# 面向无人机的轻量级 Siamese 注意力网络目标跟踪

崔洲涓[1,2*]，安军社[1]，张羽丰[1,2]，崔天舒[1,2]

[1]中国科学院国家空间科学中心复杂航天系统电子信息技术重点实验室，北京 100190；

[2]中国科学院大学，北京 100049

**摘要** 随着无人机技术在军事、民用等领域的广泛运用，高精度、低功耗智能无人机跟踪系统的需求也日益增多。针对无人机跟踪任务中目标尺度变化大、视野角度多变、遮挡等问题，提出了一种基于轻量级 Siamese 注意力网络的无人机实时跟踪算法。首先，选取易于部署在嵌入式设备中的轻量级卷积神经网络 MobileNetV2 作为特征提取主干网络；接着，设计通道空间协同注意力模块，增强模型的适应能力与判别能力；然后，搭载区域建议网络，通过互相关获取前景背景分类和边界框回归响应图；最后，加权融合多层响应图，调整候选区域筛选策略，计算得到更加准确的跟踪结果。在无人机跟踪数据集上的仿真实验结果表明，相对于当前主流算法 SiamRPN，该算法跟踪精度提升了 3.5%，能更好地应对复杂多变的场景。同时，在 NIVIDA RTX 2060 GPU 上，跟踪速度达到 60 frame/s。

**关键词** 机器视觉；目标跟踪；Siamese 网络；MobileNet；通道注意力；空间注意力；协同注意力

**中图分类号** TP391；V279 **文献标志码** A **doi**：10.3788/AOS202040.1915001

# Light-Weight Siamese Attention Network Object Tracking for Unmanned Aerial Vehicle

Cui Zhoujuan[1,2*], An Junshe[1], Zhang Yufeng[1,2], Cui Tianshu[1,2]

[1] *Key Laboratory of Electronics and Information Technology for Space Systems, National Space Science Center, Chinese Academy of Sciences, Beijing* 100190, *China*;

[2] *University of Chinese Academy of Sciences, Beijing* 100049, *China*

**Abstract** With the widespread use of unmanned aerial vehicle (UAV) technology in military, civilian, and other fields, the demand for high-precision, low-power intelligent UAV tracking systems is also increasing. Aiming at the problems of scale variation, out-of-view, and occlusion in UAV tracking tasks, a real-time tracking algorithm for UAV based on light-weight Siamese network was proposed. Firstly, the lightweight convolutional neural network MobileNetV2, which is easy to be deployed in embedded devices, is selected as the feature extraction backbone network. Secondly, the channel spatial coordination attention module is designed to enhance the adaptive and discriminative ability of the model. Thirdly, the region proposal network is equipped, and the foreground background classification and boundary box regression response map are obtained through correlation. Finally, the weighted fusion multilayer response map is calculated and proposal region screening strategy is adjusted to obtain more accurate tracking results. Simulation experimental results on the UAV tracking dataset show that the tracking accuracy is improved by 3.5% compared to the current mainstream algorithm SiamRPN, and the algorithm can better cope with complex and changeable scenes. Meanwhile, on the NIVIDA RTX 2060 GPU, the tracking speed achieves 60 frame/s.

**Key words** machine vision; object tracking; Siamese networks; MobileNet; channel attention; spatial attention; coordination attention

**OCIS codes** 150.0155; 150.0135

## 1 引 言

凭借灵敏的反应速度、平稳的悬停能力、轻巧的身型等优点，无人机开始广泛应用于民用、商业、军事领域。随着计算机视觉、人工智能领域的蓬勃发展，在高性能计算技术的推动下，面向无人机应用的目标

**收稿日期**：2020-05-13；**修回日期**：2020-05-29；**录用日期**：2020-06-11

**基金项目**：中国科学院复杂航天系统电子信息技术重点实验室自主部署基金(Y42613A32S)

* **E-mail**：constance669@126.com

跟踪算法展现了巨大的应用前景。在无人机跟踪任务中，由于平台的运动，导致拍摄的视频图像序列视野角度多变、目标尺寸变化大、分辨率低、遮挡等情况更为频繁。因此，构建一个面向无人机应用的高效稳健目标跟踪算法具有重要的研究意义及应用价值。

以核循环结构(CSK)[1]算法、核相关滤波(KCF)[2]算法为代表的传统相关滤波类算法以优异的跟踪速度广泛应用于无人机跟踪任务中。然而方向梯度直方图(HOG)等手工设计的浅层特征需要有针对性地构造旋转不变性、尺度不变性、光照不变性等特性，在面对无人机跟踪的多变场景中表现不够稳健。近年来，深度学习方法在图像分类等计算机视觉任务中展现了巨大的潜力，深度卷积网络以其强大的泛化能力与迁移能力逐步被引入目标跟踪任务中，涌现出一批精度高、稳健性好的算法，诸如多层卷积特征相关滤波(HCFT)[3]算法、多域卷积网络(MDNet)[4]算法、连续卷积相关滤波(C-COT)[5]算法、高效卷积相关滤波(ECO)[6]算法等。利用离线预训练的深度网络进行特征提取，提升了目标的特征表达能力，进而大幅提升了跟踪精度。但是预训练网络结构通常较为庞大，特征维度的升高直接影响了算法的跟踪速度，多数算法即使在GPU下也难以达到实时，更无法适应无人机的应用需求。深度学习方法在跟踪领域的融合不单单局限于作为特征提取网络嵌入传统跟踪框架中，更多是直接训练端到端的跟踪网络，比较有代表性的是基于Siamese框架的跟踪算法。从实例搜索(SINT)[7]算法、全卷积孪生网络(SiamFC)[8]算法开创性地将Siamese网络引入目标跟踪任务，到相关滤波端到端网络(CFNet)[9]算法在网络中加入相关滤波层，再到区域候选网络SiamRPN[10]算法创造性地将区域候选网络(RPN)引入到跟踪领域，形成Siamese-RPN跟踪框架，将原来的相似度计算问题转化为回归以及分类问题，通过大规模数据集进行端到端的离线训练。越来越多的研究者开始基于Siamese框架进行目标跟踪算法的研究[11-12]，相对于基于预训练深度网络的目标跟踪算法，在提升跟踪精度的同时，速度也得到了质的飞跃。

由于无人机跟踪任务中更容易受到尺度变化、遮挡等因素的干扰，这对算法的稳健性提出了更高的要求。另外，无人机平台计算资源相对有限，这为算法的实时性增加了难度。为了更好地满足无人机跟踪任务的需求，本文提出一种基于Siamese-RPN跟踪框架的无人机跟踪算法，主要包括以下三方面的工作：1)构建以轻量级神经网络MobileNetV2[13]为主干网络的Siamese跟踪框架，针对跟踪任务的特点对网络模型进行改进，使其具有更强的特征表达能力；2)引入多种注意力机制，从通道、空间以及协同三个层面提升关键特征的筛选能力，使离线训练的模型在线跟踪时具有更优越的适应能力以及判别能力；3)融合多层前景背景分类和边界框回归的响应图，调整边界框的筛选策略，以获取更精准的定位。在目标跟踪通用数据集中，进行算法性能的验证，与当前主流算法相比，本文算法能够更好地适应尺度变化、视野角度变化、遮挡等多种无人机跟踪场景，取得良好的跟踪精度。

## 2　轻量级Siamese网络跟踪算法

算法的整体框架如图1所示，主要包括融合通道空间协同注意力的轻量级Siamese网络(Siam-CSCAM)和多层区域建议网络(MLRPN)两个部分。Siam-CSCAM的主干网络采用特征表达能力更强、便于在嵌入式设备移植的轻量型网络MobileNetV2，在网络模型中添加灵活轻便的通道空间协同注意力模块(CSCAM)。两路图像序列输入Siam-CSCAM，模板分支与检测分支分别提取出特征图，再分别输入至MLRPN。MLRPN包含上层的分类分支以及下层的回归分支，将Siam-CSCAM中模板分支与检测分支输出的特征图分别在分类分支以及回归分支中进行互相关操作。

