多尺度视频变换器在自主驾驶中的类别无关分割

Leila Cheshmi* **Mennatullah Siam**

Abstract

确保自动驾驶的安全是一项复杂的挑战，需要处理未知物体和不可预见的驾驶场景。在这项工作中，我们专注于开发多尺度视频转换器，这些转换器能够仅使用运动线索来检测未知物体。视频语义/全景分割通常依赖于训练期间看到的一组封闭已知类别，忽略了未知或新颖的类别。最近利用多模态大型语言模型（*LLMs*）进行视觉定位计算成本高昂，因为它们充分利用了 *LLM* 的能力，特别是在使用更大变体以实现更高精度时以及在考虑像素级输出时。我们设想这样的模型与为安全目的设计的轻量级高效模型之间存在相互作用，如我们的模型。为了应对实时关键安全场景中的这一限制，引入了视频类不可知分割（*VCAS*）任务，允许识别此类未见物体。我们提出了一种高效的视频转换器，该转换器被训练用于视频类不可知分割，并且不使用光流作为输入。我们的方法依赖于一种新颖的多阶段多尺度查询-记忆解码和特定比例随机丢弃标记以确保效率和准确性。它在解码层之间保持详细的时空特征，使用一个共享的学习型内存模块。与传统的基于查询的多尺度解码器压缩特征（存在丢失精细空间细节的风险）不同，我们的以记忆为中心的设计保留了多个比例下的高分辨率信息，提高了分割质量。此外，我们方法避免了计算光流带来的额外开销，支持实时性能。我们在多个视频分割基准上评估了我们的方法，包括 *DAVIS' 16* 等通用基准和城市驾驶场景，如 *KITTI* 和 *Cityscapes*。与传统的多尺度基线相比，我们的方法在 *GPU* 内存使用和运行时间效率方面始终有显著优势。我们的结果突显了利用视频转换器实现实时、稳健的密集预测，在关键安全机器人应用中的一个有前途的方向。

1 介绍

视频分割 [1, 2] 是计算机视觉中的一项基础但具有挑战性的任务，涉及在视频帧之间识别具有特定语义或物理属性的对象。它在自动驾驶中起着至关重要的作用，在那里理解动态场景对于安全导航至关重要。视频分割方法通常根据推理过程中人类参与的程度进行分类：交互式、半自动和全自动。全自动视频对象分割（AVOS）的方法包括基于光流的两流模型 [3, 4, 5] 和更近期的基于变换器的融合方法 [3, 4, 6]。语义 [7]、实例 [8] 和全景 [9] 分割及其在视频中的扩展 [10, 11] 已经得到了广泛的研究。其中一个最突出的方法，Mask2Former[12]，引入了一种依赖于掩码分类损失和使用了掩码注意力和学习到的多尺度表示的变压器解码器的通用分割技术。

*Corresponding author. Email: leila.cheshmi@ontariotechu.net

Preprint. Under review.

然而，这些模型中的许多是在具有已知对象类别的封闭集数据集上训练的。这导致了在自动驾驶等关键安全应用中存在一个主要限制，即无法分割预定义标签空间之外的未知物体（例如，从移动卡车掉落的碎片、不寻常的移动障碍物或罕见动物）。最近的视觉语言模型 [13, 14, 15] 通过文本提示展示了强大的零样本/少样本识别、定位和推理能力。最近发布的模型 Sa2VA[14] 结合了 SAM-2 与多模态大型语言模型（MLLMs），用于密集视觉定位。然而，这些模型仍然依赖于语言描述，偏向于静态的每帧信息，理解时间顺序的能力较弱 [16, 17]，并且需要更高的计算资源。我们的方法消除了对语言监督的需求，直接基于运动模式检测未知物体，减少了计算开销，使其更适合实时应用。我们认为在机器人应用程序中，将存在这样的通用计算密集型 MLLMs 与旨在提高可靠性和安全性的轻量级高效技术之间的相互作用，如我们所提出的。

AVOS 方法无需手动初始化，可直接从原始视频流中推断出对象掩码。至关重要的是，AVOS 为视频类无关分割 [18] 提供了一个基础，这项任务明确要求对场景中的物体进行分割，而不考虑它们的语义类别。通过利用单目视频中的外观、运动和几何线索，这一范式能够检测未知对象。为了克服效率方面的局限以及仅关注某些特定语义类别的问题，最近的研究探索了替代策略 [19]。我们专注于将视频类无关分割扩展为包含更长的时间上下文，在此情况下我们的视频变换器接收多个 RGB 帧的输入片段，即超过两帧，并且不依赖显式的光流作为输入。

