

« Object Counting »

Count TR

(1) 掩码图像重建进行自监督预训练

(2) 下游计数任务微调

(3) Mosaic 数据增强合成训练图像

(4) 视觉编码器 } 目标样本
查询图像 → 特征表 → 视觉解码器

Query Image → Image Encoder → Query

重VIT-ENC

100

$\rightarrow \text{FIM}(\cdot) \rightarrow \text{Decoder} \rightarrow \text{Density Map}$

exemplars → Image Encoder → key Value
(E_{key}) (E_{val})

(few) 'shot' |
ΦCNN-ENC

$\Phi_{CNN \cdot ENC}$

Feature Interaction Module

D LOCA

Low-shot object counting network with iterative prototype Adaptation

LOCA ⌈ Feature extraction = 提取编码后图像特征

prototype extraction = bboxes → 固体原型

prototype matching = 目标原型 & 图像特征

density regression = 预测密度图

〈深度相关运算

相以圖

Input → encoder → f^E → q_v^o → \oplus → \tilde{r}_v → max → \tilde{r} → decoder → destiny map

$g_i^o = s \times s \times d$ n object prototypes

n similarity maps

Re n similarity maps

R Response map

131

□ SPDCN

Scale-Prior Deformable Convolution Network

《 Scale-Prior Deformable Convolution for Exemplar-Guided
class-Agnostic Counting》

(1) SPDCN scale

R度信息集成到计数网络主干中

(2) 主干网络：尺度先验的可变形卷积

(3) 给定样本尺度 学习尺度嵌入

(4) 嵌入向量调整 可变形卷积偏移量

(5) 尺度敏感的广义损失 根据目标尺度调整损失函数

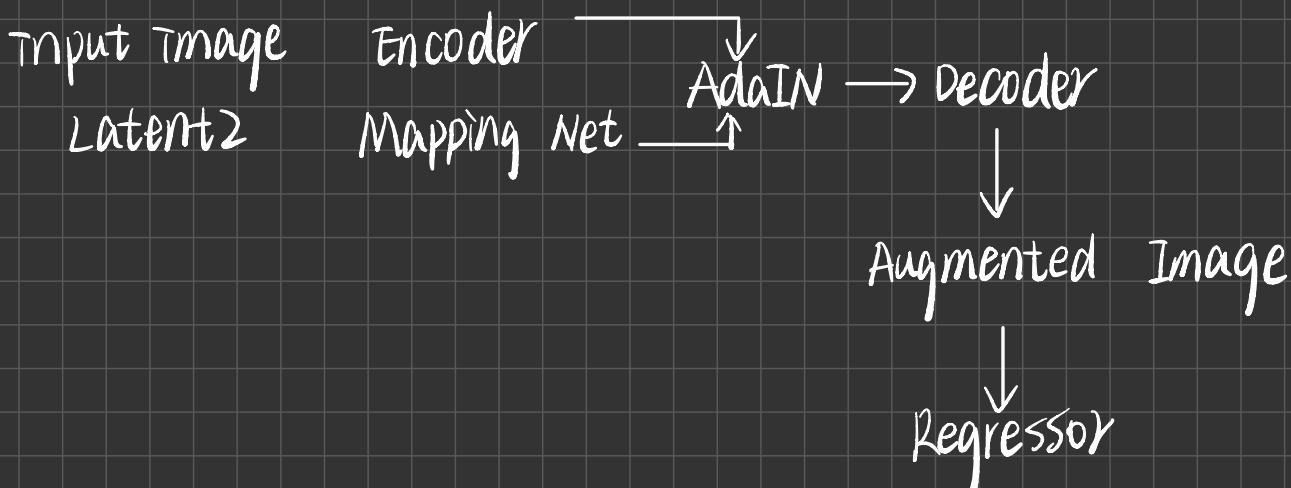
(4)

□ VCN

Vicinal Counting Network

VCN |
生成器 学习计数 & 增加现有训练数据
计数回归器

| 生成器：图像 及 随机噪声向量 车辆输入 → 输入图像的增强版本
| 计数回归器：原始图像 & 增强图像中的物体进行计数



15) SAM Counting

Can SAM Count Anything? An Empirical Study on SAM Counting

(1) SAM (ViT-H) 对图像特征编码 Image]

(2) 目标边界框 提示 生成 参考目标的分割掩码]

