

一种目标检测的方法和装置

相关申请的交叉引用

5 本申请要求享有 2023 年 12 月 19 日提交的发明名称为“一种目标检测的方法和装置”的中国申请专利 No.202311747933.2 的优先权，在此全文引用上述中国专利申请公开的内容以作为本申请的一部分或全部。

10 技术领域

 本公开涉及人工智能技术领域，尤其涉及一种目标检测的方法和装置。

背景技术

15 纯视觉三维目标检测方法是对输入的图像信息，通过模型的特征挖掘，最终输出场景中目标物体的三维检测结果。近年来随着基于注意力机制的序列模型 Transformer 模型的发展，越来越多的方法将注意力机制运用到三维目标检测之中。尽管注意力机制可以有效地提取全局特征，但其计算复杂度往往显著高于常规卷积神经网络（CNN，
20 Convolutional Neural Network），因此很难在自动驾驶这一要求车端实时计算的场景中取得较快的计算速度。

 在实现本公开过程中，发明人发现现有技术中至少存在如下问题：

 现有基于注意力机制的目标检测方法计算复杂度高，计算速度慢，
25 难以实现实时的目标检测。

发明内容

 有鉴于此，本公开实施例提供一种目标检测的方法和装置，能够
 通过对注意力机制的 QKV 向量进行质量得分评估，然后从中筛选部分
30 携带最多有效信息的 QKV 向量，基于对 QKV 数量的动态筛选策略，

在不降低特征分辨率、不减少有效特征的情况下显著减少了注意力机制的计算量，提升了计算速度，从而达到可以在车端实时的进行目标检测的效果。

5 为实现上述目的，根据本公开实施例的一个方面，提供了一种目标检测的方法，包括：

 对输入图像进行特征提取得到特征矩阵，并将所述特征矩阵在平面维度拉平得到查询内容向量组；

 对所述查询内容向量组进行第一得分评估，并根据第一得分评估结果从所述查询内容向量组中提取关键查询内容向量组；

10 对预先生成的查询向量组进行第二得分评估，并根据第二得分评估结果从所述查询向量组中提取目标查询向量组；

 根据所述关键查询内容向量组和所述目标查询向量组执行注意力机制的计算，得到目标检测结果。

15 根据本公开的一个或多个实施例，对所述查询内容向量组进行第一得分评估，并根据第一得分评估结果从所述查询内容向量组中提取关键查询内容向量组，包括：通过引入额外的模型分支得到查询内容向量评估图，并使用所述查询内容向量评估图对所述查询内容向量组进行第一得分评估；根据第一得分评估结果对所述查询内容向量组中的查询内容向量按照得分从高到低进行排序；从排序后的查询内容向量组中选取前第一指定个数的查询内容向量，得到所述关键查询内容向量组。

20 根据本公开的一个或多个实施例，还包括：采用基于二分类交叉熵的分类损失函数监督信号，对所述查询内容向量评估图进行监督。

 根据本公开的一个或多个实施例，对预先生成的查询向量组进行第二得分评估，并根据第二得分评估结果从所述查询向量组中提取目标查询向量组，包括：通过引入额外的模型分支得到查询向量评估向

30

量，并使用所述查询向量评估向量对预先生成的查询向量组进行第二得分评估；根据第二得分评估结果对所述查询向量组中的查询向量按照得分从高到低进行排序；从排序后的查询向量组中选取前第二指定个数的查询向量，得到所述目标查询向量组。

5

根据本公开的一个或多个实施例，还包括：采用基于平均绝对误差的损失函数对所述查询向量评估向量进行监督。

10

根据本公开的一个或多个实施例，所述额外的模型分支基于卷积神经网络来实现。

根据本公开实施例的另一方面，提供了一种目标检测的装置，包括：

15

图像特征提取模块，用于对输入图像进行特征提取得到特征矩阵，并将所述特征矩阵在平面维度拉平得到查询内容向量组；

第一向量提取模块，用于对所述查询内容向量组进行第一得分评估，并根据第一得分评估结果从所述查询内容向量组中提取关键查询内容向量组；

20

第二向量提取模块，用于对预先生成的查询向量组进行第二得分评估，并根据第二得分评估结果从所述查询向量组中提取目标查询向量组；

注意力机制计算模块，用于根据所述关键查询内容向量组和所述目标查询向量组执行注意力机制的计算，得到目标检测结果。

25

根据本公开的一个或多个实施例，所述第一向量提取模块还用于：通过引入额外的模型分支得到查询内容向量评估图，并使用所述查询内容向量评估图对所述查询内容向量组进行第一得分评估；根据第一得分评估结果对所述查询内容向量组中的查询内容向量按照得分从高到低进行排序；从排序后的查询内容向量组中选取前第一指定个数的查询内容向量，得到所述关键查询内容向量组。