### 2.1　融合通道空间协同注意力的轻量级Siamese网络

#### 2.1.1　基于Siamese网络的目标跟踪

基于Siamese网络的算法将目标跟踪任务转化为相似性度量问题，将视频序列第一帧 $\boldsymbol{z}$、后续帧中的候选区域 $\boldsymbol{x}$ 分别作为模板分支、检测分支的输入图像，通过权值共享的特征提取网络 $\varphi(\cdot)$ 映射到特征空间，学习度量函数 $f(\boldsymbol{z},\boldsymbol{x})$ 来比较模板图像和候选区域搜索图像之间的相似度，返回响应图。分数越高，二者相似度越高。

$$f(\boldsymbol{z},\boldsymbol{x})=\varphi(\boldsymbol{z})*\varphi(\boldsymbol{x})+b\cdot\boldsymbol{I}, \tag{1}$$

式中：$*$ 代表互相关运算；$b\cdot\boldsymbol{I}$ 表示在响应图中每个位置的取值。

#### 2.1.2　轻量级网络MobileNet

深度学习网络模型在视觉跟踪任务中应用的效果越来越好，伴随而来的是神经网络结构逐渐复杂，体积逐渐增大，对硬件资源的需求也逐渐增多。神经网络大都是在具有强大的浮点运算能力、性能优越的服务器上运行，普通PC难以承担如此繁重的

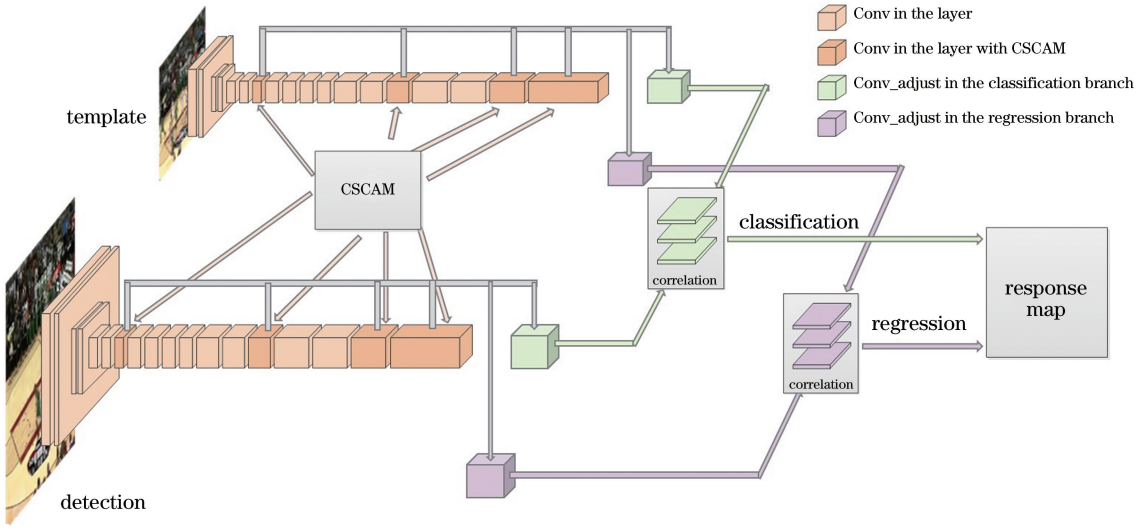


图 1　融合通道空间协同注意力的轻量级 Siamese 网络

Fig. 1　Framework of Siamese network with channel spatial coordination attention module

计算压力，资源受限的移动平台更是难以部署。因此，深度学习领域的研究者们也开始致力于在保证模型准确率的同时，促使神经网络向小型化、高速化发展。一个研究方向是对训练好的复杂模型进行压缩得到小模型；另一个研究方向是直接设计训练小模型，近年来出现的许多具有代表性的轻量级网络模型，如 SqueezeNet、ShuffleNet、NasNet、MnasNet 和 MobileNet 等，加速了神经网络模型在移动终端、嵌入式设备的应用。

MobileNetV1[14]是谷歌提出的轻量级 CNN 网络，设计了深度可分离卷积，将标准卷积的运算过程分离为深度(DW)卷积与点(PW)卷积。DW 卷积为输入特征图的每个通道分配一个单独的卷积核进行卷积运算；PW 卷积使用 1×1 卷积对 DW 卷积的运算结果进行标准卷积运算。深度可分离卷积通过分解将总计算量降低至标准卷积的 $1/N+1/D_K^2$。同时，由于将几乎所有的计算都集中于 1×1 卷积操作，所以利用现有卷积实现算法时不需要在内存中重新排序，直接加快计算速度。

MobileNetV2 相对于 MobileNetV1 引入了倒残差和线性瓶颈，卷积模块如图 2 所示。倒残差结构与 ResNet 中的残差结构相似，细节稍有不同。残差结构是先通过 1×1 卷积将特征图的通道数缩减，使得后续 3×3 标准卷积的计算量减少，再经过 1×1 卷积扩增并和输入相加。倒残差结构用 3×3 DW 卷积代替 3×3 标准卷积，大幅度降低计算量，因此可以在 DW 卷积之前增加一层 1×1 PW 卷积提升通道数，进而提升网络模型效果。在 3×3 DW

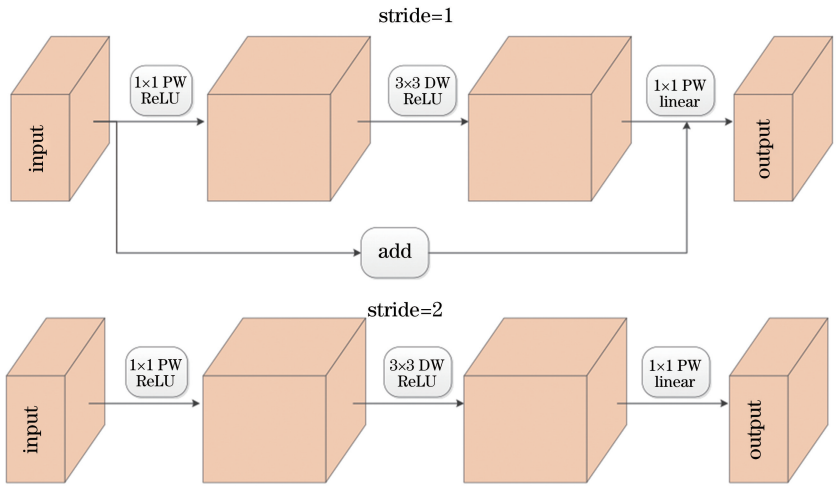


图 2　MobileNetV2[13]卷积模块框图

Fig. 2　Convolutional blocks of MobileNetV2

卷积之后，再通过瓶颈设计经由 1×1 PW 卷积降低维度并和输入相加。

为了便于后续注意力模块的设计与分析，从输入输出协同工作角度，将模块的卷积过程从通道和空间两个层面进行抽象可视化，如图 3 所示，用圆点表示卷积的输入输出，二者之间用线条连接，线条表示二者之间的依赖关系。第一层 1×1 PW 卷积增加了通道维度，中间层 3×3 DW 卷积在通道与空间中独立地执行，最后一层 1×1 PW 卷积减少了通道维度。

