文章编号:1007－6735(2024)02－0120－09　DOI: 10.13255/j.cnki.jusst.20230517003

基于显式运动建模的视频伪装目标检测

肖　涛[1]，　章　超[2,3]，　傅可人[1,4]
（1. 四川大学 计算机学院，成都 610065；2. 四川警察学院，泸州 646000；3. 智能警务四川省重点实验室，泸州 646000；
4. 四川大学 视觉合成图形图像技术国家级重点实验室，成都 610064）

摘要：目前的视频伪装目标检测方法通常采用隐式运动建模或直接输入存在噪声的离线光流图来获取运动线索，这会影响模型性能。为了解决这一问题，提出一种新的基于显式运动建模的视频伪装目标检测框架，称为 SMHNet。首先，该框架将显式运动建模与伪装目标检测联合在同一个框架中进行学习。然后利用特征双向更新模块实现两个分支的双向交互更新，相互补充、优化和纠错，输出光流估计结果和目标检测图。此外，为了解决缺少光流真值图这一问题，采用自监督策略对显式运动建模分支进行监督。在两个数据集上的对比实验结果表明，SMHNet 有效地提高了视频场景中伪装目标检测的性能。

关键词：视频伪装目标检测；显式运动建模；光流；自监督
中图分类号：TP 391　　**文献标志码：**A

Explicit motion handling for video camouflaged object detection

XIAO Tao[1],　ZHANG Chao[2,3],　FU Keren[1,4]
(1. *School of Computer Science, Sichuan University, Chengdu 610065, China*; 2. *Sichuan Police College, Luzhou 646000, China*;
3. *Intelligent Policing Key Laboratory of Sichuan Province, Luzhou 646000, China*; 4. *National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University , Chengdu 610064, China*)

Abstract: Earlier video camouflaged object detection methods often exploit motion cues by implicit motion modeling or directly inputting offline optical flow maps with noise, which affects model performance. To address the effective utilization of motion cues, an explicit motion modeling framework for video camouflaged object detection, called SMHNet, was proposed. First, an explicit motion modeling branch and a camouflaged object detection branch were jointly learned in the same framework. Then, the two branches were updated bidirectionally using a bidirectional feature updating module. The two branches performed mutual optimization and error correction to output optical flow estimation results and object detection maps. In addition, to address the lack of ground truth optical flow, a self-supervised strategy was adopted to supervise the explicit motion modeling branch. Comparison experiments on two benchmark datasets show that SMHNet effectively improves the performance of video camouflaged object detection.

收稿日期: 2023−05−17
基金项目: 国家自然科学基金资助项目（62176169）；智能警务四川省重点实验室资助项目（ZNJW2022KFMS001）
第一作者: 肖　涛（1998−），女，硕士研究生．研究方向：伪装目标检测．E-mail：xtt4483@163.com
通信作者: 章　超（1986−），男，副教授．研究方向：计算机视觉、图像处理．E-mail：galoiszhang@163.com

Keywords: *Video camouflaged object detection*; *explicit motion handling*; *optical flow*; *self-supervision*

伪装目标检测（camouflaged object detection, COD）旨在检测和分割那些与背景具有高度内在相似性的“隐匿/隐藏物体”。该技术在医学分割[1]、工业检测[2]、军事[3-4]等领域有着广泛应用。随着深度学习技术的兴起，目标检测领域取得了较大的发展。然而，伪装目标的检测仍然面临着许多挑战，例如模糊的目标边界、难以定位的目标位置、过小的伪装目标、过于杂乱的背景等。

伪装目标检测任务可以分为基于图像的 COD[5-8] 和基于视频的 COD[9-12] 两类。基于图像的 COD 方法通常借助静态线索来实现检测。但是，由于伪装目标物体往往在纹理、颜色或边缘上与背景高度相似，仅使用这些静态的外观线索来识别目标十分困难。有趣的是，自然界存在着一种有助于解决这个问题的现象——捕食行为。通常情况下，捕食者可以很容易地发现那些移动的猎物。这是因为猎物的运动打破了自己的伪装状态。受此启发，基于视频的伪装目标检测方法利用视频提供的时序信息获取可靠的运动线索，来更好地分割出伪装在背景中的目标。根据利用运动线索的方式，视频伪装目标检测（video camouflaged object detection, VCOD）方法可以分为两类：第一类[11-12] 利用现成的运动估计模型生成离线光流，并将其直接输入到网络中，作为识别伪装物体的运动线索。然而，运动估计本身也是一项极具挑战性的任务，尤其是估计伪装目标的运动时，由于伪装目标很难与背景区分开来，所以现成的运动估计模型生成的光流是有噪声，甚至存在错误的，这会误导视频伪装目标检测，进而限制检测性能。第二类[9] 通过构建隐式运动建模与目标检测的联合框架，来获取运动线索。然而，这种策略缺乏对所学运动线索的显式评估和约束，这在一定程度上限制了模型的能力。