30

5 根据本公开的一个或多个实施例，所述第二向量提取模块还用于：
通过引入额外的模型分支得到查询向量评估向量，并使用所述查询向
量评估向量对预先生成的查询向量组进行第二得分评估；根据第二得
分评估结果对所述查询向量组中的查询向量按照得分从高到低进行排
序；从排序后的查询向量组中选取前第二指定个数的查询向量，得到
所述目标查询向量组。

10 根据本公开实施例的又一方面，提供了一种电子设备，包括：一个
或多个处理器；存储装置，用于存储一个或多个程序，当所述一个
或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理
器实现本公开实施例所提供的目标检测的方法。

15 根据本公开实施例的再一方面，提供了一种计算机可读介质，其
上存储有计算机程序，所述程序被处理器执行时实现本公开实施例所
提供的目标检测的方法。

20 上述的非惯用的可选方式所具有的进一步效果将在下文中结合具
体实施方式加以说明。

20

附图说明

附图用于更好地理解本公开，不构成对本公开的不当限定。其中：
图 1 是现有的基于注意力机制进行目标检测的实现原理示意图；
图 2 是本公开实施例的基于注意力机制进行目标检测的实现原理
25 示意图；
图 3 是根据本公开实施例的目标检测的方法的主要步骤示意图；
图 4 是根据本公开实施例的目标检测的装置的主要模块示意图；
图 5 是本公开实施例可以应用于其中的示例性系统架构图；
图 6 是适于用来实现本公开实施例的终端设备或服务器的计算机
30 系统的结构示意图。

具体实施方式

以下结合附图对本公开的示范性实施例做出说明，其中包括本公开实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。
5 因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本公开的范围和精神。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

需要说明的是，本申请公开的技术方案中，所涉及的用户个人信息的采集、收集、更新、分析、处理、使用、传输、存储等方面，均
10 符合相关法律法规的规定，被用于合法的用途，且不违背公序良俗。对用户个人信息采取必要措施，防止对用户个人信息数据的非法访问，维护用户个人信息安全、网络安全和国家安全。

注意力机制的计算量正比于其查询向量 Query（以下简称为 Q）、
15 被查向量 Key（以下简称为 K）、内容向量 Value（以下简称为 V）这三个权重矩阵的长度，因此为了对基于注意机制的模型进行加速，需要对 QKV 的长度进行缩减，从而成倍提升运算速度。但是 QKV 长度的缩减（即向量维度的缩减）往往意味着特征分辨率的缩小与特征信息的损失，因此极有可能带来模型性能的下降。
20

为了解决此问题，本公开提出一种动态筛选注意力机制 QKV 的策略，在不降低特征分辨率、不减少有效特征数量的前提下显著减少 QKV
25 长度，减少计算量，在不损失精度（甚至提升精度）的前提下实现模型计算速度的显著提升，达到车端可以实时进行目标检测的效果。

图 1 是现有的基于注意力机制进行目标检测的实现原理示意图。
如图 1 所示，现有技术中，输入图像首先通过常规卷积神经网络(CNN)
提取特征，得到特征矩阵 F。F 的高度为 H，宽度为 W，通道数为 C。
30 将 F 在平面维度拉平，得到 H*W 个维度为 C 的特征向量，作为注意力

机制的 K 和 V，KV 的长度为 $H*W$ 。同时预先生成的查询向量组 Q 的长度为 N。Q 和 KV 一同输入基于注意力机制的序列模型 Transformer 的解码器中，执行注意力机制的计算，得到解码器的输出。解码器输出经过后续的神经网络即可得到最终的目标检测结果。在此过程中，

