基于隐式运动处理的视频伪装物体检测

* 程雪莲 [1], * 熊欢 [3], † 范登平 [4], 钟怡然 [6,7], Mehrtash Harandi[1,8], Tom Drummond[1], 戈宗元 [1,2,5]
[1]Faculty of Engineering, Monash University, [2]eResearch Centre, Monash University
[3]Mohamed bin Zayed University of Artificial Intelligence, [4]CVL, ETH Zurich,
[5]Airdoc Research Australia, [6]SenseTime Research, [7]Shanghai AI Laboratory, [8]Data61, CSIRO

(a) 第 T 帧 (b) 真值（GT） (c) 本文方法结果 (d) 光流变换结果 (e) 光流

Figure 1: 隐式的 (本文方法) 与显式的 (光流方法) 运动处理结果的对比。(c) 本文方法得到的第 T 帧的伪装物体检测的结果（由第 $T-1$ 帧和第 T 帧计算得到）。(d) 光流法得到的第 T 帧的伪装物体检测的结果（由第 $T-1$ 帧使用预先计算的光流进行变换得到，其中，光流如 (e) 所示）。可以看出，基于显式的方法（即光流法）进行伪装物体检测，会导致准确性显著下降（降低 10.79%），本文将对其进行进一步讨论。

摘要

本文提出一个新的视频伪装物体检测（*Video Camouflaged Object Detection*，VCOD）框架。该框架利用视频帧间的短期动态与长期一致性信息，来进行伪装物体检测。伪装物体的基本特性是，它们通常会展现出与背景相似的模式，使得难以从静态图像中将其识别出来。但当伪装物体在运动的时候，就很容易被发现。因此，有效地处理视频中的时序动态信息成为了解决 *VCOD* 任务的关键。然而，当前 *VCOD* 方法通常基于单应性变换（*homography*）或者光流来表征运动，其检测误差会随着运动误差与分割误差而累积。本文方法将运动估计与目标分割统一到一个优化框架当中。具体来讲，本文构建了一个稠密相关性体积（*dense correlation volume*），来隐式地捕捉相邻帧之间的运动，并且用最终的分割真值来同时监督优化运动估计与分割结果。此外，为了保证视频序列中的时序一致性，本文同时引入了一个时空 *Transformer* 来优化短期预测的结果。通过在 *VCOD* 基准上所进行的大量实验，证明了本文架构的有效性。本文还构建了一个大规模 *VCOD* 数据集，***MoCA-Mask***，其中包括像素级手工抠图的真值掩模，并搭建了一个包含已有方法的全面的评测基准，***VCOD benchmark***，以促进视频伪装物体检测领域的发展。数据集链接为：*https://xueliancheng.github.io/SLT-Net-project*。

本文为 CVPR2022 论文 [4] 的中文翻译版。

1. 引言

视频伪装物体检测（Video Camouflaged Object Detection，VCOD）是找出视频中在外观上与背景展现出极高相似性的物体的任务。尽管拥有广泛的应用场景（例如：监控与安防 [25]、自动驾驶 [33, 5]、医学图像分割 [12, 43]、蝗虫检测 [18] 与机器人 [29]），伪装物体检测（Camouflaged Object Detection，COD）仍是一项有挑战的任务，因为伪装物体往往连人眼都难以分辨。因此，在计算机视觉领域内，与视频目标检测（Video Object Detection，VOD）[48, 1]、视频显著性目标检测（Video Salient Object Detection，VSOD）[16] 以及视频运动分割（Video Motion Segmentation，VMS）[17, 47] 等任务相比，VCOD 仍是一个未被充分研究的问题。

在大多数计算机视觉任务中（例如：实例分割 [52] 和显著性检测 [50]），都假设物体有清晰的边界。因此，在解决此类问题时可依赖于图像层面的信息，并且融

合运动信息，还可以进一步提升目标任务的效果。相反，在伪装物体检测中，物体的边缘是模糊且难以分辨的。这不仅使得从图像中进行检测具有挑战，还导致了难以准确估计视频中的光流与运动线索 [38, 53, 37]。

缺乏清晰的边界，即伪装物体的外观与背景非常相似。这意味着两个主要的难点：**1)** 物体边缘往往无缝地混合在背景中，只有当物体运动的时候才能被发现；**2)** 物体通常有与环境相似的、重复的纹理。因此，想要通过帧间像素点的移动来估计运动（正如光流法所实现的），将难以预测且容易出错。考虑到第一个难点，要想解决 VCOD 任务，神经网络需要在运动信息的帮助下，有效地发现伪装物体与背景之间存在的细微差别。此外，如图1所示，从第二个难点可以得知，运动信息本身是有噪声的，并且是不精确的。因此，将 VOD、VSOD 和 VMS 技术直接或组合地应用于 VCOD 任务，往往会得到很差的结果。

本文提出了 **SLT-Net**，一个用于解决 VCOD 任务的新模型。该模型利用短期动态与长期一致性信息，在视频中检测伪装物体。具体来讲，本文使用一个短期动态模块来隐式地捕捉连续帧之间的运动。使用一个全程相关性金字塔策略（full-range correlation pyramid strategy）来隐式地表征运动，而非使用光流来显式地表征。使用相关性金字塔的主要动机是，即使是当前最先进的光流算法也会在伪装物体检测的任务中失败，并且其误差会在整个视频中累积。而本文的模型可以在仅有分割掩膜监督信息的情况下，同时对运动估计（隐式地）和分割预测进行优化。为了给出稳定的估计，本文进一步引入一个长期优化模块，用来降低短期动态模块中的误差累积。

SLT-Net 是由 Transformer 与 CNN 组件组合而成的混合模型。具体来讲，本文使用 Transformer 架构来编码特征，构建相关性金字塔（correlation pyramid）。该架构除了具有设计灵活的特点，Transformer 提取的特征还包含了全局上下文信息，其中包含长程依赖，以及更少的归纳偏差 [41, 32]。使得在运动估计中，更容易分辨被识别物体。

