Fine-Grained Object Classification via Self-Supervised Pose Alignment

Xuhui Yang[1], Yaowei Wang[1*], Ke Chen[2,1*], Yong Xu[1,2,3], Yonghong Tian[1]
[1]Peng Cheng Laboratory [2]South China University of Technology
[3]China Communication and Computer Network Laboratory of Guangdong
{yangxh, wangyw}@pcl.ac.cn, {chenk, yxu}@scut.edu.cn, tianyh@pcl.ac.cn

CVPR2022

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

作者介绍

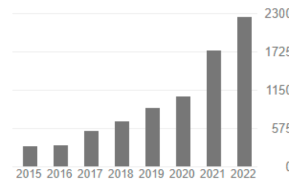
Yonghong Tian, Professor, IEEE Fellow

School of CS and School of ECE at Peking University, Pengcheng Laboratory
Verified email at pku.edu.cn - Homepage

Machine Learning   Neuromorphic Vision   Multimedia Big Data

Cited by   VIEW ALL

	All	Since 2017
Citations	8756	7182
h-index	44	39
i10-index	144	115



yaowei wang

Pengcheng Laboratory
在 bit.edu.cn 的电子邮件经过验证

multimedia analysis

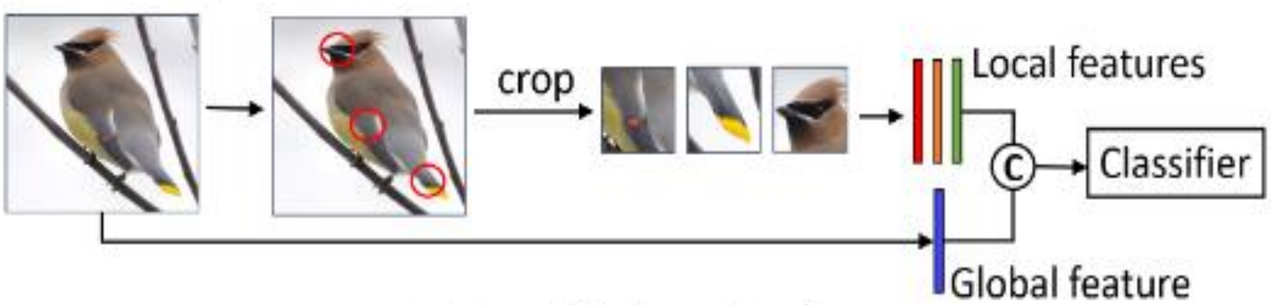
Deep relative distance learning: Tell the difference between similar vehicles H Liu, Y Tian, Y Wang, L Pang, T Huang Proceedings of the IEEE Conference on Computer Vision and Pattern ...	619	2016
Conformer: Local features coupling global representations for visual recognition Z Peng, W Huang, S Gu, L Xie, Y Wang, J Jiao, Q Ye Proceedings of the IEEE/CVF International Conference on Computer Vision, 367-376	136	2021
PanGu-α: Large-scale Autoregressive Pretrained Chinese Language Models with Auto-parallel Computation W Zeng, X Ren, T Su, H Wang, Y Liao, Z Wang, X Jiang, ZZ Yang, K Wang, ... arXiv preprint arXiv:2104.12369	62	2021

