**操纵模板像素以进行孪生视觉跟踪的模型自适应**

在本文中，我们表明可以通过简单地在孪生网络中操纵模板图像的像素来处理视觉目标跟踪中具有挑战性的模型自适应任务。对于不包含在离线训练集中的目标，模板图像像素的稍加修改将改善离线训练的孪生网络的预测结果。流行的对抗样本生成方法可用于执行模板像素修改以进行模型调整。与当前的模板更新方法（旨在合并先前帧的目标特征）不同，我们专注于在第一帧中使用目标的ground truth进行初始自适应。我们的模型自适应方法是即插即用的，因为它不会改变其基本跟踪器的总体架构。据我们所知，这项工作是在孪生跟踪器中直接操纵模板像素以进行模型调整的首次尝试。在最近的基准测试中进行的大量实验表明，我们的方法比其他一些最新的跟踪器具有更好的性能。

目标跟踪是指仅在给定其初始状态的情况下，在视频序列中定位指定运动目标的任务。最近，孪生网络[1]，[2]已被应用于该任务，并来了目标跟踪性能的显着提高。孪生跟踪器将视觉目标跟踪问题公式化为学习目标模板和搜索区域之间的互相关相似性。然后，通过计算最高的视觉相似度，从搜索图像区域中找到目标，从而进行跟踪。尽管最近取得了成功，但是对于不包含在离线训练集中的目标，所学习的孪生网络相似性度量方法不一定是可靠的，从而导致泛化性差[3]。最近的一些工作旨在使模型自适应当前的目标表观。例如，TADT [4]根据反向传播的梯度识别每个卷积滤波器的重要性，并基于用于表示目标的激活来选择目标感知特征。但是，TADT的特征提取器已在ImageNet [5]上进行了预训练，而不是在大规模视觉跟踪数据集上进行了预训练。这限制了其功能在目标跟踪任务上的表示能力。GradNet [6]利用渐变的判别信息，并通过前馈和反馈操作更新了孪生网络中的模板。但是，额外的子网会增加计算成本，并且容易过度拟合。UpdateNet [7]学习结合先前框架中的目标功能。但是，它不使用ground truth信息来自适应地调整第一帧的模板特征。在这项工作中我们表明，在视觉目标跟踪中具有挑战性的模型自适应任务可以通过简单地在孪生网络中操纵模板图像的像素来处理。给定一个目标跟踪器，我们的算法仅使用第一帧中的目标地面真相，仅在几次梯度下降迭代中修改模板像素。对于不包含在离线训练集中的目标，我们认为对模板图像像素进行略微修改可以改善离线训练的孪生网络的预测结果。我们使用对抗样本生成方法来实现此目的，因为它通常用于稍微修改输入图像，从而对网络的预测结果产生影响。需要注意的是，对抗性样本生成的目的是使网络的预测变得更糟，而我们希望孪生网络的预测变得更好。所提出的模型自适应方法可以与SiamFC ++ [9]等各种孪生跟踪器集成。孪生网络的参数保持不变，以保留离线训练的嵌入空间的生成能力。我们对4个跟踪基准进行了全面的实验：VOT2018 [10]，TrackingNet [11]，GOT-10 k [8]和OTB2015 [12]。我们的方法在以超过80 FPS的速度运行时可以达到最先进的结果（见图1）。我们对4个跟踪基准进行了全面的实验：VOT2018 [10]，TrackingNet [11]，GOT-10 k [8]和OTB2015 [12]。我们的方法在以超过80 FPS的速度运行时可以达到最先进的结果（见图1）。我们对4个跟踪基准进行了全面的实验：VOT2018 [10]，TrackingNet [11]，GOT-10 k [8]和OTB2015 [12]。我们的方法在以超过80 FPS的速度运行时可以达到最先进的结果（见图1）。