虽然相关性金字塔策略可以有效地捕捉运动，实现伪装物体检测。但是，由于它的计算复杂性高，无法方便地扩展到长视频序列中。为了解决这个问题，本文采用一个包含时空 Transformer 的 sequence-to-sequence 架构，通过长期一致性来优化并预测视频中连续帧。本文发现该架构相比标准的 ConvLSTM 模型 [45, 54]，能得到更准确的预测结果。

此外，作为一个未被深入研究的问题，VCOD 缺乏大规模的数据集来进行系统地评测。为了促进该领域的发展，本文基于 MoCA（Moving Camouflaged Animals）[19]，重新整理并构建了一个大规模的 VCOD 数据集，**MoCA-Mask**，该数据集包含 87 个视频序列，共计 22,939 帧，并配有像素级的真值掩模。MoCA-Mask 囊括大量的挑战，例如：复杂的背景、微小的、完美伪装的物体。本文对数据集中的每个视频序列，每隔 5 帧，给出标注、包围盒、以及稠密的分割掩模。本文还提供了该领域第一个全面的评测基准，包含对已有的 VCOD 以及相关方法的测评。本文的贡献点概括如下：

- 本文提出了一个新的 VCOD 框架，它可以有效地建模视频中的短期动态与长期一致性。其中，伪装物体运动与分割可以同时被优化。
- 本文构建了第一个大规模 VCOD 数据集，**MoCA-Mask**，以促进 VCOD 领域的发展。以及一个全面的评测基准，为后续 VCOD 研究提供便利。
- 在 VCOD 任务上，本文的方法达到了当前最好的结果，超过之前的 SOTA 模型 [46] 达 9.88%。

2. 相关工作

COD：如果没有先验知识，即使是人也会很容易遗漏掉伪装物体。然而，一旦被告知图像中存在有伪装物体，人们便会仔细浏览整张图片来找出它。受到该事实的启发，ANet [20] 整合了分类流，来作为对伪装物体和分类流的感知。基于类似的想法，SINet [11] 与 PFNet [28] 通过先粗略定位伪装物体，再精细化分割，来解决该问题。SINet-v2 [9] 在此基础上，引入了反向引导，来挖掘互补区域。MGL [49] 使用两个基于图的模块，来将边缘细节整合到分割流里。通过建模伪装物体相比背景显著的部分，Lv 等人 [26] 在重新标注的 NC4K 数据集上，引入两个新的任务，称为伪装目标排序和伪装目标定位。

VSOD：为了检测视频中的显著性目标，DLVS [39] 引入了全卷积网络（fully convolutional networks，FCN）来进行像素级的显著性预测。DSR3 [21] 利用一个端到端的 3D 神经网络来生成视频序列，其中整合了 3D 卷积模块和循环优化模块，来预测显著性图。时空 CRF [22]、金字塔空洞 ConvLSTM [34] 在设计中，建模了帧与帧

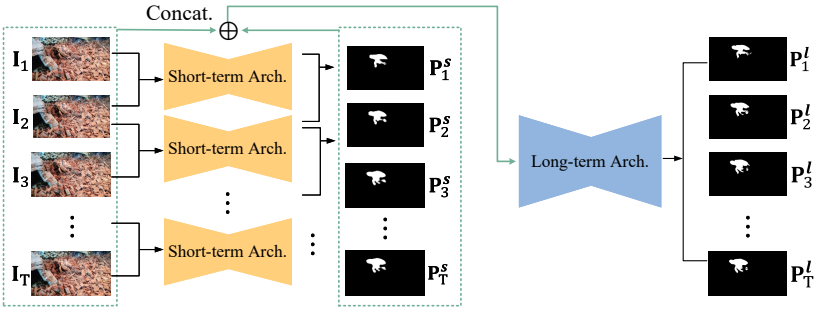

Figure 2: SLT-Net 的总体框架。SLT-Net 包含一个短期检测模块和一个长期优化模块。短期检测模块输入一对连续帧，输出参考帧的伪装物体掩模。长期优化模块输入由短期检测模块生成的 T 步预测，以及相应的参考帧，输出最终的预测结果。

之间的时序信息。FGRN [24]、RCRNet [46] 采用了额外的流引导（flow-guided）网络来提升时序一致性。之后，SSAV [13] 特别专注于显著性转移现象，并且构建了一个全面的 VSOD 评测基准。FSNet [16] 利用了外观和运动线索的相互约束，得到了比许多已有方法更好的效果。

VMS： VMS 任务专注于发现视频中存在的运动物体。传统方法通常是先提取流场中的运动边界，然后用外观特征来优化原始估计，以解决 VMS 问题 [30]。或者将运动和外观线索通过一个融合架构 [14] 结合起来。另一类工作直接利用光流作为输入，基于监督学习 [36] 或无监督学习 [47] 的方式，训练一个 CNN 网络，并生成像素级的运动标签。

VCOD： 不同于 VMS，伪装物体的视觉线索通常被认为没有运动线索有效。之前的工作主要依赖于单应性变换或者光流来检测运动模式。Bidau 等人提出了使用稠密光流 [2, 3] 来近似不同运动模型的方法，从而分割环境中的运动物体。尤其在 [2] 中，作者提出了一个两阶段的分割算法，首先对摄像头的旋转进行补偿，然后将光流角分割为物体与背景。每个运动模型是根据光流方向随时间而更新的，初始的运动是启发式的。在 [3] 中，作者使用一个网络来分割角度场（angle field），而非原始光流。文献 [19] 提出了一个视频注册和运动分割网络，以及一个更大的伪装数据集（MoCA），该数据集每 5 帧用包围盒标注一次。通过光流的显式的对齐方法建立了相邻帧之间的空间关联。然而，光流估计可能并无法足够准确地提供有效的对齐，尤其是在包含有快速移动物体的动态场景中。

3. 本文方法

本文 SLT-Net 模型的输入是一段包含有伪装物体的视频片段，输出是视频每一帧的像素级二值化伪装物体掩模。具体来讲，T 帧的视频片段用 $\{\mathbf{I}^{\mathrm{t}}\}_{t=1}^{\mathrm{T}}, \mathbf{I}^{\mathrm{t}} \in \mathbb{R}^{3\times H\times W}$ 来表示，其中 H, W 是帧的高与宽。本文的网络被用来估计出 t 时刻，视频帧 $\mathbf{I}^{\mathrm{t}}$，的一个二值化的掩模 $\mathbf{M}^{\mathrm{t}} \in \{0,1\}^{H\times W}$。

