来自Transformer的降维打击：ReID各项任务全面领先，阿里&浙大提出TransReID

原创 CV开发者都爱看的 极市平台 2021-02-09 22:00:00 手机阅读

收录于话题
#Transformer

作者丨Happy
审稿丨邓富城
编辑丨极市平台

极市导读

Transformer在ReID领域的第一次全面探索！为更好的利用ReID的数据特性与Transformer的信息嵌入特征，本文提出了两种模块改进SIE与JPM，将ReID的提升到了新的高度。 >>加入极市CV技术交流群，走在计算机视觉的最前沿

TransReID: Transformer-based Object Re-Identification

Shuting He[1,2]*, Hao Luo[1], Pichao Wang[1], Fan Wang[1], Hao Li[1], Wei Jiang[2]
[1]Alibaba Group, [2]Zhejiang University
{shuting_he,jiangwei_zju}@zju.edu.cn {michuan.lh,pichao.wang,fan.w,lihao.lh}@alibaba-inc.com

本文是阿里巴巴与浙江大学在Transformer+ReID方面的一次突破性的探索，在多个ReID基准数据集上取得了超过CNN的性能。为更好的利用ReID的数据特性与Transformer的信息嵌入特征，本文提出了两种模块改进SIE与JPM，将ReID的提升到了新的高度。比如**将MSMT17的指标从60.8%mAP提升到了69.4%mAP，将DukeMTMC-ReID的指标从78.6%mAP提升到82.6%mAP，将Occluded-DuKe的指标从43.8%mAP提升到59.4%mAP，将VehicleID的指标从84.7%mAP提升到85.2%mAP**。

Abstract

本文对 Vision Transformer 在目标重识别任务上的应用进行了探索。通过几种域适应，提出了一种强基线 ViT-BoT 作为骨干网络，它在多个ReID基准数据集上取得了CNN相当甚至更好的效果。考虑到ReID数据的特性，本文设计了两个这样两个模块：

- 将camera或viewpoint信息作为嵌入到到Transformer架构中，因此ViT可以消除不同camera或viewpoint导致的偏差；
- 设计一个与全局分之并行的Jigsaw分支以利用双分支学习框架促进模型的训练。该分支被设计用于学习更鲁棒的特征表达，同时对图像块进行置换以辅助transformer的训练。

基于上述两个模块，本文提出了首个用于ReID任务的Transformer架构：TransReID。在多个ReID基准数据集(包含行人、车辆等)，所提方法取得SOTA性能。下图给出了MSMT17数据集上不同方案的性能对比，可以看到TransReID取得显著的性能提升。

