FGVC调研报告

Paper Reading by Yiwei Sun

2022.12.19

智能多媒体内容计算实验室
Intelligent Multimedia Content Computing Lab

- FGVC概述
- 特征工程时期的识别算法
- 深度学习技术的崛起
- 步入弱监督时代

FGVC概述

细粒度视觉分类：对图片中的对象进行分类，不过所属类别的粒度更为精细。

起源：该任务的研究最早源于2006年的论文《Subordinate class recognition using relational object models》，其中研究了对不同款式摩托车的分类。这篇论文也正式提出了“局部”在FGVC中的重要意义。

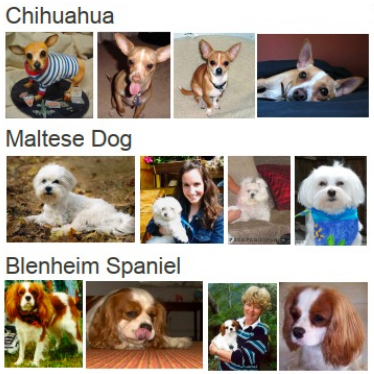
如何提取有区分度的局部信息，是这十几年来，所有科研工作者在做的研究。

国内	国外
IMCC	Stanford Vision and Learning Lab
多媒体信息处理研究室（彭宇新）	Berkeley Artificial Intelligence Research
JD AI Research（梅涛）	Computational Cognition, Vision, and Learning research group

FGVC概述

FGVC相关数据集

Recog.	*Oxford Flowers* [45]	2008	Flowers	8,189	102					✓
	CUB200-2011 [13]	2011	Birds	11,788	200	✓	✓		✓	✓
	Stanford Dogs [42]	2011	Dogs	20,580	120	✓				
	Stanford Cars [43]	2013	Cars	16,185	196	✓				
	FGVC Aircraft [44]	2013	Aircrafts	10,000	100	✓		✓		
	Birdsnap [41]	2014	Birds	49,829	500	✓	✓		✓	
	Food101 [47]	2014	Food dishes	101,000	101					
	NABirds [1]	2015	Birds	48,562	555	✓	✓			
	Food-975 [50]	2016	Foods	37,885	975				✓	
	DeepFashion [38]	2016	Clothes	800,000	1,050	✓	✓		✓	
	Fru92 [46]	2017	Fruits	69,614	92			✓		
	Veg200 [46]	2017	Vegetable	91,117	200			✓		
	iNat2017 [2]	2017	Plants & Animals	857,877	5,089	✓		✓		
	Dogs-in-the-Wild [27]	2018	Dogs	299,458	362					
	RPC [5]	2019	Retail products	83,739	200	✓		✓		
	Products-10K [48]	2020	Retail products	150,000	10,000	✓		✓		
	iNat2021 [3]	2021	Plants & Animals	3,286,843	10,000			✓		

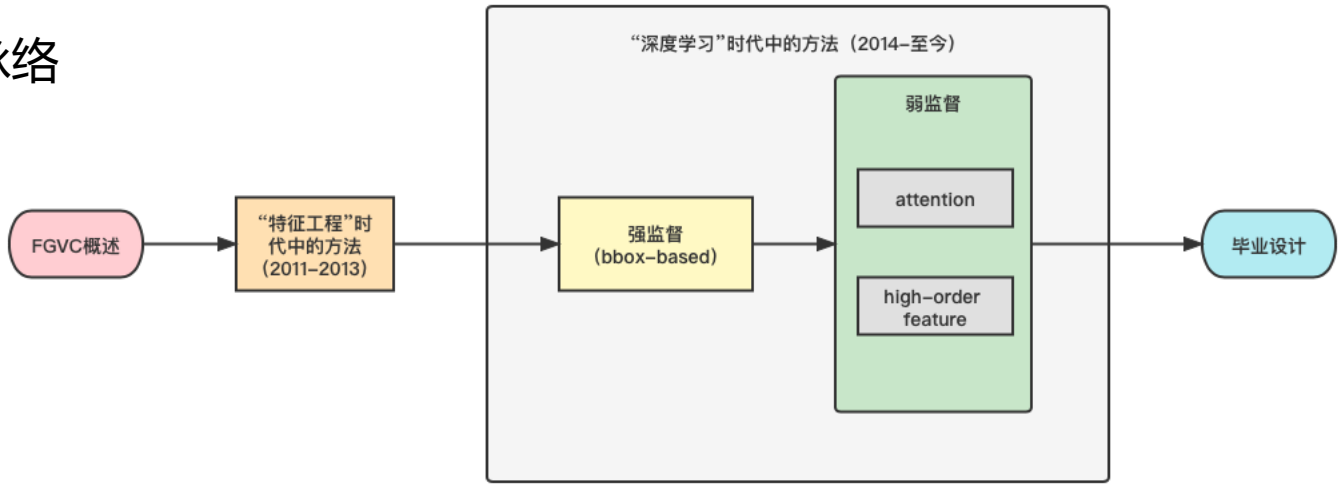


FGVC概述

FGVC的应用：

1. 食品计算：根据食物的图像判断菜品，进一步完成自助结账、营养计算等功能。
2. 智慧交通：对车辆进行有效分类，为交通管理提供更精细化的数据。
3. 动植物分类：根据用户提供的图像，判断动植物种类，辅助生态行业的非专家从业者进行工作。

Pre的脉络



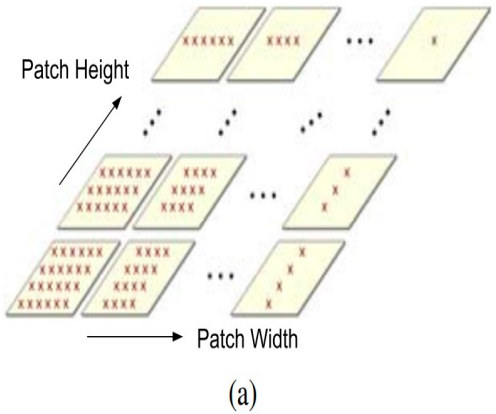
- FGVC概述
- 特征工程时期的识别算法
- 深度学习技术的崛起
- 步入弱监督时代

