문맥 인지 적응형 마스크 선택을 통한 일관된 참조 비디오 객체 분할

서준혁, 윤영우, 최원준, 이동규*

경북대학교

ssam2s@knu.ac.kr, dbsduddn@knu.ac.kr, sangju@knu.ac.kr, dglee@knu.ac.kr

Context-Aware Adaptive Mask Selection for Consistent Referring Video Object Segmentation

Jun-Hyeok Seo, Young-Woo Youn, Wonjun Choi, Dong-Gyu Lee*

Kyungpook National University

요 약

본 논문에서는 참조 비디오 객체 분할에서 프레임 간 시간적 일관성과 객체 수 정보에 기반한 문맥 인지 적응형 마스크 선택 기법을 제안한다. 기존의 참조 비디오 객체 분할 작업은 연속 프레임에서 시간적 일관성을 유지하는데 어려움을 겪는다. 이를 해결하기 위해 본 논문에서는 CLIP 기반 키프레임 선정, 마스크 일관성 점수 계산, 텍스트 기반 객체 수 판별, 그리고 시간적 마스크 보간으로 구성된 새로운 프레임워크를 제안한다. 제안하는 프레임워크는 벤치마크 데이터셋을 이용한 성능 비교에서 우수한 성능을 보임으로써 제안한 프레임워크가 참조 비디오 객체 분할 작업에 효과적임을 보여준다.

Ⅰ. 서 론

최근 딥러닝 기반 영상 이해 분야에서는 자연어로 지칭된 특정 객체를 영상 내에서 정확히 분할하는 참조 비디오 객체 분할(Referring Video Object Segmentation, RVOS)이 활발히 연구되고 있다[1, 2]. RVOS는 비디오 시퀀스와 함께 주어진 텍스트 설명을 기반으로, 해당 설명에 부합하는 객체의 픽셀 단위 마스크를 예측하는 문제로 정의된다. 그러나 기존 RVOS 연구들은 장면 내 시각적 폐색, 객체의 형태 변화, 배경 변화뿐 아니라, 프레임 간 시간적 일관성을 유지하는데 어려움을 겪는다. 특히, 프레임별로 독립적으로 예측된 마스크는 인접 프레임 간 경계 위치나 형태가 불안정해지는 문제가 발생하며, 이는 전체 비디오에서의 객체 추적 품질을 저하한다. 또한, 텍스트로 구성된 하나의 참조 표현이 복수 객체를 지칭하는 경우, 단일 객체 기반의 분할 방법은 정확도가 저하되는 한계를 가진다. 최근에는 대규모 비전-언어 모델(LVLM)과 사전학습된 범용 분할 모델을 결합하여 이러한 문제를 완화하려는 시도가 제안되었다. 예를 들어, Sa2VA[3]나 HyperSeg[4]와 같은 최신 프레임워크는 시각-언어 정렬을 통해 텍스트와 영상 간 의미적 대응을 강화하고, 범용 마스크 예측을 가능하게 한다. 그러나 이러한 방법들 역시 모델별 특성 차이로 인해 동일 장면에서도 결과 품질이 상이하며, 이를 효과적으로 결합하고 선택하는 방법은 충분히 연구되지 않았다.

이에 본 연구에서는 문맥 인지 적응형 마스크 선택 방법을 기반으로 한 RVOS 방법을 제안한다. 제안 기법은 다음과 같은 세 가지 핵심 요소로 구성된다. (1) CLIP[5] 유사도 기반 키프레임 선택과 양방향 마스크 비교를 통해 가장 일관성이 높은 마스크를 선택하는 일관성 인지형 마스크 선택, (2) 대규모 언어 모델을 활용해 참조 표현의 단·복수를 판별하고 객체 수에 따라 마스크를 조합하는 텍스트 기반 객체 수 판별, (3) 인접 프레임 정보를 활용하여 비어 있는 예측을 보완하는 시간적 마스크 보간이다.

본 논문에서는 대표적인 벤치마크 데이터셋에 대한 기존 연구들과의 객관적인 성능 비교를 진행하고, 제안하는 핵심 요소에 대한 절제 실험을 통해 제안하는 방법이 효과적임을 증명한다.