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

基于定位的细粒度识别

- 定位网络：检测图像中的**判别性区域**，裁剪出局部特征
- 识别网络：综合局部-全局特征进行识别

- 流程如下图所示

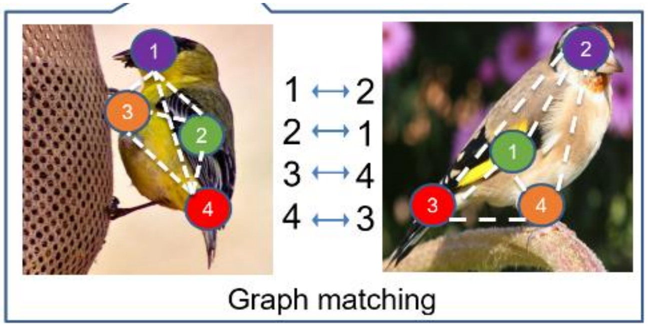


- 弱监督定位：分类越准说明定位越准，互相促进。

问题

- 细粒度识别类间差异小，类内差异大
- 为什么类内差异大？ 姿态多样

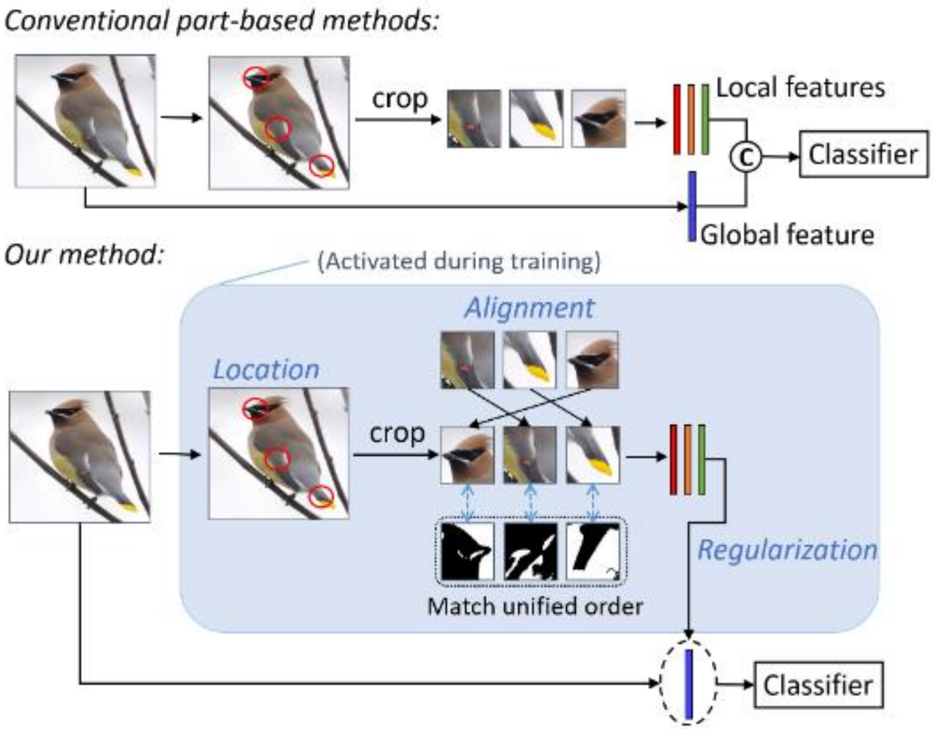
- 本文提出网络应该对姿态不敏感
  - 准确的局部特征定位
  - 实现自监督的局部特征对齐



