面向小物体检测的知识协同循环注意网络

西安交通大学 易凯 郑南宁

西安交通大学人工智能与机器人研究所，陕西西安，710049
西安交通大学视觉信息处理与应用国家工程实验室，陕西西安，710049
（本文英文完整版已提交国际期刊 IEEE Transactions on Image Processing 在审）

郑南宁 教授 中国工程院院士 IEEE Fellow

中文摘要：准确的交通标志检测（TSD）可以帮助智能系统根据交通规则做出更好的决策。TSD 作为一种典型的小目标检测问题，在高级驾驶辅助系统（ADAS）和无人驾驶中具有突出作用。然而，虽然深度神经网络已经在多个任务上实现了人类甚至超人的性能，但由于其自身的局限性，小目标检测仍然是一个悬而未决的问题。为了有效解决这一问题，在本文中，我们提出了一个脑启发的网络，命名为 KB-RANN。注意机制是我们大脑的一个基本功能，我们使用一种新的递归注意神经网络，以逐步求精的方式来提高检测精度。此外，我们结合领域特定的知识和直觉知识，以提高效率。实验结果表明，该方法比物体检测的常用方法有更好的性能。更重要的是，我们成功地将我们的方法移植到我们设计的嵌入式系统上，并成功地部署在自主研发的无人驾驶车上。

英文摘要：Accurate Traffic Sign Detection (TSD) can help intelligent systems make better decisions according to the traffic regulations. TSD, regarded as a typical small object detection problem in some way, is fundamental in Advanced Driver Assistance Systems (ADAS) and self-driving. However, although deep neural networks have achieved human even superhuman performance on several tasks, due to their own limitations, small object detection is still an open question. In this paper, we proposed a brain-inspired network, named as KB-RANN, to handle this problem. Attention mechanism is an essential function of our brain, we used a novel recurrent attentive neural network to improve the detection accuracy in a fine-grained manner. Further, we combined domain specific knowledge and intuitive knowledge to improve the efficiency. Experimental result shows that our methods achieved better performance than several popular methods widely used in object detection. More significantly, we transplanted our method on our designed embedded system and deployed on our self-driving car successfully.

关键词：小目标检测；感知与视觉；深度学习

一、 引言

在强大的设计良好的深度神经网络的帮助下，在目标检测领域中取得了极大的进展[1][2]。对于无人驾驶而言，提供实时和准确的目标的位置和类别信息是极为重要的。然而，目前检测精度和速度之间存在两难。例如，基于区域提议的方法（如 Faster RCNN[1]）可以获得高召回率和更好的准确度，但是发现得到区域建议是很耗时的。同时，基于回归的方法（如 SSD[2]）可以进行实时检测，但通常不能达到满意的精度。

设计更为时间高效而不影响精度的目标检测模型是当前目标检测领域的趋势。这种思想驱动的一个典型分支便是神经网络的压缩和加速。其中一个极为优秀的工作是 SqueezeNet[3]，它实现了与 AlexNet（一种广泛使用的图像分类模型[4]）相媲美的精度，而前者比后者速度快 50 倍。该模型使用了几种先进的策略来设计卷积网络层（例如，3×3 过滤器 和 1×1 过滤器的结合，减少输入通道的数量，延迟下采样）和一个新的 Fire Module 来设计更为强大的神经网络。利用 SqueezeNet 作为前向神经网络，[5]提出了一种不损失精度的全卷积神经网络，称为 SqueezeDet。这种方法在 KITTI 这一应用于无人驾驶评测的目标检测数据集上实现了领先的性能。

国家级大学生创新创业训练计划支持项目（项目批准号）
作者简介：易凯，（1996 -），男，湖北武汉人，软件工程专业，大三，研究方向为脑认知启发的人工智能，机器视觉，机器学习基础理论，计算心理学