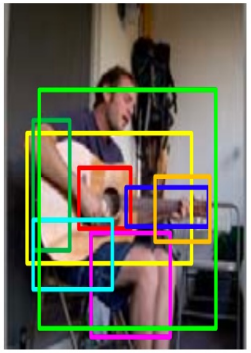
特征工程时期的识别算法

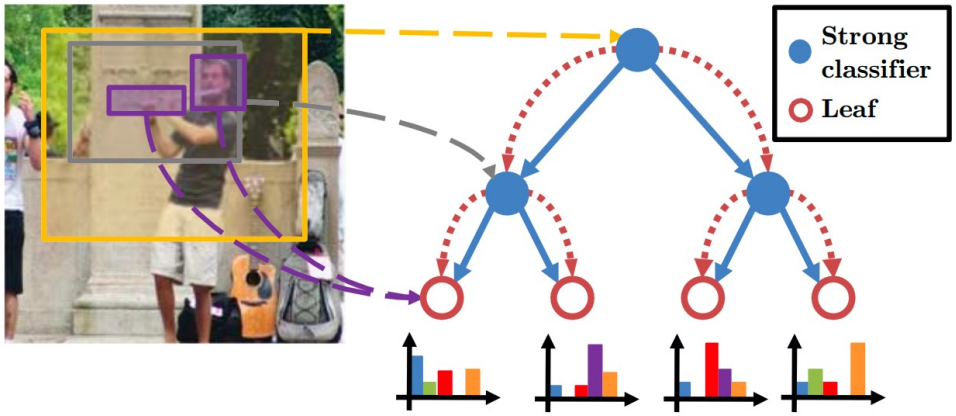
特征工程：对数据预处理、特征抽取、特征构造、特征选择的研究。

研究重点：通过一些算法直接得到相对鲁棒的局部特征是很重要的。

《Combining Randomization and Discrimination for Fine-Grained Image Categorization》





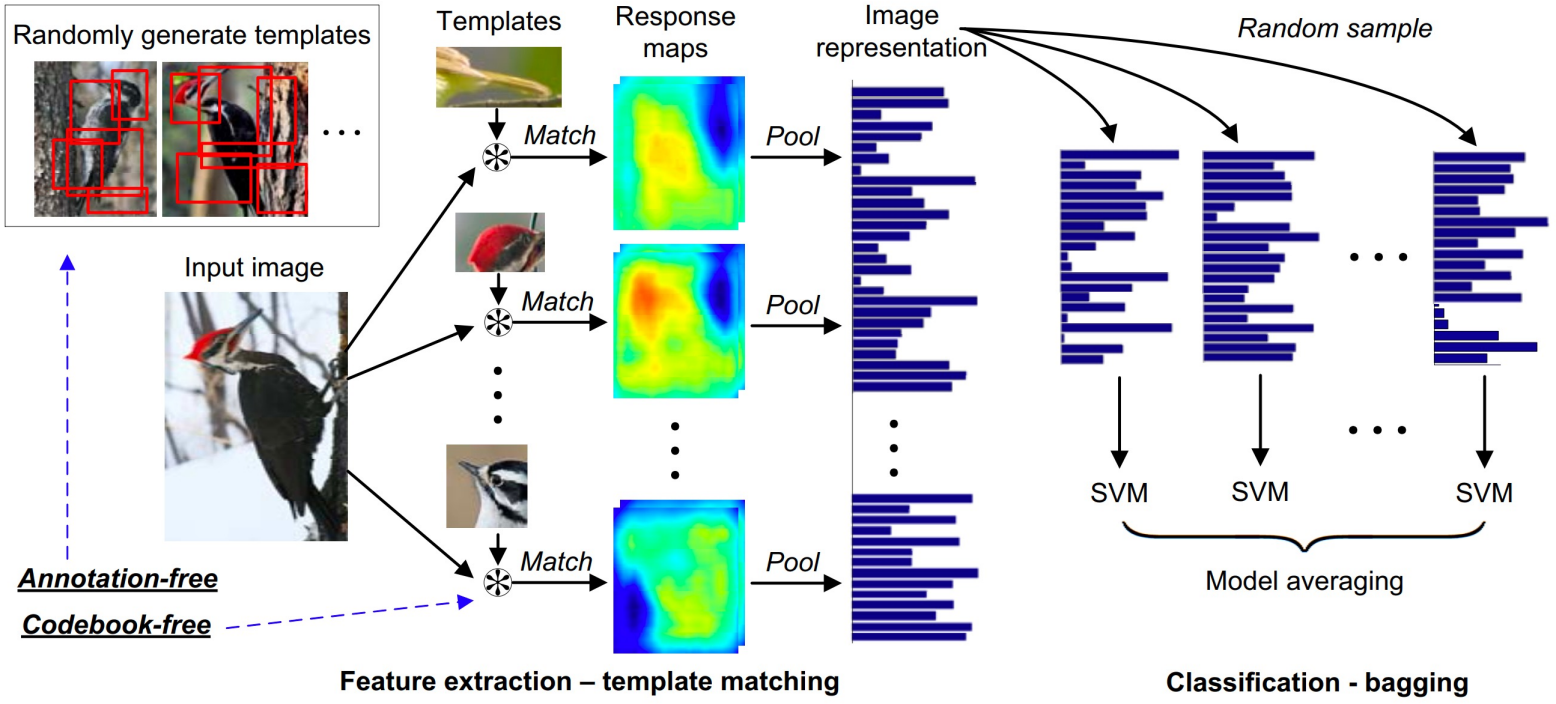


Dense Sampling

决策树（随机森林）

特征工程时期的识别算法

《A Codebook-Free and Annotation-Free Approach for Fine-Grained Image Categorization》