Ⅱ. 본론

본 논문에서는 문맥 인지 적응형 마스크 선택 방법 기반의 RVOS 프레임워크를 제안한다. 이는 주어진 영상과 참조 표현에 대하여, 시간적 일관성과 객체 수 정보를 고려하여 최적의 분할 마스크를 생성하는 것을 목표로 한다. 이를 위하여 전체 파이프라인은 (1) 일관성 인지형 마스크 선택, (2) 텍스트 기반 객체 수 판별, (3) 시간적 마스크 보간의 세 단계로 구성된다. 제안하는 프레임워크는 그림 1과 같이 구성되며, 비디오 시퀀스 $V=\{f_1, f_2, ..., f_T\}$와 참조 표현 E를 입력으로 받아, 최종 마스크 시퀀스 $\hat{M}=\{\hat{m}_1, \hat{m}_2, ..., \hat{m}_T\}$를 산출한다.

RVOS 문제에서 프레임별로 독립적인 마스크 예측은 인접 프레임 간 경계 불일치와 형태 변형을 초래할 수 있다. 본 연구에서는 이러한 문제를 완화하기 위하여 최첨단 모델들의 결과를 앙상블하여 시간적 안정성이 가장 높은 마스크를 선택한다. 우선, CLIP 기반 텍스트-이미지 유사도 함수를 정의하여 각 프레임 f_t에 대한 CLIP 임베딩과 주어진 참조 표현 E와의 코사인 유사도를 산출한다. 여기서 가장 높은 유사도를 갖는 프레임을 키프레임 f_k로 설정한다. 키프레임의 계산 과정은 아래 수식 (1)과 같다.

$$s_t = \frac{\phi_{img}(f_t) \cdot \phi_{text}(E)}{\| \phi_{img}(f_t) \| \, \| \phi_{text}(E) \|} \quad (1)$$

여기서 ϕ_{img}와 ϕ_{text}는 각각 CLIP의 이미지와 텍스트 인코더를 의미한다. 이후, 모델 A와 B가 각각 예측한 마스크의 픽셀 값이 1인 위치의 평균 좌표를 산출하여 중심점을 계산하고, 이전에 선택된 프레임과 현재 프레임의 중심점의 유클리드 거리를 측정한다. 이전 프레임의 중심점을 c_t라고 할 때, 중심점 거리는 아래 수식 (2)로 계산되며, 거리가 더 작은 마스크가 채택된다.

$$d_t = \| c_{t+1} - c_t \|_2 \quad (2)$$

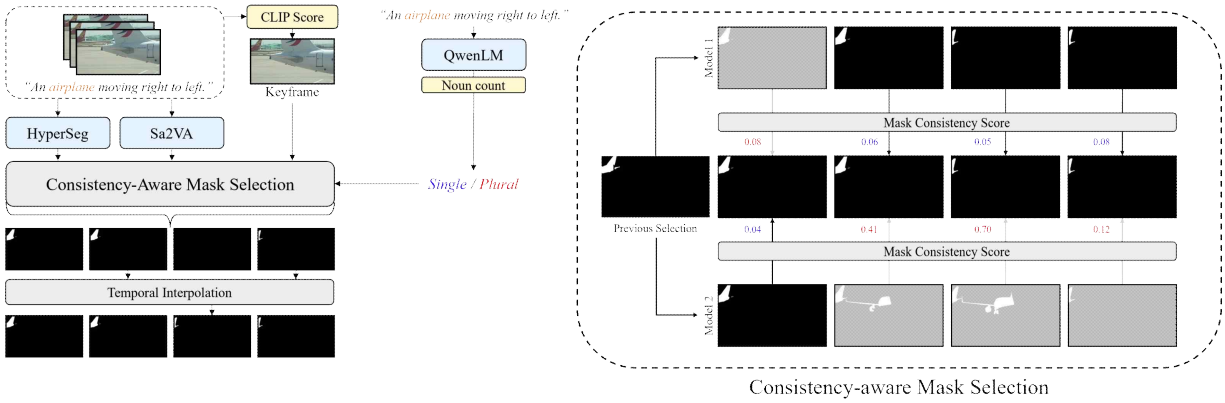

그림 1. 제안하는 프레임워크의 구조

참조 표현이 복수 객체를 지칭하는 경우, 단일 객체 분할 전략은 성능 저하를 초래한다. 이를 해결하기 위해 Qwen[6]을 이용하여 입력 텍스트의 명사 수를 판별하고, 단·복수 여부에 따라 마스크 선택 방식을 달리한다. 단수 객체의 경우 일관성 점수를 기반으로 한 단일 마스크를 선택하며, 복수 객체의 경우 두 모델이 예측한 마스크를 결합하여 다중 객체를 동시에 포함하도록 한다. 일부 프레임에서는 예측에 실패해 마스크가 비어 있는 경우, 시간적 불연속성이 발생한다. 이를 방지하기 위해, 해당 프레임의 이전 및 이후 프레임에서 예측된 마스크의 평균을 취하여 보간된 마스크를 생성한다. 이를 통해 연속적인 객체 추적과 시각적 일관성을 확보하도록 한다.