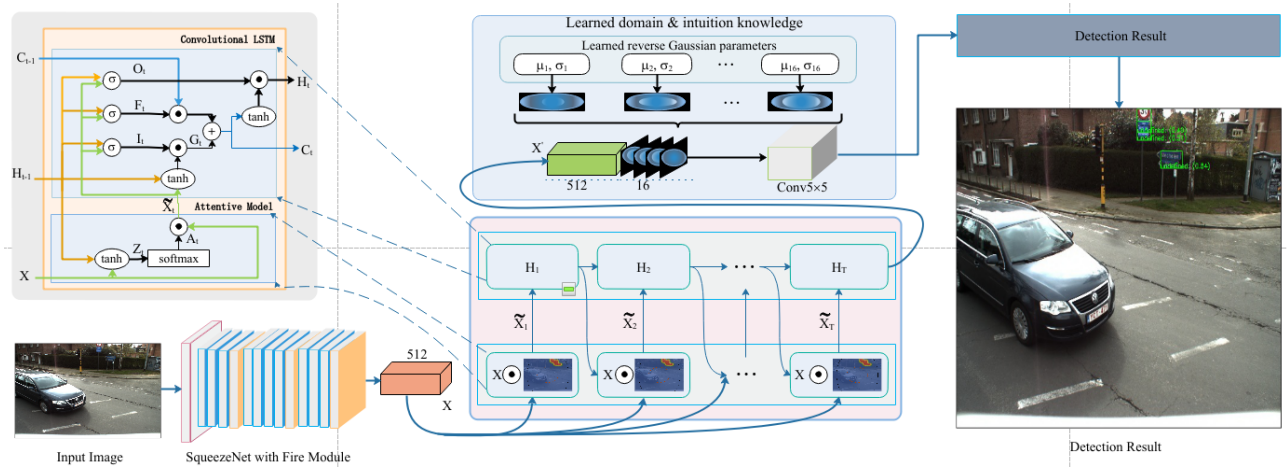

图 1 我们提出的 KB-RANN 的算法流程图。我们的方法得到输入图像的特征映射，然后将其反馈到一个新的递归注意神经网络，以一种逐步求精的方式改进特征映射的特征。学习到的领域知识和直觉知识是可行驶区域的中心偏置。为了提高对小交通标志的检测，引入了交通标志总是在于可行驶区域的中心偏置的猜想。采用特征映射法对级联卷积网络进行反馈后，利用 softmax 回归得到最终的检测结果

然而在实时物体检测任务上已经取得了可喜的成果，但是如何检测具有不同尺寸的物体，特别是小物体，仍然是一个悬而未决的问题。目前处理这一问题的方法大致可分为两个分支。一个典型的分支是设计多尺度神经网络，以提取不同级别的特征，以适应不同大小目标的检测。另一种常用的方法是利用解卷积来放大深层的特征图[6][7]。这些方法减轻了目标多样性的问题，但仍然不能很好地检测小目标。在本文中，我们提出使用认知机制启发的神经网络，它在小目标检测方面表现出更好的性能。

人脑由多个模块子系统组成，具有独特的相互作用方式。注意力机制是认知的一个重要功能。我们人类以复杂的方式使用不同类型的知识来执行棘手的任务。为了实现更好的决策，无人驾驶系统迫切需要一种很好地检测小物体的方法。在这些认知机制的启发下，我们将直觉先验知识和领域知识相结合，同时应用注意力机制，提出了一种基于知识的递归注意力神经网络。

TSD 是一个棘手的任务，满足上述所有特征。一方面，智能系统总是需要提前准备好接近交通标志。这是因为交通标志是用来提供有用的信息，以帮助驾驶方便和安全。智能系统需要部分依靠交通标志进行决策。更重要的是，系统在接收到检测结果后需要做出正确的决策。因此，通过相机得到的有用的交通标志总是非常小。然而，正如我们所提到的，小物体检测仍然是一个悬而未决的问题。此外，现实世界环境十分复杂，雨天、大雾、大雪等恶劣天气条件对交通标志的检测精度有很大影响。我们使用递归注意力神经网络（RANN）以逐步求精的方式提升检测精度与注意力定位。另一方面，驾驶条件与我们人类的认知有许多相似之处。通常，无人驾驶场景中中央偏置区域是可行驶区域。在这篇文章中，我们假设交通标志总是位于偏置选定的可驾驶区域。实验表明，该假设可以提高 5 点的检测精度。

自然科学中交通标志的检测与识别一直是众多研究者关注的焦点。这类问题的解决方法大致可分为两类：基于区域的方法和基于组件连接的方法。对于第一种方法，利用包括纹理和颜色的局部特征来定位文本区域。对于后者，通过应用诸如颜色对比度、强度变化等信息来单独地分割文本字符。在拥挤的场景中准确地检测出小的道路标志或恶劣的照明条件是至关重要的。然而，在一些流行的数据集[8][9]中，那些最先进的方法[9][10]在处理拥挤的场景中的小交通标识识别问题方面并不是很好。