特征工程时期的识别算法

早期方法的特征：

1. 指导思想明确但随机性强；
2. 得到的特征庞大且冗余，因此介入集成学习的方法。

除此之外，还有一些需要人参与的算法（Crowdsourcing），此处不过多赘述。

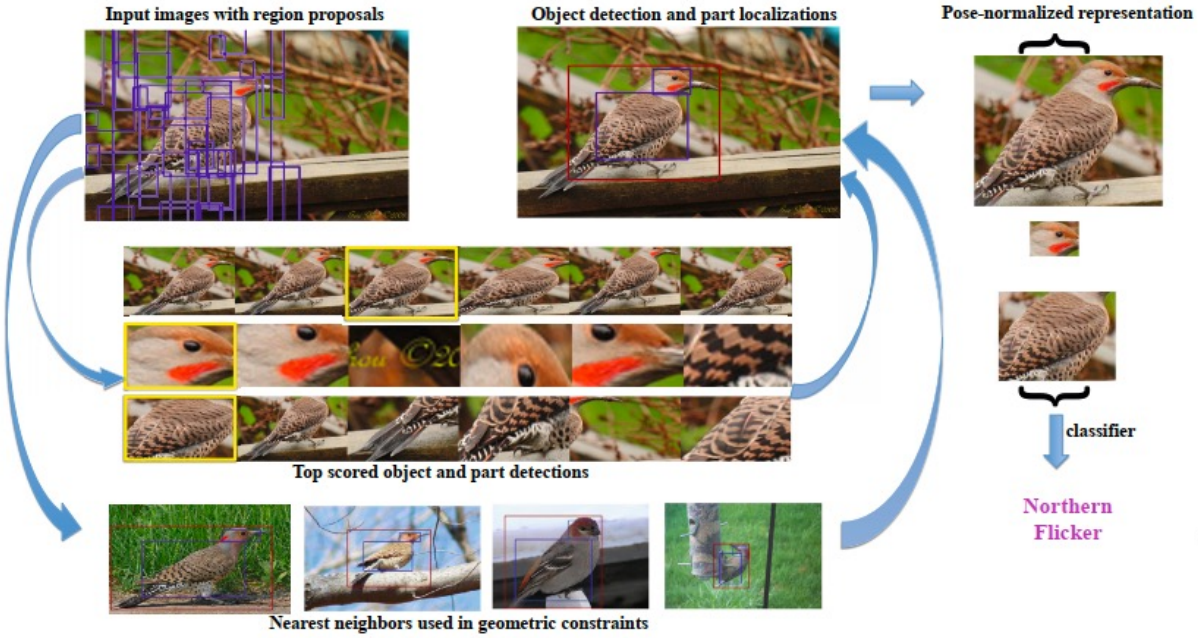
Method	Performance on CUB 200
《Combining Randomization …》	19.20%
《A Codebook-Free …》	44.73%

- FGVC概述
- 特征工程时期的识别算法
- 深度学习技术的崛起
- 步入弱监督时代

深度学习技术的崛起

随着CNN模型的诞生以及其在目标检测任务中的成功应用，基于框的FGVC方法也逐渐流行起来，即检测包含局部的框，之后提取特征进行分类。

《Part-based RCNN》



1. 自下而上的框提取方法；
2. 训练M+1个检测器，筛选合适的对象框和局部框；
3. 通过几何约束，进一步筛选。

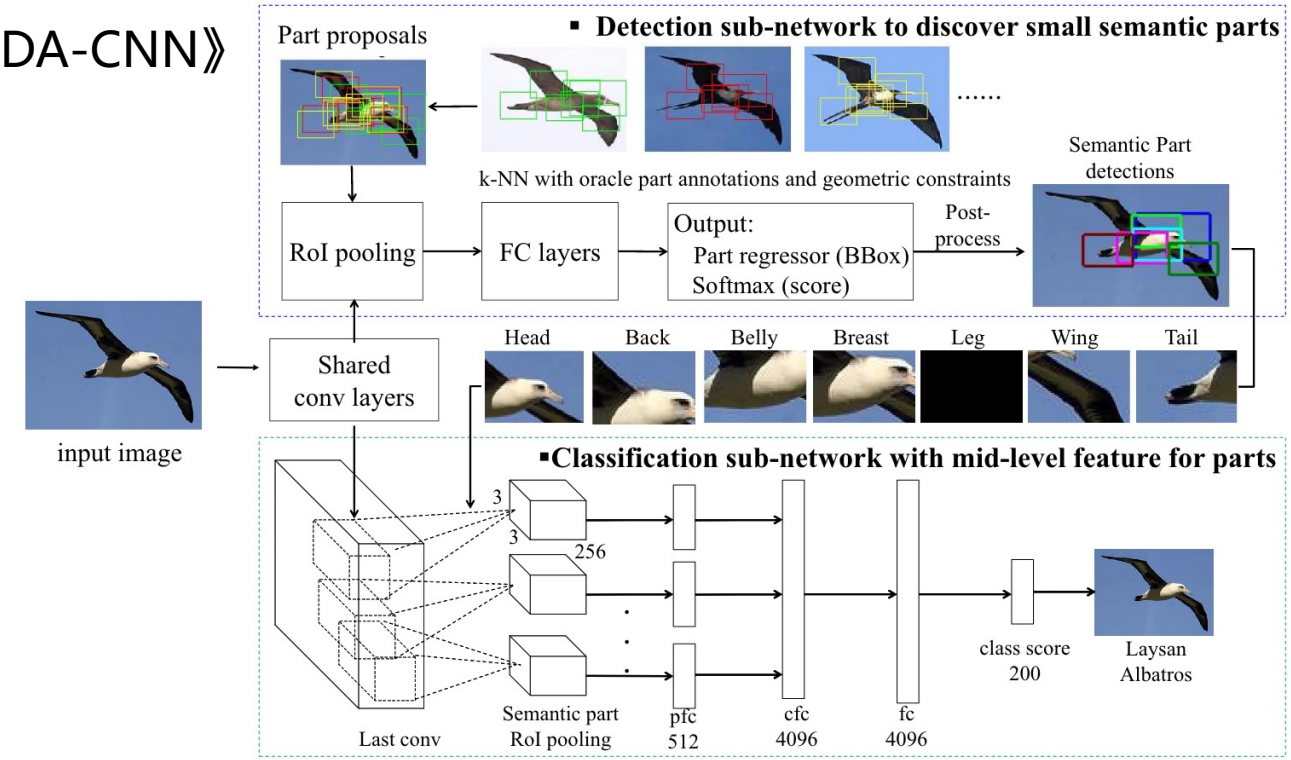
$$X^* = \arg\max_X \Delta(X) \prod_{i=0}^{n} d_i(x_i)$$

$$\Delta_{\text{box}}(X) = \prod_{i=1}^{n} c_{x_0}(x_i)$$

$$\Delta_{\text{geometric}}(X) = \Delta_{\text{box}}(X) \left(\prod_{i=1}^{n} \delta_i(x_i) \right)^{\alpha}$$

深度学习技术的崛起

《SPDA-CNN》



1. 自上而下的框提取方法；
2. 利用KNN得到K个近似图像，将其中预定义的框作为测试图像的候选框（包含结构信息）。

缺点：1. 属于强监督的范畴，需要局部注释信息。
2. 框过大、框中包含杂乱的背景。

- FGVC概述
- 特征工程时期的识别算法
- 深度学习技术的崛起
- 步入弱监督时代

步入弱监督时代

有监督？无监督？半监督？弱监督？

FGVC中的弱监督：在深度学习框架下，不依赖于局部标注而仅仅使用图像级的类别标签进行训练。

这样的研究在2015年就已经开展，据我所知，最早的一篇论文为《The Treasure beneath Convolutional Layers：Cross-convolutional-layer Pooling for Image Classification》

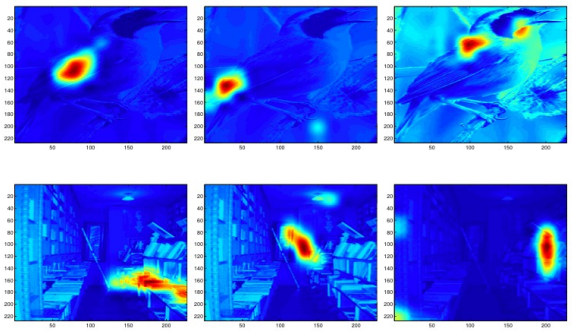


Figure 4: Visualizing of some feature maps extracted from the 5th layer of a DCNN.