5 KV 的长度 $H*W$ 与 Q 的长度 N 显著影响了注意力机制的计算量，往往无法取得较快的运算速度。

为了加速注意力机制计算，本公开实施例提出一种对 KVQ 进行动态筛选的策略，通过引入额外的模型分支，分别对 KV 和 Q 进行评估，

10 以实现 KV 和 Q 的动态筛选，以筛选出携带最多有效信息的 QKV 向量，显著减少注意力机制的计算量，提升计算速度，实现了车端实时的目标检测。

图 2 是本公开实施例的基于注意力机制进行目标检测的实现原理示意图。如图 2 所示，在本公开的实施例中，输入图像首先采取与现有技术相同的方法得到长 $H*W$ 的原始向量组 KV。具体地，输入图像通过常规卷积神经网络（CNN）提取特征，得到特征矩阵 F。F 的高度为 H，宽度为 W，通道数为 C。将 F 在平面维度拉平，得到 $H*W$ 个维度为 C 的特征向量，作为注意力机制的 K 和 V，KV 的长度为 $H*W$ 。

15

与此同时，本公开的目标检测模型引入额外的分支得到对 KV 进行得分评估的矩阵 KV 评估图（Key & Value Score）。该 KV 评估图的元素数量与原始 KV 向量组的长度相同，都是 $H*W$ ，且其中每个元素的值表示对应位置的 K/V 所携带有效信息的得分，得分越高，信息越丰富，越有利用价值。在所有的原始 KV 向量组中，有相当一部分位置对应输入图像中的无效（或低效）区域，如天空等。这些区域对于最终的目标检测收益极低，但却会带来极大的运算量。KV 评估图中某位置的得分越低，说明其越有可能是上述无效（或低效）区域。为了只让少数高效区域参与计算，对 KV 评估图进行 TOP K1 筛选，即筛选出得分前 K1 高的区域（举例来说，若 $K1=100$ ，则筛选出得分前 100 高的

20

25

30

位置)。具体来说,在本公开的模型中, $K1$ 的取值例如为 $H*W/4$,即将KV向量组的数量缩减到原先的 $1/4$ 。得到TOP $K1$ 得分的索引位置后,只对原始KV向量组的TOP $K1$ 索引位置对应的元素进行保留,得到筛选后的 $K' V'$ 向量组,长度为 $K1$ 。在此筛选过程中,特征分辨率并没有降低,而是通过对KV质量进行评估,只保留带有关键特征的“前景”区域,剔除背景区域,因此造成的信息损失量极小,不会造成模型精度的下降。甚至,这种方式会让模型更加关注有价值的键前景区域(如马路)而忽略可能造成干扰的背景区域(如天空),在实际训练过程中反而会提升模型精度,同时显著减少KV数量,提升计算速度。

与对KV向量组的筛选同理,本公开的目标检测模型引入额外的分支得到长度为 N 的对Q进行得分评估的Q评估向量(Query Score),该向量的元素数量与原始Q查询向量组的长度相同,每个位置的得分是对每个Q向量质量的评估。同样对Q评估向量进行TOP $K2$ 索引筛选($K2$ 在本实施例中具体设置为 $0.5N$,即减少Q向量组一半的数量),并只保留原始Q向量组TOP $K2$ 的索引位置,就得到动态筛选后的查询向量组 Q' ,长度为 $K2$ 。

最终,长度为 $K1$ 的 $K' V'$ 向量组与长度为 $K2$ 的 Q' 向量进入Transformer的解码器进行注意力机制计算,并最终输出目标检测结果。由于 $K1$ 只有 $H*W$ 的 $1/4$, $K2$ 只有 N 的 $1/2$,因此注意力机制的计算速度会获得显著增加,且不带来模型精度的下降,实现了实时的车端目标检测。

另外,为了保证目标检测模型的预测精度,KV评估图与Q评估向量需要能客观公正且准确地给出各个索引位置QKV的质量得分。为此需要引入额外的损失函数对二者进行监督。首先对于KV评估图,采用基于Focal loss(Focal loss是基于二分类交叉熵的一种损失函数)的分类损失函数监督信号。对于真值监督信号的制作,本公开将目标物体