为了去有效解决小物体检测的问题，我们提出了新颖的基于知识构建的循环注意力神经网络 KBRANN。总的来说，该论文的贡献可以概括为如下四个方面：

a. 据我们所知，这是第一次尝试应用认知机制的基本原理去解决交通标识检测问题。

b. 复杂天气条件情况在真实世界中极为常见，为了去很好地检测对应情境下的小物体，基于广泛使用的长短期记忆网络（LSTM），我们提出了全新的循环注意力网络。该网络能够以一种迭代逐步求精的方式来处理交通表示的不同层次的特征。

c．我们使用逆高斯先验分布去处理交通标识检测问题。这种方法结合领域知识（驾驶员的注意力在可行驶区域，该区域总是在于中央区域的偏置，而交通标识的位置往往是该可行驶区域的偏置）和直觉知识（人类的注意力区域遵循高斯分布，我们能够将交通标识的注意力视为逆高斯分布）。

d. 我们成功将所提出的算法移植到了自主研发的嵌入式系统。此外，我们在自主研发的 Pioneer I 无人驾驶车上部署了该算法同时在真实交通场景进行了测试，测试结果表明 KB-RANN 能够以可观的速度具有鲁棒性地对交通标识进行检测。

二、相关工作

在本文中，我们使用领域/直觉知识和注意力机制来帮助交通标志的检测。在这一章中，我们将从三个方面介绍相关的工作：交通标识检测、具有注意力机制的递归神经网络和基于知识的深度学习系统。第一部分将介绍交通标志检测领域中的几种主流方法。第二部分将研究基于注意的神经网络。第三部分将分析利用神经网络应用领域/直觉知识解决棘手问题的现状。

（一）交通标识检测

交通标识检测在 ADAS 以及无人驾驶领域是极为重要的任务。当前该领域取得了很大的进展，其中主流的方式是提取交通标识的形状和颜色特征。对于这些方法，基于不同的颜色和形状的方法被用来最小化环境对测试图像的影响。之前的研究中提出了许多使用形状和颜色信息的方法来进行兴趣区域检测（例如，区域增长[12]，YCbCr 颜色空间变换[13]和颜色索引[14]）。由于在面对光照和天气变化时颜色信息是不可靠的，因此引入了基于形状的算法。目前流行的基于形状的方法是相似性检测 [15]，具有 Haar 类特征的边缘[16]和距离变换匹配[17]。

近年来，深度学习在许多领域得到了广泛的应用，并在很大程度上提高了特定领域处理任务的效率。与基于颜色和形状的方法相比，深度学习方法能够自动化地从标注数据中学习到层次化特征。目前，基于局部特征的方法在多个交通标识检测和分类数据集上具有很好的性能，例如德国交通标志检测和分类数据集[18]。然而，由于交通条件的多样性，这些方法在现实世界条件下不能很鲁棒。为了测试我们提出的 KB-RNN 的性能，我们在比利时交通标志检测数据集（BTSD）和真实交通场景上进行了测试，结果表明，与几种最先进的检测方法[5][1]相比，我们的方法获得了更优的性能。

（二）具有注意机制的递归神经网络

递归神经网络（RNNs）已被广泛应用于序列问题的处理。从我们人类大脑的角度来看，我们特别注意那些对我们真正重要的单词，或者注视图片中的特定位置。具有这种注意力机制的 RNN 可以实现相同的行为，专注于给定信息的一部分。

有几个分支的想法与 RNN 的应用注意处理类似上述问题。目前，有几类典型的应用注意力机制来增强 RNN 表征能力的方法。其中一个重要分支是基于内容的注意力方法，这意味着我们只需四处看看，通过扩展传统的循环神经网络来自动学习注意力。具有注意力能力的循环神经网络产生一个查询，描述它想要关注的内容。每个项目都是带有查询的点积来计算得分，描述与查询匹配的程度。分数被放入一个 softmax 中来计算注意力分布。[19]提出了一种基于注意力的循环神经网络来处理输入以传递关于每个看到的单词的信息，然后对于 RNN 产生输出以集中在单词上，以便它们在机器翻译中变得相关。[20]使用一个 RNN 处理音频，然后在它上面有另一个 RNN 跳跃性连接，在处理语音识别问题时产生了相关的部分。

注意力 RNNs 的另一个分支是创建一个具有注意机制的更优设计的 RNNS 体系结构。在图像字幕领域 [21] 首先通过一个卷积神经网络提取输入图像的多个特征，然后利用 RNNS 生成图像的描述。当它在描述中生成每个词时，RNN 集中于对图像相关部分的解释。[22] 利用递归网络单元迭代地选择到所选择的图像子区域，以渐进的方式进行显著细化。[23]提出了精心设计的注意 LSTM 体系结构，以细化从卷积网络中提取的特征映射。基于 注意力长短期神经网络，我们设计了新的循环注意力神经网络，这种结构使小物体检测的 mAP 提升了三个百分点。

