西安电子科技大学学报
*Journal of Xidian University*
ISSN 1001-2400,CN 61-1076/TN

# 《西安电子科技大学学报》网络首发论文

题目：一种添加残差注意力机制的视觉目标跟踪算法
作者：成磊，王玥，田春娜
收稿日期：2019-12-10
网络首发日期：2020-09-22
引用格式：成磊，王玥，田春娜．一种添加残差注意力机制的视觉目标跟踪算法．西安电子科技大学学报.
https://kns.cnki.net/kcms/detail/61.1076.TN.20200922.1629.004.html

# 一种添加残差注意力机制的视觉目标跟踪算法

成 磊[1]，王 玥[1]，田春娜[1]

(1. 西安电子科技大学 电子工程学院，陕西 西安 710071)

**摘要：**由于传统的卷积神经网络结构不能有效地发挥其强大的特征学习和特征表达能力,故提出一种改良的特征提取网络用于视频目标跟踪。在传统特征提取网络的基础上，引入残差网络形式的注意力机制和特征融合策略，同时在网络模型的训练阶段引入基于区域重叠率的损失函数，使得算法模型获得更好的定位效果。实验结果表明，改进算法可以长时间准确地跟踪目标，并且该方法具有泛化能力，对其他基于深度学习的跟踪算法有借鉴意义。

**关键词：**注意力机制；卷积神经网络；残差网络；目标跟踪

**中图分类号**：TP39 **文献标识码**：A

# Residual Attention Mechanism for Visual Tracking

*CHENG Lei*[1], *WANG Yue*[1], *TIAN Chunna*[1]

(1. School of Electronic Engineering, Xidian University, Xi'an 710071, China)

**Abstract:** In recent years, with the development of training data and hardware, a large number of tracking algorithms based on deep learning have been proposed. Compared with the traditional tracking algorithm, tracking algorithms based on deep learning have a great developing potential. However, the traditional convolutional neural network structure cannot effectively perform its powerful feature learning and representation abilities in a tracking task. In this paper, an improved feature extraction network is proposed for video target tracking. Based on the traditional feature extraction network, an attention mechanism and a feature fusion strategy in the form of residual network are introduced. At the same time, a loss function based on the regional overlap rate is introduced in the training stage of the network model, which makes the algorithm produce a better positioning effect. Experimental results show that the improved algorithm can track the target accurately for a long time. Besides, the method has a generalization ability, which can be used for reference for other tracking algorithms based on deep learning.

**Key Words:** Attention Mechanism; Convolutional Neural Network; Residual Network; Object Tracking

**收稿日期：** 2019-12-10
**基金项目：**国家自然科学基金项目（61571354）
**作者简介：**成 磊(1993-)，男，西安电子科技大学硕士研究生，E-mail: lcheng_123@163.com
**通信作者：**田春娜(1980-)，女，教授，博士，Email: chnatian@xidian.edu.cn

目标跟踪技术可以在图像序列中估计目标的运动轨迹，并能够在各种外界或内部因素的干扰下稳定地锁定目标位置。在模式识别和计算机视觉相关领域，视觉目标跟踪技术是一项应用广泛且极具挑战性的工作。智能化的视觉目标跟踪可为我们的日常生活、出行交通等方面提供便利，因而具有非常广泛的实际应用价值，并且相关理论研究可促进人工智能理论的发展[1]。目前主流的跟踪算法大致分为三类：相关滤波算法、深度学习算法、以及两者之间的结合。

基于相关滤波的跟踪方法将输入特征回归为目标高斯分布来训练一个滤波模板，在后续跟踪中通过寻找预测分布中的响应峰值来定位目标的位置。相关滤波器在运算中巧妙应用快速傅立叶变换大幅提升跟踪速度，许多研究人员对相关滤波方法进行扩展，如 Henriques 等[2]提出的核相关滤波算法；Danelljan 等[3]提出的加尺度估计的相关滤波器；宋建锋等[4]针对红外单目标跟踪问题,提出一种多特征的相关滤波目标跟踪算法；王欣远等[5]提出一种融合极限学习机和相关滤波器的鲁棒性目标跟踪算法，解决相关滤波跟踪方法中跟踪结果容易陷入局部最优值以及因引入深度学习带来的特征提取过程过慢的问题。

基于深度学习尤其是卷积神经网络的跟踪方法在国内外目标跟踪领域发展十分迅猛。其跟踪方法大致可以分为基于分类的卷积神经网络目标跟踪和基于回归的卷积神经网络目标跟踪。基于分类的方法充分利用卷积神经网络优秀的特征提取与分类能力，对每一个正负样本进行二分类，比较经典的方法有 Nam 等人提出多域卷积神经网络跟踪[6]，以及提出的树形结构模型[7]跟踪算法，通过构建多个分类器，实现目标跟踪精度的提升;基于回归和相似度匹配的方法利用第一帧或者上一帧图像与候选区域进行回归或相似度计算，最终得到目标位置信息。比较经典的方法有 Held 等[8]提出的基于回归网络的通用目标跟踪算法（Generic Object Tracking Using Regression Networks, GOTURN），Bertinetto 等[9]提出一种全卷积孪生网络（Fully Convolutional Siamese Tracker，SiamFC）用于视频目标跟踪。特别是随着孪生网络的提出，大量基于孪生网络框架的跟踪算法相继出现，打破了相关滤波在目标跟踪领域的垄断地位，可以说真正出现了一种可以和相关滤波相匹敌的目标追踪网络。基于孪生网络的后续改进算法很多，例如，Guo 等[10]提出动态孪生网络，在线学习目标外观变化并且抑制背景信息；Li 等[11]以孪生网络为基础，结合候选区域生成网络，构建一种高性能跟踪网络模型；Wang 等[12]在孪生网络基础上添加多种注意力机制；Zhu 等[13]增强了样本数据，并提出了干扰感知模型应用于孪生网络等。