**拟议算法**

在本节中，我们将通过直接操纵模板像素为孪生跟踪器提供一种新的模型自适应方法。我们首先回顾与基于模板匹配的流行跟踪器的跟踪过程，这与所提出的方法密切相关。根据[13]，我们将目标跟踪公式化为基于置信度的回归问题，该问题学习函数sθ：Y×X→R，并根据给定的输出-输入对预测标量置信度得分sθ（y，x）∈R（ y，x）。最终估计值f（x）= y ∗预测如下：f（x）= arg maxy∈Ysθ（y，x），（1）其中x是输入图像。y通常表示目标目标的中心2D图像坐标。当前，有两种流行的基于模板匹配的范例：判别相关滤波器（DCF）方法和孪生跟踪方法。在基于DCF的方法中，在跟踪过程中训练一个圆形相关滤波器wθ以预测目标置信度得分：sθ（y，x）=（wθ∗φ（x））（y），（2）其中φ（x）是从搜索中提取的特征图片x。与DCF相比，孪生跟踪器采用两流体系结构。一个流基于根据地面真实边界框从第一帧裁剪的模板图像z提取目标特征φθ（z）。另一流接收大的搜索图像x作为输入，并输出搜索特征φθ（x）。两个输出互相关以预测目标置信度得分：sθ（y，x）=（φθ（z）\*φθ（x））（y）。（3）基于DCF的跟踪器和孪生跟踪器均具有利用大规模可视跟踪数据集自行训练特征提取器φ（·）或嵌入网络φθ（·）的优势。这样，可以增强特征在目标跟踪任务上的表示能力。与孪生跟踪器相反，DCF从目标表观的示例补丁中学习过滤器wθ，以将其与背景区分开。尽管使用了循环相关运算提高了跟踪效率，但其边界效应和复杂的优化却无法在计算速度和跟踪性能之间做出良好的权衡。孪生跟踪器在这方面做得更好，尽管在互相关中学习的相似性度量对于离线训练集中未包含的目标不一定是可靠的，从而导致泛化不佳。在这封信中 我们旨在设计一种新的孪生跟踪方法，该方法能够充分利用当前视频的特定信息以进行模型调整，例如基于DCF的跟踪器，尽管其中的目标未包含在离线训练集中。这是通过利用第一帧中的注释信息执行模型调整来实现的。注意Equ。（2）和等式。（3）彼此之间有一些相似之处，主要区别在于相关的核项：DCF的核是在线学习的wθ，而孪生网络的核是φθ（z）。为了使孪生网络具有模型自适应能力，我们需要使用当前视频的第一帧地面真相标注来自适应φθ（z）。有两种适应φθ（z）的设计选择：更改φθ（·）或更改z。然而，改变φθ（·）可能会导致繁琐的元学习设置，从而无法确保离线训练的嵌入空间的生成能力[14]，[15]。相比之下，我们的解决方案是以简单的方式通过更改z来执行孪生跟踪器的模型自适应，即仅使用第一帧的目标ground truth，仅在几次梯度下降迭代中修改模板像素。与当前的孪生跟踪器模型自适应方法相比，该方法具有以下优点。首先，我们从不修改孪生网络的参数来保留离线训练的嵌入空间的表示能力。其次，与当前模板更新方法[7]，[16]不同，后者旨在结合先前帧的目标特征，我们专注于在第一帧使用目标ground truth进行初始自适应。最后，我们的模型调整方法是可插拔的，因为它不会改变基本跟踪器的总体架构。在下一个小节中，我们将展示如何使用流行的对抗样本生成方法来执行用于模型自适应的模板像素操作。