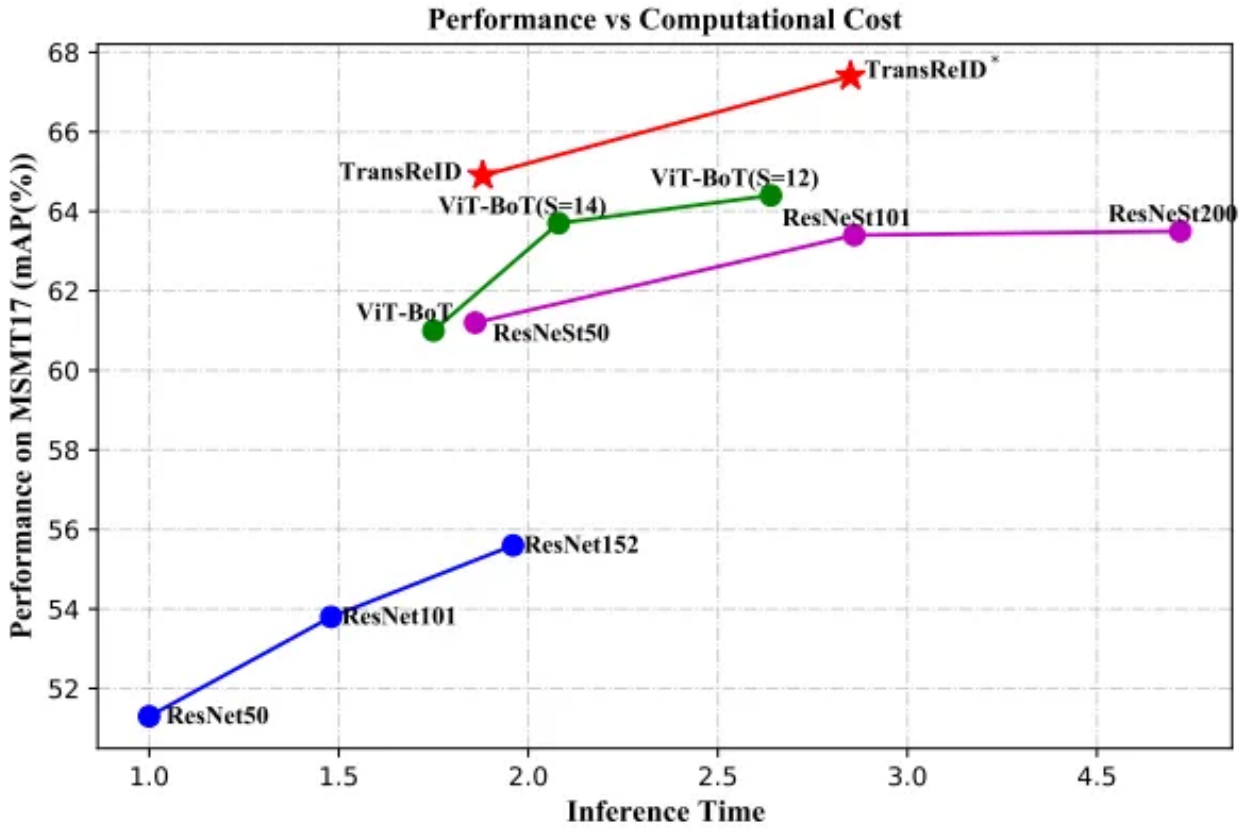


Figure 1: The comparision among TransReID, ViT-BoT, ResNet and ResNeSt on MSMT17. The computational cost of ResNet50 is taken as the baseline for inference time comparison.