基于两者之间的结合方法例如 Danelljan 等[14]基于相关滤波方法，提出了一个新的理论框架，在连续空间域中学习可区分性的卷积算子，自然地融合多通道特征，实现目标精确定位，以及后续改进版本[15]等，此类方法主要利用相关滤波结合多层深度特征融合的方法，用以提升跟踪性能。此外，Valmadre 等[16]将相关滤波转换为神经网络的一层，推导前向和后向传播的公式，实现网络的端到端训练，该算法在轻量级的网络中表现出很好的跟踪性能。

近年来，基于深度学习尤其是卷积神经网络的视觉目标跟踪算法性能提升很大，但仍然面临着诸多挑战。例如，传统的特征提取网络在快速移动、模糊、遮挡、旋转、光照变化等复杂的场景条件下，很难做到长时间稳定地跟踪目标。因此，选择更加合理的提取特征网络结构对于基于深度学习的视觉目标跟踪算法效果的影响十分关键。文章提出一种添加残差注意力机制的视频跟踪算法，在跟踪模型中加入一种残差注意力机制的特定网络，生成目标区域显著性图，引导跟踪器关注对结果有帮助的部分。此外，在进行特征提取的网络结构中，通过级联的方式融合网络深层特征与浅层特征，进一步丰富目标位置信息；在模型训练阶段使用了区域重叠率损失函数优化网络输出结果。在 VOT2016 数据集[17]上的实验结果表明，两者对于算法性能均有显著提升。

## 1 GOTURN 算法原理

### 1.1 算法框架

GOTURN 是一种基于卷积神经网络的单目标实时跟踪算法，其框架如图 1 所示。

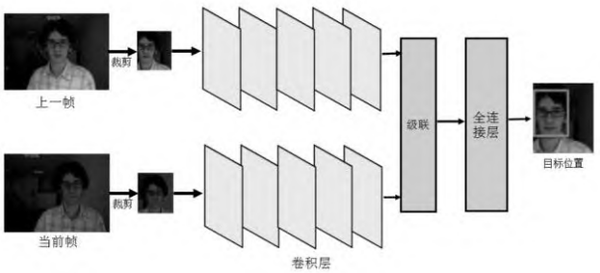


图 1　GOTURN 算法框架

该网络通过神经网络学习相邻帧之间的位移关系，根据上一帧的目标位置回归出当前帧的位置坐标。具体流程为：在当前帧和前一帧分别裁剪出目标区域之后，送入网络进行特征提取，将这些特征级联并输入到全连接层，经过训练的全连接层学习到一个特征比较函数，可通过比较前一帧与当前帧目标特征找到目标的位移信息，从而输出目标的相对位置。在网络的特征提取工作中，作者使用了在 ImageNet 上预训练的 Caffe_model 的前五层作为图像特征的提取结构，具体结构如表 1 所示。

**表 1　GOTURN 特征提取网络结构**

| 层名 | 核大小 | 步长 | 通道数 |
|---|---|---|---|
| Conv1 | 11×11 | 4 | 96 |
| Relu | | | |
| Max pool | 3×3 | 2 | |
| LRN | | | |
| Conv2 | 5×5 | 1 | 256 |
| Relu | | | |
| Max pool | 3×3 | 2 | |
| LRN | | | |
| Conv3 | 3×3 | 1 | 384 |
| Relu | | | |
| Conv4 | 3×3 | 1 | 384 |
| Relu | | | |
| Conv5 | 3×3 | 1 | 256 |
| Relu | | | |
| Max pool | 3×3 | 2 | |

### 1.2 数据处理

作者 Held[8]等通过对视频序列中的目标进行研究发现，图像当前帧目标与上一帧的目标的位置和尺度变化服从拉普拉斯分布。为了让网络学习到更为准确的分布规律，作者对原始的训练的数据集进行了人工扩充，具体分析如下。

$$\begin{aligned} C_x' &= C_x + w\cdot\Delta x \\ C_y' &= C_y + h\cdot\Delta y \\ w' &= w\cdot\gamma_w \\ h' &= h\cdot\gamma_h \end{aligned} \tag{1}$$

其中，$C_x, C_y$ 代表上一帧目标中心位置的横坐标和纵坐标，$w, h$ 分别表示为宽和高，$C_x', C_y', w', h'$ 代表当前帧目标的中心位置坐标和宽高，$\Delta x, \Delta y, \gamma_w, \gamma_h$ 代表目标位置和大小变化的程度。通过实验发现，这四个变化因子都符合拉普拉斯分布，拉普拉斯分布的概率密度函数为：

$$f\left(x|u,b\right)=\frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right) \tag{2}$$