与图像特征相乘取平均产生 参考目标的特征向量

(3) 点网格 对内容进行分割 输出掩码

与图像特征相乘取平均生成掩码的特征向量

$\Rightarrow \text{final} = \frac{\text{余弦相似度}}{\text{参考相似度}} \begin{cases} \text{预测掩码的特征向量} \\ \text{参考目标的} \end{cases}$

16) CACViT

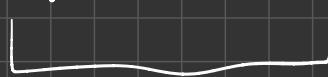
Class-Agnostic Counting Vision Transformer

<<Vision Transformer off the shelf: A Surprising Baseline for Few-shot

Class-Agnostic Counting>>

CACViT } 视觉Transformer 自注意力机制 } 特征提取
| } VIT | } 特征正则化
| } 自注意力: query Image & exemplar 特征提取
| } cross Attention: ~ 正则化

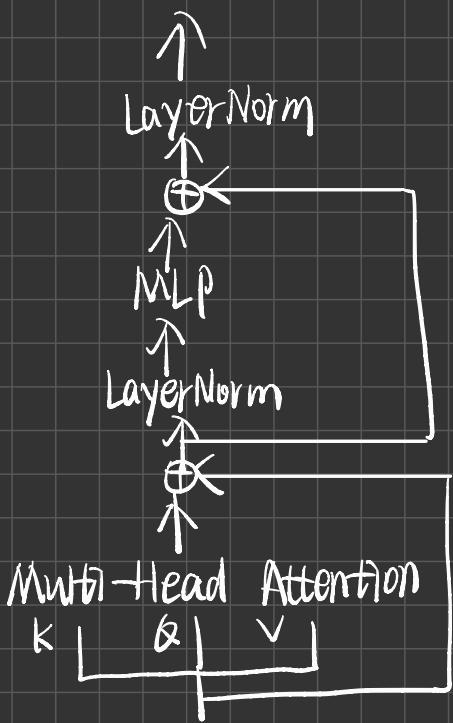
(1) Query Image X & Exemplar Z



cat

(2) ViT-block X L

Extract and Match



(7) DAVE

Detect and verify Paradigm

« A Detect and Verify Paradigm for Low-shot Counting »



Backbone

Detection stage

Prototype extraction

R Decoder

G NMS C \cup FFM BC

Feature pooling

cosine similarity

affinity matrix

verification stage

clustering

B^P Output detections Detection mask

Output density DAVE outputs

(8) SSD

Spatial Similarity Distribution

« Learning Spatial Similarity Distribution for Few-shot Object Counting »

SSD

特征提取

特征双增强

相似度金字塔计算

相似度学习：中心枢轴 4D 卷积 提取特征

归一化解码器

Feature Extraction

(1) 冻结的 ResNet-50 { query image
exemplar 金字符特征

特征交叉增强

增加特徴相似度

	Image	Query Features	Pyramid
FCE			
SLM			
F^q_i F^s_i	4D convolution		

二、无参考计数 Reference-less Counting

直接从图像中提取显著的目标区域进行计数

无需标记及模板

底棲魚：無法指定計數目標 对所有目标计数

U) Lower Count (LC)

Object Counting and Instance Segmentation with Image-level Supervision

LC 框架 | 低数量级图像监督

仅需本机汎用子4个

图像分类分支：是否存在目标类别

Pseudo GT Generation

| 密度分歧 = 双测数量

Image // ResNet50 // Conv // BN- ReLU- Conv //

Object category map // Peak Map

[2] RLC Reduced Lower-Count

汉进 <<Towards Partial Supervision for Generic Object Counting in Natural Scenes>>

① 泛化

② 权重调制层 (从训练好的卷积核迁移未标注类别)

