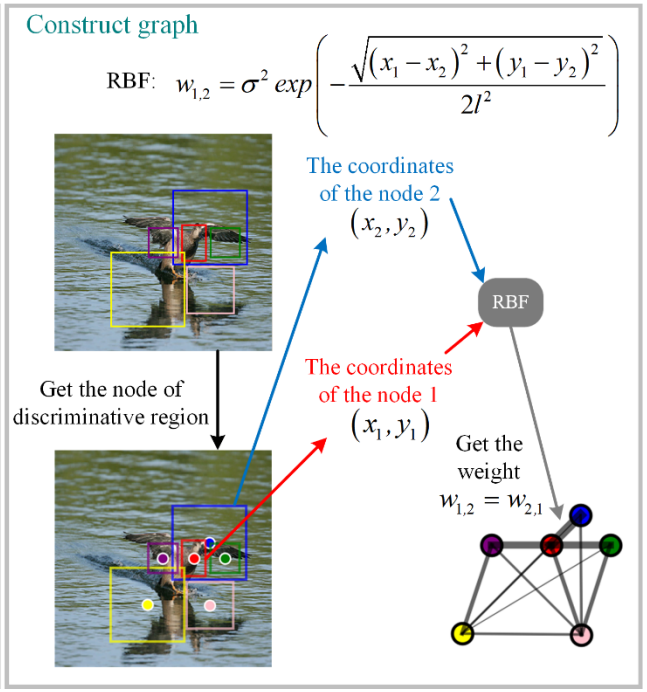
And only consider feature interaction in discriminative regions.



Method







Contributions:

(1) We propose a simple framework to mine the posture information in fine-grained classification images, our framework is able to combine easily with different backbones to good effect.

(2) We design a novel learning strategy. For the posture mining part, the loss of the Deep Navigator and the loss of message passing communicate with each other to make the model learn how to mine the posture information. For the classification part, we use RCE loss function which can effectively learn the inter-class differences of the samples.

(3) PMRC can be trained end-to-end without bounding-box/part annotations. We achieve state-of-the-art on commonly used benchmark.

Method

Message passing:

$$h_{N_{(i)}}^{(l+1)} = aggregate(\{w_{ji} \cdot h_j^{(l)}, \forall j \in N(i)\}) \quad (1)$$

$$h_i^{(l+1)} = sigmoid(W_1 \cdot concat(h_i^{(l)}, h_{N_{(i)}}^{(l+1)})) \quad (2)$$

$$h_i^{(l+1)} = h_i^{(l+1)} / \left\| h_i^{(l+1)} \right\|_2 \quad (3)$$

$$score = W_2 \cdot (\frac{1}{|V|} \sum_{v \in V} h_v) \quad (4)$$

Deep Navigator loss:

$$L_{navigator} = \sum_{(m,n)|C_m<C_n} max\{1 - (I_n - I_m), 0\} \quad (5)$$

Message passing loss:

$$\{s_i\}_{i=1}^{M} = \{W_2 \cdot h_i\}_{i=1}^{M} \quad (6)$$

$$L_{message} = -\sum_{i=1}^{M} log(\frac{exp(s_i[label])}{\sum_{j=1}^{classnum} exp(s_i[j])}) \quad (7)$$

CE loss for the score of features:

$$L_{backbone} = -log(\frac{exp(raw[label])}{\sum_{j=1}^{classnum} exp(raw[j])}) \quad (8)$$

RCE for the whole graph classification:

$$score' = \frac{exp(-score)}{\sum_{j=1}^{classnum} exp(-score[j])} \quad (9)$$

$$L_{graph} = -R_{label}^{T} log(score') \quad (10)$$

Total loss:

$$L = \alpha L_{backbone} + \beta L_{navigator} + \gamma L_{message} + \theta L_{graph} \quad (11)$$

Result: Ablation study

Table 1. Ablation study with Top-1 accuracy of Message Passing and RCE on CUB-200-2011(%), backbone: ResNet50.