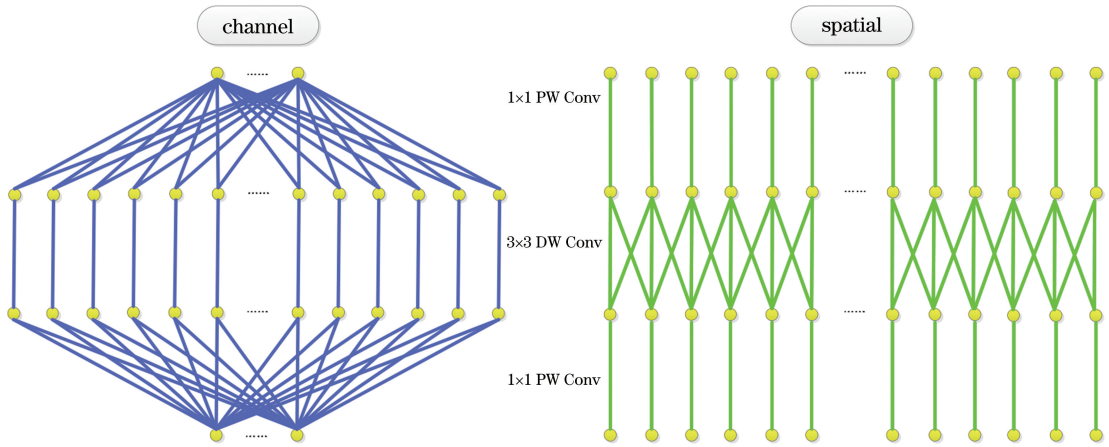


图 3　MobileNetV2 卷积过程抽象图

Fig. 3　Convolution process abstract graph of MobileNetV2

在神经网络中，诸如 ReLU 等非线性激活层通过引入非线性在高维空间有效增加特征的非线性表达，使模型具有更强的拟合能力。但 ReLU 的引入同时会带来低维数据坍塌问题，这导致了 MobileNetV1 训练后一些 DW 卷积核容易置空。原空间的低维特征通过变换 $T$ 映射至嵌入空间后，加入 ReLU 进行处理，若再通过 $T$ 的逆变换，将处理后的特征映射至原空间，信息会损失一部分。嵌入空间的维度越高，则信息损失越小；维度越低，则信息损失越大。因此 MobileNetV2 在倒残差结构中，第一层 PW、中间层 DW 使用 ReLU，最后一层 PW 卷积后使用线性激活函数替换 ReLU，称之为线性瓶颈。

相对于主流的神经网络模型，MobileNetV2 拥有更小的体积、更少的计算量、更高的精度，易于部署在诸如 FPGA、DSP 等无人机图像处理平台。而在目标跟踪任务中，特征提取网络的选择直接影响跟踪速度与性能。因此，本文选取 MobileNetV2 作为 Siamese 网络中的特征提取主干网络。同时考虑到后续互相关以及响应图的融合等操作，对其进行修改，以检测分支为例，具体网络结构如表 1 所示。

从表 1 可以看出，主要修改包括以下三个方面：

1）原始的 MobileNetV2 总步长为 32，为了适应网络应用于跟踪的精确定位，将网络的总步长限

表 1　基于 MobieleNetV2 的 Siamese 网络结构

Table 1　Architecture of Siamese network based on MobieleNetV2

| Layer name | Input | Operator | Expansion factor | Channel | Repeat time | Stride | CSCAM |
|---|---|---|---|---|---|---|---|
| Input | 255×255×3 | Conv2d | — | 32 | 1 | 2 | No |
| Layer1 | 127×127×32 | Bottleneck | 1 | 16 | 1 | 1 | No |
| Layer 2 | 127×127×16 | Bottleneck | 6 | 24 | 2 | 2 | No |
| Layer 3 | 63×63×24 | Bottleneck | 6 | 32 | 3 | 2 | Yes |
| Layer 4 | 31×31×32 | Bottleneck | 6 | 64 | 4 | 1 | No |
| Layer 5 | 31×31×64 | Bottleneck | 6 | 96 | 3 | 1 | Yes |
| Layer 6 | 31×31×96 | Bottleneck | 6 | 160 | 3 | 1 | Yes |
| Layer 7 | 31×31×160 | Bottleneck | 6 | 320 | 1 | 1 | Yes |
| Output | 31×31×320 | — | — | — | — | — | — |

制为 8，保持后四段 Layer4～Layer7 中的尺寸不变，缩小了网络模型的总步长；后四个卷积模块的分辨率保持变化一致。

2）为了提升网络的性能，在 Layer3、Layer5、Layer6、Layer7 后融入了注意力模块 CSCAM，但并未改变特征图尺寸；

3）为了便于后续分类分支与回归分支的互相关计算以及响应图的融合，在 CSCAM 的输出后，均增加一层 1×1 的卷积层 Conv_adjust，用于调节通道数。

### 2.1.3　通道空间协同注意力模块

相对于相关滤波类的跟踪算法，基于 Siamese 网络的跟踪算法为了提升速度，采取了离线训练的网络，摒弃了在线训练的环节，这就需要网络一方面需要对各种场景变化表现稳定，可以将目标代表性、本质性的特征抽象提取出来；另一方面又需要对不同目标的差异表现敏感，能够对目标的细节有所提炼。即网络需要具备强大的特征提取能力，自带判别作用。然而：1）从特征提取的层面。MobieleNetV2 等基于大规模分类数据集离线训练的通用网络，对图像的每个位置有较为平均的关注度，而跟踪任务需要根据不同的目标关注不同的特征，因此离线训练的网络并不能完全适应在线跟踪。2）从相似度判别层面。由(1)式可以看出，在互相关计算的整个过程中，不同通道、不同位置对于相似度计算的贡献是平均的，这极大限制了网络的特征提取能力与判别能力，因此需要在相似度计算中进行加权突出或筛选目标的重要信息，抑制无关的细节信息，提升网络的判别能力。

为了使离线训练的网络具有更强的特征提取能力、适应能力以及判别能力，受到人类视觉系统中注意力机制的启发，将其引入提升网络的性能。人类视觉系统的一个重要特性是不试图同时处理整个场景。相反，为了更好地捕捉视觉结构，人类利用一系列的局部关注，有选择地聚焦于显著的部分。类似地，在深度网络模型中加入注意力模块能够有效突出感兴趣的区域，为了便于直观地体现注意力模块的作用，利用类激活热力图[15](Grad-CAM)对网络进行可视化，如图 4 所示。越敏感的位置温度越高，越不敏感的位置温度越低。图 4(a)为未加注意力模块的网络模型可视化结果，图 4(b)为加入注意力模块后的网络模型可视化结果。

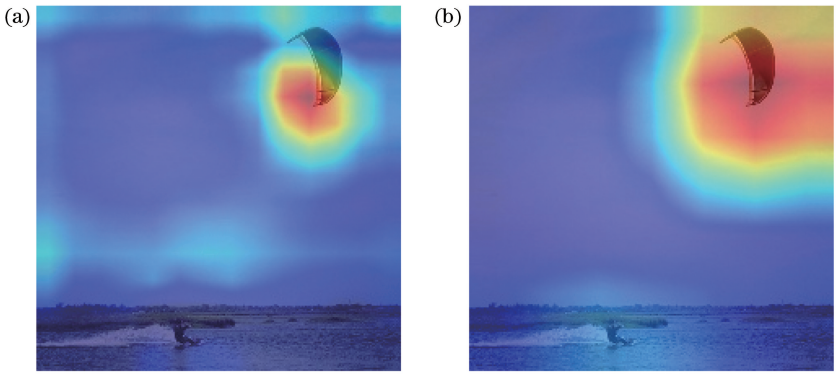


图 4　Grad-CAM 网络可视化结果。(a)无注意力模块；(b)有注意力模块

Fig. 4　Grad-CAM network visualization results. (a) No attention module; (b) with attention module

为了对不同跟踪目标特征图中不同通道、不同空间的重要性区别性地突出，同时为了更好地利用模板图像以及搜索图像的背景信息，根据 Siamese 网络的结构特点，设计通道空间协同注意力模块，如图 5 所示。

通道注意力通过对通道之间的依赖性进行建模，从语义层面学习特征之间的关联，对特征进行优化，激活与目标更相关的通道特征，去除冗余特征，使特征表达更加凝练、精确度更高。设经过 MobileNetV2 网络模板分支、检测分支提取到的特征图 $\varphi(\boldsymbol{z})\in\mathbb{R}^{C\times H_{\mathrm{T}}\times W_{\mathrm{T}}}$、$\varphi(\boldsymbol{x})\in\mathbb{R}^{C\times H_{\mathrm{D}}\times W_{\mathrm{D}}}$，首先分别经过全局平均池化、全局最大池化，前者用于凝聚空间维度，获取每个通道的全局信息，后者补充提供突显目标独有特征的更精细表达，二者联合能够提取到更为丰富的特征。然后输入全连接共享层，包括输入层、隐藏层、输出层，其中为了不过多地增加计算，隐藏层需要进行降维处理，然而伴随一定的全局信息损失。为了达到二者的平衡，本文参考卷积块注意力模块(CBAM)[16]，将隐藏层的通道数降低为输入层的 1/16。经由激活函数，输出通道注意力权重 $\boldsymbol{A}_{\mathrm{C}}\{\varphi(\boldsymbol{z})\}$、$\boldsymbol{A}_{\mathrm{C}}\{\varphi(\boldsymbol{x})\}$。最后与输入特征进行元素级乘法，得到通道注意力特征图 $\varphi_{\mathrm{C}}(\boldsymbol{z})$、$\varphi_{\mathrm{C}}(\boldsymbol{x})$。