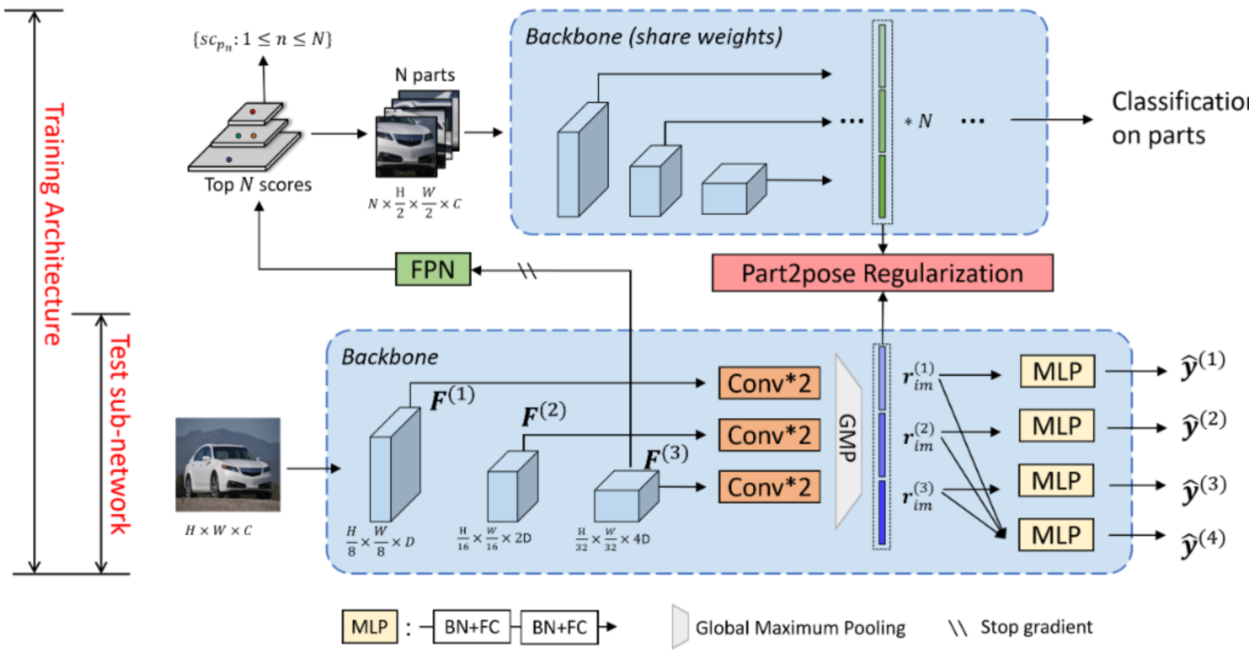
- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

Overview

- 以往方法将**局部特征**与**全局特征**简单拼接，进行预测
- 本文利用局部特征构造姿态不敏感的表示，并作为正则化约束，仅在训练过程中使用



总体结构



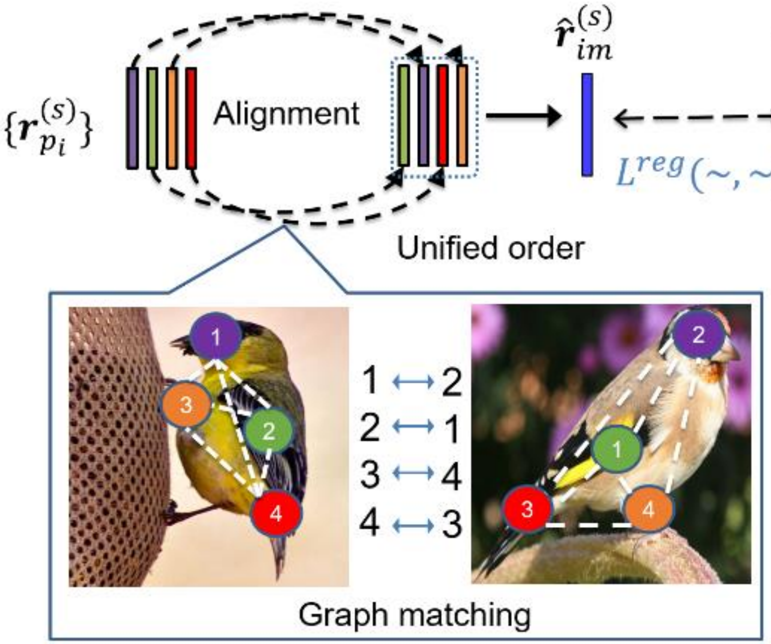
Training Architecture
Test sub-network
$\{sc_{p_n}: 1 \le n \le N\}$
Top N scores
N parts
$N \times \frac{H}{2} \times \frac{W}{2} \times C$
Backbone (share weights)
$* N$
Classification on parts
FPN
Part2pose Regularization
Backbone
$H \times W \times C$
$F^{(1)}$
$F^{(2)}$
$F^{(3)}$
$\frac{H}{8} \times \frac{W}{8} \times D$
$\frac{H}{16} \times \frac{W}{16} \times 2D$
$\frac{H}{32} \times \frac{W}{32} \times 4D$
Conv*2
Conv*2
Conv*2
GMP
$r_{im}^{(1)}$
$r_{im}^{(2)}$
$r_{im}^{(3)}$
MLP
MLP
MLP
MLP
$\hat{y}^{(1)}$
$\hat{y}^{(2)}$
$\hat{y}^{(3)}$
$\hat{y}^{(4)}$
MLP : BN+FC BN+FC
Global Maximum Pooling
\\ Stop gradient

