일반논문 (Regular Paper)
방송공학회논문지 제28권 제5호, 2023년 9월 (JBE Vol.28, No.5, September 2023)
https://doi.org/10.5909/JBE.2023.28.5.613
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

외부 메모리 어텐션 기반 준지도 비디오 객체 분할

김 지 윤[a], 홍 성 은[b)‡]

Semi-supervised Video Object Segmentation based on External Memory Attention

Jiyun Kim[a] and Sungeun Hong[b)‡]

요 약

본 논문은 준지도 비디오 객체 분할에서 어텐션 기법을 활용한 연구를 다룬다. 일반적인 준지도 비디오 객체 분할은 합성곱 신경망을 사용하므로 장거리 종속성 문제를 가지고 있다. 최근 분류, 감지, 분할 등의 딥러닝 분야에서 어텐션 기법을 적용하여 이를 완화하는 연구가 진행되고 있다. 대중적으로 셀프 어텐션 기반 접근 방식이 사용되지만, 다른 영상 간의 연관성을 모델링하기 어렵다는 한계가 있다. 본 논문에서는 장거리 종속성 문제 완화 및 다양한 영상 간의 연관성 고려를 위해 외부 메모리 어텐션 기반의 준지도 비디오 객체 분할 기법을 제안한다. 제안 기법은 두 개의 선형 레이어를 메모리로 활용하여 다양한 영상 간의 연관성을 모델링하며, 캐스케이드 연산을 통한 어텐션 과정을 거친다. 다양한 실험을 통해 DAVIS 2017 데이터셋에서 효용성을 검증하였고, 준지도 비디오 객체 분할의 대표적인 모델인 STM(Space-Time Memory) 보다 자카드지수와 경계 정확도의 평균에서 약 3.8%의 성능 향상을 보였다.

Abstract

This paper explores attention techniques in semi-supervised video object segmentation. Conventional semi-supervised video object segmentation suffers from long-range dependency due to the use of convolutional neural networks. Recent research has applied attention techniques in classification, detection, and segmentation to address this issue. While self-attention-based approaches are commonly used, they have limitations in modeling the relationships between different images. In this paper, we propose a semi-supervised video object segmentation based on external memory attention to address long-range dependencies and consider the relationships between various images. Our approach uses two linear layers as memory to model image relationships and employs cascade operations for attention. Experimental results on the DAVIS 2017 dataset demonstrate approximately 3.8% improvement in average Jaccard index and boundary accuracy compared to STM (Space-Time Memory).

Keyword : Attention Mechanism, External Memory, Semi-supervised Video Object Segmentation, Long-Term Dependency

a) 인하대학교 전기컴퓨터공학과(Department of Electrical and Computer Engineering, Inha University)
b) 성균관대학교 실감미디어공학과(Department of Immersive Media Engineering, Sungkyunkwan University)
‡ Corresponding Author : 홍성은(Sungeun Hong)
E-mail: : csehong@skku.edu
Tel: +82-2-740-1809
ORCID: https://orcid.org/0000-0003-1774-9168

※ This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2021R1F1A1054569, No. 2022R1A4A1033549).
· Manuscript July 5, 2023; Revised July 31, 2023; Accepted July 31, 2023.

Ⅰ. 서 론

비디오 객체 분할(Video Object Segmentation)[1][2][3][4]은 비디오 편집, 자율 주행 등과 같은 다양한 비디오 관련 응용에 적용될 수 있으며, 입력 비디오 시퀀스에 존재하는 특정 대상 객체를 식별하고 분할하는 것을 목표로 한다. 비디오 객체 분할에서 주로 연구되는 분야는 준지도 비디오 객체 분할[5][6][7][8]이며, 해당 분야는 비디오 객체 추적(Video Object Tracking)의 확장된 분야로서 검출 박스 형태가 아닌 객체의 경계면 분할을 요구한다. 준지도 비디오 객체 분할은 첫번째 영상에 대한 분할 영역이 주어지고 이후 영상에 대해서는 알고리즘이 분할 영역을 추론해야 하는 특성을 지닌다.