为了解决上述问题，提出一个基于显式运动建模的 VCOD 框架，提供了简单而强大的全新 VCOD 方案。搭建了伪装目标检测和显式运动建模联合学习的框架，可分为显式运动建模分支和伪装目标检测分支，通过这两个分支的联合学习，模型可以获得更加准确和鲁棒的伪装目标检测能力。同时，为了实现两个分支之间的双向交互，提出特征双向更新模块，使得两个分支实现通信、共同优化、相互促进，从而提高检测性能。此外，为缓解光流真值图难获得的问题，提出使用自监督策略来监督显式运动建模分支的学习。实验结果表明，本方法在两个 VCOD 数据集上实现了最优的性能。

1　基于显式运动建模的 VCOD 方法

1.1　方法总体框架

本研究的主要任务是检测出视频中的伪装目标。为了提高伪装目标检测精度，提出一个基于显式运动建模的 VCOD 框架，称为 SMHNet。图 1 是该方法的总体框架，其遵循典型的编码器–解码器结构。图中：GRU（gate recurrent unit）为门控循环单元；GRA（group-reversal attention）为分组反转注意；NCD（neighbor connection decoder）为近邻连接解码器。以相邻帧对 (I_t, I_{t+1}) 为输入，记 I_t 为参考帧，I_{t+1} 为相邻帧。相邻帧对经过权值共享的编码器提取特征后，得到两个多层次的特征序列。随后将相邻两帧的特征对输入到显式运动建模分支提取运动特征，将参考帧的特征输入到伪装目标检测分支，两分支通过特征双向更新模块实现交互。显式建模分支为伪装目标检测分支提供目标的运动特征信息，伪装目标检测分支为显式运动建模分支提供目标特征信息。随后显式运动建模分支迭代更新光流，两分支在迭代过程中双向交互、相互促进、共同优化，最终输出参考帧 I_t 的二进制分割掩码 $\boldsymbol{M}_t$ 和光流估计 $\boldsymbol{g}$。

1.2　特征提取

为了公平对比，参照当前最优的 VCOD 方法 SLTNet[9]，采用视觉金字塔 PVT-V2[13] 作为编码器。具体地，给定相邻帧对 (I_t, I_{t+1})，将其变换尺寸至 352×352，再通过视觉金字塔 PVT-V2 提取出不同尺度的特征序列 $\{\boldsymbol{f}_n^i \in R^{C_i \times H/2^{i+1} \times W/2^{i+1}}, n \in \{t, t+1\}, i = 1, \cdots, 4\}$，其中 W，H 和 C 分别为特征的宽、高和通道数。同时，与文献 [14] 类似，只采用有着丰富语义信息的上 3 层特征（$\boldsymbol{f}_t^2, \boldsymbol{f}_t^3, \boldsymbol{f}_t^4$）进行伪装目标检测分支中的外观建模。特别地，为了减轻网络后续操作的计算负担[7]，通过卷积将上

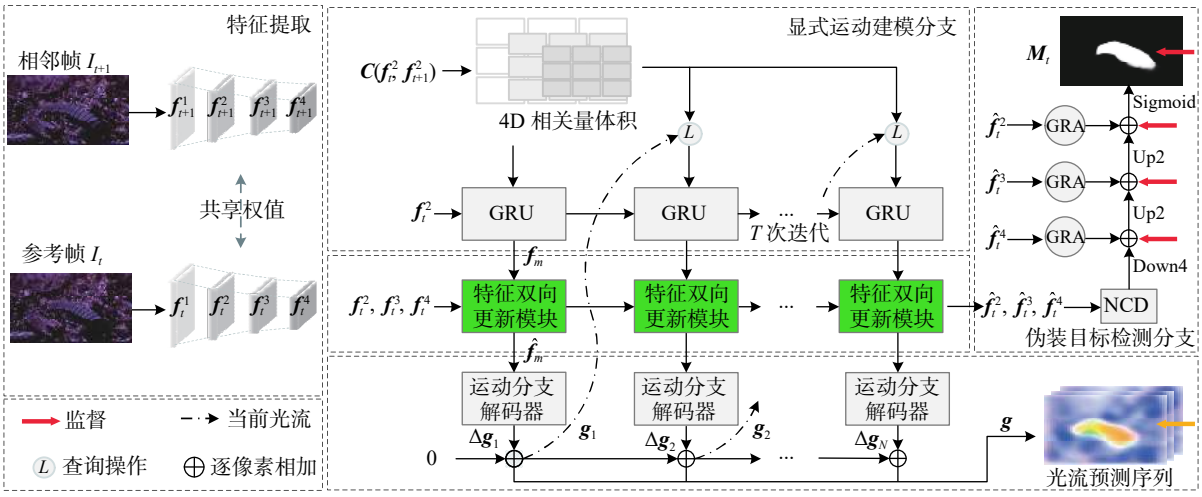

图 1 方法总体框架

Fig.1 Framework of the proposed method

2 层特征的通道数减少到 32。最终得到包含 3 个不同尺度的特征序列$\{\boldsymbol{f}_n^i, n\in\{t,t+1\}, i=2,3,4\}$，记作$\{\boldsymbol{f}_n^i\}$。