실험은 MeViS[7] 벤치마크 데이터셋을 사용하여 수행되었으며, 평가는 Jaccard Index(J), F-measure(F), 그리고 J&F 평균을 사용하였다. 실험에서는 Sa2VA-26B와 HyperSeg 두 모델을 앙상블하였다.

Methods	$J\&F$	J	F
MeViS	40.23	36.51	43.90
HyperSeg	58.95	55.32	62.59
Grounded-SAM2	59.26	54.59	63.92
Sa2VA-4B	55.10	50.74	59.46
Sa2VA-8B	59.28	54.95	63.62
Sa2VA-26B	61.38	56.53	66.23
Ours	**64.80**	**60.42**	**69.10**

표 1. 벤치마크 데이터셋에 대한 성능 결과 비교

실험 결과 본 논문에서 제안하는 방법이 기존 방법들보다 우수한 성능을 보이는 것을 확인할 수 있었다. 더 나아가, 본 논문에서 제안한 핵심 요소들의 기여를 분석하기 위해 절제 실험을 표 2와 같이 수행하였다. 그 결과, 모든 모듈을 결합했을 때 성능이 가장 높게 나타났으며, 특히 마스크 선택 기법과 객체 수 판별 모듈이 성능에 가장 큰 기여를 하는 것으로 확인되었다.

Mask Selection	Noun Counting	Interpolation	Keyframe	$J\&F$
✓	-	-	-	63.65
✓	✓	-	-	64.20
✓	✓	✓	-	64.56
✓	✓	✓	✓	**64.80**

표 2. 핵심 요소에 대한 절제 실험 결과 비교

Ⅲ. 결론

본 논문에서는 RVOS 문제에서 발생하는 시간적 불안정성과 객체 수 판별 문제를 해결하기 위해 문맥 인지 적응형 마스크 선택 기법을 제안하였다. 제안 방법은 CLIP 기반 키프레임 선정, 마스크 일관성 점수 계산, 텍스트 기반 객체 수 판별, 그리고 시간적 마스크 보간으로 구성되며, 서로 보완적으로 작용하여 높은 분할 정확도와 일관성을 확보하였다. MeViS 벤치마크 데이터셋에 대한 실험 결과, 제안 기법은 기존 최신 방법들보다 우수한 성능을 보였으며, 절제 연구를 통해 각 모듈의 독립적 기여를 검증하였다. 향후 연구에서는 본 방법을 다양한 대규모 비디오 분할 데이터셋과 실제 응용 환경에 확장 적용하고, 실시간 추론 환경에서의 효율성 향상 방안을 탐구할 예정이다. 또한, 현재는 두 개의 모델 출력을 비교·선택하는 방식이지만, 다수의 모델 출력을 동적으로 가중 결합하는 메타-앙상블 구조로 확장하는 연구도 기대할 수 있을 것으로 생각된다.

ACKNOWLEDGMENT

이 성과는 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터의 지원(IITP-2025-RS-2020-II201808, 50%)과 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. RS-2025-02214941, 50%).

참 고 문 헌

[1] Wu, Jiannan, et al. "Language as queries for referring video object segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[2] Wu, Dongming, et al. "Onlinerefer: A simple online baseline for referring video object segmentation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

[3] Yuan, Haobo, et al. "Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos." arXiv preprint arXiv:2501.04001 (2025).

[4] Wei, Cong, et al. "Hyperseg: Towards universal visual segmentation with large language model." arXiv preprint arXiv:2411.17606 (2024).

[5] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." International conference on machine learning. PmLR, 2021.

[6] Bai, Jinze, et al. "Qwen technical report." arXiv preprint arXiv:2309.16609 (2023).

[7] Ding, Henghui, et al. "MeViS: A large-scale benchmark for video segmentation with motion expressions." Proceedings of the IEEE/CVF international conference on computer vision. 2023.