其中，$\mu$ 是位置参数，$b$ 是尺度参数，这个分布的期望 $E\left(x\right)=\mu$，方差为 $D\left(x\right)=2b^2$，分布示意如图 2 所示。

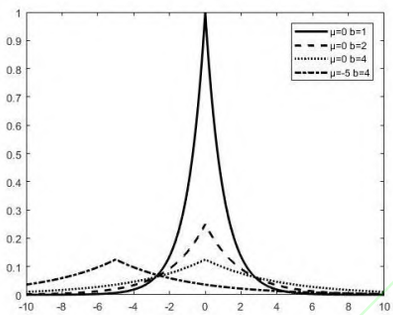


图 2 拉普拉斯分布示意图

通过图 2 可以看出，服从拉普拉斯分布的图像数据更加突出变化较小的范围，因此，在训练集中增加服从上述分布的裁剪图像，得到的扩充数据有助于跟踪器获得更好的性能。

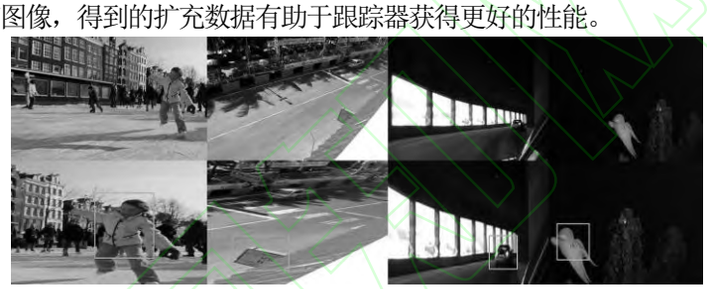

图 3 视频数据扩增

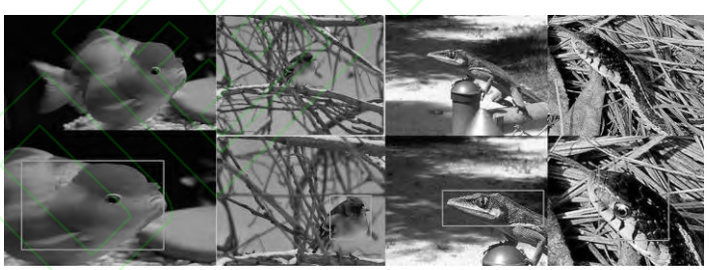

图 4 图像数据扩增

为了使跟踪器学习到更加通用的目标表示和位移特性，在实验中，作者将用于训练的视频数据和图像数据都进行了上述拉普拉斯分布的裁剪扩充。如图 3、图 4 所示，第一行为原始图像数据，第二行为裁剪扩充图像数据。

### 1.3 训练过程

该算法利用视频和静态图像训练，通过优化预测框和目标框之间的 $L_1$ 范数损失训练模型。具体来说，就是把目标真实坐标值 $Y_i$ 与网络输出值 $f\left(x_i\right)$ 的绝对差值的总和 $s$ 最小化。

$$s=\sum_{i=1}^{n}\left|Y_i-f\left(x_i\right)\right| \tag{3}$$

## 2 改进的 GOTURN 算法

针对 GOTURN 算法所存在的问题，文章提出一种添加注意力机制的目标跟踪算法。主要贡献有以下两个方面：

1) 改进了原有特征提取层的网络结构，在原有网络输入的分支上，添加注意力机制模块，再将不同层级的特征图以级联的方式进行融合，提高算法跟踪性能。

2) 在网络训练的过程中，引入基于区域重叠率的损失函数，在网络离线训练的过程中，将原有损失函数替换为基于区域重叠率的损失函数，进一步提高跟踪网络输出的精确度。

整体网络结构如图 5 所示。实验表明，改进后的算法与原有算法相比，性能显著提高，并且能够达到与现阶段部分优秀跟踪算法相当的水平。

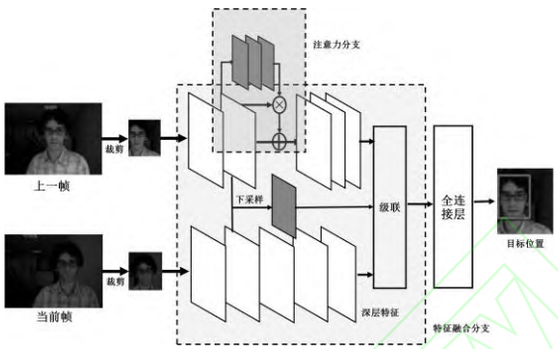


图 5　改进 GOTURN 网络结构

## 2.1 注意力机制融合

深度学习中的注意力机制原理是基于人类的视觉注意力的选择性而提出的，即从众多信息中选择出对当前任务目标更关键的信息。在计算机视觉领域，注意力机制有多种表现形式，主要可以分为空间注意力机制、通道注意力机制和多域注意力机制。为了进一步提升 GOTURN 算法的性能表现，本文通过引入残差网络形式的空间注意力机制，进一步增强深层特征的位置信息。