（三）基于知识的深度学习系统

我们人类可以使用不同的知识（例如，从自我经验的内部知识，通过与周围物体相互作用的环境知识，从全局中提取的全球知识）来学习和感知世界。受这一事实启发，近年来，以知识为基础的深度学习系统得到了极大的关注。这些方法可以简单地归类为两个分支。一种是使用直觉先验知识。[24]训练卷积神经网络来检测和跟踪对象，而无需任何标注的样本。[23]使用先验知识，即人类注视区域是我们所看到的中心偏置，以提高注视区域预测的准确性。另一个分支是使用特定领域的知识。[24]提出域约束来检测对象，而不使用标记数据。在本文中，我们将领域知识和直觉知识进行结合。我们假设无人驾驶场景下的注视区域是可行驶区域，并且交通标志总是可行驶区域的偏置。我们提出了一种新的逆高斯先验分布来形式化这个问题。与不使用该附加方法相比，使用该融合知识的目标检测精度提高了 5 个百分点。

三、 模型架构

在本章节中，我们呈现了我们的完整模型，称之为 KB-RANN。

（一）实时高精度 SqueezeNet

如今，许多预训练的卷积神经网络模型（例如，VGGNet[25] 和 ResNet[26]）在目标检测领域中占主导地位，并达到了最先进的性能。虽然这些模型提高了目标检测的效率，但它们是以时间为代价的。构建一个设计良好的预训练模型在实时目标检测中是非常重要的。SqueezeNet[3] 是一个典型的预训练模型，它具有 AlexNet 级别的精度但是具有较少的参数。同时，出于全卷积层足够强大，能够同时对物体进行分类和定位的考虑，我们在 SqueezeNet 中引入了全卷积处理模块。此外，我们使用两个额外的 Fire 模块，以提高我们的网络精度。

（二）循环注意力神经网络

注意机制是认知科学的重要组成部分。我们引入了一个精心设计的递归注意神经网络，以帮助检测交通标识。

LSTM[27]被广泛地应用于那些与时间相关的任务[28][29]。[23]利用 LSTM 的顺序特性以迭代方式处理特征，而不是利用模型来处理输入之间的时间依赖关系。在这一思想的启发下，我们提出了一种新的递归注意力神经网络。该网络由多个注意力神经网络（ANN）组成。更新规则遵循以下等式（1）：

$$
\begin{aligned}
I_t &= \sigma\left(W_i * \widetilde{X_t} + U_i * H_{t-1} + b_i\right) \\
F_t &= \sigma(W_f * \widetilde{X_t} + U_f * H_{t-1} + b_f) \\
O_t &= \sigma(W_o * \widetilde{X_t} + U_o * H_{t-1} + b_o) \\
G_t &= \tanh(W_g * \widetilde{X_t} + U_g * H_{t-1} + b_g) \\
C_t &= F_t \odot C_{t-1} + I_t \odot G_t \\
H_t &= O_t \odot \tanh(C_t)
\end{aligned}
\tag{1}
$$

C_t, H_t 是普通 LSTM 中的两类典型门。G_t是记忆门。在我们的网络结构中，特别设计I_t, F_t, O_t是注意力神经网络的内部门。

我们提出的 RANN 体系结构是通过专注于图像的不同区域的注意力机制来计算的。注意门A_t的更新规则遵从以下等式（2）：

$$
A_t = V_a * \tanh(W_a * X + U_a * H_{t-1} + b_a) \tag{2}
$$

我们所提出的 RANN（多个注意力神经网络以迭代方式的级联）能够以一种逐步精细化的方式提升检测的精度。

（三）领域知识和直觉知识的融合

人类可以以复杂的方式组合不同类型的知识来解决非常困难的问题。领域知识是一类非常重要的知

识。在无人驾驶领域，人们的目光是注视中心的偏置。通常，偏置中心是可行驶区域。在这篇文章中，我们假设交通标识总是位于可行驶区域的偏置区域。在图 2 中，左图 2(a)是原始图像，而右边的图 2(b)是反向高斯先验和领域知识的证明。对于图 2(b), 中心位置的橙色圆是我们的主要关注区域（即无人驾驶领域中的可行驶区域的注视位置），而除了可行驶区域之外的黑色圆圈是我们的方法用于检测小交通标志的焦点。右上角附近的区域是交通标志的集合。

(a) (b)

图 2 （a）为原始图像，（b）为逆高斯领域知识和直觉知识融合使用的演示

先验知识被用来处理交通标识识别问题，但几乎所有的文献都专注于提取交通标识的颜色和形状特征。据我们所知，领域知识还没有被用于处理交通标识检测。我们提出的逆高斯方法用于这一任务。此外，为了减少参数的数量和便于学习，我们约束每个高斯先验是一个 2D 高斯函数，其均值和协方差矩阵是可自由学习得到的，这使得网络完全从数据中学习它自己的先验，而不依赖于神经认知科学研究的假设。