1.3 显式运动建模分支

为了实现优越的运动建模性能，引入光流估计经典模型 RAFT[15] 作为显式运动建模分支的框架。值得注意的是，RAFT 是全监督的光流估计方法，训练依赖光流真值进行监督。而本文的显式运动建模分支在引入 RAFT 后，将其与伪装目标检测分支进行双向交互，并借助自监督策略进行训练，进一步提升了显式运动建模分支的运动估计效果，也解决了伪装场景中光流真值图获得难的问题。RAFT 通过计算两帧间所有像素对的相关量构建相关量金字塔，迭代查询金字塔，利用查询到的相关量更新光流估计。选取相邻两帧特征序列中的第 2 层用于显式运动建模，即 $(\boldsymbol{f}_t^2, \boldsymbol{f}_{t+1}^2)$，其空间维度均为 44×44。首先构建一个两帧间的 4D 相关量金字塔。相关量由两帧特征向量之间的点积计算式为

$$\boldsymbol{C}\left(\boldsymbol{f}_t^2, \boldsymbol{f}_{t+1}^2\right)_{uvnk} = \sum\nolimits_c \left(\boldsymbol{f}_t^2\right)_{uvc} \cdot \left(\boldsymbol{f}_{t+1}^2\right)_{nkc} \tag{1}$$

式中：(u, v)和(n, k)分别代表成对特征的位置索引；c 表示特征向量内的通道索引。相关量分别经过 1，2，3，4 个步长为 2 的池化层，得到相关量金字塔$\{\boldsymbol{C}1, \boldsymbol{C}2, \boldsymbol{C}3, \boldsymbol{C}4\}$，其中，各层相关量维度为：1936×1×44×44，1936×1×22×22，1936×1×11×11，1936×1×5×5。以$\boldsymbol{C}1$，$\boldsymbol{C}2$为例简单理解相关量，参考帧有 1936(44×44)个位置，$\boldsymbol{C}1$反映了每个位置与相邻帧的 44×44 个区域对应的相关量，$\boldsymbol{C}2$反映了每个位置与相邻帧的 22×22 个区域对应的相关量。特别地，该特征金字塔在整个框架中只需要计算一次，迭代过程要使用相关量，只需进行查询。

每次迭代，首先用当前光流估计值查询相关量金字塔，得到相关量。给定当前光流估计$\boldsymbol{g}(\boldsymbol{g}_1, \boldsymbol{g}_2)$，将参考帧的每个像素位置$x(u,v)$映射到相邻帧的对应点$x'(u+\boldsymbol{g}_1(u,v), v+\boldsymbol{g}_2(u,v))$。在$x'$附近定义一个整数偏移量且 L1 距离在 r 单位半径内的局部邻域，作为查询相关量金字塔的索引为

$$N(x')_r = \{x' + \mathrm{d}x | \mathrm{d}x \in \mathbb{Z}^2, \|\mathrm{d}x\|_1 \leqslant r\} \tag{2}$$

式中，依据文献 [15]，r 设定为 4。将参考帧所有像素 x映射到相邻帧得到的局部邻域坐标集合$\{N(x')_r\}$，记为 N。N 实际中可表示为维度是 1936×9×9×2 的张量形式，其中 9×9×2 即表示参考帧某个位置映射到相邻帧的局部邻域坐标。以该坐标集合 N 作为网格索引来查询相关量金字塔获得相应的相关量特征，查询公式为

$$\begin{aligned} corr = &\mathrm{Concat}(\mathrm{BS}(\boldsymbol{C}1, N), \mathrm{BS}(\boldsymbol{C}2, N), \\ &\mathrm{BS}(\boldsymbol{C}3, N), \mathrm{BS}(\boldsymbol{C}4, N)) \end{aligned} \tag{3}$$

式中：$corr$代表本次查询得到的相关量特征；$\boldsymbol{C}1$，$\boldsymbol{C}2$，$\boldsymbol{C}3$，$\boldsymbol{C}4$分别为特征金字塔不同层次的 4D 相关量；邻域 N 作为双线性采样函数的 grid 参数；Concat表示按通道串联操作；BS代表双线性采样函数和 reshape 函数的组合，其中双线性采样函数以局部邻域为索引，用于查询相关量。以查询相关量金字塔的第一层 $\boldsymbol{C}1$ 为例，对于参考帧上的某一位置，利用 9×9×2 的坐标偏移信息到 44×44

的相关量图上进行双线性插值，最后得到 9×9 个对应的相关量图。参考帧所有位置查询相关量后，输出 1936×1×9×9 的 4D 张量，经由 reshape 函数的维度二次调整，该 4D 张量被调整为 1×44×44×81，即为 BS 函数的输出。由于金字塔 BS 后的输出维度相同，因此可以直接进行串联。

其次，将当前的光流估计、查询得到的相关量和上下文特征按通道拼接，送入更新模块来提取运动特征$\boldsymbol{f}_m$。具体地，基于 GRU 的更新模块公式为

$$z_t=\sigma(\mathrm{Conv}([h_{t-1},x_t],W_z)) \tag{4}$$

$$r_t=\sigma(\mathrm{Conv}([h_{t-1},x_t],W_r)) \tag{5}$$

$$\tilde{h}_t=\tanh(\mathrm{Conv}([r_t\odot h_{t-1},x_t],W_h)) \tag{6}$$

$$h_t=(1-z_t)\odot h_{t-1}+z_t\odot\tilde{h}_t \tag{7}$$

式中：x_t表示由当前光流、光流查询 4D 相关量体积得到的相关量和参考帧层次化特征序列的第 2 层特征串联的特征；h_{t-1}表示上一次迭代输出的运动特征；[,] 表示通道串联操作；W_z，W_r，W_h为可学习的权重参数；z_t，r_t，$\tilde{h}_t$表示门控循环单元 GRU 计算过程的中间特征；h_t代表本次 GRU 循环的输出状态，则当前运动特征为$\boldsymbol{f}_m \leftarrow h_t$；$\sigma$表示 sigmoid 激活函数；tanh表示tanh激活函数；Conv 表示卷积层；⊙代表两变量按元素点乘。