的几何中心点投影至图像中并在真值图中对应位置处置 1，并在一定高斯半径内赋予衰减至 0 的浮点数得分真值。这样一来，真值 KV 评估图就具有了显著的物理意义，即距离目标物体中心点越近的像素，产生越高的质量得分。用此信号监督预测的 KV 评估图，可以保证其能客观公正地给出各个 KV 的质量得分。对于 Q 评估向量，由于目标检测中查询向量通常与物体的三维位置坐标 (x,y,z) 绑定，因此可以基于常规的 L1-Loss (平均绝对误差) 损失函数对其进行监督，距离真值目标物体越近的查询向量具有更高的得分。

图 3 是根据本公开实施例的目标检测的方法的主要步骤示意图。如图3所示，本公开实施例的目标检测的方法主要包括如下的步骤S301至步骤 S304。

步骤 S301：对输入图像进行特征提取得到特征矩阵，并将所述特征矩阵在平面维度拉平得到查询内容向量组。在本公开的一个实施例中，通过常规卷积神经网络 (CNN) 对输入图像进行特征提取，得到特征矩阵 F。F 的高度为 H，宽度为 W，通道数为 C。将 F 在平面维度拉平，得到 H*W 个维度为 C 的特征向量，作为注意力机制的被查向量 K 和内容向量 V，由 K 和 V 组成的查询内容向量组 KV 的长度为 H*W。

步骤 S302：对所述查询内容向量组进行第一得分评估，并根据第一得分评估结果从所述查询内容向量组中提取关键查询内容向量组。

根据本公开的一个实施例，对所述查询内容向量组进行第一得分评估，并根据第一得分评估结果从所述查询内容向量组中提取关键查询内容向量组，具体可以包括：通过引入额外的模型分支得到查询内容向量评估图，并使用所述查询内容向量评估图对所述查询内容向量组进行第一得分评估；根据第一得分评估结果对所述查询内容向量组中的查询内容向量按照得分从高到低进行排序；从排序后的查询内容向量组中选取前第一指定个数的查询内容向量，得到所述关键查询内

容向量组。在本公开的实施例中，所述额外的模型分支基于卷积神经网络来实现。使用所述查询内容向量评估图对所述查询内容向量组进行第一得分评估，例如是通过卷积神经网络对查询内容向量组执行卷积操作来进行第一得分评估。在具体实施过程中，在根据第一得分评估结果对查询内容向量组中的查询内容向量按照得分从高到低进行排序时，例如是对每个查询内容向量在查询内容向量组中的索引位置来进行排序的，后续在选取排序靠前的第一指定个数的查询内容向量时，也可以是选取序靠前的第一指定个数的索引位置，并根据选取的索引位置来获取对应的查询内容向量。其中，第一指定个数可根据业务场景需要进行灵活设定。

通过根据查询内容向量的得分从查询内容向量组中选取关键查询内容向量，可以只保留输入图像中带有关键特征的“前景”区域，剔除背景区域，因此造成的信息损失量极小，不会造成模型精度的下降。甚至，这种方式会让模型更加关注有价值的键前景区域（如马路）而忽略可能造成干扰的背景区域（如天空），在实际训练过程中反而会提升模型精度，同时显著减少查询内容向量组 KV 的数量，提升计算速度。

步骤 S303：对预先生成的查询向量组进行第二得分评估，并根据第二得分评估结果从所述查询向量组中提取目标查询向量组；

根据本公开的又一个实施例，对预先生成的查询向量组进行第二得分评估，并根据第二得分评估结果从所述查询向量组中提取目标查询向量组，具体可以包括：通过引入额外的模型分支得到查询向量评估向量，并使用所述查询向量评估向量对预先生成的查询向量组进行第二得分评估；根据第二得分评估结果对所述查询向量组中的查询向量按照得分从高到低进行排序；从排序后的查询向量组中选取前第二指定个数的查询向量，得到所述目标查询向量组。在本公开的实施例中，所述额外的模型分支基于卷积神经网络来实现。使用所述查询向量评估向量对所述查询向量组进行第二得分评估，例如是通过卷积神