发现feature map的通道蕴含局部响应信息，于是feature map的每一个位置和其在第k个通道上响应位置的值相乘再求和，得到第k个局部特征。

$$\mathbf{P}^t = [\mathbf{P}_1^{t\top}, \mathbf{P}_2^{t\top}, \cdots, \mathbf{P}_k^{t\top}, \cdots, \mathbf{P}_{D_{t+1}}^{t\top}]^\top$$

$$\text{where,} \quad \mathbf{P}_k^t = \sum_{i=1}^{N_t} \mathbf{x}_i^t a_{i,k}^{t+1}, \qquad (2)$$

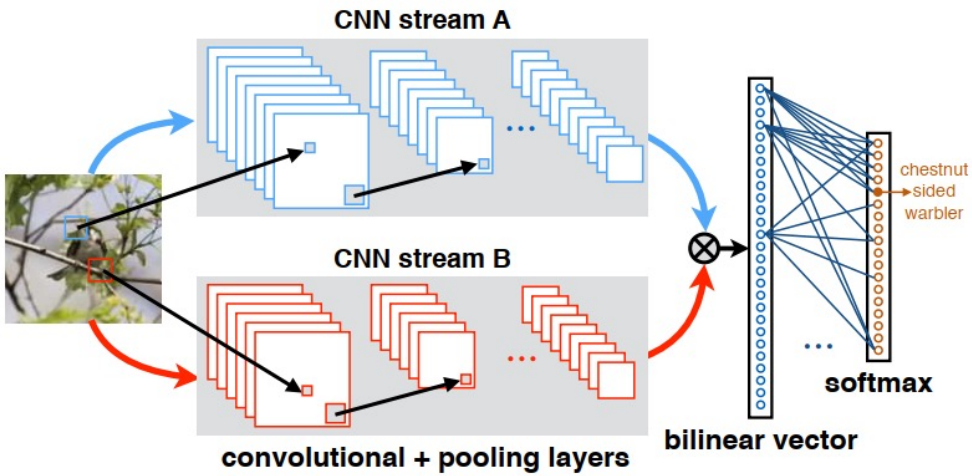
基于高阶特征的方法

我们将上述公式改写，可以得到如下式子（对于单张图像）：

$$X = [x_1, ..., x_N], N = H \times W, x \in R^C$$

$$XX^T \in R^{C \times C}$$

那个时期，称上述公式为**特征交互**，B-CNN(ICCV,2015)也由此进行了推广。



“双流”蕴含仿生思想，一支用于局部定位，一支用于特征获取。

$$XY^T \in R^{C_1 \times C_2}$$

基于高阶特征的方法

随着研究的深入，特征交互被赋予了更高端的称呼“高阶表征”。在Compact B-CNN(CVPR,2016)中证明了双线性表征和二阶核向量机是互为表里的关系。

$$B(\mathcal{X}) = \sum_{s \in \mathcal{S}} x_s x_s^T \tag{1}$$

$$\begin{aligned} \langle B(\mathcal{X}), B(\mathcal{Y}) \rangle &= \langle \sum_{s \in \mathcal{S}} x_s x_s^T, \sum_{u \in \mathcal{U}} y_u y_u^T \rangle \\ &= \sum_{s \in \mathcal{S}} \sum_{u \in \mathcal{U}} \langle x_s x_s^T, y_u y_u^T \rangle \\ &= \sum_{s \in \mathcal{S}} \sum_{u \in \mathcal{U}} \langle x_s, y_u \rangle^2 \end{aligned} \tag{2}$$

同源双线性池化得到的特征X和Y做内积相当于F((x·y))。换句话说，通过双线性池化，能够得到图像在目标空间的表征。

基于高阶特征的方法

从今天的视角来看Compact B-CNN的意义有两点：
1. 揭示了特征交互的本质是模型得到了图像在目标空间的表示；
2. 针对这样一种表示维数过高的问题，提出了不损失信息且紧凑的表示。

至此，之后的研究高阶特征的方法大都围绕两点展开：
1. 探索不同核函数对应的目标空间的表征；
2. 将这一种表征进行无损的压缩。

多项式核：

$$K(x, z) = (\gamma x \cdot z + \zeta)^p, \gamma > 0$$

《Higher-order Integration of Hierarchical Convolutional Activations for Fine-grained Visual Categorization》