Regions number	3	4	5	6	7	8
Message(RCE)	**90.9**	**91.0**	**91.3**	**91.8**	**91.5**	**91.4**
Concat(RCE)	88.6	88.9	89.3	89.1	89.0	89.1
Message(CE)	**89.2**	**89.4**	**90.1**	**90.6**	**90.4**	**90.4**
Concat(CE)	87.3	87.4	87.6	87.6	87.3	87.5

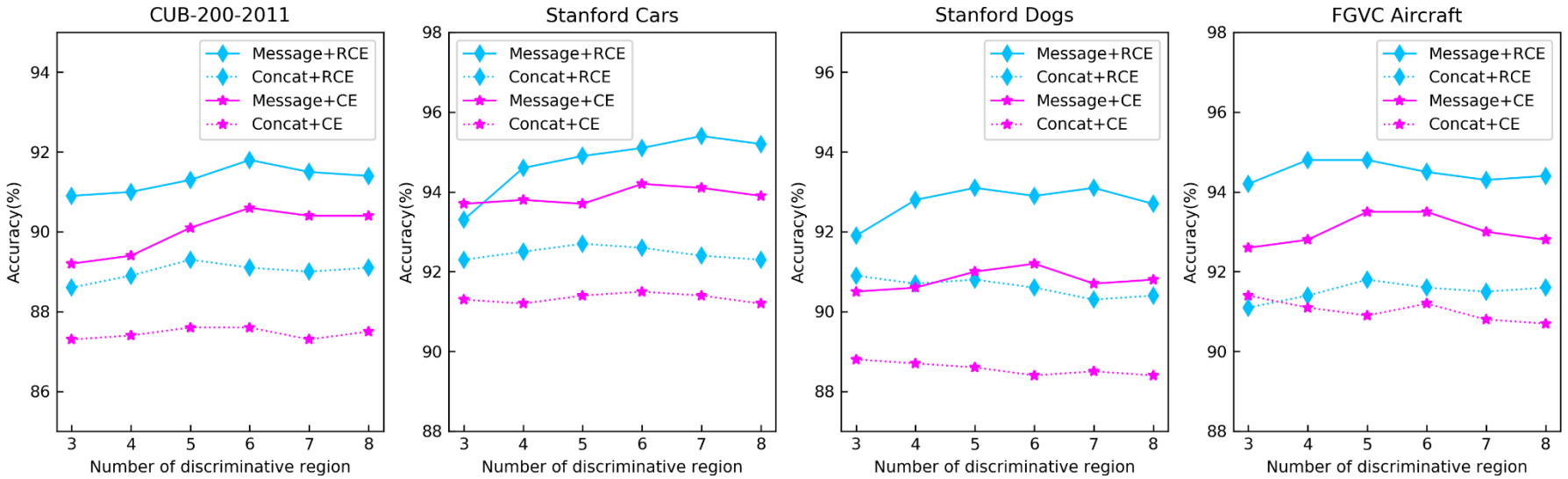
Table 2. Reasoning Speed and Accuracy Test on CUB-200-2011 (Top-1 Accuracy)

Method	backbone	Speed (fps)	Accuracy(%)
ResNet50 (RCE)	-	91	85.7
ResNet50 (CE)	-	91	84.5
Message+6 Regions (RCE)	ResNet50	85	91.8
Message+7 Regions (RCE)	ResNet50	84	91.5
Message+6 Regions (CE)	ResNet50	85	90.6
Message+7 Regions (CE)	ResNet50	84	90.4
SwinTrans (RCE)	-	49	91.2
SwinTrans (CE)	-	49	90.3
Message+6 Regions (RCE)	SwinTrans	45	94.3
Message+6 Regions (CE)	SwinTrans	45	**93.5**

(Supplementary Table 1)

Table 1. Ablation Study of Loss Propotion on CUB-200-2011, backbone: ResNet50

α	β	γ	θ	Top-1 Accuracy(%)
0.25	0.25	0.25	0.25	88.7
0.2	0.2	0.2	0.4	91.2
0.4	0.2	0.2	0.2	88.2
0.2	0.4	0.2	0.2	89.5
0.2	0.2	0.4	0.2	90.9
0.1	0.25	0.25	0.35	**91.8**