3.1. 概述

如图2所示为 SLT-Net 模型整体框架。SLT-Net 包括一个短期检测模块和一个长期优化模块。短期检测模块输入一对连续的帧，输出对于参考帧的伪装物体掩模的预测。使用一个 sequence-to-sequence 模块，用时序一致性先验，同步优化输入的视频片段的结果。本文采用两阶段（two-stage）策略来训练 SLT-Net 。首先，本文只用像素级标注来训练短期检测模块。一旦模型收敛，本文将长期优化模块连接到 SLT-Net ，固定短期检测模块，训练整个模型。

3.2. 短期网络架构

图3展示了本文的短期架构。它输入视频中的连续两帧，预测参考帧的二值化掩模。本文的模型包括三个主要模块：(1) **Transformer 编码器**，用来提取特征；(2) **短期相关性金字塔**，用来捕捉短期动态；以及 (3) **CNN 解码器**，用来预测短期分割。下面，本文详细介绍每个模块：

1. Transformer 编码器： 本文采用金字塔视觉 Transformer（Pyramid Vision Transformer，PVT [40] 的孪生结构（Siamese Structure）来从连续两帧提取特征。

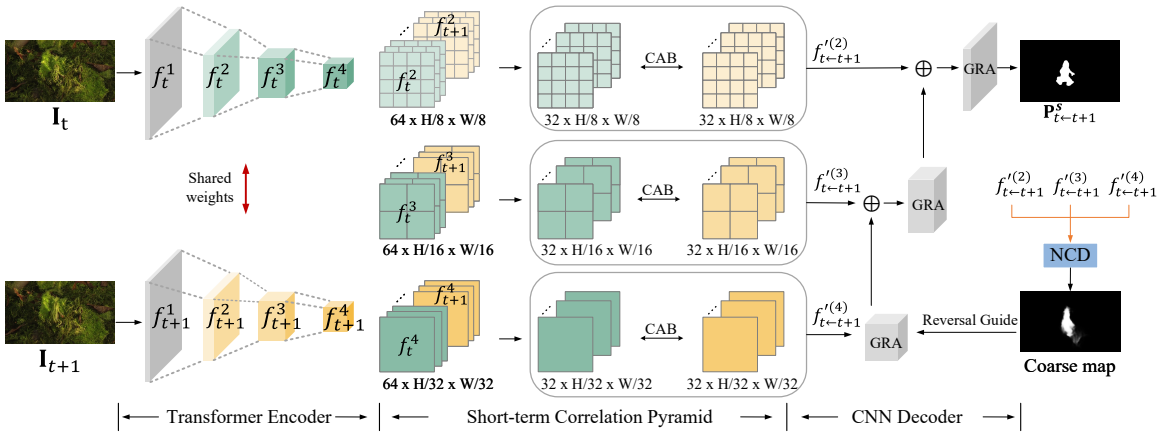

Figure 3: 短期模块的概要。网络首先使用一个 Transformer 编码器对输入帧进行特征提取，然后计算参考帧 $\mathbf{I}_t$ 和它的相邻帧 $\mathbf{I}_{t+1}$ 之间的相关性体积，并生成一个相关性体积金字塔（correlation volume pyramid）。用一个 CNN 解码器，从该金字塔得到的运动中，解码得出最终预测结果。

编码器包括四个阶段，分别从四个不同的尺度生成特征图。所有的阶段都共享相似的结构，其中包括一个 Patch 嵌入层和 Transformer Block。每个阶段的特征尺寸是 $C_i \times H/2^{i+1} \times W/2^{i+1}$, $i \in \{1,2,3,4\}$，其中 H, W, C 分别表示高、宽和通道数。在实验中，本文设置 $C = 32$。参考 [9]，本文采用三个纹理增强模块（Texture Enhanced Modules，TEM）分别作用到最后三个阶段的特征上。为了获得更具有判别力的特征表达，每个 TEM 都包括四个平行的残差分支。

2. 短期相关性金字塔：之前的工作（如 [36, 47]）显式地使用光流法，来计算连续帧的运动，作为深度神经网络的输入。然后，不准确的光流可能会导致后续帧的预测误差累积。如果想要同时用分割模块优化光流模块，那么，真实的光流是必需的。为了解决这个问题，受 [23] 的启发，本文提出一个相关性金字塔来隐式地捕捉运动信息。如图 3所示，CNN 解码器直接将相关金字塔作为它唯一的输入。这意味着，只有在给定正确的运动估计前提下，网络才能给出正确的分割估计。此外，由于形成相关性金字塔的特征的更新，要用到分割的真值（Ground Truth）。因此，可以使用分割真值来同时优化运动估计和检测结果。

如图4所示，本文给出了相关性金字塔的核心单元，相关性聚合模块（Correlation Aggregation Block，CAB）$\mathbf{C}$。给定一个视频帧特征对 $\{f_t, f_{t+1}\} \in \mathbb{R}^{C\times H'\times W'}$，4D 相关性体积 $\mathbf{C}(\mathbf{I}_t, \mathbf{I}_{t+1}) \in \mathbb{R}^{H'\times W'\times H'\times W'}$ 定义如下：

$$\mathbf{C}(\mathbf{I}_t, \mathbf{I}_{t+1})_{xyuv} = \exp\left(\sum_c \mathcal{F}_\theta(\mathbf{I}_t)_{xyc} \cdot \mathcal{F}_\theta(\mathbf{I}_{t+1})_{uvc}\right), \quad (1)$$

其中，c 是帧的特征通道维度的索引（index）。随着所有的相邻特征被相关性给配对起来，本文可以在全局尺度找出其一致性。为了降低计算复杂度，本文通过对特征进行最大池化（max-pooling）来下采样相邻帧，并保持参考帧的分辨率不变。该设计帮助模型在保持高分辨率图像细节的同时，还能学习多尺度位移。