최근 준지도 비디오 객체 분할 연구 중 저장 기반의 접근 방법들[9][10][11][12]은 높은 성능과 모델 구조의 간결함으로 인해 많은 연구가 진행되고 있다. 저장 기반 비디오 객체 분할의 대표적인 모델은 STM(Space-Time Memory Networks)[13]이며, 첫번째와 마지막 영상 뿐만 아니라 중간 영상들을 모델링에 사용함으로써 객체의 모양 변화 및 겹침 현상과 같은 문제를 효과적으로 다룰 수 있다. STM을 시작으로, STM의 저장 및 불러오기 과정을 개선하기 위한 다양한 후속 연구들이 속도 개선과 저장공간 효율 측면에서 진행되고 있다.

저장 기반의 객체 분할 기법들은 마스크 예측을 위해 합성곱 신경망 기반의 인코더를 사용하여 특징맵을 추출한다. 예를들면 대부분의 기존 기법은 ResNet18[14] 및 ResNet50[14]과 같은 합성곱 신경망을 인코더로 사용한다. 그러나 합성곱 신경망은 합성곱 연산의 지역적인 특성으로 인해, 멀리 떨어진 공간상의 정보를 연관 짓기 어려운 장거리 종속성 문제를 야기한다.

최근 분류, 감지, 분할 등의 다양한 딥러닝 분야에서는 합성곱 연산의 지역적인 특성을 극복할 수 있는 어텐션 기법[15][16][17][18]들이 확대 적용되고 있다. 그 중에서도 가장 널리 사용되는 셀프 어텐션(self-attention)[19]은 특징맵의 가중합계를 계산하기 위해 모든 위치에서 쌍별 유사성을 고려하고, 이를 통해 공간상의 장거리 종속성 문제를 완화하는 역할을 한다. 그러나 셀프 어텐션은 단일 이미지 내의 연관성에만 초점을 맞추기 때문에 비디오 내에 존재하는 다양한 영상 프레임간의 관계를 고려하기 힘들다는 한계점이 있다.

본 논문은 단일 이미지 내에서의 공간적 장거리 종속성 문제를 완화시킬 뿐만 아니라 다양한 영상 프레임 사이의 연관성을 모델링할 수 있는 외부 메모리 어텐션 기반의 비디오 객체 분할 기법을 제안한다. 외부 메모리 어텐션 기법(external memory attention)[20]에서 두 개의 선형 레이어는 외부 메모리 역할을 수행한다. 구체적으로, 한 개의 키 선형 레이어와 한 개의 밸류 선형 레이어를 통해 캐스케이드 연산을 수행하여 어텐션 과정을 거치게 된다. 적용 모델의 밸류 인코더는 단일 영상을 입력으로 받아 분할 예측을 위한 특징맵을 추출한다. 외부 메모리는 학습 중 주어지는 영상들에 대해 간접적으로 기억하고 새롭게 주어지는 영상과의 연관성을 모델링할 수 있기 때문에 객체 분할 결과의 정확성과 일관성을 향상시킬 수 있다.

본 논문의 주요 기여점을 요약하면 다음과 같다.

기존의 준지도 비디오 객체 분할 모델의 합성곱 기반 인코더에 외부 메모리 어텐션 기법 적용해서 단일 이미지 내에서의 장거리 종속성 문제를 완화하고, 서로 다른 영상 프레임간의 연관성을 효율적으로 모델링한다.

비디오 객체 분할의 대표적인 평가 데이터셋인 DAVIS 2017[2]에서 모델의 다양한 인코더 계층에 외부 메모리 어텐션을 적용한 효과를 심층 분석한다. 베이스라인 모델 대비 성능 향상이 있었으며, 대표적인 모델인 STM보다 3.8%의 성능 향상을 확인하였다.