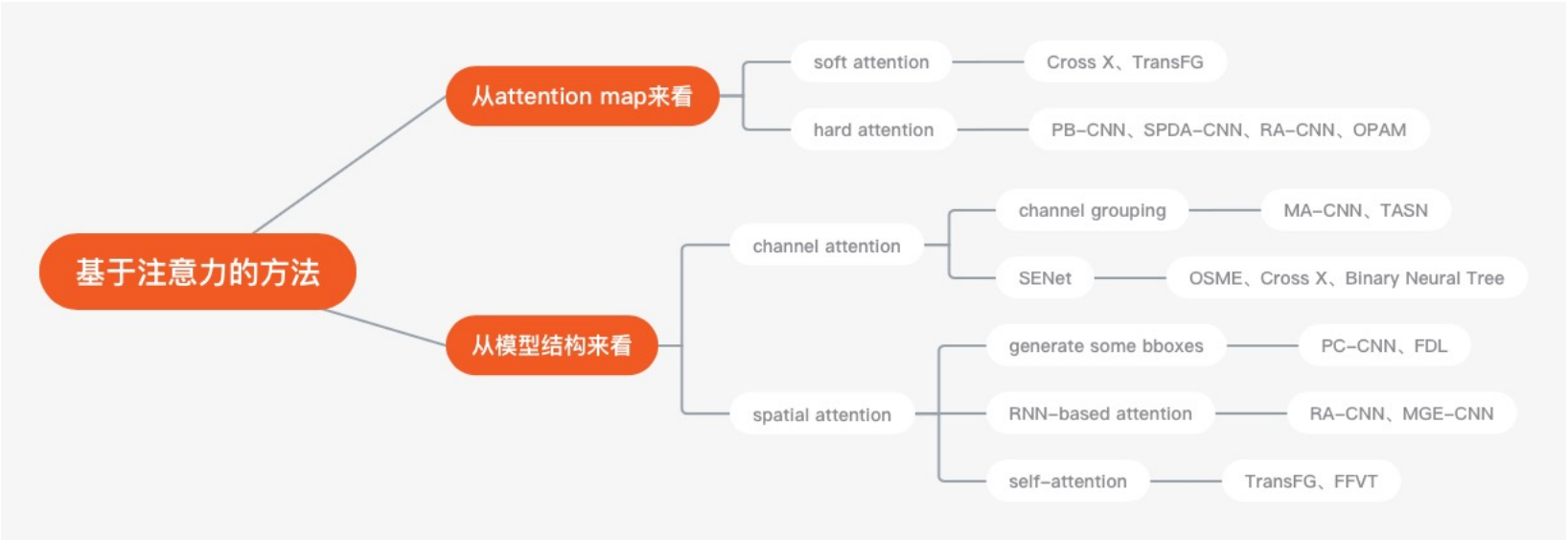
泰勒级数核/高斯核：

$$K(x, z) = \exp(-\frac{\|x - z\|^2}{2\sigma^2})$$

《Kernel Pooling for Convolutional Neural Networks》

基于注意力的方法

有的研究继承了feature map中pattern的应用研究，推动了基于注意力的方法。但是目前对注意力的方法的定义更为宽泛一些，即**引导模型关注“有意义”的局部。**

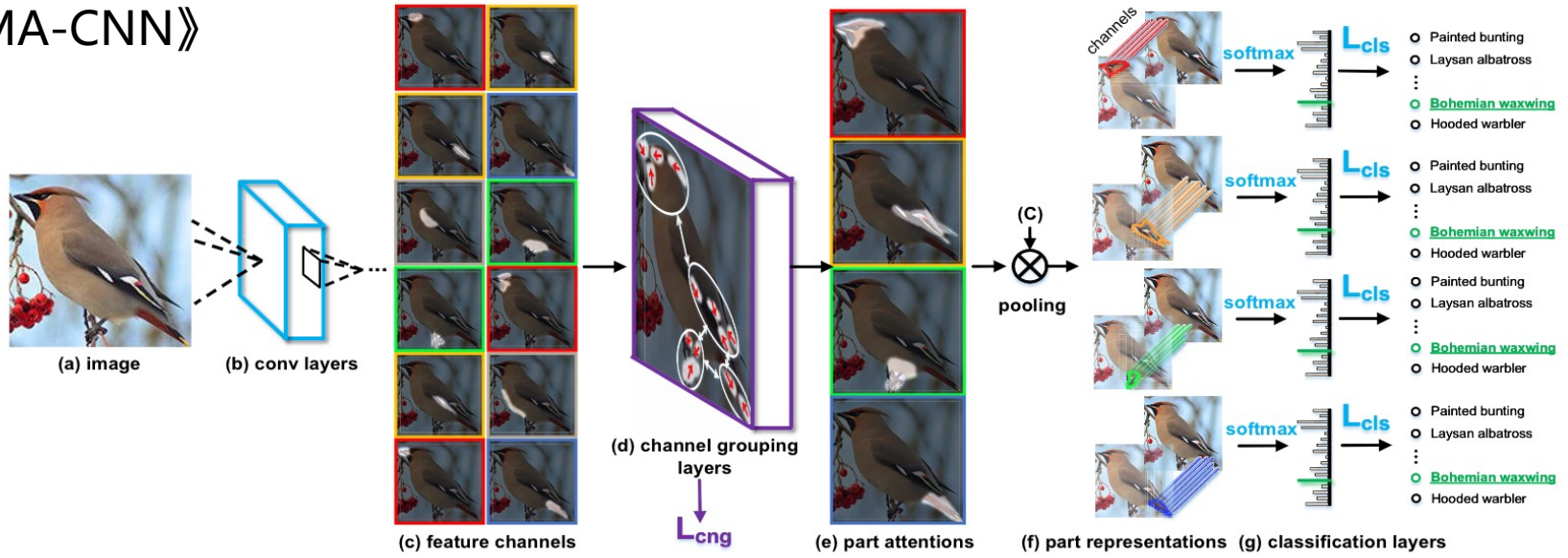


Soft attention：以一种权重的方式调整feature map（增强与抑制）。
Hard attention：从feature map上给出明确的关注区域（裁剪）。

本文主要从第二个角度介绍基于注意力的方法。

基于注意力的方法（channel-grouping）

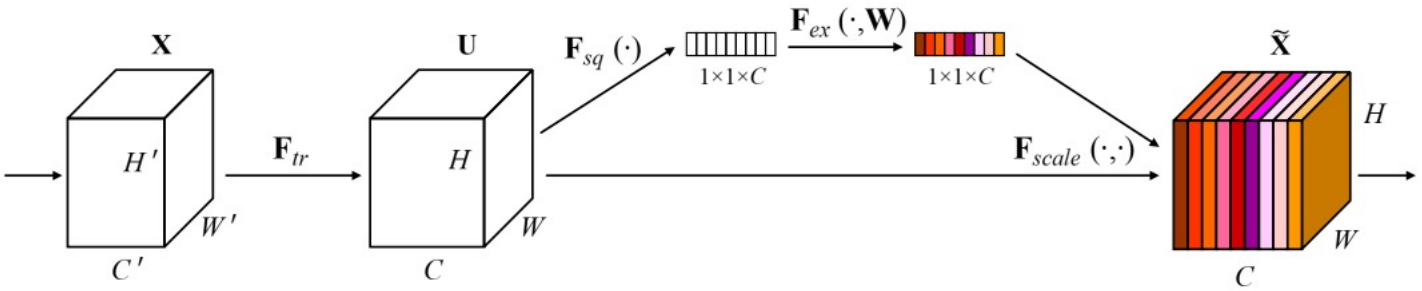
《MA-CNN》



1. 模拟聚类，通过全连接层训练得到每个通道的重要程度，之后与feature map进行weighted sum pooling得到attention map。
2. 预先定义N个FC层，用以关注不同的局部。