接下来，由于特征相关性体积在所有的空间位置上表征了参考帧与下采样相邻特征帧之间的一致性，本文对其 $\mathbf{C}(\mathbf{I}_t, \mathbf{I}_{t+1})_{xyuv}$ 沿着最后两个维度 uv，基于他们的和（sum），做归一化。归一化的相关性体积计算公式如下所示：

$$\tilde{\mathbf{C}}(\mathbf{I}_t, \mathbf{I}_{t+1})_{xyuv} = \frac{\mathbf{C}(\mathbf{I}_t, \mathbf{I}_{t+1})_{xyuv}}{\sum_u \sum_v \mathbf{C}(\mathbf{I}_t, \mathbf{I}_{t+1})_{xyuv}}. \quad (2)$$

本文用一个卷积操作 $\phi(\cdot)$ 来有选择的考虑通道维度的信息，因此，得到了优化的特征图 $\phi(\mathbf{I}_{t+1}) \in$

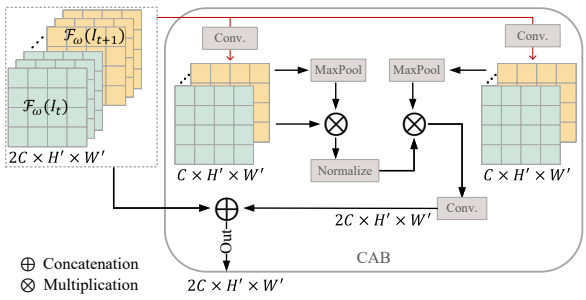

Figure 4: 相关性聚合模块（Correlation Aggregation Block，CAB）计算参考帧（绿色块）和相邻帧（黄色块）之间的特征图的归一化相关性体积。

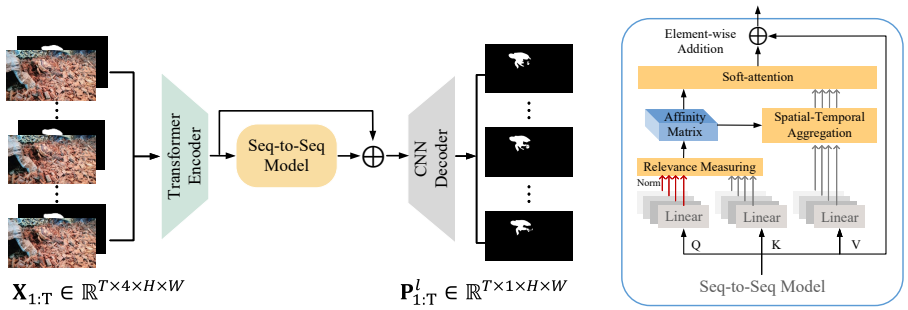

Figure 5: 长期一致性架构概览。它将该过程看作一个 seq-to-seq 的建模问题，并且通过一个 sequence-to-sequence Transformer 来优化视频帧对之间预测。

$\mathbb{R}^{C\times H'\times W'}$。具体地，汇聚特征 $f'_{t\leftarrow t+1} = \rho(\mathbf{I}_{t\leftarrow t+1}) \in \mathbb{R}^{C\times H'\times W'}$ 通过如下公式计算得到：

$$\rho(\mathbf{I}_{t\leftarrow t+1}) = \tilde{\mathbf{C}}(\mathbf{I}_t, \mathbf{I}_{t+1})\phi(\mathbf{I}_{t+1}). \tag{3}$$

图4只展示了一个尺度上的一个相关性。为了让模型学到更多细节信息，本文通过整合基于 Transformer 编码器所提取的多尺度特征，构建了相关性金字塔 $\{\mathbf{C}^i\}, i \in \{2,3,4\}$。

3. CNN 解码器： 如 [9] 所述，近邻连接解码器相比通常的连接解码器（如稠密连接或者短连接）更加可靠。此外，[9] 所使用的分组反向注意力（Group-Reversal Attention，GRA）策略可以在物体边界处提供更加准确的分割结果。基于此，本文直接将来自短期相关性金字塔的特征，即 $\{f'^{(i)}_{t\leftarrow t+1}\} \in \mathbb{R}^{C\times H/2^{i+1}\times W/2^{i+1}}, i \in \{2,3,4\}$，输入到 GRA 中，生成优化的特征图。近邻连接解码器（Neighbor Connection Decoder，NCD）用来生成粗糙的图，该图可以提供伪装物体大致定位的反向引导。如此，就可以汇聚来自 CNN 解码器的低层特征和来自相关金字塔的高层特征。

学习策略： 本文通过如下的损失函数来训练短期阶段：

$$\mathcal{L} = \mathcal{L}^w_{ce} + \mathcal{L}^w_{iou}. \tag{4}$$

带权重的交叉熵损失 $\mathcal{L}^w_{ce}$ 增加了较难像素的权重，以突出其重要性。带权重的交并比（IoU）损失 $\mathcal{L}^w_{iou}$ 更注重难的像素，而非赋予每个像素相同的权重。读者可以参考 [42] 来获取关于这两个损失函数更详细的定义。

3.3. 长期一致性架构

给定一个来自短期架构的序列，$\mathbf{I}_{1:T} = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_T\}$ 与像素级的预测 $\mathbf{P}^s_{1:T} = \{\mathbf{P}^s_1, \mathbf{P}^s_2, \ldots, \mathbf{P}^s_T\}$，本文将长期一致性优化过程定义为一个 seq-to-seq 问题。图5给出长期一致性的架构。本文使用同样的骨干网络，即 Transformer 编码器和 CNN 解码器模块，作为短期架构。因为该架构已经在伪装数据集做过了预训练，使得大大加速长期框架的训练过程。

对于输入序列的每一帧，本文将色彩帧 $\mathbf{I}_t$ 与其在通道维对应的预测 $\mathbf{P}^s_t, t \in [1:T]$ 进行拼接，然后堆叠每个序列中拼接的帧，形成一个 4D 张量 $\mathbf{X}_{1:T} \in \mathbb{R}^{T\times 4\times H\times W}$。网络将 $\mathbf{X}_{1:T}$ 作为输入，输出最终的预测序列 $\mathbf{P}^l_{1:T} \in \mathbb{R}^{T\times 1\times H\times W}$。