神经网络对查询向量组执行卷积操作来进行第二得分评估。在具体实施过程中，在根据第二得分评估结果对查询向量组中的查询向量按照得分从高到低进行排序时，例如是对每个查询向量在查询向量组中的索引位置来进行排序的，后续在选取排序靠前的第二指定个数的查询向量时，也可以是选取序靠前的第二指定个数的索引位置，并根据选取的索引位置来获取对应的查询向量。其中，第二指定个数可根据业务场景需要进行灵活设定。

通过根据查询向量的得分从查询向量组中选取目标查询向量，可以只保留查询信息得分较高的查询向量 Q，因此造成的信息损失量极小，不会造成模型精度的下降。甚至，这种方式会让模型更加关注有价值的查询信息，在实际训练过程中反而会提升模型精度，同时显著减少查询向量组 Q 的数量，提升计算速度。

步骤 S304：根据所述关键查询内容向量组和所述目标查询向量组执行注意力机制的计算，得到目标检测结果。在本公开的实施例中，由于输入到模型中的 KVQ 的长度均进行了缩减，因此注意力机制的计算速度会获得显著提高，且不带来模型精度的下降，从而实现了实时的车端目标检测。

根据本公开的一个实施例，采用基于二分类交叉熵的分类损失函数监督信号，对所述查询内容向量评估图进行监督，采用基于平均绝对误差的损失函数对所述查询向量评估向量进行监督。为了保证模型预测精度，使得 KV 评估图与 Q 评估向量能客观公正且准确地给出各个索引位置 QKV 的质量得分，本公开实施例中引入了额外的损失函数对 KV 评估图与 Q 评估向量进行了监督。

对于 KV 评估图，采用基于 Focal loss 的分类损失函数监督信号。对于真值监督信号的制作，将目标物体（用于计算损失函数时，要人工标注真实存在的物体，目标物体就是这些标注得来的真值物体）的

几何中心点投影至图像中并在真值图中对应位置处置 1，并在一定高斯半径内赋予衰减至 0 的浮点数得分真值。这样一来，真值 KV 评估图就具有了显著的物理意义，即距离目标物体中心点越近的像素，产生越高的质量得分。用此信号监督预测的 KV 评估图，可以保证其能客观公正地给出各个 KV 向量的质量得分。对于 Q 评估向量，由于目标检测中查询向量通常与三维位置坐标 (x,y,z) 绑定，因此可以基于常规的 L1-Loss 对其进行监督，距离真值目标物体越近的查询向量具有更高的得分。

通过设计并引入额外的模型分支，得到对注意力机制的 KV 评估图与 Q 评估向量，准确的描述 QKV 向量的质量得分，通过对其进行筛选，得到得分最高，携带最多有效信息的 QKV 向量，显著减少注意力机制的计算量，提升计算速度，且可以提高模型的预测精度，达到可以在车端实时的进行目标检测的效果。

图 4 是根据本公开实施例的目标检测的装置的主要模块示意图。如图 4 所示，本公开实施例的目标检测的装置 400 主要包括图像特征提取模块 401、第一向量提取模块 402、第二向量提取模块 403 和注意力机制计算模块 404。

图像特征提取模块 401，用于对输入图像进行特征提取得到特征矩阵，并将所述特征矩阵在平面维度拉平得到查询内容向量组；

第一向量提取模块 402，用于对所述查询内容向量组进行第一得分评估，并根据第一得分评估结果从所述查询内容向量组中提取关键查询内容向量组；

第二向量提取模块 403，用于对预先生成的查询向量组进行第二得分评估，并根据第二得分评估结果从所述查询向量组中提取目标查询向量组；

注意力机制计算模块 404，用于根据所述关键查询内容向量组和所述目标查询向量组执行注意力机制的计算，得到目标检测结果。

5 根据本公开的一个实施例，第一向量提取模块 402 还可以用于：
通过引入额外的模型分支得到查询内容向量评估图，并使用所述查询
内容向量评估图对所述查询内容向量组进行第一得分评估；根据第一
得分评估结果对所述查询内容向量组中的查询内容向量按照得分从高
到低进行排序；从排序后的查询内容向量组中选取前第一指定个数的
查询内容向量，得到所述关键查询内容向量组。