基于注意力的方法（channel-SENet）

SENet通过学习的方式来获取每个通道重要程度，然后据此提升有用的特征并抑制相对无用的特征。



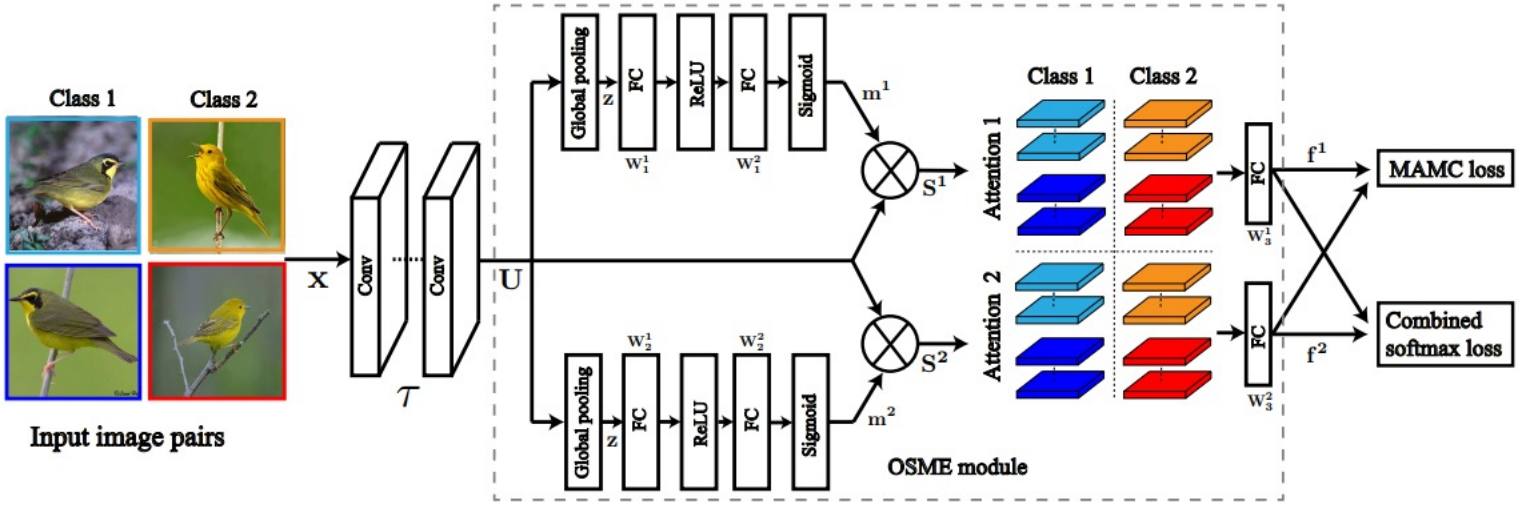
1）Squeeze（压缩）。顺着空间维度来进行特征压缩，将每个二维的特征通道变成一个实数，这个实数某种程度上具有全局的感受野，

2）Excitation（激发）。通过参数来为每个特征通道生成权重，其中参数被学习用来显式地建模特征通道间的相关性。

3）Reweight（缩放）。将Excitation的输出的权重每个特征通道的重要性，然后通过乘法逐通道加权到先前的特征上，完成在通道维度上的对原始特征的重标定。

基于注意力的方法（channel-SENet）

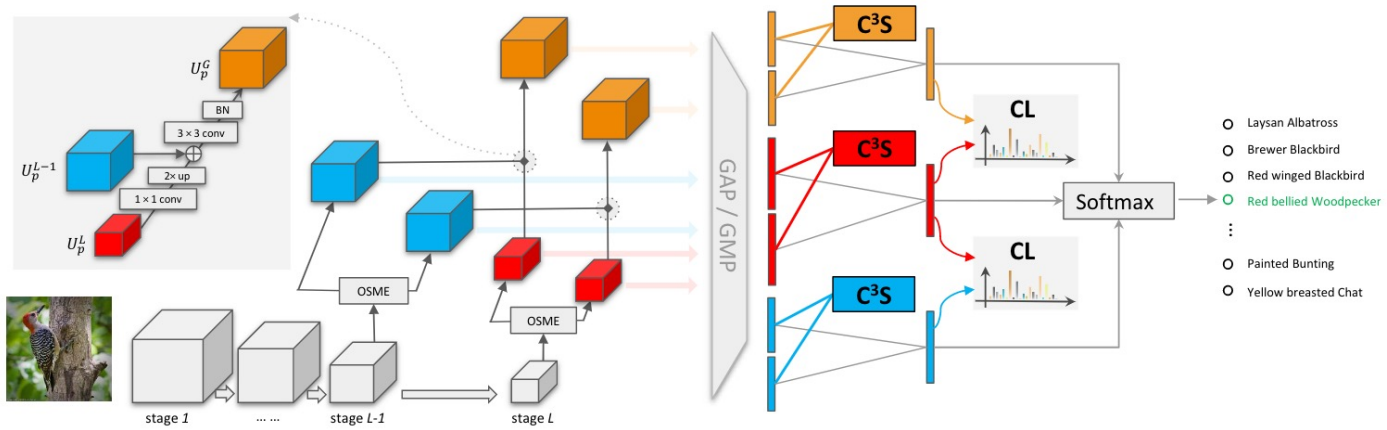
《Multi-Attention Multi-Class Constraint for Fine-grained Image Recognition》



1. 希望不同的SENet Module关注不同的局部特征；
2. 采用度量学习对两个分支的学习过程进行约束：同类同分支的特征之间距离近、不同类不同分支的特征之间距离远，同类不同分支/不同类同分支介于两者之间。