有两类 seq-to-seq 的建模框架：一类是使用 ConvLSTM 来建模时序信息，另一类使用基于 Transformer 的建模网络。本文实现了两种架构，并且在4.4小节对比了它们的结果。据本文所知，使用 Transformer 结构可以得到更好的结果，所以，本文选择 Transformer 来作为序列建模的网络，来引入长期一致性约束。

如图5中右侧所示，为 seq-to-seq 建模网络的细节。对于每一个目标像素，为了减少构建稠密时空相似度矩阵的复杂度，本文选择一个固定数量的关联性度量块来构建受约束邻域内的相关度矩阵。训练中，本文采用一个混合损失函数 [10]：

$$\mathcal{L}_{hybrid} = \mathcal{L}^w_{ce} + \mathcal{L}^w_{iou} + \mathcal{L}_e, \tag{5}$$

其中，$\mathcal{L}_e$ 是增强对齐损失。这样的混合损失可以引导网络去学习像素级、物体级以及图像级的特征。

4. 实验

本节在 CAD 数据集和本文提出的 MoCA-Mask 数据集上，对本文所提的框架，进行了系统地评估。本文同时给出了一个 VCOD 任务的全面的评测基准，来促进该领域发展。

Table 1: MoCA-Mask 数据集上的定量结果，“w/” 与 “w/o” 分别表示使用或者不使用本文的伪标签。每组表现最好的结果**加粗**显示。注意，MG [47] 使用没有标签的无监督学习方式进行训练。

Models	MoCA-Mask w/o pseudo labels						w/ pseudo labels					
	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$M \downarrow$	mDic	mIoU	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$M \downarrow$	mDic	mIoU
EGNet [51]	0.547	0.110	0.574	0.035	0.143	0.096	0.546	0.105	0.573	0.034	0.135	0.090
BASNet [31]	0.561	0.154	0.598	0.042	0.190	0.137	0.537	0.114	0.579	0.045	0.135	0.100
CPD [44]	0.561	0.121	0.613	0.041	0.162	0.113	0.550	0.117	0.613	0.038	0.147	0.104
PraNet [12]	0.614	0.266	0.674	0.030	0.311	0.234	0.568	0.171	0.576	0.045	0.211	0.152
SINet [11]	0.598	0.231	0.699	0.028	0.276	0.202	0.574	0.185	0.655	0.030	0.221	0.156
SINet-v2 [9]	0.588	0.204	0.642	0.031	0.245	0.180	0.571	0.175	0.608	0.035	0.211	0.153
PNS-Net [15]	0.544	0.097	0.510	0.033	0.121	0.101	0.576	0.134	0.562	0.038	0.189	0.133
RCRNet [46]	0.555	0.138	0.527	0.033	0.171	0.116	0.597	0.174	0.583	0.025	0.194	0.137
MG [47]	0.530	0.168	0.561	0.067	0.181	0.127	0.547	0.165	0.537	0.095	0.197	0.141
SLT-Net (Ours)	**0.631**	**0.311**	**0.759**	**0.027**	**0.360**	**0.272**	**0.656**	**0.357**	**0.785**	**0.021**	**0.397**	**0.310**

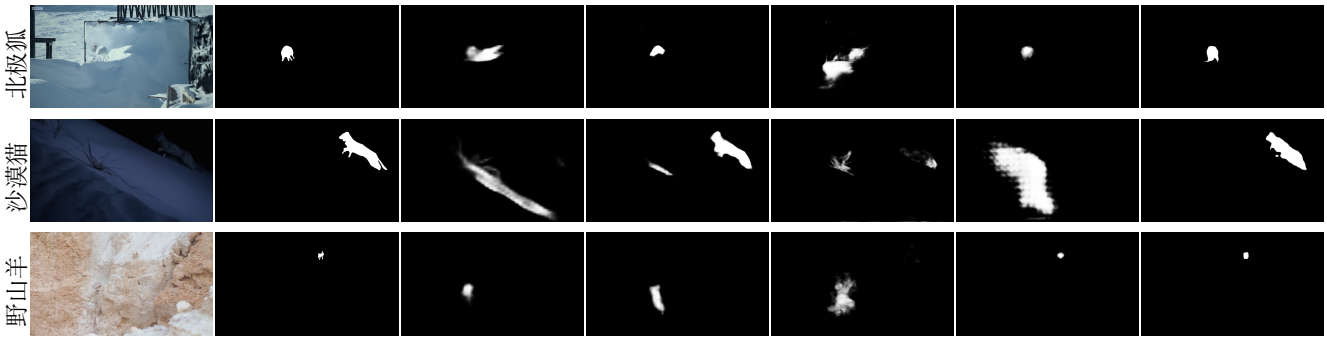

Figure 6: 在本文的 MoCA-Mask 评测基准上的定性实验。本文的模型可以在各种有挑战的情景中，给出更加准确的伪装物体预测。即：不清晰的外观（北极狐）、低光照情况（沙漠猫）与微小物体（野山羊）。

4.1. 数据集

COD10K： 本文在 COD10K [9] 上预训练了所有的基于静态图片的方法与基于视频的方法的编码器。该数据集是当前最大的 COD 数据集，其中包括 5,066 张伪装图像（3,040 张用于训练，2,026 张用于测试），分为 5 个超类和 69 个子类，同时提供高质量的抠图级标注。

CAD： CAD（Camouflaged Animal Dataset）是一个关于伪装物体的小数据集，首先由 [2] 提出。它总共包括 9 段从 YouTube 视频中提取的短视频序列。包含每 5 帧一次的手工标注真值掩模图。本文同时提供了基于双向一致性检查策略 [35] 的伪真值掩模，使得该数据集可以被进一步深入研究。

MoCA-Mask： 原始的 MoCA（Moving Camouflaged Animals）数据集 [19] 包括 37,000 帧图像，来自 141 个 YouTube 视频序列，分辨率和采样率在大多数情况下是 720×1280 和 24fps。该数据集包含在自然场景中运动的 67 种动物，但是有一些是非伪装的动物。原始数据集的真值是包围盒，而非稠密分割掩模。这使得评估 VCOD 分割性能变困难。为此，本文将该数据集重新整理为 *MoCA-Mask*，并且构建了一个包含全面的评价标准的评测基准。