10 根据本公开的又一个实施例，目标检测的装置 400 还可以包括评
估向量监督模块（图中未示出），用于：采用基于二分类交叉熵的分
类损失函数监督信号，对所述查询内容向量评估图进行监督。

15 根据本公开的又一个实施例，第二向量提取模块 403 还可以用于：
通过引入额外的模型分支得到查询向量评估向量，并使用所述查询向
量评估向量对预先生成的查询向量组进行第二得分评估；根据第二得
分评估结果对所述查询向量组中的查询向量按照得分从高到低进行排
序；从排序后的查询向量组中选取前第二指定个数的查询向量，得到
所述目标查询向量组。

20 根据本公开的又一个实施例，评估向量监督模块（图中未示出）
还可以用于：采用基于平均绝对误差的损失函数对所述查询向量评估
向量进行监督。

25 根据本公开的又一个实施例，所述额外的模型分支基于卷积神经
网络来实现。

30 根据本公开实施例的技术方案，通过对输入图像进行特征提取得
到特征矩阵，并将特征矩阵在平面维度拉平得到查询内容向量组；对
查询内容向量组进行第一得分评估，并根据第一得分评估结果从查询
内容向量组中提取关键查询内容向量组；对预先生成的查询向量组进

行第二得分评估，并根据第二得分评估结果从查询向量组中提取目标
查询向量组；根据关键查询内容向量组和目标查询向量组执行注意力
机制的计算，得到目标检测结果的技术方案，通过对注意力机制的 QKV
向量进行质量得分评估，然后从中筛选部分携带最多有效信息的 QKV
5 向量，基于对 QKV 数量的动态筛选策略，在不降低特征分辨率、不减少有效特征的情况下显著减少了注意力机制的计算量，提升了计算速度，从而达到可以在车端实时的进行目标检测的效果。

图 5 示出了可以应用本公开实施例的目标检测的方法或目标检测
10 的装置的示例性系统架构 500。

如图 5 所示，系统架构 500 可以包括终端设备 501、502、503，网络
504 和服务器 505。网络 504 用以在终端设备 501、502、503 和服务
器 505 之间提供通信链路的介质。网络 504 可以包括各种连接类型，
15 例如有线、无线通信链路或者光纤电缆等等。

用户可以使用终端设备 501、502、503 通过网络 504 与服务器 505
交互，以接收或发送消息等。终端设备 501、502、503 上可以安装有
各种通讯客户端应用，例如车辆导航应用、网页浏览器应用、搜索类
20 应用、即时通信工具、邮箱客户端、社交平台软件等（仅为示例）。

终端设备 501、502、503 可以是具有显示屏并且支持网页浏览的
各种电子设备，包括但不限于智能手机、平板电脑、膝上型便携计算
机和台式计算机等等。

服务器 505 可以是提供各种服务的服务器，例如对用户利用终端
设备 501、502、503 所浏览的网站提供支持的后台管理服务器（仅为
示例）。后台管理服务器可以对接收到的目标检测请求等数据进行对
输入图像进行特征提取得到特征矩阵，并将所述特征矩阵在平面维度
30 拉平得到查询内容向量组；对所述查询内容向量组进行第一得分评估，

并根据第一得分评估结果从所述查询内容向量组中提取关键查询内容向量组；对预先生成的查询向量组进行第二得分评估，并根据第二得分评估结果从所述查询向量组中提取目标查询向量组；根据所述关键查询内容向量组和所述目标查询向量组执行注意力机制的计算，得到目标检测结果等处理，并将处理结果（例如目标检测结果--仅为示例）反馈给终端设备。

需要说明的是，本公开实施例所提供的目标检测的方法一般由服务器 505 执行，相应地，目标检测的装置一般设置于服务器 505 中。

应该理解，图 5 中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器的。

下面参考图 6，其示出了适于用来实现本公开实施例的终端设备或服务器的计算机系统 600 的结构示意图。图 6 示出的终端设备或服务器仅仅是一个示例，不应对本公开实施例的功能和使用范围带来任何限制。