受 Mask2Former 的多尺度变换解码器启发，近期的一些多尺度视频变换架构 [20, 21] 提出了无类别和实例感知的变体，无需使用每段掩码分类损失。这些方法可以很容易地扩展到语义分割，并且在训练过程中不需要实例、全景或跟踪标注。这特别适用于获取此类注释成本高昂的领域，例如医学、机器人技术或遥感领域。受此启发，我们的工作重点在于改进多尺度解码机制，以进一步提高其效率同时保持强大的性能。我们提出了一种多阶段多尺度查询-内存变换解码器，该解码器通过共享内存注意力增强压缩查询，并直接在高分辨率的时空特征图上操作。此外，我们还提出了一种计算高效的机制，使用特定规模的随机丢弃标记来减少其 GPU 内存占用和推理时间。

2 方法

架构概述。我们的工作提出了一种端到端的多尺度视频变换器，旨在用于无类别特定分割的视频处理，并可在无需光流输入的情况下应用于自动驾驶环境中。我们的架构总结如图 1所示，该架构以包含 T 帧的原始 RGB 图像片段为输入，并输出相应的类别无关分割掩码。它包括一个依赖于 Video-Swin [22] 的卷积主干网络，用于提取时空特征 $\mathbb{X} = \{X_l\}_{l=1}^{L}$，共 L 个尺度级别。每个特征图 X_l 然后被缩减并展平，$\bar{\mathbb{X}} = \{\bar{X}_l\}_{l=1}^{L}$，其中，$\bar{X}_l \in \mathbb{R}^{TH_lW_l \times D}$、$H_l$ 和 W_l 是空间维度，而 D 是特征维度。最粗略的尺度表示通过一个应用自注意力机制的变压器编码器，并通过特征金字塔网络（FPN）[23] 与多尺度表示融合 ($\bar{\mathbb{F}} = \mathcal{E}(\bar{\mathbb{X}})$。这些特征被输入一个多阶段多尺度查询-记忆解码器变换器 $O_L = \mathcal{D}(\bar{\mathbb{F}})$，该变换器保留了空间信息。输出 O_L 表示在进行这种多尺度交换后具有最高分辨率的密集特征，这确保了输入时空特征之间的区域级和像素级交互更佳。接下来是一个轻量级分割头 $M = \mathcal{H}(O_L)$，它直接操作这些时空特征以生成每个帧的密集二进制掩码 M。分割头包含多个 3D 卷积层。

多阶段多尺度查询-记忆 Transformer 解码器。在本节中，我们详细说明了我们的变换解码器的操作，$\mathcal{D}$。我们通常使用可学习查询的紧凑表示形式，$\mathcal{O}^{\text{query}}$，这些查询可以作为输入图像/片段的主要部分的视觉摘要。此操作被称为*多尺度查询解码*。假设一个自注意力后跟交叉注意力操作使用多头注意力，$\mathcal{A}(Q, K, V)$，分别将输入，Q, K, V 作为查询、键和值。我们通过