4.2. 评测基准

指标： 本文采用如下指标来评估像素级掩模：(1) MAE (M)，用于测量预测值和真值之间的像素级差异。(2) 增强对齐指标（Enhanced-alignment measure）E_ϕ [8]，同时评测像素级别的匹配度和图像级别的统计量。这个指标自然地适合来评估伪装物体检测的整体和局部精确度。注意，实验报告的是 E_ϕ 平均值。(3) S 指标 (S_α) [7]，用来衡量区域和物体的结构相似度。(4) 带权重的 F 指标 F_β^w [27]，比传统的 F_β 更能提供可靠的评价结果。(5) 平均 Dice，用于度量两组数据之间的相似度。(6) 平均 IoU，用来度量两个掩模的重合度。

基准模型： 本文选择 9 个先进的基准模型，包括 **I.** 6 个基于图像的方法，即：EGNet [51]、BASNet [31] 、CPD [44]、PraNet [12]、SINet [11] 与 SINet-v2 [9]，以及 **II.** 3 个基于视频的方法，即：PNS-Net [15]、RCR-Net [46] 与 MotionGroup [47]。

实验设置： 本文主要和当前表现最好的基于图像的和

视频的基准模型进行对比。网络架构、输入分辨率、模态、预处理以及后处理都不相同，但是本文尽可能保证对比的公平性。对于基于图像的基准模型，本文采用 [11, 9] 提到的相同的数据预处理方法。具体来讲，图片在进行数据增强（随机翻转、旋转以及颜色增强）以后，将尺寸变换为 352×352。在训练阶段，本文将随机椒盐噪声加到真值图上。由于 EGNet [51] 需要额外的边缘信息来训练，本文采用与其论文中同样的预处理方式来得到边缘图。这些额外的数据包含在本文重新整理过后的 MoCA-Mask 数据集中。

大多数的基于视频的方法，如 PNS-Net [15] 与 RCRNet [46]，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，然后连接上时序模块来处理视频数据集。本文参考该训练策略，对于所有方法都在 COD10K [9] 训练集上进行预训练（除了 MotionGroup [47]，因为它并不包含静态模型）。实验中，加载在 COD10K 数据集预训练的权重，可以进一步提升模型在 MoCA-Mask 数据集的表现。相比 COD10K 图像数据集，MoCA-Mask 视频数据集更具挑战，因为其中存在摄像头运动、图像模糊、小尺寸动物、更小的身体结构，例如纤细的四肢和躯干。在某些视频片段中，动物只占整个画面的很小的一部分，这导致它们很难被发现（例如图6中所示的野山羊）。考虑到上面几点，本文提供了基于如下设置的结果：(a) 在 COD10K 训练的模型；(b) 在 MoCA-Mask 精调的模型，其中参数权重在 COD10K 预训练；(c) 在 CAD 整个数据集和 MoCA-Mask 的测试集上评估模型。

4.3. 结果

MoCA-Mask 数据集上的性能： 如表1所示，本文的方法明显优于对比方法。其中，在 S_α 指标上以 9.88% 的差距，超过当前最好的对比模型 RCRNet [46]，在 F_β^w 指标上以 92.97% 超过了 SINet [11] 模型。如图6，给出本文的方法与基准模型的定性对比。本文的模型可以在很多有挑战性的情况（例如物体有纤细的躯干或复杂的外观纹理、模糊或者突变运动）中，更加准确地定位与分割伪装物体。

CAD 数据集上的性能： 如表2所示，本文评估了不同方法在 CAD 数据集上的跨数据集的泛化性。本文的方法再次在所有 6 个评测指标上优于对比方法。进一步展现出本文方法的鲁棒性。如图7所示，本文的方法可以得到更加清晰的边界与更细粒度的可视化细节。这受益于模型在特征空间所构建的像素级相关性对。

Table 2: CAD 数据集上的定量结果。加粗表示最好的结果。本文模型在所有指标上都比对比模型更好。

Models	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$M \downarrow$	mDic	mIoU
EGNet [51]	0.619	0.298	0.666	0.044	0.324	0.243
BASNet [31]	0.639	0.349	0.773	0.054	0.393	0.293
CPD [44]	0.622	0.289	0.667	0.049	0.330	0.239
PraNet [12]	0.629	0.352	0.763	0.042	0.378	0.290
SINet [11]	0.636	0.346	0.775	0.041	0.381	0.283
SINet-v2 [9]	0.653	0.382	0.762	0.039	0.413	0.318
PNS-Net [15]	0.655	0.325	0.673	0.048	0.384	0.290
RCRNet [46]	0.627	0.287	0.666	0.048	0.309	0.229
MG [47]	0.594	0.336	0.691	0.059	0.368	0.268
SLT-Net (Ours)	**0.696**	**0.481**	**0.845**	**0.030**	**0.493**	**0.401**

Table 3: 在 MoCA-Mask 数据集上的 SLT-Net 的短期与长期模块的消融实验。加粗表示最好的结果。

Backbone	Short-term	Long-term	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$M \downarrow$
√			0.648	0.330	0.748	0.025
√	√		**0.662**	0.350	0.766	0.021
√	√	√	0.656	**0.357**	**0.785**	**0.021**

Table 4: 不同的时序信息处理策略。本文将 RCRNet [46] 的编码器替换为本文的基于 Transformer 的编码器，来评估不同处理策略所带来的性能的提升。“T” 表示 Transformer 编码器，“S” 表示单帧，“V” 表示视频输入，Δ 表示提升程度。加粗表示最好。

Model	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$M \downarrow$
RCRNet-TS	0.597	**0.206**	**0.618**	0.043
RCRNet-TV	**0.606**	0.204	0.617	**0.040**
RCRNet-Δ	1.51%	-0.97%	-0.16%	6.98%
SLT-Net-TS	0.648	0.330	0.748	0.025
SLT-Net-TV	**0.656**	**0.357**	**0.785**	**0.021**
SLT-Net-Δ	1.23%	8.18%	4.94%	16.00%