The relationship between accuracy and the number of discriminative regions. Under the condition of setting different number of discriminative regions, the model recognition effects of introducing message passing network and RCE training are tested respectively

Result: Comparisons with existing approaches

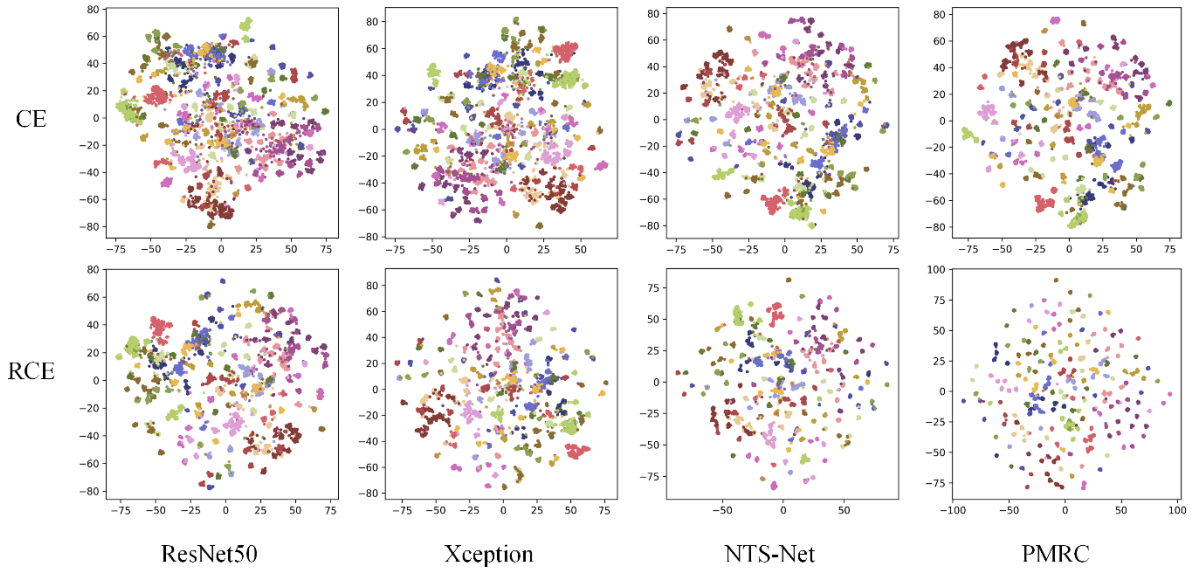
Table 3. Comparison of Different Methods on CUB-200-2011, Stanford Cars, FGVC Aircraft and Stanford Dogs. (Top-1 Accuracy(%))