Ⅱ. 관련 연구

1. 준지도 비디오 객체 분할

준지도 비디오 객체 분할[9][10][11][12]은 첫번째 프레임에서 주어진 대상 객체 영역을 이후 프레임에 효과적으로 전파하는 것을 목표로 한다. 초기 연구에서는 프레임을 한 번에 하나씩 처리하며, 새로운 프레임이 도착할 때마다 분할된 객체를 업데이트하는 방식을 사용되었으며, 이러한 기법은 추론 과정이 매우 느린 한계점이 있었다. 보다 효율적인 모델 학습 및 추론을 위해 빠른 온라인 학습 알고리즘[21][22][23], 임베딩 학습[24][25][26], 시공간 매칭[15][27][28][29]등의

연구가 최근까지도 활발히 진행되고 있다.

본 논문은 준지도 비디오 객체 분할 분야에서 최근 큰 주목을 받고 있는 저장 기반 모델링 기법에 초점을 맞춰서 연구를 진행하였다. 대표적인 저장 기법 기법인 STM은 저장 공간을 구축하여 추론된 객체의 정보를 지속적으로 업데이트하는 방식을 사용한다. STM은 대상 객체에 대한 저장 공간을 구축하고, 구축된 저장 공간과 모든 쿼리 영상을 매칭시키는 과정을 거친다. 최근 STM의 개선을 위해 데이터 증강 기법[9][12], 불러오기 프로세스 개선[19][30], 저장 공간의 크기 제한[11][31] 등 다양한 후속 연구가 진행되고 있다. 본 논문에서는 실험 모델로 직접적인 이미지 대 이미지 대응을 통해 기존의 STM을 개선한 STCN(Space-Time Correspondence Network)[32]을 베이스라인 모델로 사용한다. 해당 모델은 서로 다른 영상 프레임에서의 부분 영역간 유사도 맵(affinity matrix)을 계산하기 위해 STM에서 사용되는 저장 및 불러오기 과정 대신, RGB 영상들의 관계에만 기반한 단일 유사도 맵을 사용한다. STCN이 합성곱 연산 기반의 인코더를 학습에 사용함으로써 발생하는 장거리 종속성 문제를 완화하고 인접 프레임간의 연관성을 모델링하기 위해서 본 연구는 외부 메모리 어텐션을 적용한다.

2. 어텐션 메커니즘

합성곱 신경망은 전통적인 기계학습 기반 특징 추출 기법에 비해 높은 특징 추출 능력을 보이지만, 합성곱 연산의 지역적인 특성으로 인해 공간내 장거리 종속성 문제를 가지고 있다. 이러한 문제를 완화하기 위해 어텐션 메커니즘[15][16][17][18]에 대한 연구가 이루어지고 있으며, 셀프 어텐션이 최근 계산 효율성과 기존 기법들 대비 높은 성능 향상으로 인해 널리 사용되고 있는 추세이다. 셀프 어텐션은 임베딩된 공간에서 모든 위치 간의 유사도를 계산하고 이를 가중치로 활용해서 장거리 종속성 문제를 해결할 수 있다. 그러나, 이 기법은 단일 이미지 내의 관계에만 초점을 두기 때문에 다양한 영상 간의 관계를 고려하기 어렵다. 본 논문에서 사용하는 준지도 비디오 객체 분할 모델은 ResNet을 인코더로 사용하고 있으며, 연속된 영상 프레임을 다루기 때문에 단일 이미지내 장거리 종속성 및 다양한 영상 간의 관계를 고려할 수 있는 방법론이 필요하다. 이를 해결하기 위해, 본 논문은 정지 영상 분석에 활용되어온 외부 메모리 어텐션 기법을 비디오 객체 분할 연구에 도입하였으며, 해당 기법은 학습 가능한 독립적인 두 개의 메모리 레이어를 사용하여 다양한 영상간의 관계를 효율적으로 고려할 수 있다.