残差形式的网络结构应用十分广泛，其最初的目的，是解决在训练极深的网络时，网络性能出现退化的现象，而产生这一现象的主要原因就是梯度问题。目前比较经典的残差网络主要有 ResNet[18]及其变体网络。本文主要借鉴残差网络的结构形式，利用残差分支生成空间注意力机制的特征图，这样能够在保证原有特征信息完整的情况下，弱化背景突出前景，引导跟踪器关注对判定跟踪目标有帮助的部分。具体结构形式如图 6 所示。

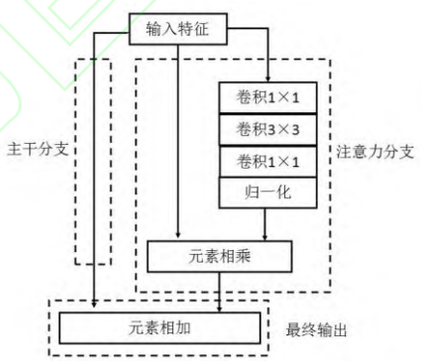


图 6　残差网络结构

首先使浅层的输入特征分别进入主干分支 $T$ 和注意力分支 $M$ 。一方面，经过注意力分支 $M$ 的输入特征经过三个卷积层后，通过 Sigmoid 层归一化，再与原始特征输入元素相乘，得到空间注意力机制特征图 $M_{i,c}(x)$ 。另一方面，主干分支 $T$ 保留原始特征输入 $T_{i,c}(x)$ 不变，与注意力分支 $M$ 的输出 $M_{i,c}(x)$ 相加，得到最终输出结果 $H_{i,c}(x)$，其中，$i$ 代表空间位置，$c$ 代表特征通道的索引。上述残差网络结构用数学方法表示为下式：

$$H_{i,c}(x)=M_{i,c}(x)+T_{i,c}(x) \tag{4}$$

由于注意力分支生成的空间注意力特征图增强了目标区域的显著性，而主干分支则完整保留了原始的输入特征信息，因此，这种残差形式的网络结构有助于提升网络的特征表述能力。

### 2.2 多级特征融合

卷积神经网络在不同层上的特征描述是不同的，在网络的顶层蕴含着更多的语义特征，这些特征对于形变和遮挡有很好的鲁棒性，但是缺乏将目标从一些相似类别场景区分开来的能力。而更底层的特征则携带着包含位置以及其它更多的可区分性的信息，它能够更好的区分前景与背景，将目标从相似的外观上区分开来。因此，为提高 GOTURN 算法的精确度和鲁棒性，可以同时利用这两个层次上的特征来进行改进。

文章采用级联操作融合深层特征与浅层特征，利用预先在 ImageNet[19]上训练的 caffenet 模型[20]的第二个卷积层和第五个卷积层进行特征融合，基本结构如图 5 特征融合分支模块所示。首先，分别选取前一帧网络分支和当前帧网络分支的浅层特征进行级联操作，再将得到的结果下采样到与深度特征相同的尺寸，最后将三个分支的深度特征级联，作为网络下一层的输入。

### 2.3 基于区域重叠率的损失函数构建

在 GOTURN 算法中，作者使用了 $L1$ 范数损失函数，并且通过实验验证分析了使用 $L1$ 损失可以显著地降低总体跟踪误差。由于 $L2$ 惩罚损失输出相对较小，所以网络不会成功地惩罚损失接近但不正确的输出，并且网络通常会输出一个稍微过大或过小的边界框。相比之下，$L1$ 惩罚损失输出较大，针对网络输出的调节更严格，因此网络训练结果较好。但是 $L1$ 损失也有一些缺点，那就是并不能全面的反映出预测输出的好坏，如图 7 所示：在 $L1$ 损失输出相同的情况下，以下三种跟踪情况好坏难以判别。

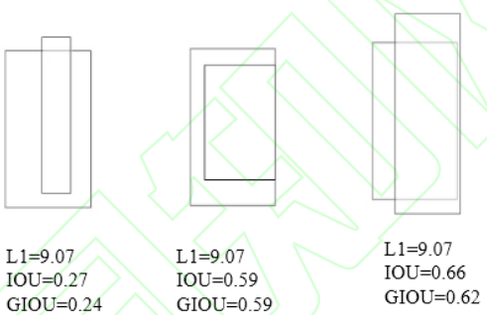


图 7 损失函数对比[21]

在目标跟踪领域，区域重叠率是最为常用的衡量指标之一。如图 8 所示，$a$ 为跟踪到的目标框，$b$ 为真实目标框，区域重叠率计算如公式 5 所示

$$I=\frac{|a\cap b|}{|a\cup b|} \tag{5}$$

区域重叠率 $I$ 通过 $a$、$b$ 之间的交集与并集的比例进行计算，经常用于评价跟踪框的好坏。当跟踪任务中某一帧的重叠率大于设定阈值时，则该帧被视为成功跟踪，总的成功帧数占所有帧数的百分比即为跟踪任务的成功率。作为目标跟踪任务的衡量指标之一，区域重叠若是用作模型训练的损失函数，则相比于 L1 范数损失，描述的更为精细。但是在跟踪任务中，当发生两个目标框没有重叠的情况时，目标的区域重叠率为 0，很近的无交集框和很远的无交集框的输出一样，这样就失去了梯度方向，无法进行优化。Hamid 提出了一种优化区域重叠率的损失函数[21]（Generalized Intersection over Union，GIOU），将其设计为一种可用于目标检测模型中的损失函数，提升了检测模型的效果。