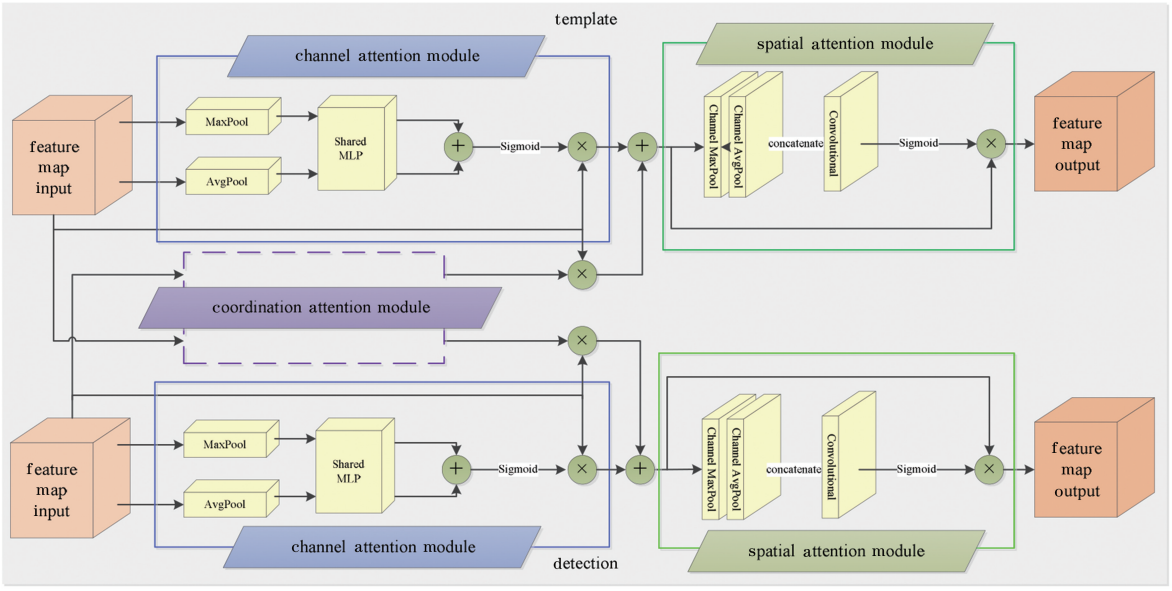


图 5　通道空间协同注意力模块

Fig. 5　Channel spatial coordination attention module

在 Siamese 网络中，模板分支与检测分支通常都是独立操作，在此设计协同注意力模块，将各自分支编码融入另外的分支，使背景信息得到充分利用。为了便于和特征进行矩阵乘法，需要对其进行整形和调整维度，输出协同注意力权重 $\boldsymbol{A}_{\mathrm{I}}\{\varphi(\boldsymbol{x})\}$、$\boldsymbol{A}_{\mathrm{I}}\{\varphi(\boldsymbol{z})\}$。经由通道注意力、协同注意力模块后，两个分支的权重进行融合，得到 $\varphi'(\boldsymbol{z})$、$\varphi'(\boldsymbol{x})$。

空间注意力更集中于位置的描述，能够构建特征图中不同位置的相互关系，对位置进行加权融合，对通道注意力进行补充。将特征图 $\varphi'(\cdot)$ 沿着通道维度压缩，分别得到通道最大池化、通道平均池化，拼接后再进行卷积操作，经由激活函数，输出空间注意力权重 $\boldsymbol{A}_{\mathrm{S}}(\cdot)$。与输入特征 $\varphi'(\cdot)$ 进行元素级乘法，得到最终注意力特征图 。

$$\varphi''(\boldsymbol{z})=\boldsymbol{A}_{\mathrm{S}}\{\varphi'(\boldsymbol{z})\}\otimes\varphi'(\boldsymbol{z}),\tag{2}$$

$$\varphi''(\boldsymbol{x})=\boldsymbol{A}_{\mathrm{S}}\{\varphi'(\boldsymbol{x})\}\otimes\varphi'(\boldsymbol{x})。\tag{3}$$

## 2.2　多层区域建议网络

区域建议网络包括分类分支与回归分支。Siamese 网络中模板分支、检测分支输出的各层特征图通过调整层的卷积操作调整为分辨率统一、通道数相同的特征图，再分别输入区域建议网络的分类分支以及回归分支中。设分类分支输入的第 $q$ 层调整后的特征图为 $[\varphi^{(q)}(\boldsymbol{z})]_{\mathrm{cls}}$、$[\varphi^{(q)}(\boldsymbol{x})]_{\mathrm{cls}}$；回归分支输入的第 $q$ 层调整后的特征图为 $[\varphi^{(q)}(\boldsymbol{z})]_{\mathrm{reg}}$、$[\varphi^{(q)}(\boldsymbol{x})]_{\mathrm{reg}}$，分别进行逐通道相关操作，减少计算成本和冗余参数。最后，再将多层特征图经由区域建议网络得到的输出进行加权融合。

$$A_{\mathrm{cls}}^{(q)}=[\varphi^{(q)}(\boldsymbol{x})]_{\mathrm{cls}}*[\varphi^{(q)}(\boldsymbol{z})]_{\mathrm{cls}},\tag{4}$$

$$A_{\mathrm{reg}}^{(q)}=[\varphi^{(q)}(\boldsymbol{x})]_{\mathrm{reg}}*[\varphi^{(q)}(\boldsymbol{z})]_{\mathrm{reg}}。\tag{5}$$

在对每个候选区域进行前景或背景分类时，由于同一目标可能同时存在于多个重叠的矩形框内，因此通常采用非极大值抑制(NMS)进行剔除，得到更为精准的位置信息。设置分类得分较高的矩形框为抑制窗口，在筛选过程中，以矩形框交并比(IoU)为指标来判断当前矩形框与抑制窗口是否重叠，超过设定 IoU 阈值时，则当前矩形框被剔除，否则被保留。

$$s_i=\begin{cases}0, & \mathrm{IoU}(b_i,b_{\max})\geqslant T\\ s_i, & \text{otherwise}\end{cases},\tag{6}$$

式中：$b_i$ 为第 $i$ 个矩形框；$s_i$ 为第 $i$ 个矩形框对应的得分，是设定的 IoU 阈值。

然而这个策略存在的问题是：分类得分高的矩形框未必是所有矩形框中最优的，如果其他的矩形框因为与抑制窗口重叠率高而被剔除，则很有可能造成跟踪漂移，因此对其进行改进，

$$s_i=\begin{cases}s_i-s_i[\mathrm{IoU}(b_i,b_{\max})], & \mathrm{IoU}(b_i,b_{\max})\geqslant T\\ s_i, & \text{otherwise}\end{cases}。\tag{7}$$

当前矩形框与抑制窗口 IoU 超过阈值时，得分会线性衰减，而不是直接置 0，从而达到保留矩形框的问题。重叠面积越大，得分衰减越大，重叠面积未超过阈值时，得分并无影响。在不增加额外计算量的同时使结果更为精准。

## 3　实验与分析

### 3.1　实验平台及参数配置

本文算法实验平台硬件配置为：CPU Intel(R) CoreTM i7-9700，基础频率 3.0 GHz，睿频加速频率 4.7 GHz，内存 16 G；GPU NVIDIA GeForce RTX-2060，内存 6 G。

训练数据选自 ImageNet VID[17] 和 Youtube-BB[18]，分别覆盖了大约 4000 个逐帧注释的视频和超过 10 万个每 30 帧注释的视频，通过特定的比例组合，前者包含更多的细粒度信息而后者包含粗粒度信息。从相同的视频序列中随机选取两帧，并将它们组合成一对模板图像和检测图像，作为 Siamese 网络的输入，学习如何测量用于视觉跟踪的一般对象之间的相似性的一般概念。使用 MobileNetV2 预训练模型初始化卷积层，采用随机梯度下降(SGD)，训练时学习率为 $10^{-4}$～$10^{-6}$。整个训练过程包含 100 多个阶段，每个阶段由 6000 对样本组成。每次计算 8 对样本的平均损失值。