如图 6 所示，计算机系统 600 包括中央处理单元（CPU）601，其可以根据存储在只读存储器（ROM）602 中的程序或者从存储部分 608 加载到随机访问存储器（RAM）603 中的程序而执行各种适当的动作和处理。在 RAM 603 中，还存储有系统 600 操作所需的各种程序和数据。CPU 601、ROM 602 以及 RAM 603 通过总线 604 彼此相连。输入/输出（I/O）接口 605 也连接至总线 604。

以下部件连接至 I/O 接口 605：包括键盘、鼠标等的输入部分 606；包括诸如阴极射线管（CRT）、液晶显示器（LCD）等以及扬声器等的输出部分 607；包括硬盘等的存储部分 608；以及包括诸如 LAN 卡、调制解调器等网络接口卡的通信部分 609。通信部分 609 经由诸如因特网的网络执行通信处理。驱动器 610 也根据需要连接至 I/O 接口 605。

可拆卸介质 611，诸如磁盘、光盘、磁光盘、半导体存储器等等，根据需要安装在驱动器 610 上，以便于从其上读出的计算机程序根据需要被安装入存储部分 608。

5 特别地，根据本公开公开的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本公开公开的实施例包括一种计算机程序产品，其包括承载在计算机可读介质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信部分 609 从网络上被下载和安装，
10 和/或从可拆卸介质 611 被安装。在该计算机程序被中央处理单元（CPU）601 执行时，执行本公开的系统限定上述功能。

 需要说明的是，本公开所示的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是一——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM 或闪存）、
15 光纤、便携式紧凑磁盘只读存储器（CD-ROM）、光存储器件、磁存储器件、或者上述的任意合适的组合。在本公开中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本公开中，计算机可读的信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读的信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机
20 可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限

于：无线、电线、光缆、RF 等等，或者上述的任意合适的组合。

附图中的流程图和框图，图示了按照本公开各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，上述模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的，框图或流程图中的每个方框、以及框图或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

描述于本公开实施例中所涉及到的单元或模块可以通过软件的方式实现，也可以通过硬件的方式来实现。所描述的单元或模块也可以设置在处理器中，例如，可以描述为：一种处理器包括图像特征提取模块、第一向量提取模块、第二向量提取模块和注意力机制计算模块。其中，这些单元或模块的名称在某种情况下并不构成对该单元或模块本身的限定，例如，注意力机制计算模块还可以被描述为“用于根据所述关键查询内容向量组和所述目标查询向量组执行注意力机制的计算，得到目标检测结果的模块”。

作为另一方面，本公开还提供了一种计算机可读介质，该计算机可读介质可以是上述实施例中描述的设备中所包含的；也可以是单独存在，而未装配入该设备中。上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被一个该设备执行时，使得该设备包括：对输入图像进行特征提取得到特征矩阵，并将所述特征矩阵在平面维度拉平得到查询内容向量组；对所述查询内容向量组进行第一得分评估，并根据第一得分评估结果从所述查询内容向量组中提取关

键查询内容向量组；对预先生成的查询向量组进行第二得分评估，并根据第二得分评估结果从所述查询向量组中提取目标查询向量组；根据所述关键查询内容向量组和所述目标查询向量组执行注意力机制的计算，得到目标检测结果。

5

根据本公开实施例的技术方案，通过对输入图像进行特征提取得到特征矩阵，并将特征矩阵在平面维度拉平得到查询内容向量组；对查询内容向量组进行第一得分评估，并根据第一得分评估结果从查询内容向量组中提取关键查询内容向量组；对预先生成的查询向量组进行第二得分评估，并根据第二得分评估结果从查询向量组中提取目标查询向量组；根据关键查询内容向量组和目标查询向量组执行注意力机制的计算，得到目标检测结果的技术方案，通过对注意力机制的 QKV 向量进行质量得分评估，然后从中筛选部分携带最多有效信息的 QKV 向量，基于对 QKV 数量的动态筛选策略，在不降低特征分辨率、不减少有效特征的情况下显著减少了注意力机制的计算量，提升了计算速度，从而达到可以在车端实时的进行目标检测的效果。

10

15

上述具体实施方式，并不构成对本公开保护范围的限制。本领域技术人员应该明白的是，取决于设计要求和因素，可以发生各种各样的修改、组合、子组合和替代。任何在本公开的精神和原则之内所作的修改、等同替换和改进等，均应包含在本公开保护范围之内。