4.4. 消融实验

本文在 MoCA-Mask 数据集上进行消融实验。具体来讲，本文深入研究了：短期和长期模块的功能性分析，sequence-to-sequence 框架选择以及本文的伪掩模。

短期和长期模块： 本文从两个方面来评估短期和长期模块的有效性。首先，如表3所示，本文在 MoCA-Mask 数据集上对短期和长期模块进行了消融实验。通过添加短期模块，模型的性能在 S_α 上提升了 2.16%，在 F_β^w 上提升了 6.06%，在 E_ϕ 上提升了 2.41%，在 M 上提升了 16.00%，在 mDic 上提升了 4.53%，在 mIoU 上提升了 4.84%。通过添加长期模块，性能进一步在 F_β^w 提升了 2%，而在 S_α 有 0.91% 的微弱下降。

然后，本文在当前先进的 VSOD 模型 RCRNet [46] 上，用本文的基于 Transformer 的编码器来替换它原来

Figure 7: CAD 数据集上的定性实验。可以看出，本文的模型可以给出更精细的细节（蝎子），并且在运动中断时也可以表现的很好（青蛙）。这受益于特征体积的稠密关联对（dense correspondence pair of the feature volume）。

Table 5: 在 MoCA-Mask 测试集上，不同长期架构的消融实验。输入分辨率是 256×448。

Arch.	Variant	Params	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$M \downarrow$
ConvLSTM		179.03 MB	0.651	0.348	0.767	**0.021**
Transformer		82.30 MB	**0.656**	**0.357**	**0.785**	**0.021**

的编码器，用来比较这两种方法对于时序信息的处理策略，结果如表4所示。在时空一致性建模层面，在评估指标上，能看出操作所带来的指标的提升与下降，在 S_α 指标上提升 1.51%，在 F_β^w 上下降 0.97%，在 E_ϕ 上降低 0.16%，在 M 上提升 6.98%。

Transformer 与 ConvLSTM 的对比：本文用两种不同的方法来实现长期架构，分别是基于 Transformer 的方法和基于 ConvLSTM 的方法。对于 ConvLSTM 方法，本文采用了 [6] 提出的一个比较新的 ConvLSTM 模型变体，但是本文将其中 VGG 风格的 CNN 编码器解码器结构，用本文的 Transformer 风格的骨干网络代替。从表5中能看到，Transformer 变体相比 ConvLSTM 模型在 4 个指标上都更优，并且参数量更少。

伪掩模：如表1所示，尽管生成的伪标签包含有一些噪声，但是他们可以利用时序信息来抑制标签噪声，使得在视频方法中有更好的表现。对于静态图片的基准模型，几乎所有的方法都被标签噪声严重地影响，导致表现差于没有伪掩模的情况。这也证明了，在 VCOD 问题中，运动估计的误差是不能被忽略的，本文将其与分割误差同时优化，以取得更好的效果。

Table 6: 从头开始训练与 MoCA-Mask 上预训练对比。

Model	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$M \downarrow$
从头开始训练	0.655	0.351	0.764	0.024
预训练	**0.656**	**0.357**	**0.785**	**0.021**

在 MoCA-Mask 数据集上从头开始训练：从完整性考量，本文提供了模型使用预训练权重与不使用预训练的准确性的对比，如表6所示。该实验展示了本文的模型是否使用预训练的差别是微小的，仅在 S_α 上有 0.15% 的下降。

Table 7: 与 [19] 在 DAVIS16 数据集上的对比实验

Model	$\mathcal{J}_{Mean} \uparrow$	$\mathcal{J}_{Recall} \uparrow$	$\mathcal{F}_{Mean} \uparrow$	$\mathcal{F}_{Recall} \uparrow$
[19]	65.3	77.3	65.1	74.1
SLT-Net	77.96	95.49	78.65	92.08

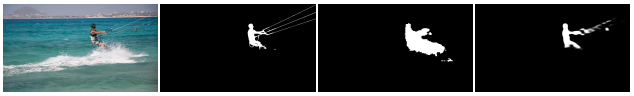

Figure 8: DAVIS16 数据集上的定性结果。MG [47] 模型无法正确给出预测结果，它错误地包含了溅起的水花。主要是因为不准确的光流估计（光流无法在模型训练时进行优化）。

泛化性：本文的模型可以应用于更一般的视频物体检测任务，例如视频实例分割。除了表1展示的与 MG [47] 的细节对比，本文还与 [19] 在 DAVIS16 数据集（表7）上进行对比并展现出优越性。

5. 结论

本文提出了一个用于视频伪装物体分割的方法 SLT-Net。具体来讲，本文使用短期模块来隐式捕捉连续帧之间的运动，使得本文可以在一个框架下同时优化运动的估计和分割。本文还提出了一个基于 sequence-to-sequence Transformer 的长期模块，来保证视频序列上的时序一致性。为了促进 VCOD 领域的发展，本文重新构造了名为 **MoCA-Mask** 的新数据集，包括了 87 段高质量视频序列，共计 22,939 帧。是 VCOD 领域目前最大的像素级标注数据集，使得物体级别评测成为可能。对比了当前最先进的基准模型，本文提出的网络在两个 VCOD 评测基准上均取得了最优的表现。

更广泛的影响：伪装物体检测可以被用于检测和保护稀有物种，防止野生动物贩卖，医疗应用（例如：息肉或者肺部感染的检测），以及搜救工作等。值得说明的是，MoCA-Mask 数据集并不包含任何军事或敏感场景。除了上述重要的使用场景，本文的文章还向着，从带噪声的运动信息中进行视频内容理解的方向，迈出了扎实的一步。

References

[1] Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context r-cnn: Long term temporal context for per-camera object detection. In *CVPR*, 2020.

[2] Pia Bideau and Erik Learned-Miller. It' s moving! a probabilistic model for causal motion segmentation in moving camera videos. In *ECCV*, 2016.

[3] Pia Bideau, Rakesh R Menon, and Erik Learned-Miller. Moa-net: self-supervised motion segmentation. In *ECCV Workshops*, 2018.