Ⅲ. 제안 기법

제안하는 기법을 설명하기에 앞서 관련 연구 배경인 셀프 어텐션과 외부 메모리 어텐션 기법을 설명한다. 셀프 어텐션은 어텐션의 특수한 경우로, 컴퓨터비전 분야의 많은 연구들[15][16][17][18]에서 적용되고 있다. 셀프 어텐션의 핵심은 특징맵 내 요소 간의 유사성 계산을 통하여 장거리 종속성을 포착하는 것이다. 구체적으로 하나의 특징맵으로부터 쿼리(Q), 키(K), 밸류(V) 3가지의 특징맵을 추출하여 이들 간의 문맥적인 관계성을 파악하는 과정으로 식은 다음과 같다.

$$\begin{aligned} Q &= X \times W_Q \\ K &= X \times W_K \\ V &= X \times W_V \end{aligned} \tag{1}$$

식(1)에서, X는 입력 벡터 시퀀스, Q, K, V는 각각 쿼리, 키, 밸류를 나타내며, W_Q, W_K, W_V는 쿼리, 키, 밸류에 대한 학습 가능한 모델 파라미터를 나타낸다. 위의 Q, K, V를 통해 셀프 어텐션이 진행되는 과정은 다음과 같다.

$$Self\ Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

식(2)를 살펴보면, 먼저 쿼리와 키를 행렬 곱한 후 해당 행렬의 값을 키 차원의 제곱근으로 나누어 주어 행 단위로 소프트맥스 함수를 적용하여 스코어 행렬을 생성한다. 이후 해당 스코어 행렬과 밸류 행렬을 곱한다. 결과적으로 셀프 어텐션은 기존의 컨볼루션 신경망이 가지고 있었던 공간내 장거리 종속성 문제를 완화시킬 수 있다.

셀프 어텐션의 여러 장점에도 불구하고 상대적으로 높은 계산 복잡도를 가지며 서로 다른 영상 간의 상관관계를 고려하기 힘들다는 단점을 가진다. 이와 같은 단점을 극복하기 위해 외부 메모리 어텐션 기법에 제안되었다. 해당 기법

은 두 개의 선형 레이어와 두 개의 정규화 레이어를 통한 학습 가능한 공유 메모리를 기반으로 한다. 셀프 어텐션 달리 선형 복잡성을 가지기 때문에 계산 효율이 좋고, 다른 영상 간의 상관 관계를 고려할 수 있다. 외부 메모리 어텐션은 입력 특징맵의 스케일에 민감하므로, 다음과 같이 열과 행을 각각 정규화하는 두 단계의 과정을 거친다.

$$External\ Memory\ Attention = Norm(FM_k^T)\,M_V \quad (3)$$

위의 식에서 M_k, M_v는 각각 독립적인 학습 가능한 공유 메모리 역할을 하는 선형 레이어를 의미하며, Norm의 경우 소프트맥스와 L1 정규화 두 개의 정규화를 나타낸다. 그림 1의 a)는 단일 영상 내의 쿼리, 키, 밸류를 통하여 영상 내의 연관성을 고려하는 셀프 어텐션 구조이며, b)는 다양한 샘플들간의 연관성을 고려할 수 있도록 설계된 외부 메모리 어텐션 구조를 보여준다.

본 연구는 준지도 객체 분할을 위해 STCN[15]을 베이스라인 모델로 활용하며 그림 2는 제안하는 프레임워크의 흐름을 보여준다. 제안 프레임워크는 이미지를 입력으로 받는 키 인코더와 이미지와 마스크를 입력으로 받는 밸류 인코