最后，更新模块提取得到的运动特征$\boldsymbol{f}_m$随即被送入特征双向更新模块，生成双向更新后的运动特征。更新的运动特征$\hat{\boldsymbol{f}}_m$被输入到运动分支解码器。该解码器采用两个卷积层，卷积核大小为 3。然后得到本次迭代光流的更新量$\Delta\boldsymbol{g}$。因此，当前的流量估计被计算为$\boldsymbol{g}\leftarrow\Delta\boldsymbol{g}+\boldsymbol{g}$。经过 T 次迭代，网络输出了一系列的光流估计结果，如图 1 右下方所示。

1.4　特征双向更新模块

特征双向更新模块主要用于双向更新显式运动建模分支的运动特征与伪装目标检测分支的外观特征。其详细设计如图 2 所示，引入文献 [16] 提出的空间注意力（SA）模块和通道注意力（CA）模块来有选择性地引导运动特征与目标特征进行双向交互。

特征双向更新模块主要分为两个阶段：聚合和分配。聚合阶段对应图 2 中的绿色箭头，实现多尺度的外观特征与运动特征的融合；分配阶段对应图 2 中的紫色箭头，将更新后的运动特征广播回各尺度的外观特征。在聚合阶段，3 个不同尺度的参考帧外观特征$\{\boldsymbol{f}_t^i\}$通过 SA 模块加强对目标空间的关注，与运动特征$\boldsymbol{f}_m$串联拼接后输入到卷积层，再借助 CA 模块进一步提取有效特征信息，得到最终更新的运动特征$\hat{\boldsymbol{f}}_m$。由于注入了多尺度的伪装目标的外观特征，更新后的运动特征更注重于伪装物体的运动。具体地，更新公式如下所示：

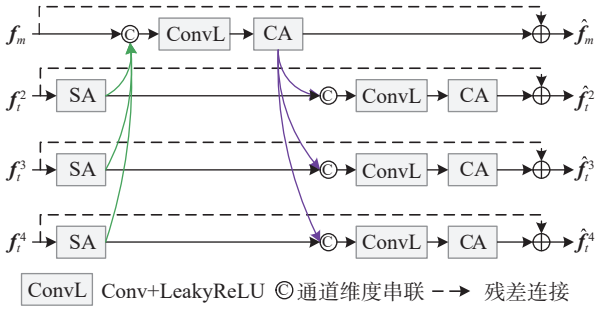

图 2　特征双向更新模块

Fig.2　Feature bidirectional update module

$$\dot{\boldsymbol{f}}_m=\mathrm{Concat}\left(\boldsymbol{f}_m,\mathrm{SA}\left(f_t^2\right),\mathrm{UP}\left(\mathrm{SA}\left(\boldsymbol{f}_t^3\right)\right),\mathrm{UP}\left(\mathrm{SA}\left(\boldsymbol{f}_t^4\right)\right)\right) \tag{8}$$

$$\hat{\boldsymbol{f}}_m=\boldsymbol{f}_m+\mathrm{CA}\left(\mathrm{ConvL}\left(\dot{\boldsymbol{f}}_m\right)\right) \tag{9}$$

式中：SA，CA 分别为输入与输出张量维度相同的空间注意力模块和通道注意力模块；UP 表示上采样模块，用于将不同层次的特征维度与运动特征$\hat{\boldsymbol{f}}_m$对齐；ConvL由输出通道为 3 的卷积层和 LeakyReLU 层构成；$\dot{\boldsymbol{f}}_m$表示中间特征。

在分配阶段，将更新的运动特征广播回各尺度的外观特征，使其进一步的信息融合。更新后的运动特征$\hat{\boldsymbol{f}}_m$和各层级特征向量序列$\{\boldsymbol{f}_t^i\}$分别拼接，再使用通道注意力机制，得到再次更新后的层次化特征序列。具体地，按照如下公式更新不同层次的参考帧特征为

$$\hat{\boldsymbol{f}}_t^i=\boldsymbol{f}_t^i+\mathrm{CA}\left(\mathrm{ConvL}\left(\mathrm{Concat}\left(\mathrm{DB}\left(\hat{\boldsymbol{f}}_m\right),\mathrm{SA}\left(\boldsymbol{f}_t^i\right)\right)\right)\right) \tag{10}$$

式中：$\mathrm{i}\in\{2,3,4\}$，代表参考帧层次化特征序列中的层数；Concat 表示在通道维度进行串联；DB 表示下采样模块，用于将运动特征$\hat{\boldsymbol{f}}_m$与不同层次的特征维度对齐；$\hat{\boldsymbol{f}}_t^i$表示更新后的不同层次参考帧特征。