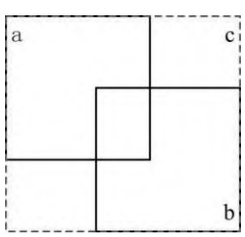


图 8 跟踪框、目标框以及最小包围框

如式（6）所示，$a$、$b$ 分别代表跟踪框和目标框，$c$ 代表 $a$、$b$ 之间最小的包围框，$G$ 代表优化区域重叠率，从公式可以看出，由于 $c$ 的引入，当原有两个目标框没有交集，即区域重叠率为 0 时，GIOU 仍然能够衡量跟踪效果的好坏。因此，在 GOTURN 模型训练过程中，我们使用重叠率损失函数进行模型的优化。

$$\begin{aligned} G &= I - \frac{\left|c / \left(a \cup b\right)\right|}{\left|c\right|} \\ L_I &= 1 - I \\ L_G &= 1 - G \end{aligned} \tag{6}$$

由于在模型训练的初期过程中，预测输出框与标注框的偏差非常大，为使模型加快收敛，改进算法在模型训练的初级阶段，仍然使用 $L1$ 损失函数，当模型训练迭代 45 万次之后，使用 GIOU 损失函数对模型输出做进一步更为精细的调整，实验结果表明，当总迭代次数为 50 万次时，模型训练结果达到最优。

## 3　实验结果与分析

将改进的算法与原有 GOTURN 算法在 VOT2016 基准数据集上进行对比实验，这些实验序列包括目标快速运动、遮挡、旋转、光照变化等挑战。算法采用 caffe 深度学习框架来实现，实验环境是：MATLAB2016b，处理器：Intel i7 8700K，内存：64G，操作系统：Ubuntu16.04。实验利用 VOT2016 数据集的 60 个视频序列对本文算法进行评估。评估所用的评价指标为 VOT 中的精确度和鲁棒性。精确度为成功跟踪的平均重叠精度，数值越大，精确度越高。鲁棒性为跟踪失败的平均次数，数值越小，稳定性越好。此外，我们将实验结果与经典跟踪算法：KCF、STRUCK[22]、DFT[23]、DAT[24]、EBT[25]、ACT[26]、DPT[27]和 MLDF[28]进行对比，对比算法的跟踪结果来自 VOT 数据库网站。

训练时，卷积层部分直接是利用 ImageNet 的预训练参数，不进行微调，以防止过拟合。网络训练的基础学习率为 le-5，其它参数都参照 CaffeNet 中的默认参数。为保证实验结果的可靠性，文章均采用 ImageNet 和 ALOV300++数据集[29]来重新训练网络模型。

### 3.1　消融实验

为了验证改进算法的有效性，首先对添加不同改进策略的算法进行性能测试，实验分为五大类：第一类为 GOTURN 原有算法，第二类为添加注意力机制的算法 GOTURN_atten，第三类为添加多层特征融合策略算法 GOTURN_concat，第四类为使用添加基于区域重叠率损失函数训练的算法 GOTURN_giou，第五类为融合以上多种策略的算法 GOTURN_all。文章主要分析两个跟踪指标：精确度和鲁棒性。当然每个视频都包含有一些属性：遮挡、光照变化、运动变化、尺度变化、摄像机运动，跟踪器在这几个属性上分别测量了精确度和鲁棒性并求了平均，以此作为测度指标，实验数据如表 2 所示，表中对排名前三的实验数据用粗体标注数字表示。

实验结果表明，添加注意力机制后，GOTURN_atten 比基线跟踪算法 GOTURN 的精确度提升了 1.96%，鲁棒性提升了 3.12，性能提升明显，主要由于注意力机制网络的增加，使得改进的特征提取网络对于图像的目标区域作用更加显著，同时对于目标背景区域进行弱化，从而使算法的精度和鲁棒性有所提高；添加多层特征融合策略后，GOTURN_concat 比基线跟踪算法 GOTURN 的精确度提升 0.06，鲁棒性下降 0.88，一方面，添加多层特征融合策略的跟踪器融合了图像区域的浅层特征，使得图像特征的位置信息更加丰富，因此目标定位更加准确，精度有所提高。另一方面，由于更深层次的特征表达具有更高级的语义信息，使得对于目标外观变化等场景更具鲁棒性，浅层特征的融合使得算法的鲁棒性有所降低；使用基于区域重叠率的损失函数对网络进行进一步的调整之后，GOTURN_giou 比基线跟踪算法 GOTURN 的精确度提升 0.97，鲁棒性下降 5.59，表明了基于区域重叠率的损失函数使网络的输出更加精细，因此精度更高，而鲁棒性的下降也反映出该损失函数并不能提升网络本身判别能力。

当同时添加注意力机制以及多层特征融合策略并在模型训练的后期阶段使用基于区域重叠率的损失函数后，GOTURN_all 精确度相对于基线算法 GOTURN 提升了 2.76%，鲁棒性提高 1.4653，通过综合不同