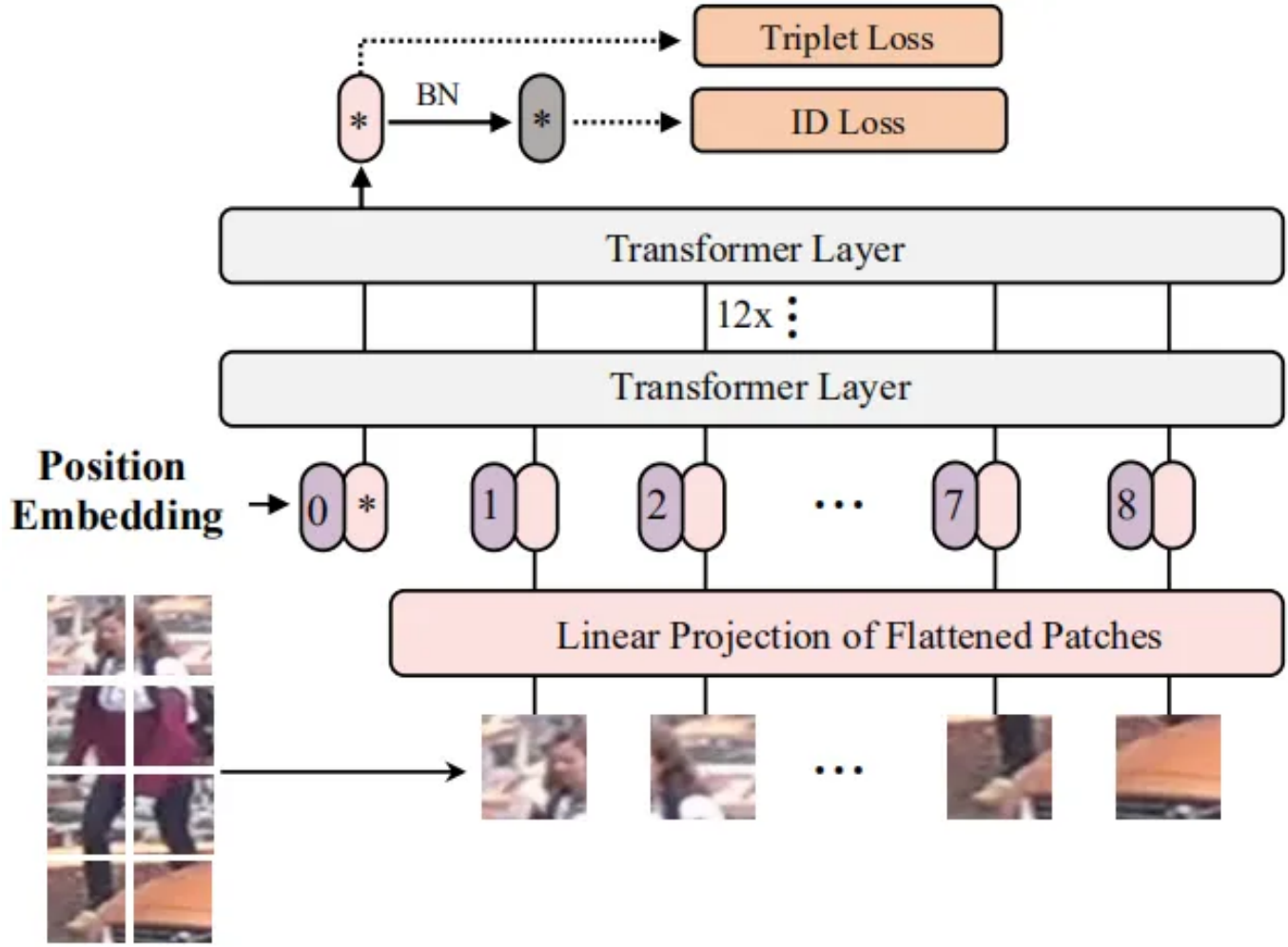
总而言之，本文有这样几点贡献：

- 首次提出一种纯Transformer架构用于ReID，并构建了一种强基线ViT-BoT(添加了几种改进)取得了与CNN相当的性能；
- 将Side Information Embedding引入到统一框架中以编码目标ReID的不同类型的边界信息，通过实验证实：SIE可以降低不同camera或者viewpoint导致的特征偏差；
- 提出JPM，它有助于更鲁棒特征表达的学习；
- 所提TransReID在多个ReID基准数据集(包含MSMT17、Market1501、DukeMTMC-ReID、Occluded-Duke、VeRi-776、VehicleID)上取得了SOTA性能。

Method

接下来我们将从两个方面来介绍本文所提方案，首先：我们针对ReID任务进行一些简单的适应性改进，得到了基线网络ViT-BoT；然后在ViT-BoT基础上，提出了融合 `Side Information Embedding` (SIE)与 `Jigsaw Patch, Module` (JPM)的TransReID。

ViT-BoT



上图给出了本文所提的ViT-BoT的结构示意图。由于原始的ViT是针对图像分类任务所设计，不能直接用于ReID任务，为此，我们对其进行了几点适应性调整。

Overlapping Patches

在预处理阶段，ViT需要将图像块拆分为N个不重叠块，这就会导致块的局部近邻结构信息无法较好的保持；相反，我们提出**采用滑动窗口形式生成重叠块**。假设滑动窗口的步长为S像素，每个块的尺寸$P=16$，那么重叠部分的形状为$(P-S)\times P$。基于上述定义，如果输入图像的尺寸为$H\times W$，那么所得到的图像块数量如下：

$$N=N_H\times N_W=\lfloor\frac{H+S-P}{S}\rfloor\times\lfloor\frac{W+S-P}{S}\rfloor$$

从上式可以得出：**重叠区域越大，所提图像块数量越多**。而更多的块通常可以带来更好的性能，但同时也会造成更高的计算量。为更好的区分，$\mathrm{ViT\text{-}BoT}_{s=12}$表示S=12；而S=P时则忽略下角标。

Position Embedding

位置嵌入ρ_i则编码图像块p_i的位置信息，它有助于Transformer的Encoder编码空间信息。注：由于ReID任务的图像分辨率不同于ImageNet，故ImageNet上位置嵌入无法直接应用，本文提出了采用双线性插值辅助ViT-BoT处理任意输入尺寸。