### 3.2　轻量级注意力模块有效性对比实验

为了验证注意力模块的功能性，在 OTB-2015[19] 上进行了对比实验。为了直观地分析注意力模块对算法性能的改进，将 11 种不同属性下的跟踪精度与绘图成功率汇总以图表形式呈现，分别如图 6、图 7 所示。其中横轴为 OTB-2015 中 11 种不同的视频属性，不同的颜色代表不同的注意力模块组合，从图中可以看出，在 11 种不同的视频属性中，添加 CSCAM 的性能均比未添加注意力模块有了较

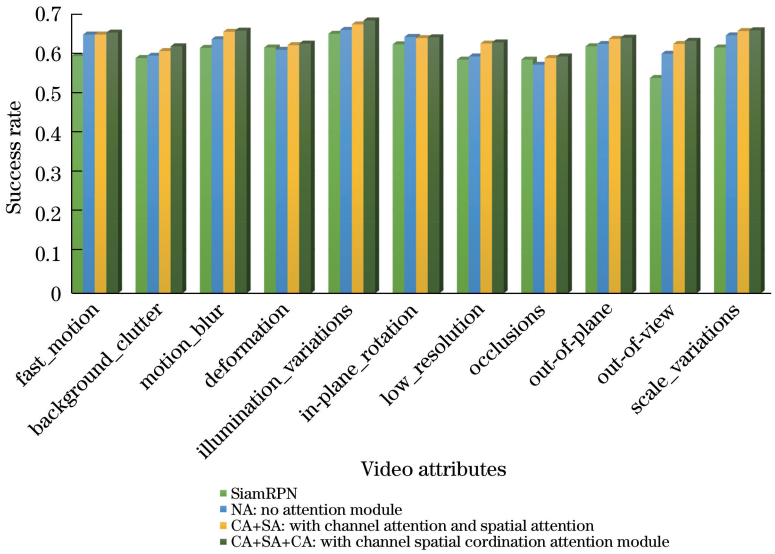


图 6　不同注意力模块组合在 OTB-2015 中成功率对比图

Fig. 6　Success rate comparison of different attention module combinations on OTB-2015

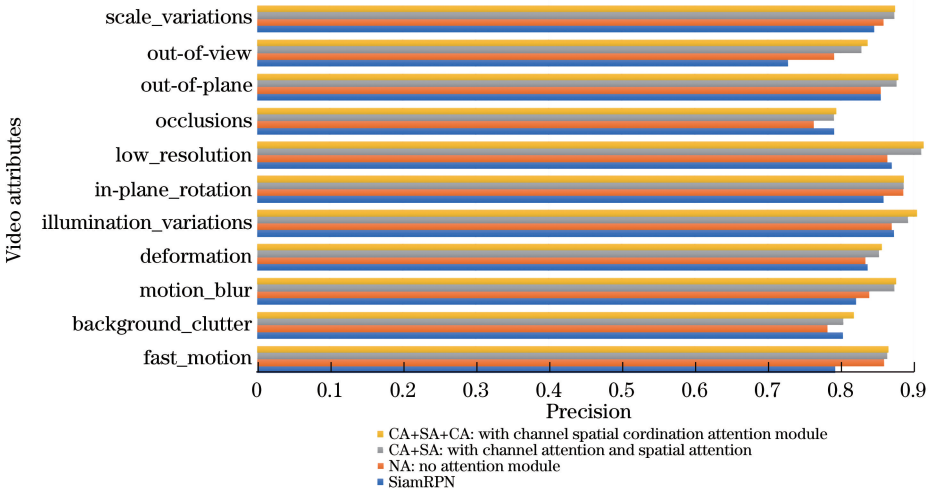


图 7　不同注意力模块组合在 OTB-2015 中跟踪精度对比图

Fig. 7　Precision comparison of different attention module combinations on OTB-2015

大的提升，特别是在背景复杂、运动模糊、形变、低分辨率、遮挡、超出视野等情况下。这证明注意力模块的融合，使得算法的判别能力更强，能够更好地适应环境的变化。

### 3.3　基于 UAV123 的实验

将具有代表性的 9 种算法（本文算法、ECO[6]、SiamRPN[10]、SRDCF[20]、ASLA[21]、SAMF[22]、DSST[23]、DCF[2]、KCF[2]），在囊括 123 组视频序列、覆盖 12 种属性的高分辨率无人机跟踪数据集（UAV123）[24]上评估算法的性能。UAV123 数据集包括 3 个子集，其中子集 1 包含 103 个用专业级无人机拍摄的不同物体的序列，跟踪高度在 5～25 m 之间，视频序列的帧速率在 30～96 frame/s 之间，分辨率在 1280×720～4096×2160 之间；子集 2 包含 12 个从安装在小型低成本无人机（UAV）上的摄像头捕获的序列，由于视频传输带宽有限，这些序列的质量和分辨率较低，并且包含合理数量的噪声；子集 3 包含 8 个合成序列，由 UAV 模拟器捕获[24]。以一次通过评估 OPE（One Pass Evaluation）[19]作为跟踪算法准确性能的评价标准。

#### 3.3.1　定性分析

选取九种算法在部分视频序列上的表现进行定性分析，跟踪效果如图 8 所示，每个视频涵盖三种以上的视频属性，每种颜色的矩形框代表不同算法的跟踪框，本文算法用红色表示。

图 8(a) car6_5 序列中，目标发生尺度变化，由于相机的移动导致目标的长宽比不断变化，偶尔还伴随着相机镜头受到路灯的遮挡，基于 Siamese-RPN 框架的网络具备了多尺度检测的能力，因此本文算法以及 SiamRPN 算法可以持续地适配目标的变化。

图 8(b) car17 序列中，目标快速移动，导致目标尺度、拍摄视角急剧变化，候选区域过小则无法快速捕捉目标，过大则会带来背景的干扰，这对特征提取能力提出了更严峻的挑战，可以看出，未提取深度特征的算法已经漂移。

图 8(c) person9 序列中，随着目标的移动，偶有超出视野的情况发生，部分算法受到周围相似物的干扰，直接出错，本文算法可以在目标重新出现时摒除相似物影响，迅速定位目标。

图 8(d) person1_s 序列中，目标快速移动，偶有被障碍物遮挡，同时光照不断发生变化，复杂的场景对算法提出了极大的挑战，本文算法的特征提取网络具备较强的判别能力与适应能力，能够较好地提取特征。