我们所提出的模型可以通过以下方程（3）来学习每个先验图的参数：

$$f^*(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\left(\frac{(x-\mu_x)^2}{2\sigma_x^2}+\frac{(y-\mu_y)^2}{2\sigma_y^2}\right)\right) \tag{4}$$

我们可以通过以下方程（4）计算逆高斯分布：

$$f^*(x,y) = 1 - f^*(x,y)$$

我们将 N 个反向高斯特征映射与由卷积网络提取的 W 特征映射相结合。在本研究的系列试验中，我们将 N 设为 16，W 设为 512。因此，在级联之后，我们得到具有 528 个通道的混合特征映射。通过比较几种典型的模型，不难得到领域和直觉知识的注入被证明是有效的，我们可以从 KB-CNN 的结果（表 2）看到。

四、 实验设计

（一）训练原则

如[30]所说明的，一步训练策略将局部丢失和分类损失一起训练，可以加快网络而不会丢失太多的精度，受到这种想法的启发，我们定义了一种多任务损失函数，其形式如下（5）所示：

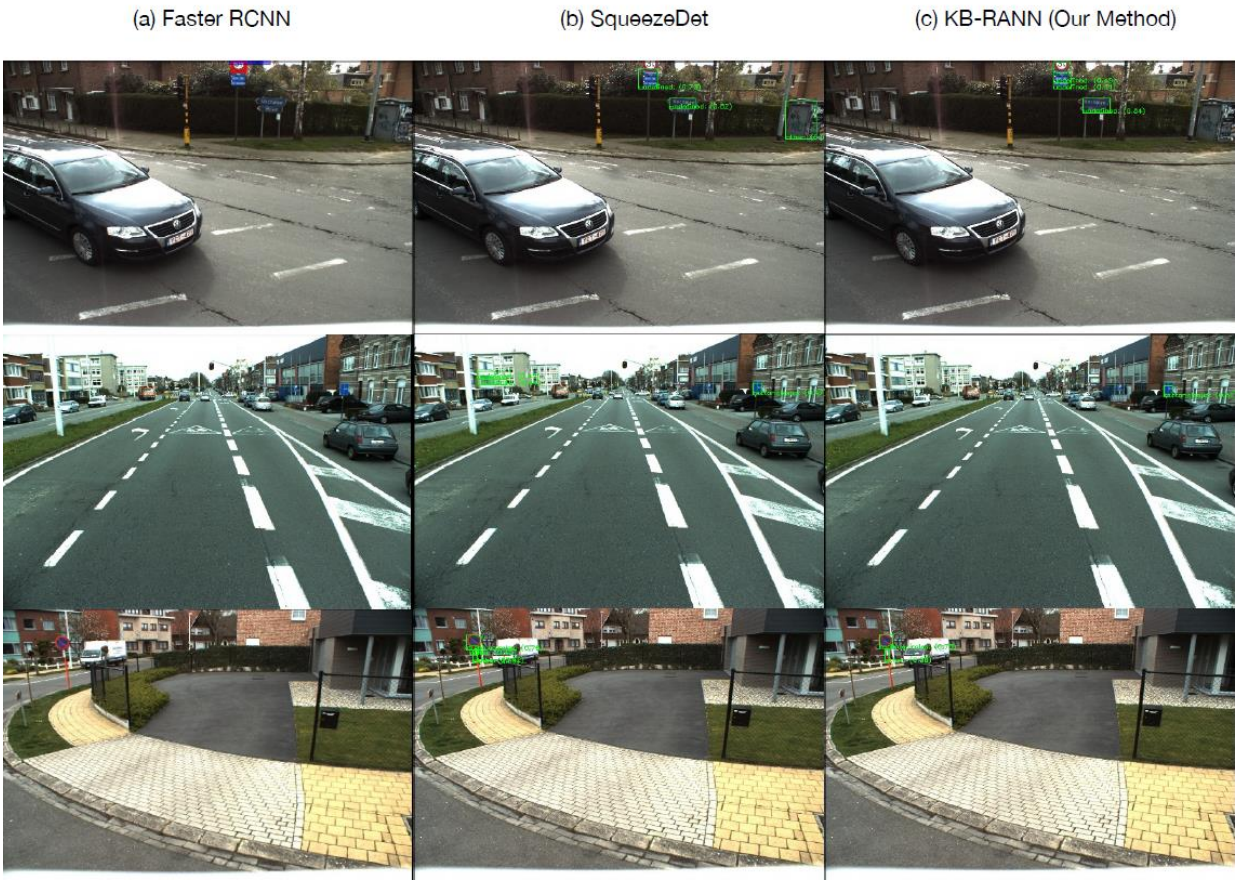

图 3：三种被测方法检测结果之间的比较，包括 Faster RCNN [1], SqueezeDet 和 KB-RANN. (a) Faster RCNN 的检测结果，其漏检率很高。我们从第实验结果的第一列可以看到，这种方法几乎不能从背景挑选出小的交通标识。(b) 是 SqueezeDet [5] 的检测结果。在我们测试期间，我们发现该方法导致的错检可能性很高，它容易挑选出那些没有交通标识的区域。(c) 是我们提出的方法 KB-RANN 的检测结果，它达到了相对可观的检测结果。更甚的是，KB-RANN 检测结果的置信度值比之前提到的两种方法都要高