Graph Matching for Part Alignment

- 每张图像的N个局部特征构成一个graph
- 对N做全排列，选出相似度得分最高的

$$M_{ij} = <\boldsymbol{r}_{p_i}, \boldsymbol{r}_{p_j}>,$$

$$\hat{M} = \underset{M'}{\operatorname{argmax}}\, vec(M')^T vec(M),$$

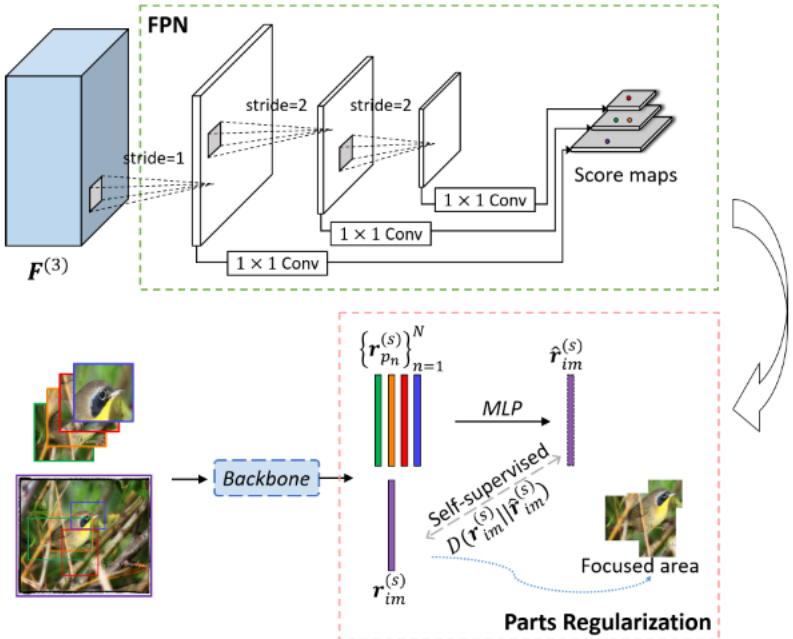


Contrastive Feature Regularization

- 局部特征组合得到的特征应该与全局特征接近，即希望模型忽略背景信息

- 局部特征拼接后通过MLP与全局特征计算KL散度

$$L^{reg} = \sum_{s=1}^{S} \ell_{kl}(\boldsymbol{r}_{im}^{(s)}, \phi([\boldsymbol{r}_{p_1}^{(s)}; \boldsymbol{r}_{p_2}^{(s)}; \ldots; \boldsymbol{r}_{p_N}^{(s)}])),$$



Curriculum Supervision

- 课程学习是指模拟人类学习的过程，由易到难进行训练。
- 本文认为将标签软化后更容易学习，因此在训练过程中，让软化的标签逐步接近one-hot标签。
- 即 $\alpha: \frac{1}{k} \rightarrow 1$
- 效果：