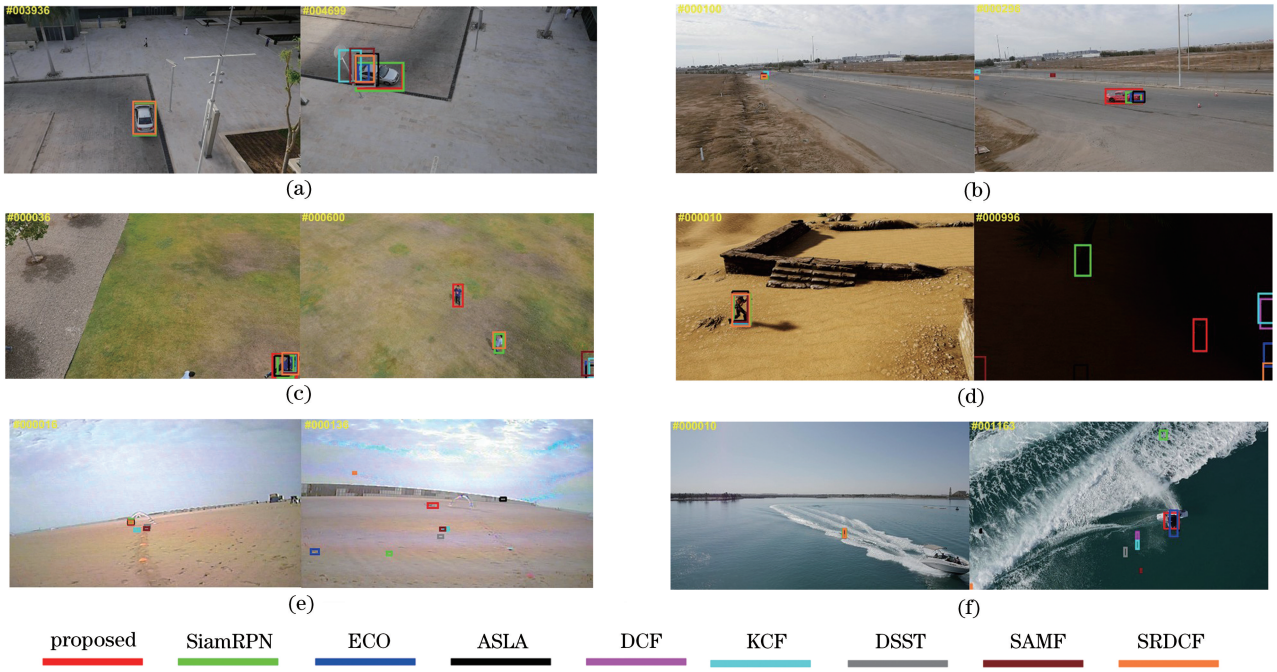


图 8　九种跟踪算法在不同视频序列上的定性结果显示。(a) car6_5；(b) car17；(c) person9；(d) person1_s；(e) uav4；(f) wakeboard6

Fig. 8　Qualitative results of the nine tracking algorithms on different video sequences. (a) car6_5; (b) car17; (c) person9; (d) person1_s; (e) uav4; (f) wakeboard6

图 8(e)uav4 序列分辨率较低，特征提取难度加大，且随着目标的移动，甚至出现花图，部分算法在视频初始已发生漂移，本文算法一直稳定跟踪目标。

图 8(f)wakeboard6 序列目标自身不断旋转且持续快速运动，背景以及拍摄角度均在持续变化，只有提取深度特征的算法可以稳定跟住目标，但部分算法可能受到背景干扰发生漂移。

也存在个别问题序列，如图 9 所示。uav1_1 序列是整个数据集中包含属性最全面的视频，分辨率较低，且随着目标的移动，不断遇到遮挡、超出视野等情况发生，部分算法在跟踪开始就已发生漂移，本文算法一直保持在稳定的水平，直到 1336 帧开始，目标超出视野，再到 1350 帧出现时，本文算法由于受到背景中相似物的干扰，跟错目标，直至视频结束。说明本文算法在超出视野之后再跟踪，需要适当调整检测机制。

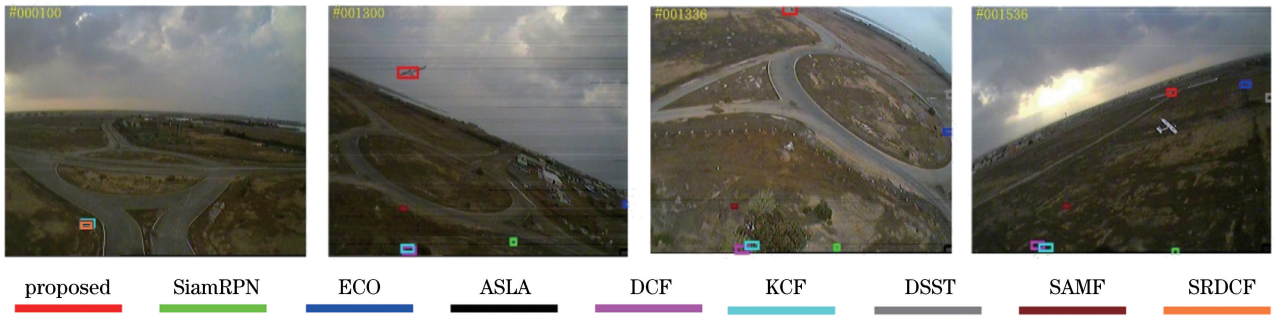


图 9　问题序列 uav1_1 结果图

Fig. 9　Results of problem sequence on uav1_1

3.3.2　定量分析

跟踪算法的评估指标主要包括中心位置误差和覆盖率。前者是指跟踪结果与真实目标的中心位置之间的欧氏距离；后者是指跟踪结果与真实目标的重叠率，分别体现在精度图以及成功率图中，通过设定一定阈值对跟踪结果进行判定。

1) 九种算法在 UAV123 上的成功率曲线以及精度曲线如图 10 所示。本文算法成功率为 0.604，跟踪精度达到 0.803，相对于 SiamRPN 算法分别提高了 4.7%、3.5%。证明本文算法在 Siamese 网络框架的基础上，提取了更深的网络特征，同时融入了注意力模块，使网络提取到适应能力更强的特征，提升了算法的总体精度与稳健性。

2) 针对不同的视频属性类别，九种算法的跟踪成功率定量分析结果如图 11 所示。本文算法在所有属性中均排名第一。

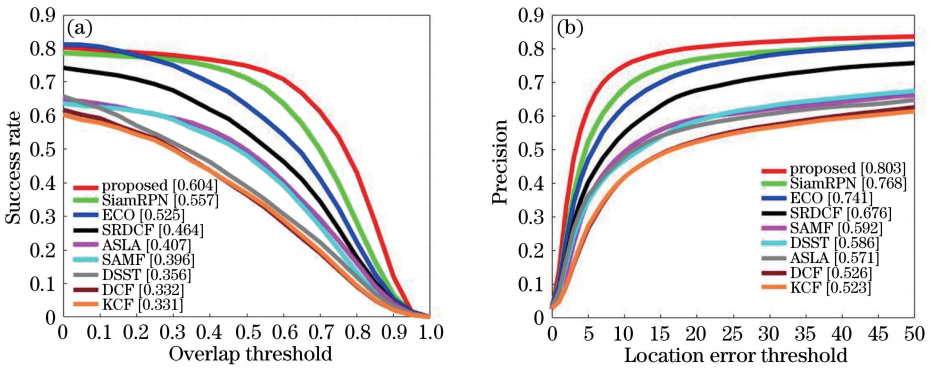


图 10　跟踪算法基于 OTB-2015 的结果。(a)成功率曲线图；(b)跟踪精度曲线图

Fig. 10　Results of the tracking algorithms on OTB-2015. (a) Success plot; (b) precision plot

3) 针对不同的视频属性类别，九种算法的跟踪精度定量分析结果如图 12 所示。本文算法在尺度变化、宽高比变化、完全遮挡、部分遮挡、超出视野等 9 个属性中排名第一。

4) 在 UAV123 中，尺度变化与宽高比变化两种属性的视频占比最高，分别占 89%、55%，也恰好是面向无人机的目标跟踪中面临最多的挑战场景，因此以 car6_5、wakeboard6 为例进行进一步地分析。如图 13 所示，尺度变化范围为 0～4.5，宽高比变化范围为 0.5～3.5，属于变化较为剧烈的视频序列，中心位置误差(CLE)基本保持在 20 个像素点的阈值以内。在尺度变化值首次超过3.5时，中心位