1.5　伪装目标检测分支

仅依靠外观特征，伪装目标难以从视频中检测出来。联合框架让伪装目标检测分支和显式运动建模分支通过特征双向更新模块进行双向交互，实现运动特征与外观特征的双向更新，从而

提升伪装目标的检测性能。

编码器提取的参考帧的外观特征与显式运动建模分支输出的目标运动特征经过特征双向更新模块交互后，可得到更新后的目标特征和运动特征。将更新后的目标特征输入到解码器中即可获得目标检测结果。如文献[5]所述，与传统的解码器相比，NCD可以简单有效地传播特征，而GRA可以获得更好的伪装物体的细节。因此，在伪装目标检测分支中引入NCD和GRA来进一步提升伪装目标检测性能。先将多尺度目标特征输入到NCD中以获得粗略的检测图，再通过GRA模块逐步细化粗略的检测图，经过Sigmoid激活函数，输出最终的伪装目标检测图。

2 监督和损失函数

显式运动建模分支和伪装目标检测分支相辅相成，相互促进。对两分支输出的光流估计结果和伪装目标检测结果进行联合优化，其总损失为

$$L_{\text{total}} = L_{\text{pred}} + L_{\text{flow}} \tag{11}$$

式中：L_{total}为总损失；L_{pred}为目标检测损失；L_{flow}为光流估计损失。

对于伪装目标检测分支，采用文献[17]中的混合损失函数，通过使用IoU(intersection over union)损失、二进制交叉熵(bce)损失和增强对齐损失(e-loss)来指导目标检测流，定义为

$$L_{\text{pred}} = L_{\text{IoU}} + L_{\text{bce}} + L_{\text{e-loss}} \tag{12}$$

对于显式运动建模分支，考虑到视频伪装场景中缺乏光流真值图，采用自监督的策略监督光流估计流学习。参考文献[18]，根据第k次迭代的光流结果对参考帧I_t进行映射和变换，得到I_{t+1}的重建帧，表示为$\acute{I}_{t+1}(k)$。通过计算I_{t+1}和$\acute{I}_{t+1}(k)$之间的距离对显式运动建模分支进行监督，而该分支最终输出多次迭代的光流序列。为了专注于多次迭代后的光流结果，引入权重指数衰减策略来引导运动估计。L_{flow}定义为

$$L_{\text{flow}} = \sum_{k=1}^{T} \gamma^{T-k} D\left(I_{t+1}, \acute{I}_{t+1}(k)\right) \tag{13}$$

式中：T为显式运动建模分支的迭代次数；D为像素级光度损失SSIM (structural similarity)[19]；$\acute{I}_{t+1}(k)$表示参考帧I_t基于光流估计进行扭曲后的重建帧；γ表示衰减系数。设定γ=0.8，T=6。

3 实验结果分析

3.1 数据集与评估指标

为公平比较，在最大的VCOD数据集MoCA-Mask[9]的训练集上训练，并在两个流行的VCOD数据集上测试并对实验结果进行评价分析，即整个CAD[20]数据集和MoCA-Mask的测试集。参考文献[9]，评价指标采用伪装目标检测领域中的常用指标，即S-measure(S_α)[21]，加权F-measure(F_β^w)[22]，MAE(mean absolute error, M)[9]，用于测量目标预测图和ground truth之间的相似性的指标Dice系数(dice similarity coefficient)[9]和用于测量目标物体ground truth与分割物体之间的区域相似度的平均交并比IoU[9]。

3.2 实验设置

大多数基于视频的方法，采用多阶段的训练流程，模型首先使用静态图像数据集进行预训练，再连接时序模块在视频数据集上进行训练。相较于静态数据集COD10K[5]，MoCA-Mask[9]视频数据集更具有挑战性，存在着摄像头运动、小物体运动、光照暗等情况。加载在COD10K数据集上预训练的权重，可以进一步提升模型在MoCA-Mask数据集上的表现。因此，采用与文献[9]相同的两阶段训练策略来训练基于视频的方法：在静态图片数据集COD10K上训练网络骨干，然后在MoCA-Mask的训练集上微调。

SMHNet实验方法基于PyTorch框架，训练在NVIDIA 3090 GPU上实现，并采用Adam优化器通过余弦退火策略进行优化，最大和最小学习率以及最大迭代次数分别设置为1×10^{-5}，1×10^{-6}和20。网络输入图像尺寸为352×352×3。训练时运用颜色增强，随机旋转方法进行数据增强，Batch_size设置为6，Epoch为30。

3.3 实验结果对比与分析

3.3.1 定量结果对比与分析

为说明所提出方法的有效性，将SMHNet与7个前沿方法进行了定量比较，包括3个基于图像的方法(SINet[6]、SINet-V2[5]、DGNet[7])和4个基于视频的方法(RCRNet[23]、MG[12]、PNS-Net[24]、SLTNet[9])。参考文献[9]的对比策略，同样基于图像的方法均在MoCA数据集上重新训练，基于视频的方法均进行了两阶段训练。