20

权 利 要 求 书

1. 一种目标检测的方法，包括：
- 对输入图像进行特征提取得到特征矩阵，并将所述特征矩阵在平面维度拉平得到查询内容向量组；
- 对所述查询内容向量组进行第一得分评估，并根据第一得分评估结果从所述查询内容向量组中提取关键查询内容向量组；
- 对预先生成的查询向量组进行第二得分评估，并根据第二得分评估结果从所述查询向量组中提取目标查询向量组；
- 根据所述关键查询内容向量组和所述目标查询向量组执行注意力机制的计算，得到目标检测结果。
2. 根据权利要求 1 所述的方法，其中，对所述查询内容向量组进行第一得分评估，并根据第一得分评估结果从所述查询内容向量组中提取关键查询内容向量组，包括：
- 通过引入额外的模型分支得到查询内容向量评估图，并使用所述查询内容向量评估图对所述查询内容向量组进行第一得分评估；
- 根据第一得分评估结果对所述查询内容向量组中的查询内容向量按照得分从高到低进行排序；
- 从排序后的查询内容向量组中选取前第一指定个数的查询内容向量，得到所述关键查询内容向量组。
3. 根据权利要求 2 所述的方法，进一步包括：
- 采用基于二分类交叉熵的分类损失函数监督信号，对所述查询内容向量评估图进行监督。
4. 根据权利要求 1 所述的方法，其中，对预先生成的查询向量组进行第二得分评估，并根据第二得分评估结果从所述查询向量组中提取目标查询向量组，包括：
- 通过引入额外的模型分支得到查询向量评估向量，并使用所述查

询向量评估向量对预先生成的查询向量组进行第二得分评估；

根据第二得分评估结果对所述查询向量组中的查询向量按照得分从高到低进行排序；

从排序后的查询向量组中选取前第二指定个数的查询向量，得到所述目标查询向量组。

5

5. 根据权利要求 4 所述的方法，进一步包括：

采用基于平均绝对误差的损失函数对所述查询向量评估向量进行监督。

10

6. 根据权利要求 2 或 4 所述的方法，其中，所述额外的模型分支基于卷积神经网络来实现。

7. 一种目标检测的装置，包括：

15 图像特征提取模块，用于对输入图像进行特征提取得到特征矩阵，并将所述特征矩阵在平面维度拉平得到查询内容向量组；

第一向量提取模块，用于对所述查询内容向量组进行第一得分评估，并根据第一得分评估结果从所述查询内容向量组中提取关键查询内容向量组；

20 第二向量提取模块，用于对预先生成的查询向量组进行第二得分评估，并根据第二得分评估结果从所述查询向量组中提取目标查询向量组；

注意力机制计算模块，用于根据所述关键查询内容向量组和所述目标查询向量组执行注意力机制的计算，得到目标检测结果。

25

8. 根据权利要求 7 所述的装置，其中，所述第一向量提取模块还用于：

通过引入额外的模型分支得到查询内容向量评估图，并使用所述查询内容向量评估图对所述查询内容向量组进行第一得分评估；

30 根据第一得分评估结果对所述查询内容向量组中的查询内容向量

按照得分从高到低进行排序；

从排序后的查询内容向量组中选取前第一指定个数的查询内容向量，得到所述关键查询内容向量组。

- 5 9. 根据权利要求 7 所述的装置，其中，所述第二向量提取模块还用于：

通过引入额外的模型分支得到查询向量评估向量，并使用所述查询向量评估向量对预先生成的查询向量组进行第二得分评估；

- 10 根据第二得分评估结果对所述查询向量组中的查询向量按照得分从高到低进行排序；

从排序后的查询向量组中选取前第二指定个数的查询向量，得到所述目标查询向量组。

- 15 10. 一种电子设备，包括：

一个或多个处理器；

存储装置，用于存储一个或多个程序，

当所述一个或多个程序被所述一个或多个处理器执行，使得所述一个或多个处理器实现如权利要求 1-6 中任一所述的方法。

- 20 11. 一种计算机可读介质，其上存储有计算机程序，所述程序被处理器执行时实现如权利要求 1-6 中任一所述的方法。