Feature Learning

给定拆分图像块，另一个可学习嵌入(如class token)将嵌入到上述块信息中，最后一个编码层的class token将作为图像的全局特征表达。假设最终的class token表示为f，其他的输出表示为$P_o=\{p_{o1},p_{o2},p_{o3},\cdots,p_{oN}\}$。损失函数定义如下：

$$\mathcal{L}+T=log[1+exp(\|f_a-f_p\|_2^2-\|f_a-f_n\|_2^2)]$$

TransReID

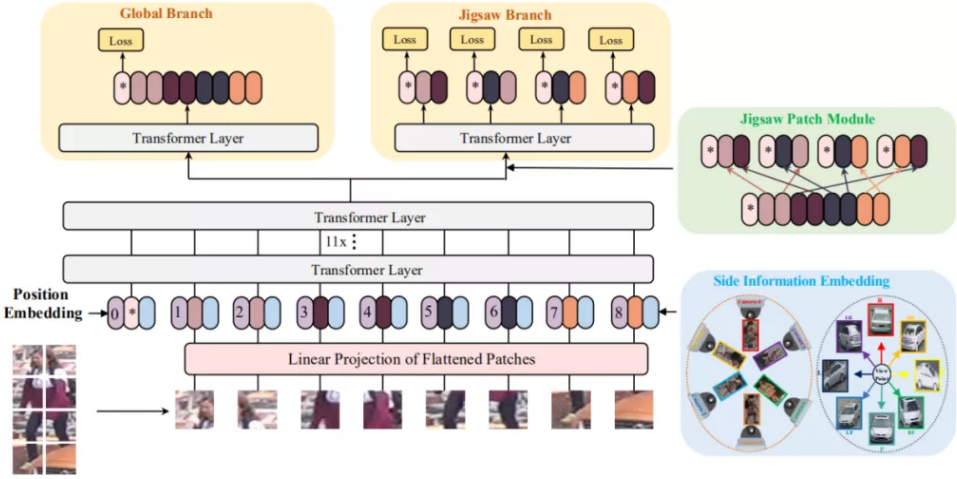


Figure 3: The framework of proposed TransReID. Side Information Embedding (light blue) encodes non-visual information such as camera or viewpoint into embedding representations. It is input into the transformer encoder together with patch embedding and position embedding. The last layer includes two independent transformer layers. One is standard to encode the global feature. The other one contains the Jigsaw Patch Module (JPM) which shuffles all patches and regroups them into several groups. All these groups are input into a shared transformer layer to learn local features. Both the global feature and local features contribute to ReID loss.

尽管前述所设计的ViT-BoT可以在目标ReID任务上取得很好的效果，但它并未充分利用ReID数据的特性。为更好的探索边界信息与细粒度信息，我们提出了SIE与JPM，并将所提框架称之为TranReID，其结构信息见上图。

Side Information Embedding

在目标重识别领域，一个极具挑战的问题是：**不同相机、视觉及其他因素导致的视觉偏差**。为克服上述问题，基于CNN的方案通常需要修改网络结构或者调整损失函数以利用这些非可视化信息(比如相机ID，视角等)。

Transformer则及其善于融合这类边界信息，故而Transformer非常适用于ReID任务。类似与位置嵌入，我们采用可学习层编码这些边界信息。具体的说，如果图像的相机ID为C，那么它对应的相机嵌入可以表示为$\mathcal{S}(C)$，不同位置嵌入会岁图像块变化，相机嵌入则对所有块相同。另外，如果视角信息V可知，我们同样可以将其编码到所有块中。

接下来，我们就需要考虑如何集成这两种不同类型的信息了。最简单的一个想法：直接进行想家，即$\mathcal{S}(C)+\mathcal{S}(V)$。但这种方式可能导致信息抵消。我们提出采用联合编码方式：$\mathcal{S}(C,V)$。也就是说，如果有C_N个相机ID，V_N个视角标签，那么$\mathcal{S}(C,V)$总计有$C_N \times V_N$个不同的值。那么第i个块的输入定义如下：

$$E^i = \mathcal{F}(p_i) + \rho_i + \lambda\mathcal{S}(C,V)$$

Jigsaw Patch Module

接下来，我们对ViT-BoT的最后一层调整为双并行分之结构(采用两个独立Transformer层分别用于学习全局特征与局部特征)。假设最后一层的输入隐含特征表示为$Z_{l-1}=[z_{l-1}^0; z_{l-1}^1, z_{l-1}^2, \cdots, z_{l-1}^N]$。全局分支为标准transformer层，它将输入编码为$Z_l=[f_g; z_l^1, z_l^2, \cdots, z_l^N]$，注：$f_g$表示全局特征。为学习更细粒度的部件级特征，一种直接的方式将输入特征分为k组，然后送入transformer层中学习k个局部特征，表示为$\{f_l^1, f_l^2, \cdots, f_l^k\}$。已有研究表明：token、主要受近邻token影响，因此近邻块的组合只能观察到有限的连续区域信息。