由表1可得，方法SMHNet在几乎所有的评

价指标上都取得了优异的表现。相较于采用隐式建模策略的最优 VCOD 方法 SLTNet[9]，SMHNet 在 MoCA-Mask 测试集上的性能，在S_α上提升了 4.28%；在F_β^w上提升了 11.9%；在 M 上提升了 22.22%；在 Dice 上提升了 9.17%；在 IoU 上提升了 12.87%。相较于采用直接输入光流策略的最优 VCOD 方法 MG[12]，方法 SMHNet 在 MoCA-Mask 测试集上的性能提升也十分明显，例如：在S_α上提升了 24.15%；在指标 M 上提升了 36.36%。

3.3.2　定性结果对比与分析

图 3 给出了方法 SMHNet 和 3 个基准模型的定性结果对比，展示了在同一个视频中的不同帧的分割结果。SMHNet 相较于其他基准模型，可以在同一视频片段中准确地定位与分割出伪装在雪地中的北极狐，并且保持了一定的长期一致性。

图 4 给出了 SMHNet 和 3 个基准模型在不同视频片段的单帧上的定性结果对比。对比第 1 行，SMHNet 可以在低光照情况下清晰定位并分割出沙漠猫；对比第 2 行，SMHNet 可以分割出微小的野山羊；而第 3 行、第 4 行中目标与背景难以分辨的青蛙和北极狐，也能够被 SMHNet 清晰分割。可以看出，SMHNet 可以在光照条件不佳、目标微小、不清晰的外观等各种富有挑战的情况下，给出更准确的伪装物体检测。

3.3.3　消融实验结果

为验证模块的有效性，进行了各组件的消融实验，如表 2 所示。静态分支表示 SMHNet 的静态部分，仅保留特征提取和解码器组件。运动分支代表显式运动建模分支的相关操作。#A 作为评估组件的基线；#B 表示 SMHNet 去掉特征双向更新模块中的分配阶段的交互（即去掉图 2 中的绿色箭头）；#C 表示 SMHNet 去掉特征双向更新模块中的聚合阶段的交互（即去掉图 2 中的紫色箭头）；#D 表示去掉特征双向更新模块中的注意力模块，直接不加处理地拼接运动特征和外观特征；#E 为本文提出的 SMHNet。

对比表 2 中的#A，#B，#C，可以看出运动特征和目标特征的交互对伪装目标检测性能的提升是显著的。#B，#C 和#E 结果对比表明，运动特征和目标特征双向交互策略比只进行单向的交互更有效。对比#D 和#E，凸显了注意力机制的必要性，注意力机制可以有选择地引导特征的双向更

表 1　在 MoCA-Mask 和 CAD 数据集上的定量比较

Tab.1　Quantitative comparison on the MoCA-Mask and CAD datasets

模型	MoCA-Mask[9]					CAD[20]				
	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	Dice↑	IoU↑	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	Dice↑	IoU↑
SINet[6]	0.598	0.231	0.028	0.276	0.202	0.636	0.346	0.041	0.381	0.283
SINet-V2[5]	0.588	0.204	0.031	0.245	0.180	0.653	0.382	0.039	0.413	0.318
DGNet[7]	0.581	0.184	0.024	0.222	0.156	0.686	0.416	0.037	0.456	0.340
RCRNet[23]	0.555	0.138	0.033	0.171	0.116	0.627	0.287	0.048	0.309	0.229
MG[12]	0.530	0.168	0.067	0.181	0.127	0.594	0.336	0.059	0.368	0.268
PNS-Net[24]	0.544	0.097	0.033	0.121	0.101	0.655	0.325	0.048	0.384	0.290
SLTNet[9]	0.631	0.311	0.027	0.360	0.272	0.696	0.481	0.030	0.493	0.401
SMHNet	**0.658**	**0.348**	**0.021**	**0.393**	**0.307**	**0.707**	**0.508**	**0.028**	**0.527**	**0.429**