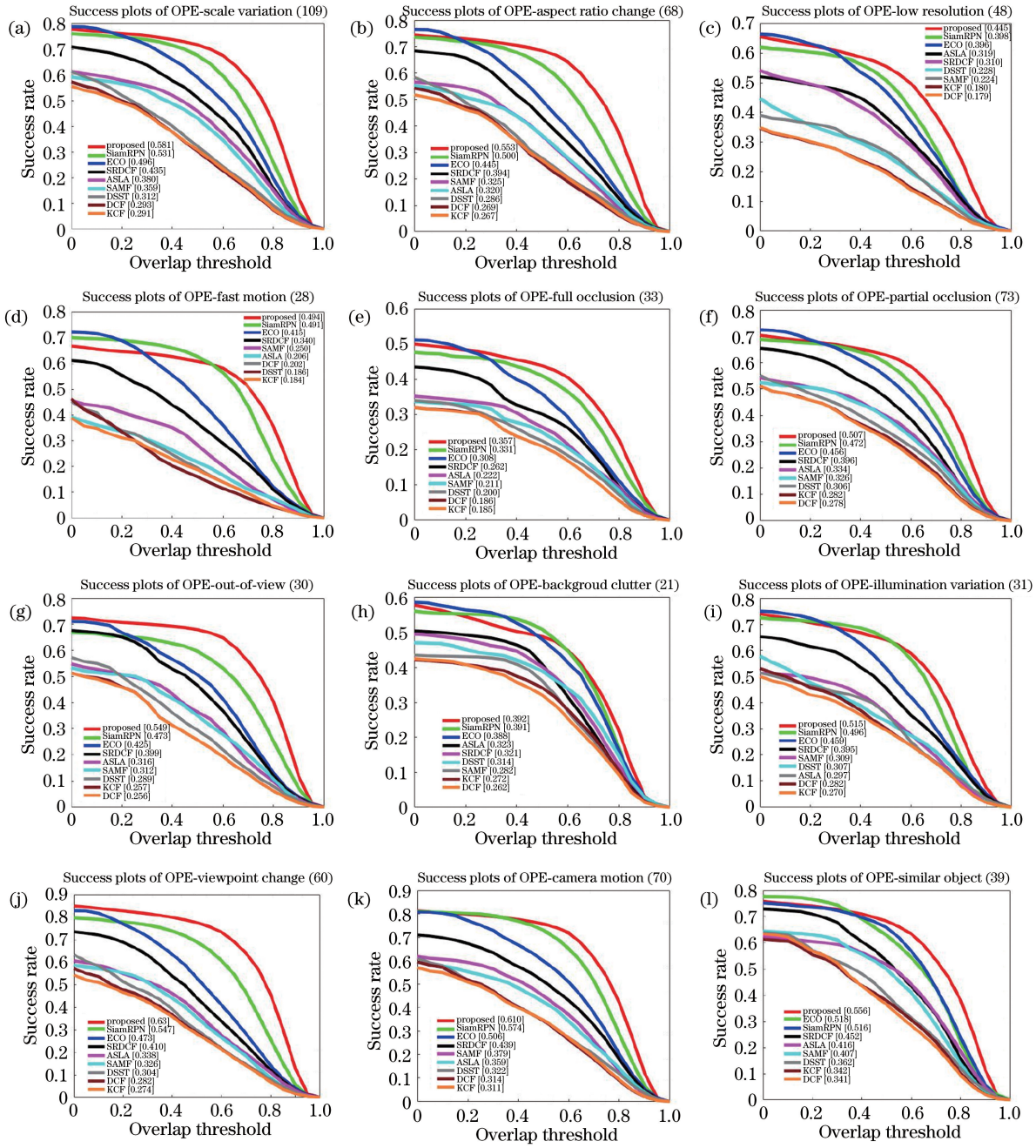


图 11　不同属性视频跟踪成功率曲线图。(a)尺度变化;(b)宽高比变化;(c)低分辨率;(d)快速运动;(e)完全遮挡;(f)部分遮挡;(g)超出视野;(h)背景干扰;(i)光照变化;(j)视角变化;(k)相机移动;(l)相似物体

Fig. 11　Tracking success plots of different attributes videos. (a) Scale variation; (b) aspect ratio change; (c) low resolution; (d) fast motion; (e) full occlusion; (f) partial occlusion; (g) out-of-view; (h) background clutter; (i) illumination variation; (j) viewpoint change; (k) camera motion; (l) similar object

置误差达到峰值,但由于算法极强的适应能力,当尺度变化达到最大值,超过 4 时,中心位置误差反而回落,未发生漂移。

5) 相较于 ECO 算法,本文算法跟踪精度在背景干扰场景仍有一点差距。然而基于深度网络特征的 ECO 算法在 GPU 上的速度为 8 frame/s[3],本文算法的平均速度达到 60 frame/s,在跟踪速度上有了大幅提升。

综上,本文算法在 UAV123 的 12 种属性的视频序列中表现稳定,通过设计通道空间协同注意力模块,以及多层响应图融合取得了良好的跟踪效果,在提高跟踪精度与稳健性的同时,保证了实时的跟踪速度,可以更好地适应尺度变化、视野角度变化、遮挡等无人机跟踪场景。

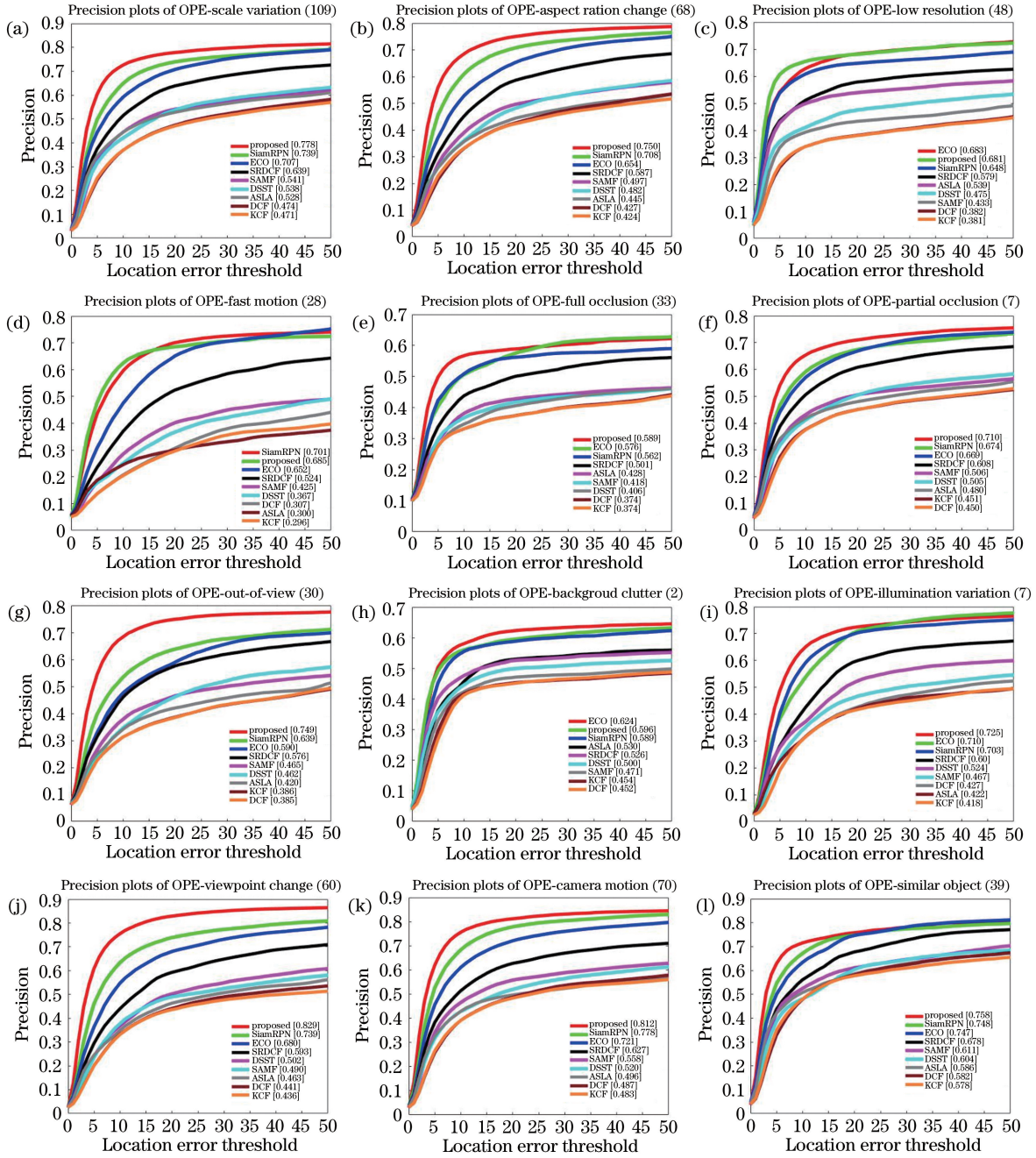


图 12　不同属性视频跟踪精度曲线图。(a)尺度变化；(b)宽高比变化；(c)低分辨率；(d)快速运动；(e)完全遮挡；(f)部分遮挡；(g)超出视野；(h)背景干扰；(i)光照变化；(j)视角变化；(k)相机移动；(l)相似物体