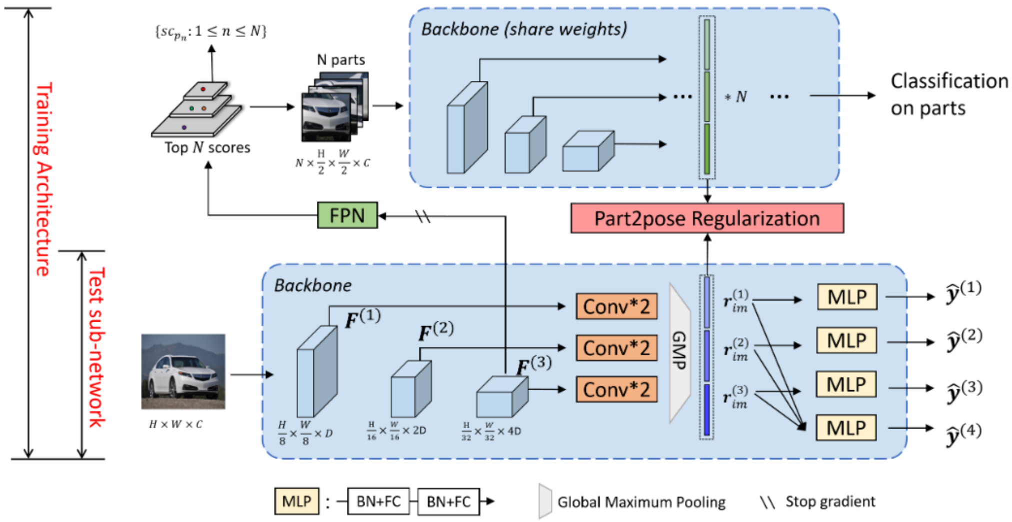
Method	CUB	CAR	AIR
(a) Baseline	85.5	92.7	90.3
(b) Baseline+CS	88.4	94.9	93.8

$$y_\alpha[t] = \begin{cases} \alpha, & t = y \\ \frac{1-\alpha}{K}, & t \neq y \end{cases},$$

$$\begin{aligned} \ell_{sce}(\hat{\boldsymbol{y}}^{(s)}, y, \alpha^{(s)}) &= \ell_{ce}(\hat{\boldsymbol{y}}^{(s)}, \boldsymbol{y}_{\alpha^{(s)}}) \\ &= \sum_{t=0}^{K-1} -\boldsymbol{y}_{\alpha^{(s)}}[t] \log(\hat{\boldsymbol{y}}^{(s)}[t]), \end{aligned}$$

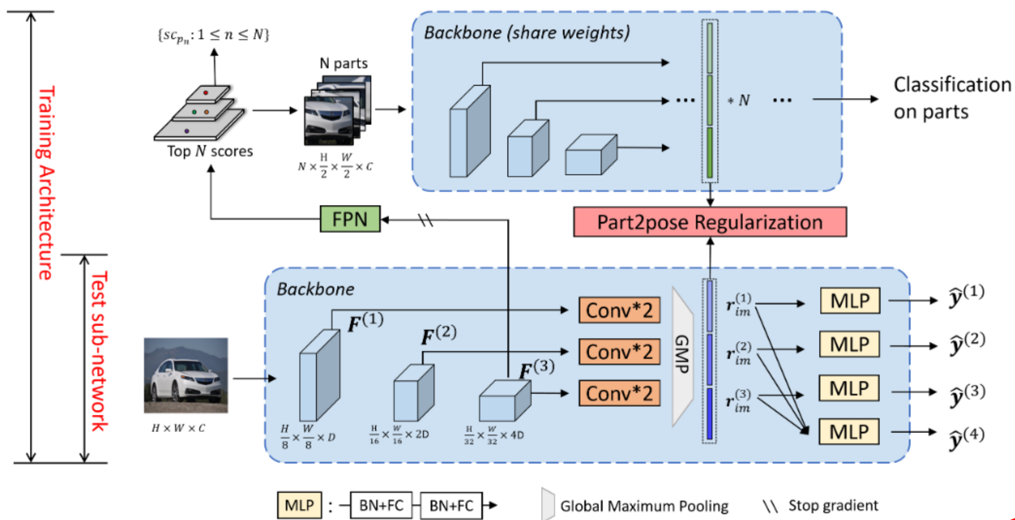
$$L_{im}^{cls} = \sum_{s=1}^{S+1} \ell_{sce}(\hat{\boldsymbol{y}}^{(s)}, y, \alpha^{(s)}).$$