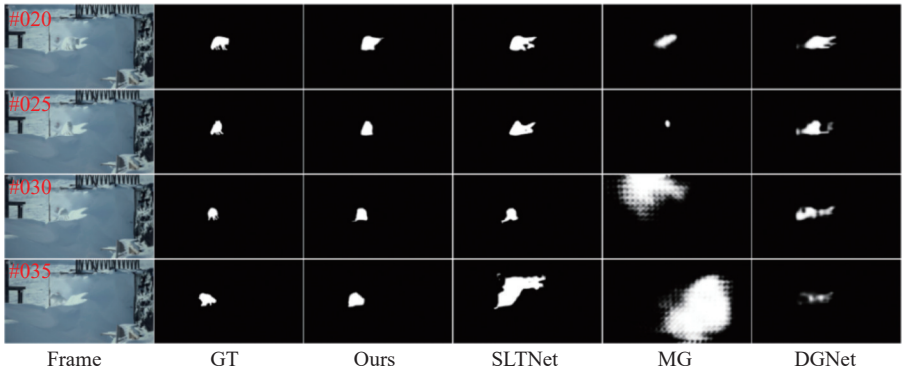

图 3　同一视频片段上的定性比较

Fig.3　Qualitative comparison on the same video clip

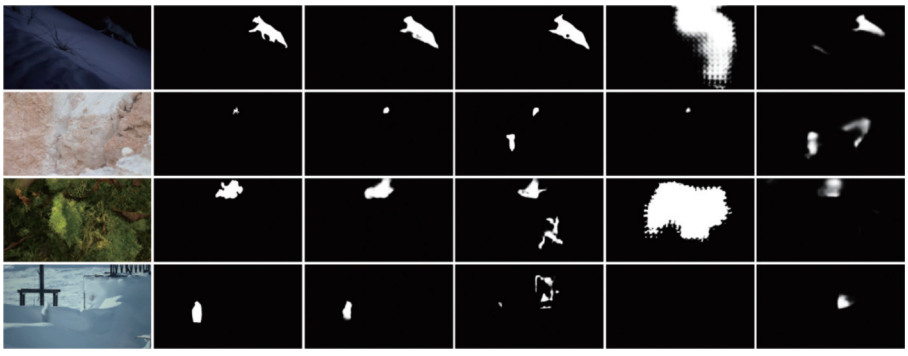

图 4 不同视频片段上的定性比较

Fig.4 Qualitative comparison on different video clips

表 2 主要组件的消融结果

Tab.2 Ablation results for the main modules

序号	静态分支	运动分支	聚合阶段	分配阶段	注意力机制	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	Dice↑	IoU↑
#A	√					0.628	0.292	0.028	0.341	0.259
#B	√	√	√		√	0.650	0.336	0.026	0.389	0.298
#C	√	√		√	√	0.645	0.329	0.025	0.373	0.291
#D	√	√	√	√		0.637	0.308	0.024	0.355	0.273
#E	√	√	√	√	√	**0.658**	**0.348**	**0.021**	**0.393**	**0.307**

新。特征双向更新模块经过选择和优化，引导运动特征和外观特征双向交互，进一步提升伪装目标检测的性能。

存在噪声和错误的光流图作为模型监督会影响模型检测性能，表 3 的结果对比反映了本文采用的自监督策略的有效性。其中，伪监督策略代表将 RAFT 离线生成的光流结果作为 SMHNet 显式运动建模分支输出的光流的真值进行监督。可以看出，采用自监督策略后，SMHNet 的性能得到了进一步提升。

此外，为了更直观对比不同监督策略对显式运动建模分支的影响，进行光流估计结果可视化，如图 5 所示。其中，(a)为采用自监督策略训练的 SMHNet 在测试集上生成的光流图；(b)为采用伪监督策略训练的 SMHNet 在测试集上生成的

表 3 学习光流的不同监督策略

Tab.3 Different supervision strategies for learning optical flow

设置	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	Dice↑	IoU↑
伪监督策略	0.652	0.332	0.024	0.381	0.294
自监督策略	**0.658**	**0.348**	**0.021**	**0.393**	**0.307**

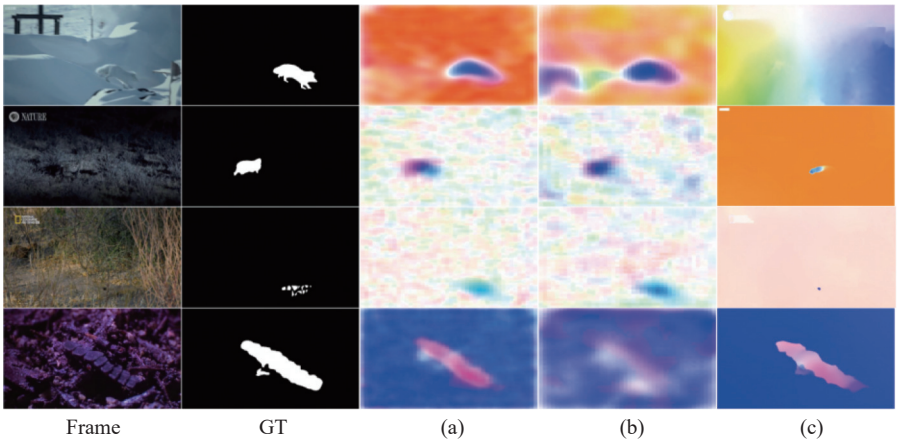

图 5 采用不同监督策略在测试集上的光流估计结果

Fig.5 Optical flow estimation results on test images using different supervision strategies

光流图；（c）为 RAFT[15] 离线模型在测试集上生成的光流图。结果表明，在某些 RAFT 不能感知伪装的运动物体的情况下（如图 5 第 1 行），采用自监督策略的 SMHNet 仍可以较好地反映目标的运动。相比之下，采用伪监督策略的 SMHNet 很容易受到 RAFT 光流的错误影响。

进一步探讨使用不同的迭代次数进行光流优化的效果，如表 4所示。与文献 [15] 一致，迭代次数增大可以细化光流估计结果，提升伪装目标检测的性能。然而，由于硬件的限制，最大迭代次数设定为 6。由表 4 可见，当迭代次数 T 增大时，伪装目标检测的指标逐步提升，并在 T=6 时，取得最佳性能。

表 4　不同迭代次数的评估结果

Tab.4　Evaluation results for different iterations

迭代次数	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	Dice↑	IoU↑
1	0.626	0.299	0.029	0.339	0.263
2	0.633	0.304	0.028	0.349	0.269
3	0.632	0.308	0.028	0.351	0.270
4	0.645	0.324	0.026	0.374	0.287
5	0.647	0.330	0.021	0.370	0.290
6	**0.658**	**0.348**	**0.021**	**0.393**	**0.307**