$$\frac{\lambda_{bb0x}}{N_{obj}}\sum_{i=1}^{W}\sum_{j=1}^{H}\sum_{k=1}^{K}I_{ijk}\left(Q_x+Q_y+Q_w+Q_h\right)+\sum_{i=1}^{W}\sum_{j=1}^{H}\sum_{k=1}^{K}\frac{\lambda_c^+}{N_{obj}}I_{ijk}Q_\gamma+\frac{\lambda_{conf}^-}{WHK-N_{obj}}\tilde{I}_{ijk}\gamma_{ijk}^2$$

$$+\frac{1}{N_{obj}}\sum_{i=1}^{W}\sum_{j=1}^{H}\sum_{k=1}^{K}\sum_{c=1}^{C}I_{ijk}l_c^G\log(p_c) \tag{5}$$

损失函数包括三个部分，分别为边界框回归，置信度回归和分类交叉熵损失。Q^*代指$(\delta_{*ijk}-\delta_{*ijk}^G)^2$（$*$分别代指x, y, γ, w 和 h. ）

第一部分是上述提到的边界框回归的损失。$(\delta x_{ijk},\delta y_{ijk},\delta w_{ijk},\delta h_{ijk})$表示在(i, j)处的第 K 个 anchor 的相对坐标。与此同时，$\delta x_{ijk}^G\ or\ (\delta x_{ijk}^G,\delta y_{ijk}^G,\delta w_{ijk}^G,\delta h_{ijk}^G)$是真实值边界框，它能够通过方程(6)进行计算：

$$\delta x_{ijk}^G=(x^G-\hat{x}_i)/\hat{w}_k,\delta y_{ijk}^G=(y^G-\hat{y}_i)/\hat{h}_k$$
$$\delta w_{ijk}^G=\log(w^G/\hat{w}_k),\delta h_{ijk}^G=log\left(h^G/\hat{h}_k\right) \tag{6}$$

第二部分是置信度的回归损失。最后一个特征图的输出是γ_{ijk}，它表示在(i, j) 处的第 K 个 anchor 的预测置信度得分。γ_{ijk}^G是真实值以及预测的边界框之间的 IoU。除此之外，我们通过$\tilde{I}_{ijk}\gamma_{ijk}^2$的权值来对那些与任务不相关的检测结果进行惩罚。其中，$\tilde{I}_{ijk}=1-I_{ijk},\lambda_{conf}^-$。此外，$\lambda_{conf}^-$被用来改变网络的连接权。

最后一部分是分类的交叉熵损失。l_c^G是二值化的标签，$p_c, c \in [1, \mathrm{C}]$是通过网络预测得到的分类分布。我们使用 softmax 回归去归一化该项得到的打分从而确保p_c在$[1, \mathrm{C}]$的范围之间。通过使用反向传统联合损失函数能够以自动化的方式进行更新。

（二）数据集以及基准

1. 数据集

在交通标识检测与识别领域有几种主流的数据集。多年间对于识别和检测存在概念混淆的情况。总的来说，识别任务旨在对目标进行分类，而检测任务需要在分类的同时进行定位。德国交通标识识别数据集(German Traffic Signs Recognition) 是种广泛使用的交通标识识别数据集，传统的非深度学习方法也能够很好地解决这一问题。与此同时，该数据集中只有 600 张图片的训练样本，不足以训练神经网络使其收敛。我们专注于另外一个称之为比利时交通标识检测（Belgium Traffic Signs Detection）的数据集上。我们接下来将对训练的细节以及在不同数据集上的结果进行介绍。

2. 基线

我们将我们提出的方法与几大目标检测领域流行的开源方法进行了对比，对比的方法包括[1]和SqueezeDet[5]。我们所使用的源代码直接来源于对应方法的原论文。我们唯一做的是重新设置 面向 BTSD 数据集的 RBG 颜色均值以及 anchor 的大小。