策略的优势，改进算法的性能得到了显著的提高，同时，不同策略所体现出来的一些弊端也得到了一定得弥补，因此整体性能表现最好。

由于实验环境存在差异，报告中提到的帧率（速度）与运行实验的计算机以及配置占用具有很强的相关性，因此帧率（速度）的测试结果仅供参考。总体来讲，由于改进的算法引入了新的网络结构，使得算法复杂度增加，算法运行速度应当有所下降，但仍能满足实时的要求。

表 2 跟踪结果比较（VOT2016 实验结果）

| 算法 | 精确度(%) | 鲁棒性(f) | 帧率(f/s) |
|---|---|---|---|
| GOTURN | 47.27 | **38.01** | **60.0573** |
| GOTURN_concat | 47.33 | 38.89 | **64.4125** |
| GOTURN_atten | **49.23** | **34.89** | 44.6003 |
| GOTURN_giou | **48.24** | 43.60 | **45.3759** |
| GOTURN_all | **49.87** | **36.15** | 11.0855 |

为了定性的比较实验结果，文章通过 VOT 的排名机制对算法性能进行分析。具体来讲就是将跟踪器在不同属性序列上的表现按照精确度和鲁棒性分别进行排名，再进行平均，得到跟踪器的综合排名，图 9 给出了定性的比较结果，其中横坐标表示算法的鲁棒性，纵坐标表示算法的精确度，越靠近右上角的跟踪器性能越好。由图中可以看出，在所比较的算法中，基线跟踪算法 GOTURN 排名最低，添加多种策略的 GOTURN_all 整体表现最好，与定量分析的结果保持一致。

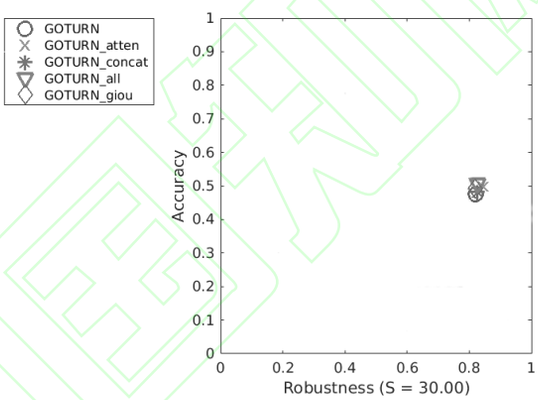


图 9 VOT2016 实验结果对比图

图 10 给出了跟踪器在部分属性视频序列上相比于目标基准框(红色框所示)的跟踪结果(绿色框所示)。在视频序列 iceskater2 中，两个运动员互相遮挡非常严重，并且在运动的过程存在快速移动以及尺度变化的情况。在视频序列 motocross2 中，摩托车手在完成特技动作的过程中，场景光照变化巨大，并且存在目标旋转情况，这些具有挑战性的视频都给目标跟踪任务带来很大困难。由于文中方法添加的注意力机制模块能够突出关注目标区域，并且融合的浅层特征会使得目标定位更加准确，从跟踪结果来看，该算法能够准确的完成跟踪任务。

(a)iceskater2

(b) motocross2

图 10　部分属性视频跟踪结果

表 3、表 4 给出了不同跟踪属性视频序列的跟踪精度和鲁棒性得分，对于精确度而言，改进后的算法在大部分属性上有所提升，尤其在摄像机运动、光照变化、移动、遮挡、尺度变化这几个方面，改进算法 GOTURN_all 的精确度分别提升了 2.93%、8.04%、1.33%、0.44%、7.04%，对于鲁棒性而言，改进的算法在大部分视频属性上也有提高。

表 3　VOT2016 数据库上精确度(%)测试实验结果

| 属性 / 算法 | 未定义属性 | 摄像机运动 | 光照变化 | 移动 | 遮挡 | 尺度变化 | 平均 | 加权平均 |
|---|---|---|---|---|---|---|---|---|
| GOTURN | 50.78 | 47.58 | 56.91 | 43.97 | **40.04** | 46.49 | 47.63 | 47.27 |
| GOTURN_concat | 49.85 | 48.42 | 61.65 | 41.56 | 38.00 | **49.24** | 48.12 | 47.33 |
| GOTURN_atten | **51.08** | **50.71** | **64.45** | **43.99** | 36.29 | **52.49** | **49.84** | **49.23** |
| GOTURN_giou | **50.99** | **49.34** | **62.13** | **44.27** | **40.69** | 47.23 | **49.11** | **48.24** |
| GOTURN_all | **50.81** | **50.51** | **64.95** | **45.30** | **40.48** | **53.53** | **50.93** | **49.87** |