4　结　论

为了提高现有 VCOD 方法中对运动线索的利用效率，提出了一个基于显式运动建模的 VCOD 框架，称为 SMHNet，以自监督的方式显式处理运动线索。首先，该框架将运动建模和伪装目标检测联合在同一个框架中，该框架包括运动建模分支和伪装目标检测分支，并由特征双向更新模块引导两个分支进行双向交互。两个分支任务相互优化、相互纠错，最终输出更优的光流估计结果和伪装目标检测结果。在两个 VCOD 数据集上的对比实验结果显示，SMHNet 取得了最优的性能。综上，本文提出的基于显式运动建模的 VCOD 方法可以提高伪装目标检测的性能，并为解决 VCOD 问题提供了新方案。

参考文献：

[1] JI G P, XIAO G B, CHOU Y C, et al. Video polyp segmentation: a deep learning perspective[J]. Machine Intelligence Research, 2022, 19(6): 531−549.

[2] HE T, LIU Y, XU C Y, et al. A fully convolutional neural network for wood defect location and identification[J]. IEEE Access, 2019, 7: 123453−123462.

[3] 武国晶, 吕绪良, 邢海宁, 等. 三维凸面分析法在迷彩伪装检测中的应用 [J]. 解放军理工大学学报 (自然科学版), 2015, 16(6): 582−586.

[4] 王烨奎, 曹铁勇, 王杨, 等. CAMOU-YOLO: 一种迷彩伪装目标检测模型 [J]. 计算机技术与发展, 2022, 32(12): 29−36.

[5] FAN D P, JI G P, CHENG M M, et al. Concealed object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(10): 6024−6042.

[6] FAN D P, JI G P, SUN G L, et al. Camouflaged object detection[C]//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020: 2774–2784.

[7] JI G P, FAN D P, CHOU Y C, et al. Deep gradient learning for efficient camouflaged object detection[J]. Machine Intelligence Research, 2023, 20(1): 92−108.

[8] 童旭巍, 张光建. 基于全局多尺度特征融合的伪装目标检测网络 [J]. 模式识别与人工智能, 2022, 35(12): 1122−1130,

[9] CHENG X L, XIONG H, FAN D P, et al. Implicit motion handling for video camouflaged object detection [C]//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, LA, USA: IEEE, 2022: 13854–13863.

[10] XIE J Y, XIE W D, ZISSERMAN A. Segmenting moving objects via an object-centric layered representation[C]// Proceedings of the 36th Conference on Neural Information Processing Systems. New Orleans, LA, USA, 2022.

[11] LAMDOUAR H, YANG C, XIE W D, et al. Betrayed by motion: camouflaged object discovery via motion segmentation[C]//Proceedings of the 15th Asian Conference on Computer Vision. Kyoto, Japan: Springer, 2020.

[12] YANG C, LAMDOUAR H, LU E, et al. Self-supervised video object segmentation by motion grouping[C]// Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada: IEEE, 2021: 7157–7168.

[13] WANG W H, XIE E Z, LI X, et al. PVT v2: improved baselines with pyramid vision transformer[J]. Computational Visual Media, 2022, 8(3): 415−424.

[14] FAN D P, JI G P, ZHOU T, et al. PraNet: parallel reverse attention network for polyp segmentation[C]//Proceedings of the 23rd International Conference on Medical Image Computing and Computer Assisted Intervention–MICCAI 2020. Lima, Peru: Springer International Publishing, 2020: 263–273.

[15] TEED Z, DENG J. RAFT: recurrent all-pairs field transforms for optical flow[C]//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer International Publishing, 2020: 402–419.

[16] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer, 2018: 3–19.

[17] 范登平, 季葛鹏, 秦雪彬, 等. 认知规律启发的物体分割评价标准及损失函数 [J].中国科学: 信息科学, 2021, 51(09): 1475-1489.

[18] Liu R, Wu Z, Yu S, et al. The emergence of objectness: learning zero-shot segmentation from videos[J]. Advances in neural information processing systems, 2021, 34: 13137-13152.

[19] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600−612.

[20] BIDEAU P, LEARNED-MILLER E. It's moving! A probabilistic model for causal motion segmentation in moving camera videos[C]//Proceedings of the 14th European Conference on Computer Vision–ECCV 2016. Amsterdam, The Netherlands: Springer International Publishing, 2016: 433–449.

[21] FAN D P, CHENG M M, LIU Y, et al. Structure-measure: a new way to evaluate foreground maps[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 4558–4567.

[22] MARGOLIN R, ZELNIK-MANOR L, TAL A. How to evaluate foreground maps[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014: 248–255.

[23] YAN P X, LI G B, XIE Y, et al. Semi-supervised video salient object detection using pseudo-labels[C]// Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE, 2019: 7283–7292.

[24] JI G P, CHOU Y C, FAN D P, et al. Progressively normalized self-attention network for video polyp segmentation[C]//Proceedings of the 24th International Conference on Medical Image Computing and Computer Assisted Intervention–MICCAI 2021. Strasbourg, France: Springer International Publishing, 2021: 142–152.

（编辑：董 伟）