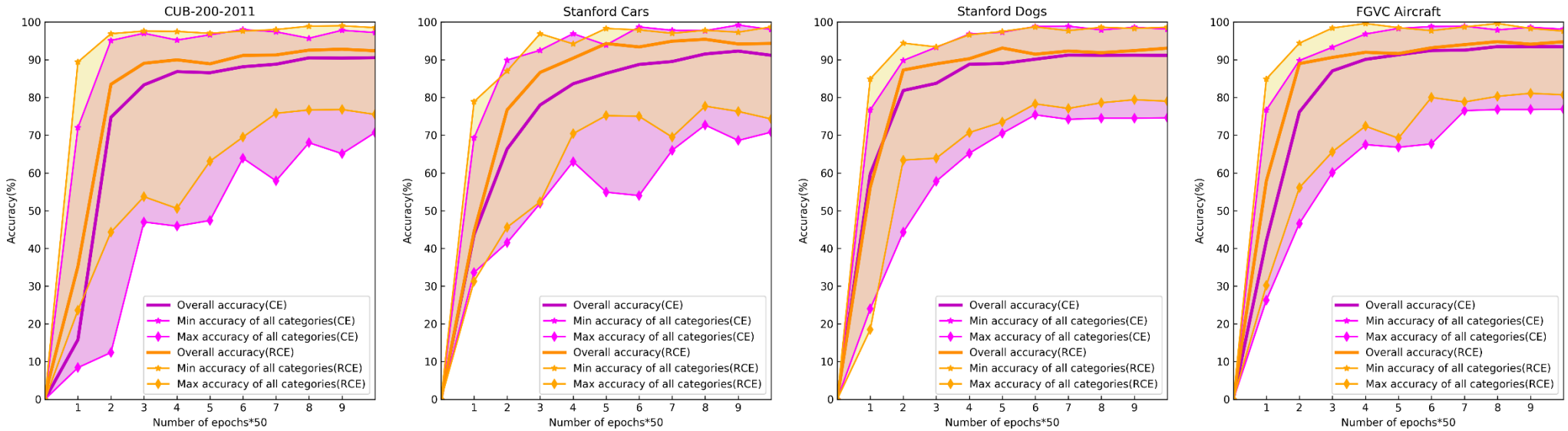
Method	Extra Supervision	CUB-200-2011	Stanford Cars	FGVC Aircraft	Stanford Dogs	speed	params
MetaFormer [8]	✓	**92.9**	**95.4**	92.8	-	-	-
DATL [17]	✓	91.2	94.5	**93.1**	**92.2**	-	-
TA-FGVC [23]	✓	88.1	-	-	88.9	-	-
PA-CNN [20]	✓	85.4	92.8	-	-	-	-
BoT [43]	✓	-	92.5	88.4	-	-	-
FCAN [27]	✓	84.3	91.3	-	88.9	-	-
MG-CNN [40]	✓	85.1	-	86.6	-	-	-
PIM [6]	×	**92.8**	-	-	-	-	-
DCAL [51]	×	92.0	95.3	93.3	-	-	-
Vit-SAC [9]	×	91.8	94.5	93.1	-	-	-
CAP [1]	×	91.8	-	**94.9**	-	-	-
TransFG [13]	×	91.7	94.8	-	**92.3**	-	-
FFVT [41]	×	91.6	-	-	91.5	-	-
CAL [33]	×	90.6	**95.5**	94.2	-	-	-
Inception-v4 [32]	×	-	95.3	-	-	-	-
API-Net [52]	×	90.0	95.3	-	90.3	-	-
DenseNet161+MM+FRL [50]	×	88.5	95.2	-	-	-	-
GCL [44]	×	88.3	95.1	93.2	90.5	-	-
NTS-Net [47]	×	87.5	91.4	93.9	-	-	-
SwinTransformer [28]	×	90.3	92.7	90.6	91.1	49fps	88M
ResNet50 [24]	×	84.5	88.6	87.2	84.7	91fps	25M
DenseNet161 [15]	×	86.6	90.4	90.9	88.3	38fps	29M
VGG16 [37]	×	77.8	83.3	85.3	81.6	68fps	138M
Our PMRC (SwinTransformer)	×	**94.3**	**96.9**	**96.7**	**95.2**	45fps	89M
Our PMRC (ResNet50)	×	91.8	95.4	94.8	93.1	85fps	26M
Our PMRC (DenseNet101)	×	91.3	95.2	94.0	92.7	36fps	30M
Our PMRC (VGG16)	×	86.3	89.1	91.3	89.9	63fps	139M