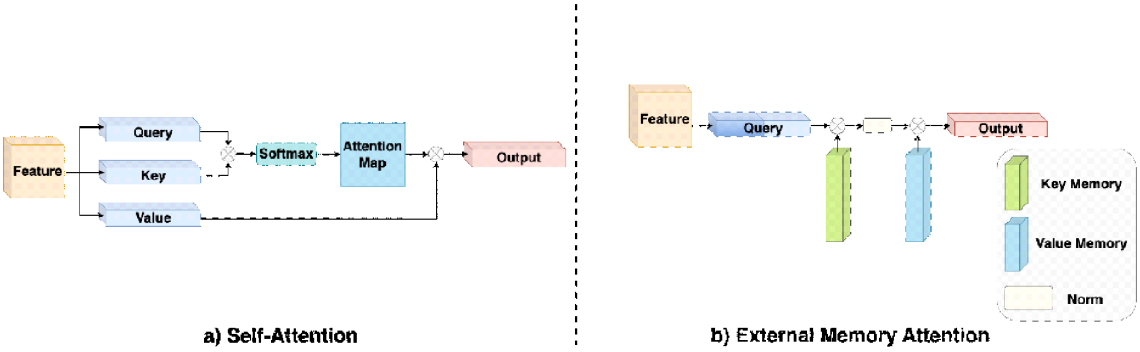

그림 1. 단일 샘플 내에서 어텐션을 거치는 a) 셀프 어텐션(self-attention)과 메모리 레이어를 통한 다양한 샘플들 사이의 연관성을 고려할 수 있는 b) 외부 메모리 어텐션(external memory attention)의 구조 비교
Fig. 1. Comparison of a) self-attention and b) external memory attention

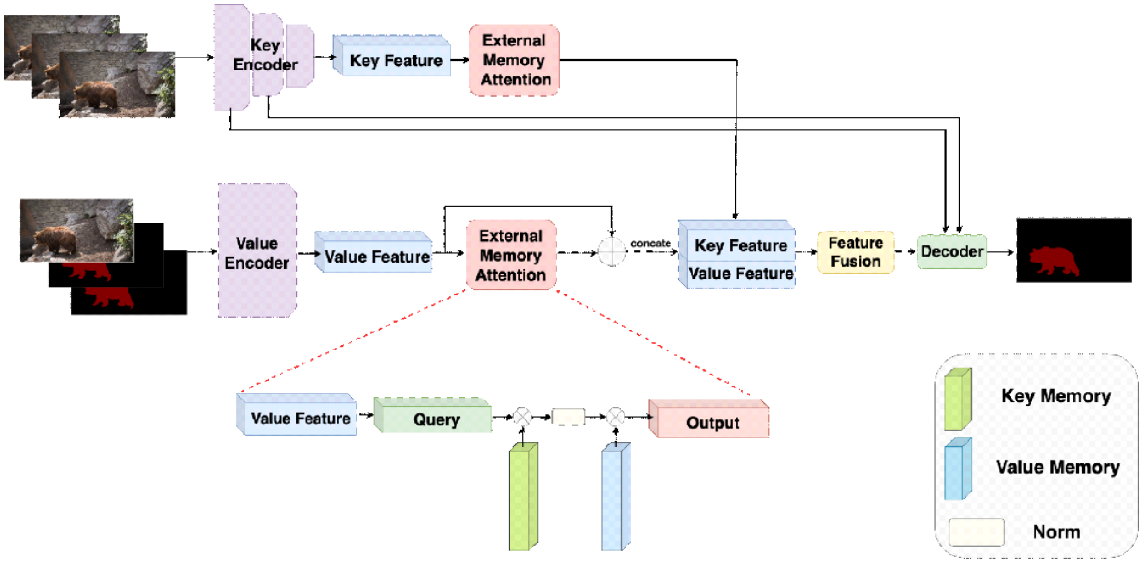

그림 2. 외부 메모리 어텐션 기반의 제안하는 준지도 비디오 객체 분할 프레임워크
Fig. 2. Proposed external memory attention-based framework

더로 각각 ResNet50과 ResNet18을 사용한다. 키 인코더는 학습 시 세 장의 연속된 RGB 이미지를 입력으로 받으며, 마스크 없이 독립적으로 특징맵을 추출한다. 세 장의 연속된 RGB 이미지들간의 대응 관계를 나타내는 특징맵을 추출하는 어렵기 때문에 키 인코더는 밸류 인코더와 달리 더 깊은 네트워크를 사용하였다. 키 인코더가 마스크에 독립적인 이유는 RGB 이미지들 간의 대응 관계를 파악하는데 있어 마스크는 주의를 분산시킬 수 있어 혼란을 줄 수 있는 요소로 작용될 수 있기 때문이다.