**A. 操纵模型像素以进行模型自适应**

乍看之下，模型自适应任务与对抗样本生成任务之间可能存在矛盾，因为这两个任务具有不同的目的。一个对抗样本[17]是输入数据的样本，该样本的输入数据经过了非常轻微的修改，旨在对机器学习系统进行攻击，这意味着导致机器学习模型做出错误的预测。然而，我们工作中模型自适应的目的是充分利用第一帧中的注释信息，以提高当前视频的跟踪性能。在下文中，我们将指出这两个任务之间存在一些相似之处，并且我们可以利用对抗样本生成方法来执行模型自适应任务。在介绍提出的方法之前，我们首先回顾流行的对抗样本生成方法。Iadv生成对抗图像的最简单方法之一是通过对干净图像的L∞邻域中的损失函数进行线性化，并使用以下闭式方程[18]来找到线性化函数的精确最大值：Iadv = I +符号∇IL（I， ytrue）（4），其中I是输入图像，像素的值是[0，255]范围内的整数。ytrue是图像I的真实标签。L（I，y）是给定图像I和标签y的用于攻击目的的神经网络的成本函数。是要选择的超参数。扩展上述方法的一种直接方法是以较小的步长多次应用它，并在每个步骤之后裁剪中间结果的像素值，以确保它们位于原始图像的附近。这导致在[17]中引入的基本迭代方法（BIM）：Iadv 0 = I，Iadv N + 1 = ClipI，{Iadv N +αsign（∇IL（Iadv N，ytrue））}}，（5）其中ClipI，{I？是对图像I执行逐像素裁剪的功能吗？，因此结果将在源图像I的L∞附近。对于特定的所需目标类别，BIM可以轻松地变成攻击者，称为迭代目标类别方法[17]：Iadv 0 = I，Iadv N + 1 = ClipI，{Iadv N-α符号（∇IL（Iadv N，ytarget））}。（6）式 （6）表明，仅需进行几次梯度下降迭代操作就可以将输入图像的像素更改为目标类ytarget的网络预测。请注意，我们的目的是在第一帧中处理模板图像的像素，以便使预测更接近于地面真实边界框。因此，我们可以使用Equ对孪生网络进行模型自适应。（6）进行一些修改：z0 = z，zN + 1 = Clipz，{zN-αsign（∇zL（zN，ybb））}，（7）其中z是第一帧中的模板图像，而ybb是从地面真相边界框生成的孪生跟踪器的标签。在下一个小节中，我们将介绍配备了建议的模型自适应模块的整体跟踪过程。B．跟踪框架所提出的模型自适应方法以即插即用的方式与SiamFC ++跟踪器[9]集成在一起。SiamFC ++基于SiamFC [2]，并根据[9]中提出的若干准则逐步完善。原始SiamFC ++网络的输入由从第一帧裁剪的模板图像z0和从第i帧裁剪的搜索图像xi组成。但是，我们希望执行模板处理，以便在对输入对（z0，x0）进行N步像素更新后获得z?。，跟踪器在（z ,? xi）上表现良好。为此，我们首先使用地面真实边界框从第一帧裁剪初始模板图像z0∈R3×128×128和初始搜索图像x0∈R3×289×289。然后，将z0和x0发送到SiamFC ++网络以获得第一帧的跟踪预测。SiamFC ++ [9]中的跟踪损耗计算如下：L = Lcls + Lquality + Lreg，（8）其中Lcls是焦点损耗[19]。Lquality是用于质量评估的二进制交叉熵（BCE）损失。Lreg是边界框回归的IoU损失[20]。相对于模板z0的梯度用于根据方程式生成z1。（7）。更新后的模板z？通过应用极少量的Equ迭代获得。（7）。