为解决上述问题，我们提出了Jigsaw Patch Module(JPM)对图像块先置换再分组。置换操作通过移动操作+块置换实现，可以描述如下：

- Step1: **The shift operation**

前每个块被移到后面，即$[z_{l-1}^1, z_{l-1}^2, \cdots, z_{l-1}^N]$移动m步变为$[z_{l-1}^{m+1}, z_{l-1}^{m+2}, \cdots, z_{l-1}^N, z_{l-1}^1, z_{l-1}^2, \cdots, z_{l-1}^m]$.

- Step2: **The patch shuffle operation**

经过前述移动后的块进一步通过置换操作(group=k)进行处理，此时隐含特征变为：$[z_{l-1}^{x_1}, z_{l-1}^{x_2}, \cdots, z_{l-1}^{x_N}], x_i \in [1, N]$.

我们将置换后的特征分成k组，JPM可以将其编码为k个局部特征$\{f_l^1, f_l^2, \cdots, f_l^k\}$，因此每个特征可以编码不同的部件，全局特征与局部特征分别采用分类损失$\mathcal{L}_{ID}, \mathcal{L}_T$进行训练，整体损失定义如下：

$$\mathcal{L} = \mathcal{L}_{ID}(f_g) + \mathcal{L}_T(f_g) + \frac{1}{k}\sum(\mathcal{L}_{ID}(f_l^i) + \mathcal{L}_T(f_l^i))$$

在推理阶段，我们对卷积特征与局部特征进行拼接得到最终的特征表达$[f_g, f_l^1, f_l^2, \cdots, f_l^k]$，如果仅仅采用$f_g$可以进一步减少计算浪，但性能也会稍微下降。

Experiments

为验证所提方案的性能，我们在四个行人ReID数据集(Market-1501, DukeMTMC-ReID, MSMT17, Occluded-Duke)与两个车辆ReID数据集(VeRi-776、VehicleID)上进行验证。下图给出了不同数据集的ID、数量等信息的简介。

Dataset	Object	#ID	#image	#cam	#view
MSMT17	Person	4,101	126,441	15	-
Market-1501	Person	1,501	32,668	6	-
DukeMTMC-reID	Person	1,404	36,441	8	-
Occluded-Duke	Person	1,404	36,441	8	-
VeRi-776	Vehicle	776	49,357	20	8
VehicleID	Vehicle	26,328	221,567	-	2

在实现方面，所有行人图像resize为256×128大小，所有车辆图像resize为256×256，数据增强包含padding(10)、RandomCrop、RandomErasing等。batch=64，每个ID4张图像。SGD优化器，初始学习率0.01，cosine学习率机制。如无特殊说明，对于行人而言，$m = 5, k = 4$；对于车辆而言，$m = 8, k = 4$。

Backbone	Training Time	MSMT17 mAP	MSMT17 R1	VeRi-776 mAP	VeRi-776 R1
ResNet50	1x	51.3	75.3	76.4	95.2
ResNet101	1.48x	53.8	77.0	76.9	95.2
ResNet152	1.96x	55.6	78.4	77.1	95.9
ResNeSt50	1.86x	61.2	82.0	77.6	96.2
ResNeSt200	4.72x	63.5	83.5	77.9	96.4
ViT-BoT	1.75x	61.0	81.8	78.2	96.5
ViT-BoT$_{s=14}$	2.08x	63.7	82.7	78.6	96.4
ViT-BoT$_{s=12}$	2.64x	64.4	83.5	79.0	96.5

上表给出了ViT-BoT在MSMT17与VeRi-776数据集上的性能对比。从表中数据可以看到：

- 具有更大骨干网络的BoT可以取得更好的性能；
- ViT-BoT可以取得与ResNeSt50相当的性能，而训练时间更少；
- 进一步降低步长s，模型的性能可以进一步的提升；
- ViT-BoT(s=12)取得了最佳的速度-精度均衡。