③ 密度分支引入类别无关的分支 估计一幅图像中所有目标数量



TODO

Movie DSAA CaOC RepRPN-Counter RCC

1 Movie

Movie: Revisiting Modulated Convolutions for Visual Counting and Beyond

视觉计数VQA

{ 图像：在预测图像中与预测相关的目标数量
预测 }

ResNet 提取图像 Feature —————> Movie

Movie 模块 = 4 个卷积核并行处理
↓ ↑ 接收预测 (密外输入)
输出相同尺寸特征图

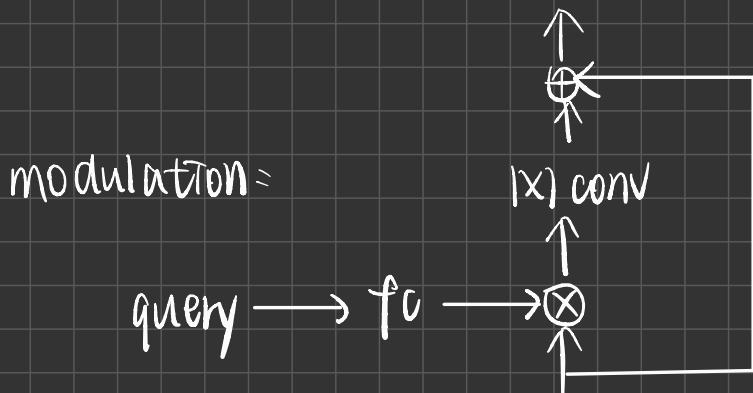
平均池化 & 两层分类器 来预测 Answer

Image query

P1

Image \rightarrow conv + pool \rightarrow res₁ ... res-4 \rightarrow our model \rightarrow pool & MLP \rightarrow answer

module = bottleneck1 bottleneck2 bottlenecks bottleneck4



□ DSAA 扩张尺度感知注意力

Dilated-Scale-Aware Attention

<< Dilated-Scale-Aware Attention ConvNet For Multi-Class Object Counting >>

给定篇标注，预测所有类别的密度图

DSAA 首先 VGG16 \longrightarrow 扩张尺度感知模块 DSAM

DSAM 不同的卷积扩张率提取不同尺度的特征

CAM | 类别注意力模块

| 减少不同类别密度图中的负关联

□ CaOC

Class-aware Object Counting 类别感知目标计数

基于检测 X

□ Rep RPN Counter

△ Exemplar Free class Agnostic Counting

△ ResNet-50 提取输入图像特征

△ Rep RPN 按照每个 anchor 位置处的 Proposal 边界框、目标得分。

P2

重复得令

△ 重复得令 proposal 中目标在图像中出现的次数

示例样本：重复次数最高的 proposal

□ RCC

Reference-less Class-agnostic Counting

无参考无类别计数

Learning to Count Anything: Reference-less Class-Agnostic Counting
with Weak Supervision

自监督 知识蒸馏 ViT-small { 教师网络 q_t
损失函数选用 $\frac{|C - \hat{C}|}{C}$ 学生网络 q_s

线性投影 = 归一化目标计数值

$$\text{损失函数选用 } \frac{|C - \hat{C}|}{C}$$

Query Image \rightarrow ViT $\xrightarrow{\downarrow}$ Linear Projection \rightarrow 67.91
backbone Features \rightarrow Unsampled Features

GCNet ✓ DSC ✓ ABC123 ✓ OmniCount Count CLIP

□ GCNet

GCNet = Probing Self-Similarity Learning for Generalized Counting Network
探测 自相关学习 混合计数网络