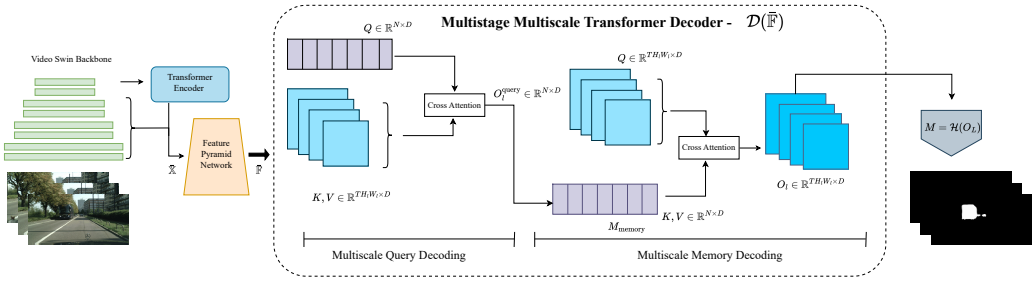

图 1: 我们提出的多阶段多尺度查询记忆变压器架构概述，用于自动驾驶中的视频类别无关分割。它使用 Video-Swin 主干网络在多个尺度上提取时空特征，$\bar{\mathbb{X}}$，并随后通过变压器编码器和 FPN，产生金字塔特征，$\bar{\mathbb{F}}$。这些特征经过我们两阶段的多尺度查询记忆解码器处理，$\mathcal{D}$。在第一阶段中，进行多尺度查询解码，其中每个尺度，l，使用一组压缩学习到的查询，O_l^{query}，来表示剪辑中的不同片段。然后是第二阶段的多尺度内存解码，它们作为共享可训练记忆用于精细化每个尺度的时空上下文，O_l，同时保留细粒度信息。每个阶段包括使用查询与键值之间的交叉注意力进行迭代多尺度交换，分别对应于 Q 和 K, V。我们仅展示最后一次交换及其输出的使用情况，并为简单起见省略了自注意力操作。最终的高分辨率特征通过分割头解码以生成密集二进制运动掩模，$M = \mathcal{H}(O_L)$。

跨注意力学习 N 紧凑的表示，$\mathcal{O}_l^{\text{query}} \in \mathbb{R}^{N \times D}$ 用于每个尺度，l，如下所示，

$$\mathcal{O}_l^{\text{query}} = \mathcal{A}(\mathcal{O}_{l-1}^{\text{query}} + \mathcal{P}, \bar{X}_l + \mathcal{P}_l^{\text{sc}} + \mathcal{P}_l^{\text{st}}, \bar{X}_l),$$

其中 $\mathcal{P}, \mathcal{P}_l^{\text{sc}}, \mathcal{P}_l^{\text{st}}$ 是可学习的查询位置嵌入、可学习的尺度级别嵌入和固定的时空位置嵌入，分别。跨注意力操作采用来自前一个尺度的可学习查询 $l-1$，以学习当前尺度的查询 l。在第一次跨注意力中，我们在训练期间随机初始化这些查询。这些表示在学习紧凑的区域级别信息方面具有优势，可以改善分割。然而，它们在跨尺度交换过程中会损失精细细节和空间精度。因此，我们选择执行一种连续操作，在检索密集的时空细节的同时利用这些学习到的紧凑表示，我们将此称为*多尺度记忆解码*。我们在每个尺度级别上的输入特征 $\bar{X}_l$ 与共享可学习记忆 $\mathcal{M}_{\text{memory}}$ 之间计算注意力。每个输出 $\mathcal{O}_l$ 是通过应用交叉注意力操作获得的，该操作同样包含位置嵌入 $\mathcal{P}, \mathcal{P}_l^{\text{sc}}, \mathcal{P}_l^{\text{st}}$，如前所述，

$$\mathcal{O}_l = \mathcal{A}\left(\bar{X}_l + \mathcal{P}_l^{\text{sc}} + \mathcal{P}_l^{\text{st}}, \mathcal{M}_{\text{memory}} + \mathcal{P}, \mathcal{M}_{\text{memory}}\right).$$

与*多尺度查询解码*形成对比，丰富的特征图 $\bar{X}_l$ 作为查询，共享学习内存 $\mathcal{M}_{\text{memory}}$ 作为键和值。此设计保持了原始的空间和时间分辨率，确保输出 $\mathcal{O}_l$ 维持其维度 $(TH_lW_l \times D)$，这对于密集预测任务至关重要。共享的已学习内存 $\mathcal{M}_{\text{memory}}$ 用先前学习的查询 O_L^{query} 进行初始化。随后，输出 $\mathcal{O}_l$ 与下一个输入 $\bar{X}_{l+1}$ 通过混合操作进行融合，

$$\bar{X}_{l+1} = \mathcal{M}(\bar{X}_{l+1}, \mathcal{O}_l) = \bar{X}_{l+1} + \mathcal{I}_{l+1}^l(\mathcal{O}_l),$$

其中 $\mathcal{I}_{l+1}^l$ 执行双线性插值以匹配从尺度 l 到尺度 $l+1$ 的空间分辨率。这种融合允许在不同尺度上迭代细化特征。这个*多阶段多尺度查询记忆模块*是多尺度变换解码的一种变体，改进了传统的单阶段多尺度查询解码的性能。在每个阶段，跨尺度的多尺度交换 $\{1, 2, ..., L\}$ 可以重复多次迭代 R，作为从粗到细的细化。为了减少计算和内存使用，我们在解码器中采用了随机丢弃标记策略，在最精细的两个尺度上特定于自注意力操作，在执行交叉注意力之前。这意味着在最精细的尺度下进行较少的计算以确保高效的内存占用和计算。基于*令牌保留比例* $r \in (0, 1]$ 随机选择一组标记子集，例如，在 $r = 0.5$ 处我们只保留一半的标记。经过细化的最精细尺度特征 O_L 用于预测。