在本节中，我们首先介绍实现细节。然后，我们将我们的方法与四个跟踪数据集上的最新跟踪器进行比较：OTB2015 [12]，VOT2018 [10]，GOT10 k [8]和TrackingNet [11]。具体来说，OTB2015 [12]包含100个序列，这些序列用不同的属性标记以对跟踪性能进行深入分析。VOT2018 [10]独特地应用了基于重置的方法，并特别选择了60种序列的各种跟踪方案进行评估。GOT-10 k [8]和TrackingNet [11]是两个最近的大规模高多样性数据集。它们都涵盖了火车和测试区中的各种目标类和场景。对于GOT-10 k，训练和测试拆分之间的目标类别没有重叠，从而提高了对看不见的目标类别进行泛化的重要性。一种。实现细节我们使用SiamFC ++ [9]作为基本跟踪器，而主干Siamese网络采用GoogLeNet [21]。除了模型调整组件，我们不执行任何更改。等式中的参数α。（7）设置为0.05。我们的方法是使用PyTorch在Python中实现的。拟议的跟踪器在NVIDIA RTX 2080Ti GPU上以超过80 FPS的速度运行。B.最新技术比较OTB2015：我们使用成功率评估OTB2015数据集上跟踪器的性能。成功率取决于预测边界框和地面实况边界框的交集相交（IOU）。我们将我们的方法与各种跟踪算法进行比较，包括ECO [22]，MDNet [23]，SiamRPN ++ [24]，ATOM [1]和SiamFC ++ \_ GoogLeNet [9]。模板更新的迭代次数设置为16。结果显示在表I中。我们的跟踪器在成功分数方面比在线跟踪器ATOM高出2.8％，这证明了我们方法的强大模型自适应能力。VOT2018：我们将我们的方法与RCO [10]，UPDT [25]，SiamRPN [26]，MFT [10]，LADCF [10]，ATOM [1]，SiamRPN ++ [24]，SiamFC ++ \_ AlexNet [9]和VOT2018上的SiamFC ++ \_ GoogLeNet [9]。使用鲁棒性和准确性度量来比较跟踪器。稳健性表示跟踪失败的次数，而准确性表示跟踪器预测与ground truth框之间的平均重叠。两种度量均合并为单个预期平均重叠（EAO）分数。模板更新的迭代次数设置为2。如图3所示，所有列出的跟踪器的性能均不如我们的算法。GOT-10 k：我们根据[8]使用平均重叠（AO）得分作为绩效指标。我们将我们的方法与CF2 [27]，ECO [22]，CCOT [28]，GOTURN [29]，SiamFC [2]，SiamFCv2 [30]，ATOM [1]，SiamFC ++ \_ AlexNet [9]和SiamFC + + \_GoogLeNet [9]在此数据集上。模板更新的迭代次数设置为2。在图2中，我们可以发现，与列出的最新跟踪器相比，该算法具有更好的跟踪性能。TrackingNet：我们将我们的方法与SiamFC [2]，ECO [22]，MDNet [23]，SiamRPN ++ [24]，ATOM [1]，SiamFC ++ \_ AlexNet [9]和SiamFC ++ \_ GoogLeNet [9]进行了比较。模板更新的迭代次数设置为32。表II显示，我们的跟踪器在精度和标准化精度方面表现最佳，同时保持了非常有竞争力的成功价值。C.消融研究对嘈杂的初始帧的鲁棒性：为了研究该方法在序列中的第一帧与下一帧相比有噪声的情况下的鲁棒性，我们通过以下方法向第一帧添加三种噪声：更改图像亮度，应用高斯模糊以及使用不准确的ground truth边界框注释。我们将γ1∈[0.5，1.5]表示为亮度变化系数，将γ2∈[0，2]表示为模糊半径，并且将γ3∈[0.75，1]表示为不准确的第一帧边界框注释和实数之间的IoU。边界框注释。我们在OTB2015上对基线跟踪器和跟踪器都运行了10次，并随机采样了γ1，γ2和γ3（请参见表III）。与SiamFC ++ \_ GoogLeNet [9]相比，我们的跟踪器在不同的噪声水平下始终表现更好，这表明我们的跟踪器对嘈杂的初始帧具有鲁棒性。基于属性的分析：为了进一步分析跟踪性能，我们还通过对OTB-2015数据集的序列进行基于属性的比较来证明算法的优势（请参见表IV）。在OTB2015中，每个序列都具有11种不同的属性，即：背景杂波（BC），变形（DEF），快动作（FM），照明变化（IV），面内旋转（IPR），低分辨率（LR） ，运动模糊（MB），遮挡（OCC），平面外旋转（OPR），视线外（OV）和缩放比例变化（SV）。与SiamFC ++ \_ GoogLeNet [9]相比，我们的跟踪器在11个属性中的10个上具有更好的性能，这证明了其在具有挑战性的跟踪场景（例如照明变化和运动模糊）中的鲁棒性。

**IV 结论**

在这封信中，我们提出了一种用于孪生追踪器的新型模型自适应方法，实现高速准确跟踪。我们表明，在视觉目标跟踪中具有挑战性的模型自适应任务可以通过使用第一帧中的目标地面真相简单地操纵模板图像的像素来处理。我们的模型调整方法是可插拔的，因为它不会改变基本跟踪器的总体架构。在四个目标跟踪基准上的大量实验结果证明了所提出的模型自适应方法的有效性。