Method	Size	MSMT17 mAP	MSMT17 R1	Market-1501 mAP	Market-1501 R1	DukeMTMC-reID mAP	DukeMTMC-reID R1	Occluded-Duke mAP	Occluded-Duke R1
CBN© [51]	256×128	42.9	72.8	77.3	91.3	67.3	82.5	-	-
OSNet [49]	256×128	52.9	78.7	84.9	94.8	73.5	88.6	-	-
MGN [37]	384×128	52.1	76.9	86.9	95.7	78.4	88.7	-	-
RGA-SC [44]	256×128	57.5	80.3	**88.4**	**96.1**	-	-	-	-
ABDNet [4]	384×128	**60.8**	**82.3**	88.3	95.6	**78.6**	**89.0**	-	-
PGFA [26]	256×128	-	-	76.8	91.2	65.5	82.6	37.3	51.4
HOReID [36]	256×128	-	-	84.9	94.2	75.6	86.9	**43.8**	**55.1**
TransReID©	256×128	64.9	83.3	88.2	95.0	80.6	89.6	55.7	64.2
TransReID*©	256×128	**67.4**	**85.3**	**88.9**	**95.2**	**82.0**	**90.7**	**59.2**	**66.4**
TransReID®	384×128	66.6	84.6	88.8	95.0	81.8	90.4	57.2	64.0
TransReID*®	384×128	**69.4**	**86.2**	**89.5**	**95.2**	**82.6**	**90.7**	**59.4**	**66.7**

Method	Size	VeRi-776 mAP	VeRi-776 R1	VehicleID R1	VehicleID R5
PRReID[11]	256×256	72.5	93.3	72.6	88.6
SAN[28]	256×256	72.5	93.3	79.7	94.3
UMTS [14]	256×256	75.9	95.8	80.9	87.0
VANet© [5]	224×224	66.3	89.8	83.3	96.0
PVEN© [25]	256×256	79.5	95.6	**84.7**	**97.0**
SAVER [16]	256×256	**79.6**	**96.4**	79.9	95.2
CFVMNet [33]	256×256	77.1	95.3	81.4	94.1
TransReID ©	256×256	79.6	97.0	83.6	97.1
TransReID®	256×256	80.6	96.9	-	-
TransReID*©	256×256	80.5	96.8	**85.2**	**97.5**
TransReID*®	256×256	**81.7**	**97.1**	-	-

Table 6: Comparison with state-of-the-art methods. The star * in the superscript means the backbone is ViT-BoT$_{s=12}$. Results are shown for person ReID datasets (left) and vehicle ReID datasets (right). Only the small subset of VehicleID is used in this paper. © and © indicate the methods are using camera IDs and viewpoint labels, respectively. ® means both are used. Viewpoint and camera information are only used wherever available. Best results for previous methods and best of our methods are labeled in bold.

上表给出了TransReID在不同ReID数据集上的性能对比。从表中信息可以看出：

1.在行人重识别方面，在MSMT17与DukeMTMC-ReID数据集上，**TransReID以显著优化优于ABDNet的性能(+8.6%/+4.0%mAP)**；在Market-1501数据集上，TransReID同样取得了与SOTA方案相当的指标。

2.在遮挡ReID方面，相比PGFA与HOReID，**TransReID取得了11.9%mAP的性能提升，切无需任何语义信息进行对齐**；甚至通过重叠块辅助，**TransReID可以取得59.2%mAP的指标。**

3.在车辆重识别方面，在VeRi-776数据上，TransReID取得了81.7%mAP，以2.1%mAP由于SAVER；在更大的数据集VehicleID上，所提方法可以取得85.2%mAP的指标。

全文到此结束，更多消融实验与分析建议各位同学查看原文。

在**极市平台**后台回复“**TransReID**”，即可获取论文下载链接。

◎作者档案

Happy，一个爱“胡思乱想”的AI行者
个人公众号：AIWalker
欢迎大家联系极市小编（微信ID:fengcall19）加入极市原创作者行列

推荐阅读

- ResNet被全面超越了，是Transformer干的：依图科技开源“可大可小”T2T-ViT，轻量版优于MobileNet
- 搞懂 Vision Transformer 原理和代码，看这篇技术综述就够了（一）
- 搞懂 Vision Transformer 原理和代码，看这篇技术综述就够了（二）

阅读原文

喜欢此内容的人还喜欢

15个目标检测开源数据集汇总

极市平台