推理方法	戴维斯'16	VCAS-KITTI	VCAS-城市景观
Multiscale Query (Baseline)	81.3	52.7	56.3
Multi-stage Multiscale Query-Memory (Ours)	**84.2**	**54.6**	**59.3**

表 1: 与基线的比较使用传统的多尺度解码跨越三个数据集。我们的方法始终优于基线，该基线是从 Mask2Former [12] 适应而来的。

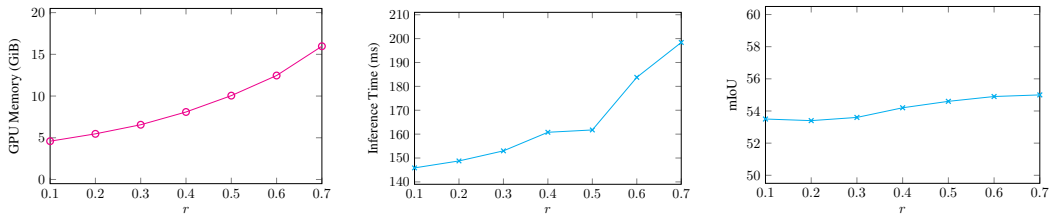

图 2: 保持比率，r，分析在 VCAS-KITTI 数据集上的表现：（左）GPU 内存消耗以 GB 为单位，（中）每帧推理时间以毫秒为单位，（右）分割质量（mIoU）。

3 实验结果

实验设置。我们在 DAVIS'16 和 VCAS [18] 数据集上评估了我们的方法，该数据集包括两个部分，分别是重新用于视频类不可知分割的 KITTI 和 Cityscapes（即 VCAS-KITTI, VCAS-Cityscapes）。所有模型都在 $T = 5$ 上训练了 15 个周期，使用 $R = 9$transformer 解码器层和 $L = 4$ 多尺度级别。每个 transformer 块有 $D = 384$ 隐藏维度和八个注意力头。我们学习 $N = 5$ 查询，从而决定了内存条目的数量。我们使用了学习率为 1×10^{-4} 和 Video-Swin [22] 预训练骨干架构。代币保留比例在主要实验中设置为 0.5。训练是在配备 40GB 内存的 NVIDIA A100 GPU 上进行的。我们比较了两种模型：(i) 多尺度查询（基线），使用传统的多尺度视频变换解码器，以及 (ii) 多阶段多尺度查询-内存（我们的）。

消融研究。
分割精度。我们提出的视频转换器在表 1中始终优于多尺度基线。结果验证了内存感知解码比压缩查询更有效地保留时空细节。两阶段解码也促进了跨尺度的分层细化，进一步提高了复杂场景中的分割准确性。

计算效率。图 2显示了变化的标记保留比率对 GPU 内存使用、推理时间和分割质量（mIoU）的影响，使用的数据集为 VCAS-KITTI。它仅影响我们变压器解码器中的自注意力机制，而在交叉注意力中使用所有标记。当标记保留比率从 0.1 增加到 0.7 时，GPU 内存消耗显著增加，表明保留更多标记的计算成本更高。同样，每帧的推理时间稳步增加。有趣的是，mIoU 曲线显示随着标记数量的增多略有提升。因此，使用较少的标记在计算效率指标上提供了明显的优势，同时对 mIoU 的影响最小。比率超过 0.7 会导致内存不足的问题。

4 结论

我们提出了一种多阶段多尺度查询-记忆转换器解码器，用于视频类别无关的分割，旨在实现自动驾驶中的实时安全。我们的以内存为中心的设计在减少计算开销的同时保留了高分辨率的时空特征。实验结果表明，在 DAVIS' 16、KITTI 和 Cityscapes 数据集上，我们的方法优于传统的多尺度基线方法。我们的方法为安全关键型机器人应用中未知物体检测建立了高效的转换器基础。我们计划通过精心策划一个训练数据集中未包含移动对象集合的评估基准来

进一步探索此类未知对象分割的基准测试工作。此外，我们计划比较实例级与像素级模型，例如我们的模型，在上述基准上的泛化能力和鲁棒性。

参考文献

[1] T. Zhou, F. Porikli, D. J. Crandall, L. Van Gool, and W. Wang, ‘‘A survey on deep learning technique for video segmentation,’’ *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7099–7122, 2023.