因有重复模式的自相关性 伪造目标样本

{ 伪目标样本 \Rightarrow 生成高保真自相关图
原始图像低级特征

作为随后计数回归的输入

计数级监督训练 端到端训练

pseudo 伪造

P3

□ 2SC

Zero-shot Counting

IN: 给定物体类别

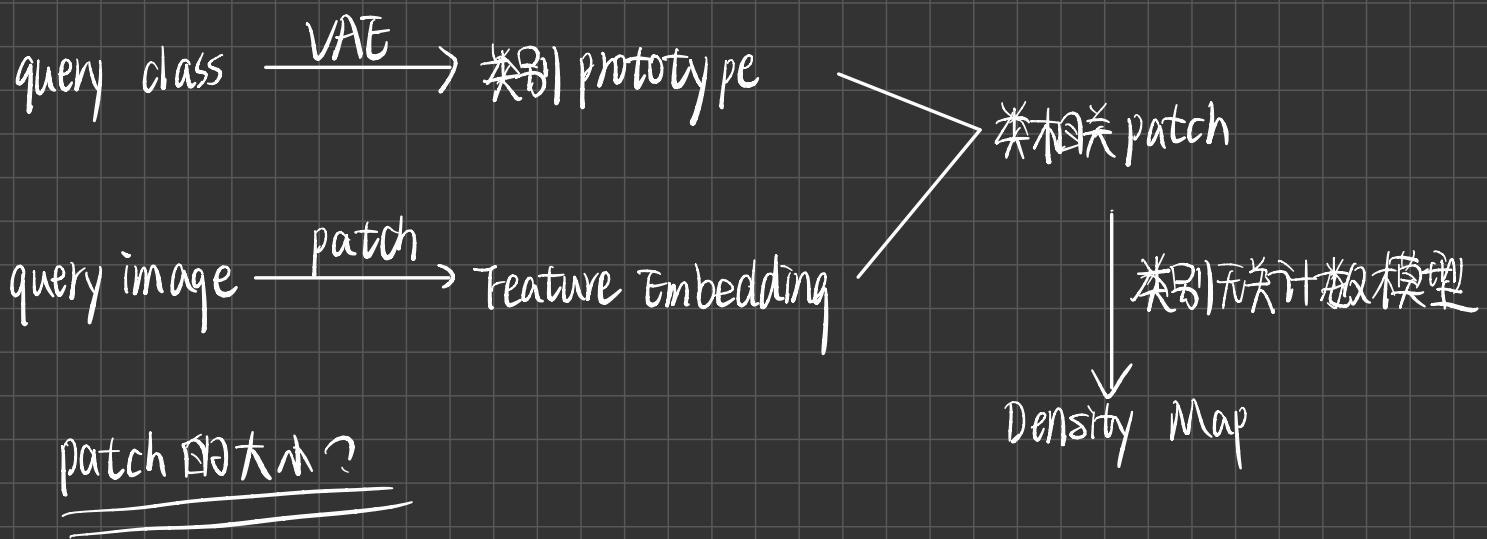
(1) for query class, 训练集的VAE 特征空间生成类别原型

(2) for query image, 抽取图像patch, 为每个 patch 提取特征嵌入

(3) 特征嵌入 & 类别原型最近 patch \Rightarrow 类相关 patch

(4) 类相关 patch $\xrightarrow{\text{类别无关计数模型}}$ density map

(5) 误差预测器 误差最小的 patch \rightarrow exemplar



for me: automatically generate exemplar

□ ABC123

A Blind Count 商用

ABC Easy as 123: A Blind Counter for Exemplar-Free Multi-class
Class-agnostic Counting

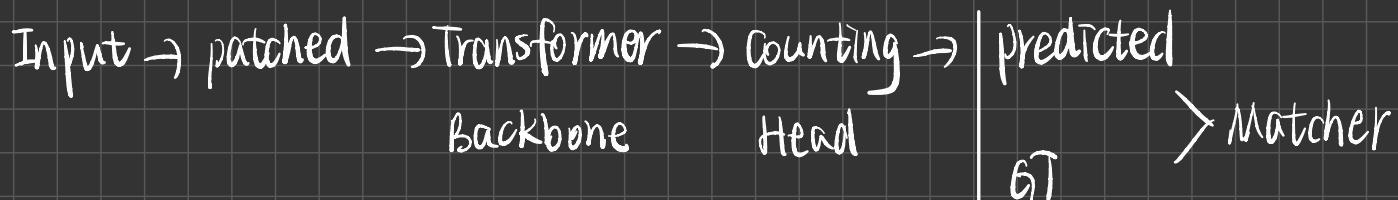
ABC123 { 预测的类别数
类别密度图

124

VIT-Small 提取图像特征

↓ m 个卷积采样头
因此 m 个类别密度图

m 个预测密度图
 m 个真实密度图 \rightarrow 最优二分匹配



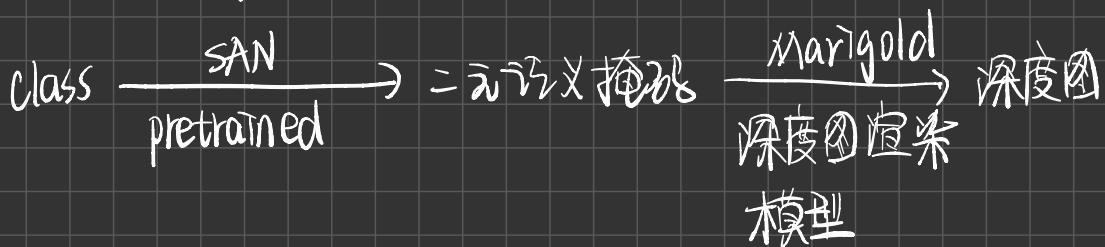
OmniCount

多标签目标计数

语义

OmniCount = Multi-label Object Counting with Semantic-Geometric Priors

无需训练



深度图细化语义掩码 重复计数问题

语义 mask $\xrightarrow{\text{基于 K 近邻 N 例校正}}$ reference points

Input \rightarrow semantic & structural encoding \rightarrow structure guided ...
refinement \rightarrow class-specific counting \rightarrow output