训练、推理



- Loss: $$L = L_{im}^{cls} + L_{parts}^{cls} + L^{rank} + \beta \cdot L^{reg},$$
- Prediction: $$\hat{y}^{(final)} = \sum_{s=1}^{S+1} \hat{y}^{(s)},$$

训练、推理



$$L^{rank} = \sum_{n=1}^{N}\sum_{n'=1}^{N} \ell_{hg}(L_{p_n}, L_{p_{n'}}) * c_{nn'}$$
$$= \sum_{n=1}^{N}\sum_{n'=1}^{N} max(0, L_{p_n} - L_{p_{n'}} + \delta) * c_{nn'},$$

with the indicator $c_{nn'}$ is defined as

$$c_{nn'} = \begin{cases} 1, & sc_{p_n} > sc_{p_{n'}} \\ 0, & sc_{p_n} \leqslant sc_{p_{n'}} \end{cases},$$

定位越准确←→分类越准确

- Loss: $$L = L_{im}^{cls} + L_{parts}^{cls} + \boxed{L^{rank}} + \beta \cdot L^{reg},$$
- Prediction: $$\hat{y}^{(final)} = \sum_{s=1}^{S+1} \hat{y}^{(s)},$$

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

模型性能

Method	Backbone	Accuracy (%)		
		CUB	CAR	AIR
B-CNN [21]	VGG	84.1	91.3	84.1
RA-CNN [9]	VGG19	85.3	92.5	88.2
MA-CNN [44]		86.5	92.8	89.9
FCAN [22]	ResNet50	84.7	93.1	-
MAMC [32]		86.3	93.0	-
DFL-CNN [35]		87.4	93.1	91.7
NTS-Net [41]		87.5	93.9	91.4
DCL [4]		87.8	94.5	93.0
TASN [46]		87.9	93.8	-
Cross-X [24]		87.7	94.6	92.6
S3N [6]		88.5	94.7	92.8
LIO [47]		88.0	94.5	92.7
BNT [15]		88.1	94.6	92.4
ASD [31]		88.6	94.9	93.5
DF-GMM [36]		88.8	94.8	93.8
PMG [8]		89.6	95.1	93.4
API-Net [48]	ResNet101	88.6	94.9	93.4
API-Net [48]	DenseNet-161	90.0	95.3	93.9
P2P-Net (ours)	ResNet34	89.5	94.9	92.6
P2P-Net (ours)	ResNet50	**90.2**	**95.4**	**94.2**

Table 1. Comparison with the state-of-the-art methods.

消融实验

- CS： curriculum supervision
- FR： feature regularization
- FC： feature concatenation
- UPA： unsupervised part alignment

Method	Accuracy (%)		
	CUB	CAR	AIR
(a) Baseline	85.5	92.7	90.3
(b) Baseline+CS	88.4	94.9	93.8
(c) Baseline+FR (w/o UPA)	89.0	94.8	92.0
(d) Baseline+FR (w/ UPA)	89.0	95.0	92.5
(e) Baseline+FC	88.4	94.7	93.8
(f) Baseline+CS+FR (w/o UPA)	90.0	95.0	93.9
(g) Baseline+CS+FR (w/ UPA)	**90.2**	**95.4**	**94.2**