3. BTSD 数据集实验结果

BTSD 数据集是一种广泛使用的交通标识检测数据集，所有的交通标识被视为 13 个不同的类别。该数据集包括 5905 张图片训练样本和 3101 张测试图片。我们在该数据集上比较了多种最为先进的方法，结果见表 2。我们使用召回率，不同 IoU 的 mAP 来评估不同的方法。可以从表中看到，我们提出的 KB-RANN 方法在不同的 IoU 情况下有着更好的检测效果。此外，RANN 以及 KB-CNN 方法与两大广泛使用的方法比较中也得到了相对更好的性能。

我们在表 2 中绘制了多种不同方法的检测结果。从该表我们可以看到，从领域以及直觉中提取出来的知识以及循环注意力神经网络确实能够提升交通标识的检测精度。我们还选择了不同方法测试得到的多张图片。从这些图片中可以看出，我们的方法在小交通标识检测上比 Faster RCNN 和 SqueezeDet 更好。

表 1：BTSD 标准数据集上的检测结果

Method	mAP(IoU = 0.3)	mAP(IoU = 0.5)
Faster RCNN	0.53	0.39
SqueezeDet	0.74	0.57
RANN	0.78	0.62
KB-CNN	0.79	0.60
KB-RANN	0.81	0.65

4. 实验设置

在实验结果 2 中，这些方法的迭代次数是 100K。我们将批处理大小设置为 32。BTSD 数据集中的原始图片大小为 1626×1236。当将这些图像馈入深度神经网络时，计算能力超出单个 NVIDIA GTX 1080 Ti 的限制。为了解决这个问题，我们将图片调整为 542×412。这个操作使交通标识变小了。在某种程度上，增加了交通标志检测的难度。

五、 嵌入式系统上的实现

一种具有前瞻性和有价值的无人驾驶方法，有两个重要的本质问题：一个是在低功耗平台上的一个可用实现，另一个是高帧率以满足实时处理的要求。因此，我们在自行设计的嵌入式系统上实现了该算法，在该平台上面向无人驾驶等工业应用的算法进行评价和优化。我们选择 NVIDIA JETSON TX2 模块作为嵌

入式系统的核心，它平衡了功率效率和计算能力。

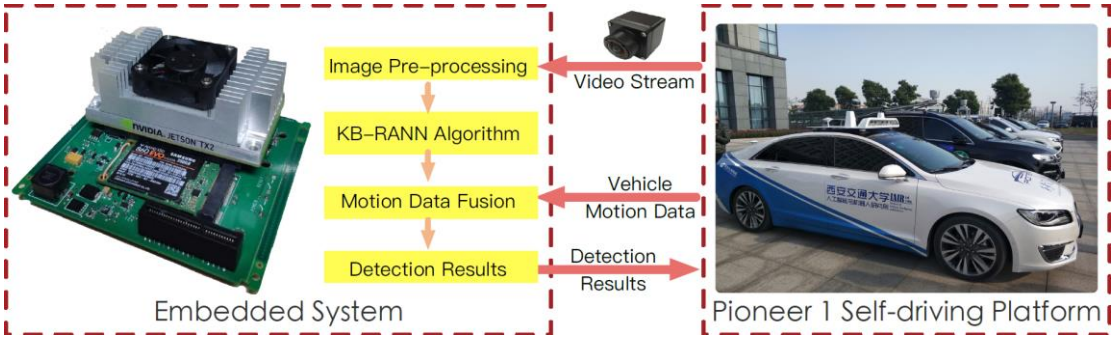

图 4：KB-RANN 算法在我们自主研发的无人驾驶系统中的实现。该图显示了基于 NVIDIA TX2 设计的嵌入式系统。左边是我们的算法处理流程，右边是我们的无人驾驶平台

此外，我们实现了上述方法（包括 Faster RCNN，SqueezeNet，以及我们提出的一系列方法），其中硬件加速和软件优化被执行。结果显示在表 2 上。我们的方法运行在我们自己设计的 10FPS 的嵌入式系统上，速度足够快，能够进行交通标识检测。然而，我们利用 TensorRT 加速了我们的实时性能，这对于缩短深度神经网络的推理时间是有效的，因为实时检测也可以用于道路障碍检测，例如行人检测。

表 2：嵌入式系统上不同检测方法的性能比较。所有的输入图片分辨率为 542×412

方法	图片大小	帧率
Faster RCNN	542× 412	0.84FPS
SqueezeDet(SqueezeNet)	542× 412	10.92FPS
RANN	542× 412	9.72FPS
KB-CNN	542× 412	10.57FPS
KB-RANN	542× 412	9.45FPS

所提出的算法在无人驾驶平台上进行了实现，如图 4 所示。TX2 模块与我们的无人驾驶平台相互作用。首先，该系统从车载摄像机获取原始视频流。然后，我们提出的 KB-RANN 算法得到图像预处理后的检测结果数据。接着，该系统融合无人驾驶平台获得的移动数据。最后，可以产生有效的检测结果，并进一步用于决策。