基于注意力的方法（channel-SENet）

《Cross-X Learning for Fine-Grained Visual Categorization》



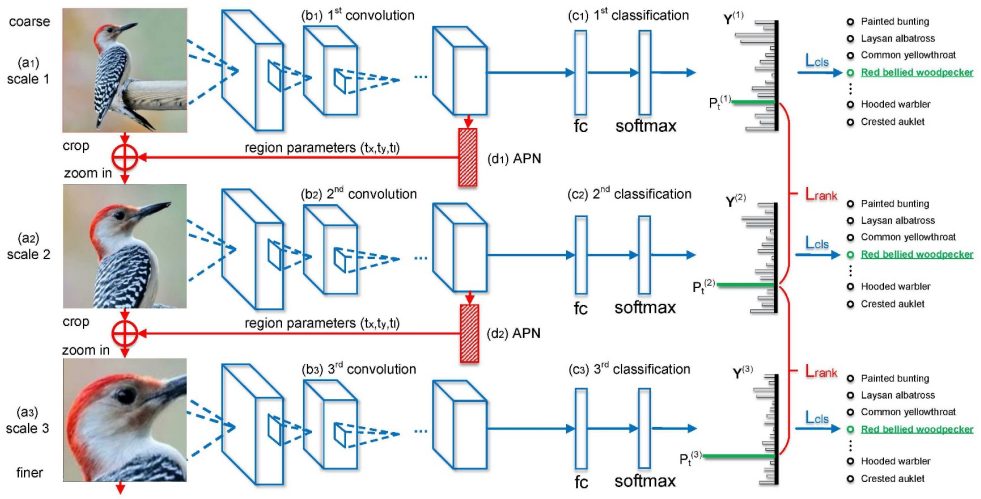
对前文结构的革新：优化了特征，引入了FPN结构，得到了不同层次的信息；优化了两个激励的学习方式，采用更为简单的特征矩阵进行约束。

总结：该类方法属于SENet在FGVC中的一次发展。其亮点在于结构清晰因此训练过程较为简单（相对于MA-CNN这样需要交替学习的模型）。瓶颈在于受制于SENet的性能。

基于注意力的方法（Spatial-RNN_based）

Spatial attention，顾名思义是在空间维度上得到需要关注的区域。基于RNN的注意力的思路是每一个时间步在给出预测的同时将得到的先验知识传递给下一个时间步。

《Look Closer to See Better ：Recurrent Attention Convolutional Neural Network for Fine-grained Image Recognition》



1. 本文发现：局部检测和细粒度特征学习是相关的，能够彼此促进。
2. 本文没有直接crop（不便于反向传播），而是采用mask的形式抑制无关区域。

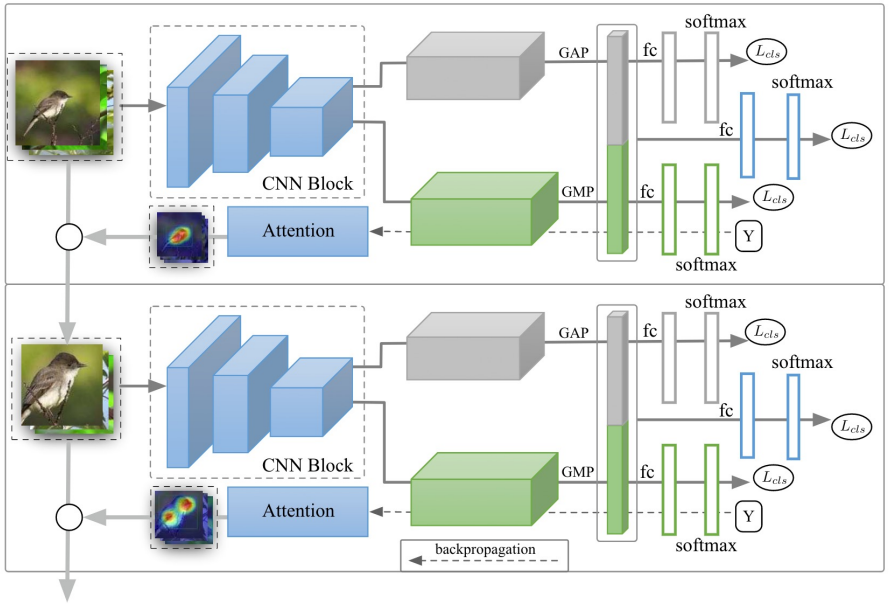
$$\mathbf{M}(\cdot) = [h(x - t_{x(tl)}) - h(x - t_{x(br)})] \cdot [h(y - t_{y(tl)}) - h(y - t_{y(br)})], \quad (5)$$

and $h(\cdot)$ is a logistic function with index k:

$$h(x) = 1/\{1 + \exp^{-kx}\}. \quad (6)$$

基于注意力的方法（Spatial-RNN_based）

《Learning a Mixture of Granularity-Specific Experts for Fine-Grained Categorization》



1. 本文的初衷是对ME方法进行优化，引入了先验知识，形成了RNN的结构。
2. 特征增强策略：深层卷积特征+GAP和浅层卷积特征+GMP。
3. 通过Grad-CAM得到热力图（通过梯度信息理解模型的关注区域），选择覆盖最大的区域进行crop。
4. 为了得到不同类型的专家，通过KL散度对不同专家的预测分布进行约束。