Discussion: RCE

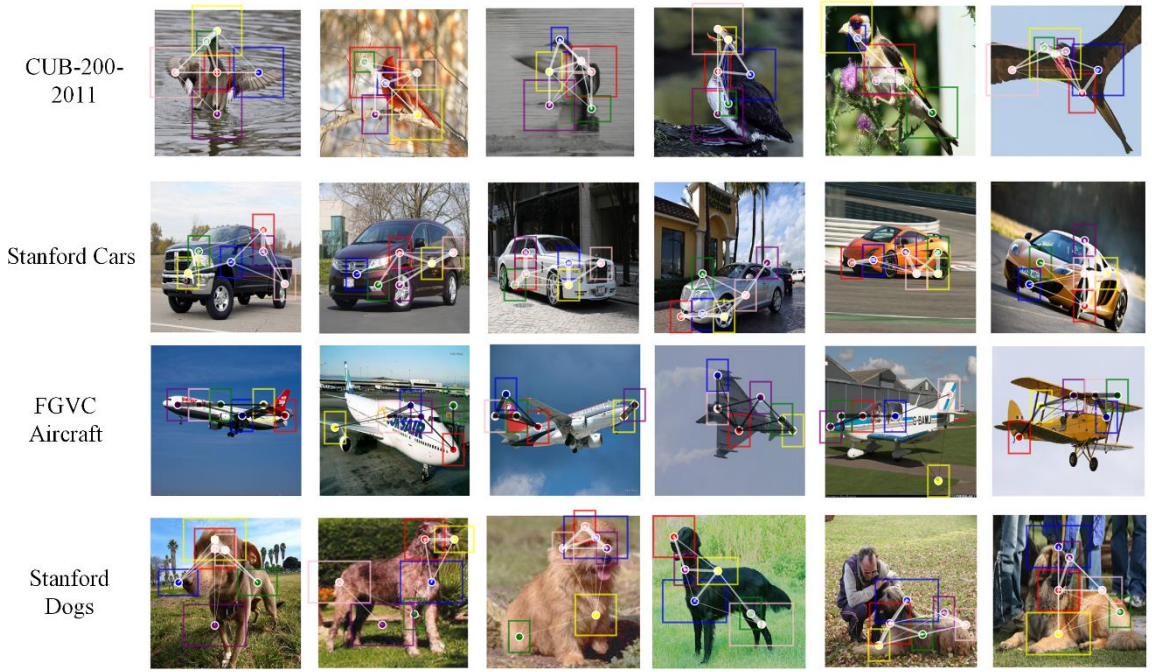
Table 4. Comparison of CE and RCE in previous fine-grained tasks with the increment of Top-1 accuracy(%).

Method	Extra.S	CUB	Cars	Aircraft	Dogs
ResNet50 [14]	×	+1.23	+0.62	+0.54	+1.16
DenseNet161 [15]	×	+0.37	+0.70	+1.09	+0.71
Xception [5]	×	+1.17	+1.52	+1.19	+1.60
Incep.V3 [38]	×	+1.11	+1.27	+1.73	+1.82
MobileNetV2 [35]	×	+1.13	+1.69	+1.04	+1.28
B-CNN [26]	×	+1.68	+0.97	+0.99	+1.42
NTS-Net [47]	×	+1.72	+1.60	+2.12	+1.45
DATL [17]	✓	-0.28	+0.27	-0.19	+0.11
DAT [30]	✓	+1.20	+1.25	+1.53	+1.25
TResNet-L-V2 [34]	✓	-0.07	+0.42	+0.18	+0.36
SAM [11]	✓	+1.12	+1.68	+1.96	+1.65
MG-CNN [40]	✓	+1.62	+1.28	+1.02	+1.41



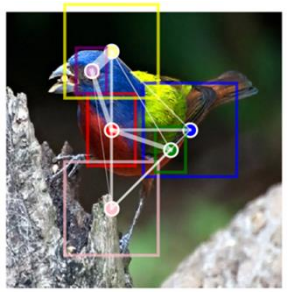


Discussion: Posture Mining

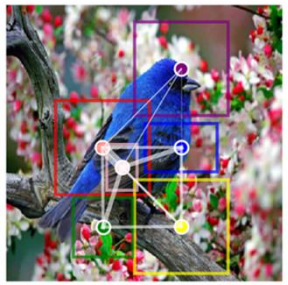


The first row to the fourth row correspond to CUB-200-2011, Stanford Cars, FGVC Aircraft, Stanford Dogs.

Bounding boxes represent the discriminative regions of the object, and graph represents the posture of the object.



Painted Bunting



Indigo Bunting

pose information is helpful for fine-grained recognition

Known facts:
Painted bunting and indigo bunting belong to bunting. The subtle differences between them are not only reflected in appearance, but also in behavior.
Painted bunting likes to feed on seeds on shrubs near the ground, and indigo bunting likes to feed on insects on trees.