表 4　VOT2016 数据库上鲁棒性(f)测试实验结果

| 属性 / 算法 | 未定义属性 | 摄像机运动 | 光照变化 | 移动 | 遮挡 | 尺度变化 | 平均 | 加权平均 |
|---|---|---|---|---|---|---|---|---|
| GOTURN | **32.0000** | **58.0000** | **4.0000** | **45.0000** | 26.0000 | **19.0000** | **30.6667** | **38.0153** |
| GOTURN_concat | **30.0000** | 60.0000 | **2.0000** | 49.0000 | 26.0000 | **20.0000** | 31.1667 | 38.8867 |
| GOTURN_atten | 34.0000 | **51.0000** | **2.0000** | **35.0000** | **22.0000** | **22.0000** | **27.6667** | **34.8899** |
| GOTURN_giou | 38.0000 | 70.0000 | **2.0000** | 47.0000 | **24.0000** | **22.0000** | 33.8333 | 43.5964 |
| GOTURN_all | **27.0000** | **55.0000** | **4.0000** | **44.0000** | **24.0000** | 23.0000 | **29.6667** | **36.5500** |

### 3.2 算法性能对比

VOT2016 数据集的实验结果如表 5、图 11 所示，对比的八个经典算法中，使用相关滤波进行跟踪的 KCF 精确度最高，总排名第三，而 GOTURN_all 算法精确度比 KCF 高 0.99%，排名第一，在鲁棒性方面，改进的算法 GOTURN_all、GOTURN_atten 相比于基线跟踪算法 GOTURN 均有所提高，而使用特征融合和区域重叠率损失的 GOTURN_concat 和 GOTURN_giou 算法有所下降，但综合整体分析，文章所提出的算法成功地改进了原有算法的特征提取结构，并且通过训练过程中使用区域重叠率的损失函数对模型做进一步精细化调整，使得改进算法性能得到显著提升，在所对比的经典算法中，改进算法也具有一定的优势。由于对比算法的跟踪结果来自 VOT 数据库网站，考虑到算法速度与运行实验的计算机以及配置占用具有很强的相关性，表 5 中并未列举运行速度的对比结果，文中所提出的算法运行速度请参考表 2。

表 5　跟踪结果比较（VOT2016）

| 性能 / 算法 | 精确度(%) | 鲁棒性(f) |
|---|---|---|
| DFT | 44.56 | 59.6116 |
| DAT | 45.82 | **28.3533** |
| EBT | 45.29 | **15.1935** |
| STRUCK2014 | 44.52 | 56.1027 |
| KCF2014 | **48.88** | 38.0820 |
| ACT | 43.72 | 42.6031 |
| DPT | 48.46 | 31.9389 |
| MLDF | 48.73 | **15.0437** |
| GOTURN | 47.27 | 38.0153 |
| GOTURN_atten | **49.23** | 34.8899 |

| | | |
|---|---|---|
| GOTURN_concat | 47.33 | 38.8867 |
| GOTURN_giou | 48.24 | 43.5964 |
| GOTURN_all | **49.87** | 36.1468 |

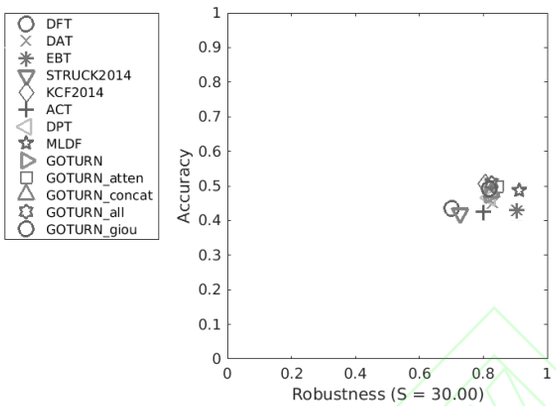


图 11 算法实验结果对比

## 4 总 结

文章深入分析了 GOTURN 算法模型的优缺点，分别从模型训练的网络结构和损失函数两方面对原有算法进行改进，提出一种添加注意力机制以及特征融合策略的目标跟踪算法，在模型训练后期阶段使用了区域重叠率损失函数对网络输出进行优化，使得算法获得更为精细的输出结果。实验结果表明，改进的算法相比与原有算法性有了较为显著的提升。在 VOT2016 标准数据集上的实验结果表明，改进的算法能够更好地解决严重遮挡、光照变化、背景杂乱等场景影响跟踪性能的问题。

### 参考文献

[1] 王海军,张圣燕. 自适应权值卷积特征的鲁棒目标跟踪算法[J]. 西安电子科技大学学报, 2019, 46(1): 117-123. WANG Haijun,ZHANG Shengyan. Robust object tracking via adaptive weight convolutional features[J]. Journal of Xidian University, 2019, 46(1): 117-123.

[2] HENRIQUES J F, CASEIRO R, MARTINS P, et al. High-Speed Tracking with Kernelized Correlation Filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014. 37(3): 583-596.

[3] DANELLJAN M, HAGER G, KHAN F S. Accurate Scale Estimation for Robust Visual Tracking [C/OL]. [2017-10-21]. http://www.cvl.isy.liu.se /research /object /visual tracking /scale visual track /ScaleTracking_BMVC14. pdf.

[4] 宋建锋,苗启广,申猛,权义宁,陈毓生. 多特征融合的相关滤波红外单目标跟踪算法[J]. 西安电子科技大学学报, 2019, 46(5): 142-147. SONG Jianfeng,MIAO Qiguang,SHEN Meng,QUAN Yining,CHEN Yusheng. Algorithm for tracking an infrared single target based on correlation filtering with multi-feature fusion[J]. Journal of Xidian University, 2019, 46(5): 142-147.