总结：优势在于模仿人的感知过程，由粗到细。缺陷在于每一次预测受到先验知识影响，存在累积误差。

基于注意力的方法（Spatial-self_attention）

ViT在视觉领域的应用十分成功，因此引入到FGVC任务中。其学习过程实现了对不同局部之间存在关系的分析。目前存在的问题主要有：

1. 如何产生有效的patch输入（patch能够包含完整的局部、patch中的冗余信息少…）
2. 如何增强class token中的细粒度信息

《Quantifying Attention Flow in Transformers》

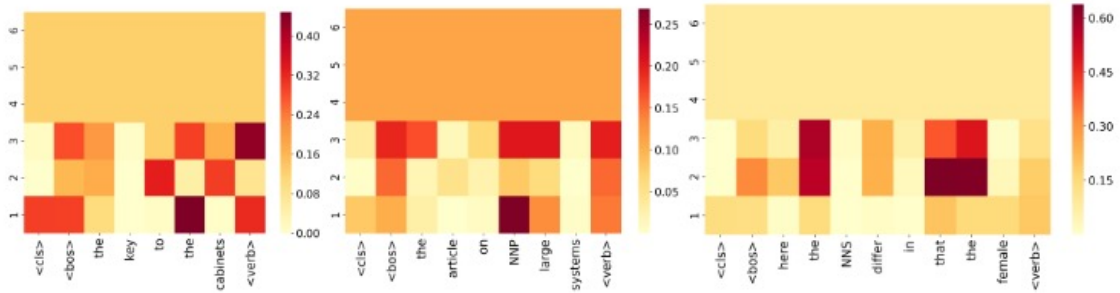
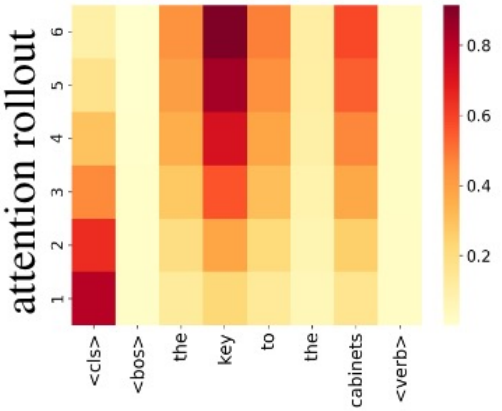
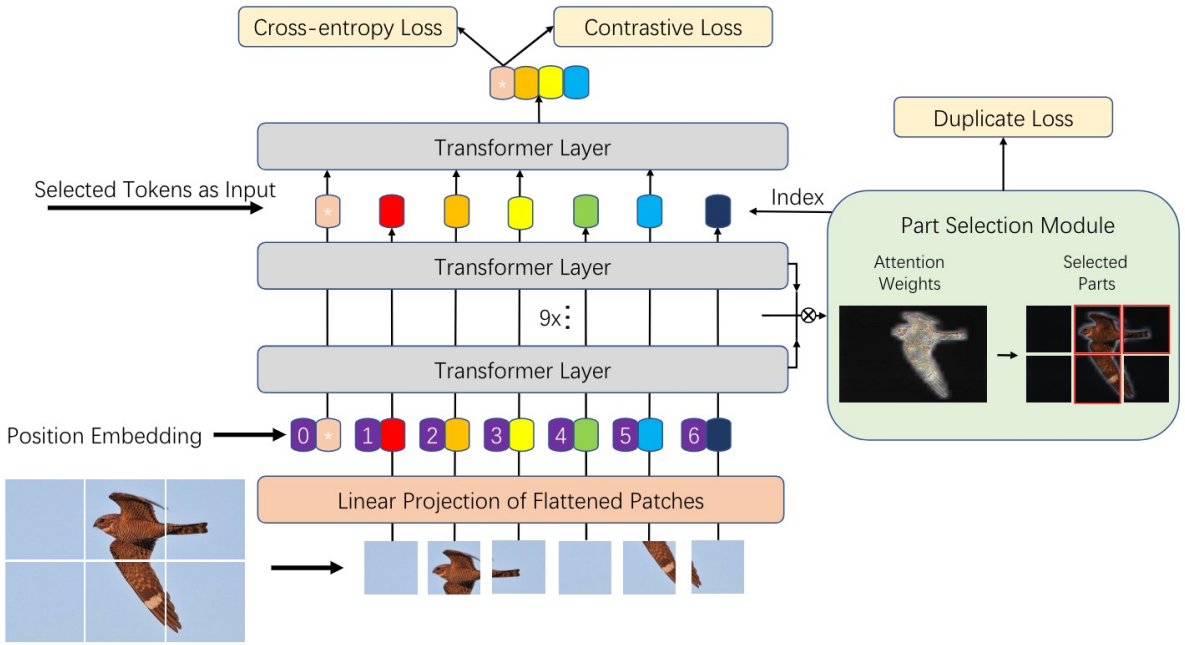


Figure 2: Raw Attention maps for the CLS token at different layers.



基于注意力的方法（Spatial-self_attention）

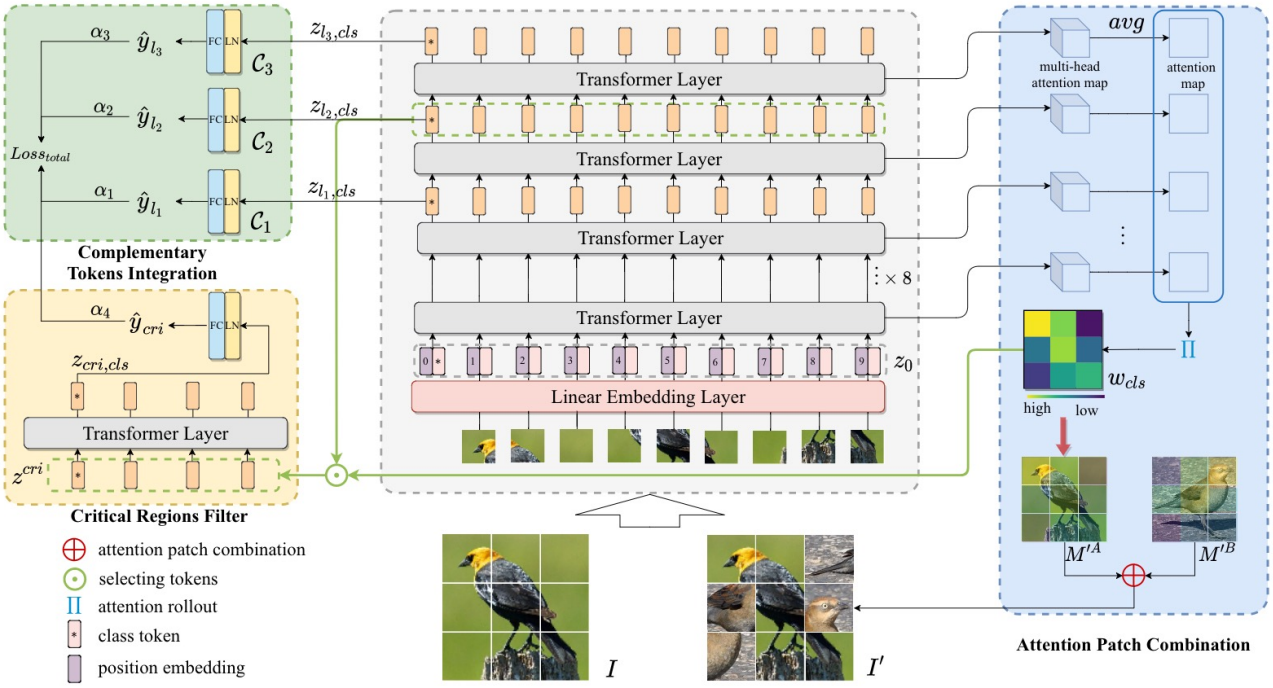


《TransFG》

1. 对图像进行密集采样，尽可能得到包含完整局部的patch；
2. 通过attention rollout得到相对重要的token，与class token一同送入最后一层。

基于注意力的方法（Spatial-self_attention）

《ViT-FOD：A Vision Transformer based Fine-grained Object Discriminator Abstract》



1. 发现不同层的class token存在互补信息，因此每一层都抽取进行分类；
2. 通过attention map得到class token的attention vector，一方面据此**提取特定的patch与其他图像进行组合**以此达成数据增广的目的，另一方面据此**选择特定的token进行分类。**