part 3 Text guided counting

Text prompt \rightarrow 计数类别

VLM \longrightarrow counting

VLM Vision Language Model

* For me = clip, pseudo, distill \Rightarrow paper b
伪迹 contrast

p5

□ Count - CLIP

Teaching CLIP to count to Ten



{ Obtaining Counting Training set
Teaching CLIP to Count

□ CLIP - Count

« CLIP Count: Towards Text-Guided Zero-Shot Object Counting

块文本 对比损失

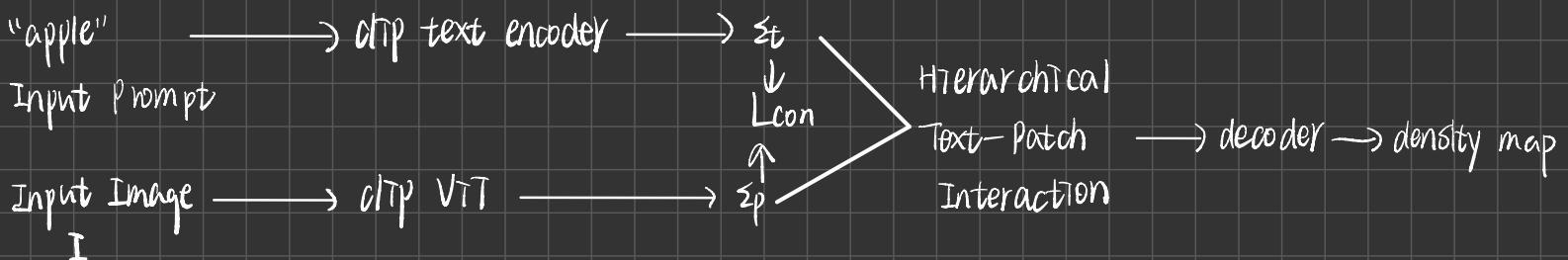
CLIP { 视觉编码器特征块 \rightarrow 对齐 align
且类别引文

层次化文本 - 块交互模块 \longrightarrow 不同分辨率 resolution

feature map

\downarrow decoder

目标密度图



3. Text - Guided Counting

VLM (Vision Language Model)

(1) Count CLIP

tR - 文本对比损失

- ① CLIP的视觉编码器与目标类别文本对齐
- ② 设计层归化文本 - 块交互模块 生成不同分辨率的特征图
- ③ 特征图解码为目标的密度图

(2) CountTX

Open-world Text-specified Object Counting

组成 ① 图像编码器 > 预训练的CLIP

② 文本编码器

③ 特征交互模块：2个Transformer解码器层

④ 解码器

图像特征计算查询向量

文本特征计算键向量及值向量

解码器 = 特征解码为单通道密度图

(3) VLCounter

Visual-language Counter

<< VLCounter = Text-aware Visual Representation for Zero-shot

Object Counting >>

VLBase { CLIP 编码器
计数解码器 }

CLIP的输入空间 { 文本 token 嵌入 } > 隐式关联 → 对目标物体定位

VLBase → VLCounter

P

①引入语义条件提示微调

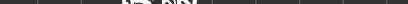
② 可学习依赖变换及语义感知跳跃连接
增强预测相似图 & 密度图中的语义化信息

4) Clip Counting

(NewIPS workshop 2023) Zero-shot Improvement of Object Counting with CLIP.

☆ Clip 文本嵌入空间 提高计数精度

① zero-shot 双本嵌入编辑方法

② 容易计数的图形中  **提取** 计数知识

(表示为嵌入空间中的线性子向)

③ 向目标嵌入中增加特定计数向量

④ 将该知识应用于目标对象

WSI Express Count

Enhancing Zero-shot Counting via Language-guided Exemplar Learning

ExpressCount } 面向语言的根极感知模块
 } 理样本目标计数模块

语言编解码器 (训练集 Clip)
视觉编解码器
基于 Transformer 的模块集成模块

通过学习文本表达式与目标示例的边界框之间的映射
来提取目标样本进行后续计数