Fig. 12　Tracking precision plots of different attributes videos. (a) Scale variation; (b) aspect ratio change; (c) low resolution; (d) fast motion; (e) full occlusion; (f) partial occlusion; (g) out-of-view; (h) background clutter; (i) illumination variation; (j) viewpoint change; (k) camera motion; (l) similar object

## 4　结　　论

本文提出一种嵌入轻量级网络 MobileNetV2 作为特征提取主干网络，融合通道空间协同注意力模块，结合区域建议网络的端到端跟踪算法，并通过实验对算法进行了验证。实验表明：1)通道空间协同注意力模块的融入，显著了提升网络模型的特征提取能力、适应能力与判别能力。基于 OTB-2015 的对比实验表明，由于从通道、空间、协同三个层面进行了注意力调整，使得网络性能在不同属性的视频中有了显著的提升，特别是在背景复杂、运动模糊、形变、低分辨率、遮挡、相似物干扰等情况下，网

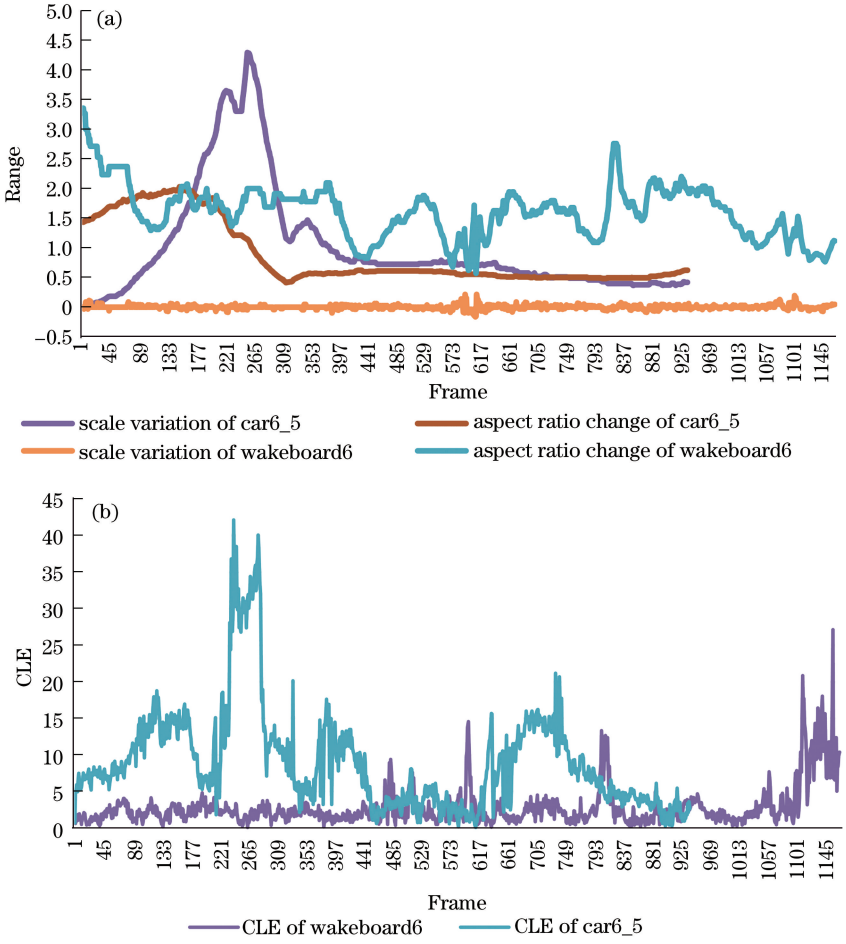


图 13　部分视频序列的定量分析。(a)尺度变化与宽高比变化;(b)中心位置误差

Fig. 13　Quantitative analysis of some video sequences. (a) Scale variation and aspect ratio change; (b) CLE

络学习到更具代表性的特征,区分性更强。2)在无人机目标跟踪数据集 UAV123 上的实验表明,与当前主流算法相比,本文算法成功率为 0.604,跟踪精度达到 0.803,在目标外观变化、相似物干扰、目标遮挡等无人机常见的复杂场景下,表现更为稳健。3)在 NVIDIA RTX 2060 GPU 下的平均跟踪速度可达到 60 frame/s 。

## 参　考　文　献

[1] Henriques J F, Caseiro R, Martins P, et al. Exploiting the circulant structure of tracking-by-detection with kernels[J]. Computer Vision-ECCV 2012, 2012: 702-715.

[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, 2015, 37(3): 583-596.

[3] Ma C, Huang J B, Yang X K, et al. Hierarchical convolutional features for visual tracking[C]//2015 IEEE International Conference on Computer Vision (ICCV). December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 3074-3082.

[4] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 4293-4302.

[5] Danelljan M, Robinson A, Shahbaz Khan F, et al. Beyond correlation filters: learning continuous convolution operators for visual tracking [J]. Computer Vision-ECCV 2016, 2016: 472-488.

[6] Danelljan M, Bhat G, Khan F S, et al. ECO: efficient convolution operators for tracking[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6931-6939.

[7] Tao R, Gavves E, Smeulders A W M. Siamese instance search for tracking [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). June 27-30, 2016, Las Vegas,

NV, USA. New York: IEEE Press, 2016: 1420-1429.

[8] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional Siamese networks for object tracking[M]//Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016: 850-865.

[9] Valmadre J, Bertinetto L, Henriques J, et al. End-to-end representation learning for correlation filter based tracking [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5000-5008.

[10] Li B, Yan J J, Wu W, et al. High performance visual tracking with Siamese region proposal network [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, USA, 2018: 8971-8980.

[11] Qiu Z L, Zha Y F, Zhu P, et al. Visual tracking algorithm based on online feature discrimination with siamese network[J]. Acta Optica Sinica, 2019, 39(9): 0915003.
仇祝令, 查宇飞, 朱鹏, 等. 基于孪生神经网络在线判别特征的视觉跟踪算法[J]. 光学学报, 2019, 39(9): 0915003.

[12] Chen Z W, Zhang Z X, Song J, et al. Tracking algorithm for siamese network based on target-aware feature selection[J]. Acta Optica Sinica, 2020, 40(9): 0915003.
陈志旺, 张忠新, 宋娟, 等. 基于目标感知特征筛选的孪生网络跟踪算法[J]. 光学学报, 2020, 40(9): 0915003.

[13] Sandler M, Howard A, Zhu M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[EB/OL]. (2019-03-21)[2020-05-13]. https://arxiv.org/abs/1801.04381.

[14] Howard A G, Zhu M L, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17)[2020-05-13]. https://arxiv.org/abs/1704.04861.

[15] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[J]. International Journal of Computer Vision, 2020, 128(2): 336-359.

[16] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[C]//European Conference on Computer Vision, 2018: 3-19.

[17] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[18] Real E, Shlens J, Mazzocchi S, et al. YouTube-BoundingBoxes: a large high-precision human-annotated data set for object detection in video[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 7464-7473.

[19] Wu Y, Lim J, Yang M H. Object tracking benchmark [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834-1848.

[20] Danelljan M, Häger G, Khan F S, et al. Learning spatially regularized correlation filters for visual tracking[C]//2015 IEEE International Conference on Computer Vision (ICCV). December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 4310-4318.

[21] Jia X, Lu H C, Yang M H. Visual tracking via adaptive structural local sparse appearance model [C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. June 16-21, 2012, Providence, RI, USA. New York: IEEE Press, 2012: 1822-1829.

[22] Li Y, Zhu J K. A scale adaptive kernel correlation filter tracker with feature integration[J]. Computer Vision-ECCV 2014 Workshops, 2015: 254-265.

[23] Danelljan M, Häger G, Shahbaz Khan F, et al. Accurate scale estimation for robust visual tracking [C]//Proceedings of the British Machine Vision Conference 2014. Nottingham. British Machine Vision Association, 2014: 1-11.

[24] Mueller M, Smith N, Ghanem B. A benchmark and simulator for UAV tracking[M]//Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 445-461.