[5] 王欣远,肖嵩,李磊,焦玲玲. 融合 ELM 和相关滤波的鲁棒性目标跟踪算法[J]. 西安电子科技大学学报, 2019, 46(1): 57-63. WANG Xinyuan,XIAO Song,LI Lei,JIAO Lingling. Robust target tracking algorithm based on the ELM and discriminative correlation filter[J]. Journal of Xidian University, 2019, 46(1): 57-63.

[6] NAM H, HAN B. Learning Multi-domain Convolutional Neural Networks for Visual Tracking // IEEE Computer Vision and Pattern Recognition. Washington, USA: IEEE, 2016: 4293－4302.

[7] NAM H, BAEK M, HAN B. Modeling and Propagating CNNs in a Tree Structure for Visual Tracking[C /OL]. [2017-10- 21]. https: //arxiv. org /pdf /1608. 07242. pdf.

[8] HELD D, THRUN S, SAVARESE S. Learning to Track at 100 FPS with Deep Regression Networks // IEEE Conference on Computer Vision. Washington, USA: IEEE, 2016: 749－765.

[9] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully Convolutional Siamese Networks for Object Tracking // European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 850－865.

[10] GUO Q, FENG W, ZHOU C, et al. Learning Dynamic Siamese Network for Visual Object Tracking[C]. IEEE International Conference on Computer Vision. 2017: 1763-1771.

[11] LI B, YAN J, WU W, ZHU Z, HU X, High Performance Visual Tracking with Siamese Region Proposal Network[C]. IEEE Computer Vision and Pattern Recognition. 2018: 8971-8980.

[12] WANG Q, TENG Z, XING J, GAO J, HU W. et al. Learning Attentions: Residual Attentional Siamese Network for High Performance Online Visual Tracking[C]. IEEE Computer Vision and Pattern Recognition, 2018: 4854-4863

[13] ZHU Z, WANG Q, LI B, et al. Distractor-aware Siamese Networks for Visual Object Tracking[J]. IEEE Computer Vision and Pattern Recognition. 2018: 101-117.

[14] DANELLJAN M, ROBINSON A, KHAN F S, et al. Beyond Correlation Filters: Learning Continuous Convolution Operators for Visual Tracking // European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 472-488.

[15] DANELLJAN M, BHAT G, KHAN F S, et al. ECO: Efficient Convolution Operators for Tracking[C/OL]. [2017-10-21], https: // arxiv. org /pdf /1611. 09224. pdf.

[16] VALMADRE J, BERTINETTO L, HENRIQUES J F, et al. End to End Representation Learning for Correlation Filter Based Tracking[C/OL]. [2017-10-21]. https: // arxiv. org /pdf /1704 .06036. pdf.

[17] KRISTAN M , LEONARDIS A , MATAS J , et al. The Visual Object Tracking VOT2016 Challenge Results[C]. IEEE International Conference on Computer Vision Workshops, 2015: 1-23.

[18] GUAN HAO，XUE XIANGYANG，AN ZHIYONG．Advances on Application of Deep Learning for Video Object Tracking[J]. Acta Automatica Sinica, 2016, 42( 6) : 834-847.

[19] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 2015, 115(3):211-252.

[20] JIA Y, SHELHAMER E, DONAHUE J, et al. Caffe: Convolutional Architecture for Fast Feature Embedding[J]. 22nd ACM International Conference on Multimedia. Orlando: Computer Science, 2014: 675-678.

[21] REZATOFIGHI H, TSOI N, GWAK J Y, et al. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression[J]. 2019.

[22] HARE S, SAFFARI A, TORR P H S. Struck: Structured Output Tracking with Kernels[C]. IEEE International Conference on Computer Vision, Barcelona, Spain, 2011: 2096-2109.

[23] SEVILLA LAURA L, LEARNED MILLER E. Distribution Fields for Tracking[C]. Computer Vision and Pattern Recognition, 2012: 16-21.

[24] WANG LI JUN, WANG XIAO GANG. et al. STCT: Sequentially Training Convolutional Networks for Visual Tracking[C]. IEEE Computer Vision and Pattern Recognition, 2016: 1373-1381.

[25] FELSBERG M. et al. Enhanced Distribution Field Tracking Using Channel Representations[C]. IEEE International Conference on Computer Vision Workshops, 2013: 121-128.

[26] POSSEGGER H, MAUTHNER T, BISCHOF H. In Defense of Color-based Model-free Tracking. IEEE Conference on Computer Vision and Pattern Recognition, 2015: 2113-2120.

[27] LUKERZIC A, ZAJC L C, KRISTAN M. Deformable Parts Correlation Filters for Robust Visual Tracking[J]. IEEE Transactions on Cybernetics, 2017:1-13.

[28] AKIN O, ERDEM E, ERDEM A, MIKOLAJCZYK K. Deformable Part-based Tracking by Coupled Global and Local Correlation filters[C]. Journal of Visual Communication and Image Representation, 2016: 763-774.

[29] AMOLD W, M SMEULDERS, DUNG M CHU, RITA CUCCHIARA. et al. Visual Tracking: An Experimental Survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(7):1442-1468.