[2] R. Yao, G. Lin, S. Xia, J. Zhao, and Y. Zhou, ‘‘Video object segmentation and tracking: A survey,’’ *ACM Trans. Intell. Syst. Technol.*, vol. 11, May 2020.

[3] T. Zhou, J. Li, S. Wang, R. Tao, and J. Shen, ‘‘Matnet: Motion-attentive transition network for zero-shot video object segmentation,’’ *IEEE transactions on image processing*, vol. 29, pp. 8326–8338, 2020.

[4] S. Ren, W. Liu, Y. Liu, H. Chen, G. Han, and S. He, ‘‘Reciprocal transformations for unsupervised video object segmentation,’’ in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15455–15464, 2021.

[5] M. Siam, C. Jiang, S. Lu, L. Petrich, M. Gamal, M. Elhoseiny, and M. Jagersand, ‘‘Video object segmentation using teacher-student adaptation in a human robot interaction (hri) setting,’’ in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 50–56, IEEE, 2019.

[6] Y. Yuan, Y. Wang, L. Wang, X. Zhao, H. Lu, Y. Wang, W. Su, and L. Zhang, ‘‘Isomer: Isomerous transformer for zero-shot video object segmentation,’’ in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 966–976, 2023.

[7] J. Long, E. Shelhamer, and T. Darrell, ‘‘Fully convolutional networks for semantic segmentation,’’ in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

[8] K. He, G. Gkioxari, P. Dollár, and R. Girshick, ‘‘Mask r-cnn,’’ in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

[9] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollár, ‘‘Panoptic segmentation,’’ in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9404–9413, 2019.

[10] L. Yang, Y. Fan, and N. Xu, ‘‘Video instance segmentation,’’ in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5188–5197, 2019.

[11] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, ‘‘Video panoptic segmentation,’’ in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9859–9868, 2020.

[12] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, ‘‘Masked-attention mask transformer for universal image segmentation,’’ in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

[13] M. R. I. Hossain, M. Siam, L. Sigal, and J. J. Little, ‘‘The power of one: A single example is all it takes for segmentation in vlms,’’ *arXiv preprint arXiv:2503.10779*, 2025.

[14] H. Yuan, X. Li, T. Zhang, Z. Huang, S. Xu, S. Ji, Y. Tong, L. Qi, J. Feng, and M.-H. Yang, ''Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos,'' *arXiv preprint arXiv:2501.04001*, 2025.

[15] Y. Li, C. Fan, C. Ge, Z. Zhao, C. Li, C. Xu, H. Yao, M. Tomizuka, B. Zhou, C. Tang, *et al.*, ''Womd-reasoning: A large-scale dataset for interaction reasoning in driving,'' *arXiv preprint arXiv:2407.04281*, 2024.

[16] O. Zohar, X. Wang, Y. Dubois, N. Mehta, T. Xiao, P. Hansen-Estruch, L. Yu, X. Wang, F. Juefei-Xu, N. Zhang, *et al.*, ''Apollo: An exploration of video understanding in large multimodal models,'' in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18891–18901, 2025.

[17] Z. Xue, M. Luo, and K. Grauman, ''Seeing the arrow of time in large multimodal models,'' *arXiv preprint arXiv:2506.03340*, 2025.

[18] M. Siam, A. Kendall, and M. Jagersand, ''Video class agnostic segmentation benchmark for autonomous driving,'' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2825–2834, June 2021.

[19] R. Inoue, M. Tsuchiya, and Y. Yasui, ''Channel-wise motion features for efficient motion segmentation,'' in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9557–9564, IEEE, 2024.

[20] R. Karim, H. Zhao, R. P. Wildes, and M. Siam, ''Med-vt: Multiscale encoder-decoder video transformer with application to object segmentation,'' in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6323–6333, 2023.

[21] M. Siam, R. Karim, H. Zhao, and R. Wildes, ''Multiscale memory comparator transformer for few-shot video segmentation,'' *arXiv preprint arXiv:2307.07812*, 2023.

[22] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, ''Video swin transformer,'' in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3202–3211, 2022.

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, ''Feature pyramid networks for object detection,'' in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.