[4] Xuelian Cheng, Huan Xiong, Deng-Ping Fan, Yiran Zhong, Mehrtash Harandi, Tom Drummond, and Zongyuan Ge. Implicit motion handling for video camouflaged object detection. In *CVPR*, 2022.

[5] Xuelian Cheng, Yiran Zhong, Yuchao Dai, Pan Ji, and Hongdong Li. Noise-aware unsupervised deep lidar-stereo fusion. In *CVPR*, 2019.

[6] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *ICML*, 2018.

[7] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A New Way to Evaluate Foreground Maps. In *ICCV*, 2017.

[8] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *IJCAI*, 2018.

[9] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE TPAMI*, 2021.

[10] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *SSI*, 2021.

[11] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, 2020.

[12] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*, 2020.

[13] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, 2019.

[14] Suyog Dutt Jain, Bo Xiong, and Kristen Grauman. Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos. In *CVPR*, 2017.

[15] Ge-Peng Ji, Yu-Cheng Chou, Deng-Ping Fan, Geng Chen, Huazhu Fu, Debesh Jha, and Ling Shao. Progressively normalized self-attention network for video polyp segmentation. In *MICCAI*, 2021.

[16] Ge-Peng Ji, Keren Fu, Zhe Wu, Deng-Ping Fan, Jianbing Shen, and Ling Shao. Full-duplex strategy for video object segmentation. In *ICCV*, 2021.

[17] Pan Ji, Yiran Zhong, Hongdong Li, and Mathieu Salzmann. Null space clustering with applications to motion segmentation and face clustering. In *ICIP*, 2014.

[18] Karthika Suresh Kumar and Aamer Abdul Rahman. Early detection of locust swarms using deep learning. In *MLCI*. 2021.

[19] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. In *ACCV*, 2020.

[20] Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *CVIU*, 2019.

[21] Trung-Nghia Le and Akihiro Sugimoto. Deeply supervised 3d recurrent fcn for salient object detection in videos. In *BMVC*, 2017.

[22] Trung-Nghia Le and Akihiro Sugimoto. Video salient object detection using spatiotemporal deep features. *IEEE TIP*, 2018.

[23] Dongxu Li, Chenchen Xu, Kaihao Zhang, Xin Yu, Yiran Zhong, Wenqi Ren, Hanna Suominen, and Hongdong Li. Arvo: Learning all-range volumetric correspondence for video deblurring. In *CVPR*, 2021.

[24] Guanbin Li, Yuan Xie, Tianhao Wei, Keze Wang, and Liang Lin. Flow guided recurrent neural encoder for video salient object detection. In *CVPR*, 2018.

[25] Ting Liu, Yao Zhao, Yunchao Wei, Yufeng Zhao, and Shikui Wei. Concealed object detection for activate millimeter wave image. *IEEE TIE*, 2019.

[26] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *CVPR*, 2021.

[27] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, 2014.

[28] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *CVPR*, 2021.

[29] Jeff Michels, Ashutosh Saxena, and Andrew Y Ng. High speed obstacle avoidance using monocular vision and reinforcement learning. In *ICML*, 2005.

[30] Anestis Papazoglou and Vittorio Ferrari. Fast object segmentation in unconstrained video. In *ICCV*, 2013.

[31] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019.

[32] Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosformer: Rethinking softmax in attention. In *ICLR*, 2022.

[33] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *CVPR*, 2019.

[34] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*, 2018.

[35] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020.

[36] Pavel Tokmakov, Karteek Alahari, and Cordelia Schmid. Learning motion patterns in videos. In *CVPR*, 2017.

[37] Jianyuan Wang, Yiran Zhong, Yuchao Dai, Stan Birchfield, Kaihao Zhang, Nikolai Smolyanskiy, and Hongdong Li. Deep two-view structure-from-motion revisited. In *CVPR*, 2021.

[38] Jianyuan Wang, Yiran Zhong, Yuchao Dai, Kaihao Zhang, Pan Ji, and Hongdong Li. Displacement-invariant matching cost learning for accurate optical flow estimation. In *NeurIPS*, 2020.

[39] Wenguan Wang, Jianbing Shen, and Ling Shao. Video salient object detection via fully convolutional networks. *IEEE TIP*, 2017.

[40] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvtv2: Improved baselines with pyramid vision transformer. *arXiv preprint arXiv:2106.13797*, 2021.

[41] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021.

[42] Jun Wei, Shuhui Wang, and Qingming Huang. F^3net: Fusion, feedback and focus for salient object detection. In *AAAI*, 2020.

[43] Yu-Huan Wu, Shang-Hua Gao, Jie Mei, Jun Xu, Deng-Ping Fan, Rong-Guo Zhang, and Ming-Ming Cheng. Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation. *IEEE TIP*, 2021.

[44] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, 2019.

[45] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, 2015.

[46] Pengxiang Yan, Guanbin Li, Yuan Xie, Zhen Li, Chuan Wang, Tianshui Chen, and Liang Lin. Semi-supervised video salient object detection using pseudo-labels. In *ICCV*, 2019.

[47] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *ICCV*, 2021.

[48] Zhao Yang, Qiang Wang, Luca Bertinetto, Weiming Hu, Song Bai, and Philip HS Torr. Anchor diffusion for unsupervised video object segmentation. In *CVPR*, 2019.

[49] Qiang Zhai, Xin Li, Fan Yang, Chenglizhao Chen, Hong Cheng, and Deng-Ping Fan. Mutual graph learning for camouflaged object detection. In *CVPR*, 2021.

[50] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, and Ling Shao. Rgb-d

saliency detection via cascaded mutual information minimization. In *ICCV*, 2021.

[51] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet:edge guidance network for salient object detection. In *ICCV*, 2019.

[52] Yiran Zhong, Yuchao Dai, and Hongdong Li. 3d geometry-aware semantic labeling of outdoor street scenes. In *ICPR*, 2018.

[53] Yiran Zhong, Pan Ji, Jianyuan Wang, Yuchao Dai, and Hongdong Li. Unsupervised deep epipolar flow for stationary or dynamic scenes. In *CVPR*, 2019.

[54] Yiran Zhong, Hongdong Li, and Yuchao Dai. Open-world stereo video matching with deep rnn. In *ECCV*, 2018.