六、 结论与未来工作

本文针对小目标检测，特别是在交通标识检测方面进行了研究。受人类大脑认知机制的启发，我们提出了一种新的基于知识的递归注意力神经网络。该方法比常用的目标检测方法具有更好的性能。此外，我们证明了从领域和直觉中提取的知识确实有效，并且递归注意机制可以帮助以逐步求精的方式更好地检测小对象。

未来的研究方向可能是将注意力机制应用到视频中的路标检测中，我们可以使用丰富的背景信息。更重要的是，如何将交通规则与直觉知识相结合，构建一种更加真实的动态小目标检测方法

致谢

本研究得到了中国国家自然科学基金（61773312 号，61790563 号），高等学校学科创新引智计划（B13043 号）的部分支持。

参考文献：

[1] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region

proposal networks. In *Advances in neural information processing systems*, pages 91 – 99, 2015.

[2] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21 – 37. Springer, 2016.

[3] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. arXiv preprint arXiv:1602.07360, 2016. [5] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., & Fu, C. Y., et al. (2015). Ssd: single shot multibox detector. 21-37.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks.In *Advances in neural information processing systems*, pages 1097 – 1105, 2012. [7]

[5] Bichen Wu, Forrest Iandola, Peter H Jin, and Kurt Keutzer. Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. *arXiv preprint arXiv:1612.01051*, 2016.

[6] J. L Starck, A Bijaoui, I Valtchanov, and F Murtagh. A combined approach for object detection and deconvolution.Astronomy and Astrophysics Supplement, 147(1):139–149, 2000.

[7] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi,and Alexander C Berg. Dssd: Deconvolutional single shot etector. *arXiv preprint arXiv:1701.06659*, 2017.

[8] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark:a multi-class classification competition. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1453 – 1460. IEEE, 2011.

[9] Radu Timofte, Karel Zimmermann, and Luc Van Gool. Multiview traffic sign detection, recognition, and 3d localisation. *Machine vision and applications*, 25(3):633 – 647, 2014.

[10] A Kumar K Sumi. Detection and recognition of road traffic signs - a survey. *International Journal of Computer Applications*, 2017.

[11] Jack Greenhalgh and Majid Mirmehdi. Recognizing textbased traffic signs. *IEEE Transactions on Intelligent Transportation Systems*, 16(3):1360 – 1369, 2015.

[12] L Priese and V Rehrmann. On hierarchical color segmentation and applications. In *Computer Vision and Pattern Recognition, 1993. Proceedings CVPR ’93., 1993 IEEE Computer Society Conference on*, pages 633 – 634, 1993.

[13] Ahmed Hechri and Abdellatif Mtibaa. Automatic detection and recognition of road sign for driver assistance system. In *Electrotechnical Conference*, pages 888 – 891, 2012.

[14] Michael J Swain and Dana H Ballard. Color indexing. International journal of computer vision, 7(1):11–32, 1991.

[15] S. Vitabile, G. Pollaccia, G. Pilato, and F. Sorbello. Road signs recognition using a dynamic pixel aggregation techniquein the hsv color space. In International Conference on Image Analysis and Processing, page 572, 2001.

[16] Benjamin Höferlin and Klaus Zimmermann. Towards reliable traffic sign recognition. *Intelligent Vehicles Symposium IEEE*, 5(3):324 – 329, 2009.

[17] Dariu Gavrila. Traffic sign recognition revisited. In *Mustererkennung 1999, 21. DAGM-Symposium*, pages 86 – 93, 1999.

[18] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 1453 – 1460. IEEE, 2011.

[19] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align andtranslate. *arXiv preprint arXiv:1409.0473*, 2014.

[20] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 4960 – 4964. IEEE, 2016.

[21] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048 – 2057, 2015.

[22] Jason Kuen, Zhenhua Wang, and Gang Wang. Recurrent attentional networks for saliency detection. *arXiv preprint*

arXiv:1604.03227, 2016.

[24] Russell Stewart and Stefano Ermon. Label-free supervision of neural networks with physics and domain knowledge. In *AAAI*, pages 2576 – 2582, 2017.

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXivpreprint arXiv:1409.1556*, 2014.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedingsof the IEEE conference on computer vision and pattern recognition*, pages 770 – 778, 2016.

[27] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[28] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625 – 2634, 2015.

[29] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE Conference on Computer Vision andPattern Recognition*, pages 4622 – 4630, 2016.

[30] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer* d*vision and pattern recognition*, pages 779 – 788, 2016.