Parts number (N)	2	3	4	5	6
CUB	88.1	89.9	**90.2**	**90.2**	90.0

可视化

- T-SNE

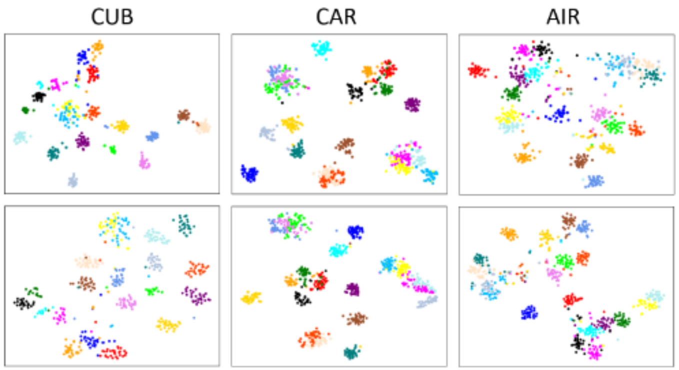


Figure 7. A T-SNE plot of learned representations on three datasets. First row: the baseline model; second row: baseline+FR (w/ UPA).

- 类内特征的均方差

Method	RMSE		
	CUB	CAR	AIR
Baseline	0.501	0.354	0.399
Baseline+FC	0.268	0.213	0.354
Baseline+FR (w/ UPA)	**0.213**	**0.179**	**0.263**

Table 3. RMSE of learned representations of different methods.

可视化

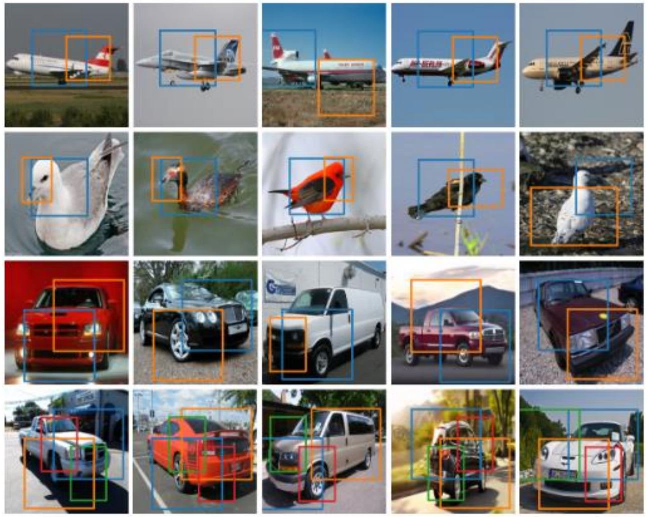


Figure 5. Discriminative parts detected by our P2P-Net.

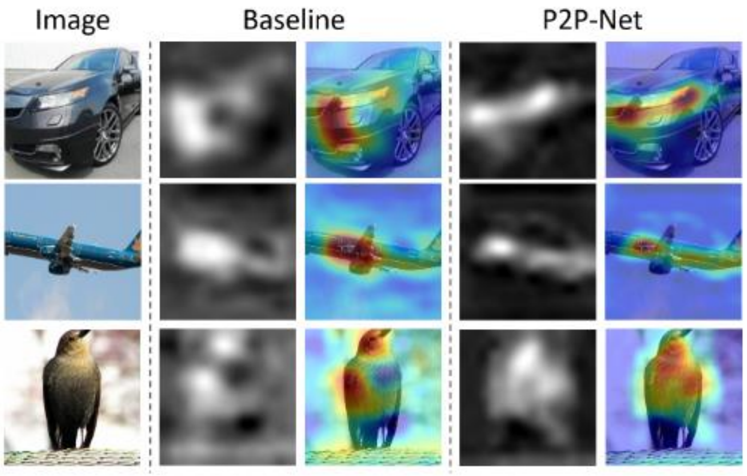


Figure 6. Class activation maps of some test samples.

- 作者介绍
- 研究背景
- 本文方法
- 实验效果
- 总结反思

总结

- 本文是首次尝试解决姿态多样性的问题
- 用图匹配方法实现自监督姿态对齐

- 向网络中加入几何结构的先验信息

Thanks!