키 특징맵이 마스크와 독립적이므로, 입력된 쿼리 이미지를 전파 중에 저장 공간에 저장하기 위한 특징맵으로 전환하기로 결정하면 "쿼리 키 특징맵"를 나중에 "메모리 키 특징맵"으로 재사용할 수 있는 장점이 있다. 이를 통해, STM 특징맵 추출의 독립성과 연관성을 개선하고, 메모리와 쿼리 간의 대칭성을 보장할 수 있게 되었다. 해당 키 특징맵에 외부 메모리 어텐션 기법을 적용하면 셀프 어텐션과 달리 독립적인 학습 가능한 메모리(M)와 특징맵 사이의 유사성을 계산하고, 이를 통해 어텐션 맵을 생성하게 된다. 이 과정을 통해 메모리는 객체 분할 작업에 필요한 다양한 영상 프레임에 존재하는 특징맵의 통계적 특성을 학습할 수 있게 된다. 이를 통해 모델은 학습을 진행할 때 독립적으로 학습 가능한 메모리를 활용하여 전체 데이터셋의 정보를 효과적으로 포착할 수 있게 되며, 새로운 입력 데이터에 대한 정보의 일관성을 높일 뿐만 아니라, 장기적인 의존성을 모델의 특징맵 개선을 통해 효과적으로 모델링할 수 있게 된다.

모델의 밸류 인코더는 한 장의 이미지와 해당하는 객체 분할 정보를 입력으로 받는다. 밸류 인코더의 마지막 레이어에서 추출된 특징맵과 키 인코더에서 추출된 특징맵을 결합하여 특징맵 융합(feature fusion) 과정을 통해 최종적인 특징맵을 추출한다. 이러한 과정을 통해 초기에 입력된 세 장의 영상에 대한 반복적인 객체 분할 예측을 수행한다. 밸류 인코더의 마지막 레이어에서 출력된 특징맵은 이미지와 마스크간의 대응 관계에 대한 정보를 표현하고 있기 때문에 특징맵 융합 과정 전에 외부 메모리 어텐션 기법을 적용하게 되면 앞서 키 인코더와 마찬가지로 메모리는 객체 분할 작업에 필요한 다양한 영상 프레임의 통계적 특성을 학습할 수 있다. 따라서, 세 장의 연속된 RGB 이미지를 입력받는 키 인코더와 달리 한 장의 RGB 이미지를 입력받는 밸류 인코더가 연속적인 이미지를 처리하여 최종 출력된 특징맵에 적용하면, 메모리 역할을 하는 선형 레이어가 첫번째 영상부터의 정보를 기억할 수 있다. 이러한 과정은 이후 예측을 위해 입력되는 두 번째 영상 및 세 번째 영상에 대한 객체 분할 예측에 긍정적인 영향을 미치게 된다.

본 모델에서는 사용되는 손실 함수는 표준 크로스 엔트로피 손실의 변형인 부트스트랩 크로스 엔트로피(Bootstrapped Cross-Entropy) 라는 손실 함수을 사용하며 다음과 같다.

$$BootstrappedCE(I,T,) = \begin{cases} Cross\,Entroph(I,T), & if < s \\ p \times mean(raw\,loss), & otherwise \end{cases} \quad (4)$$

위의 손실 함수에서 I는 모델의 예측 출력, T는 실제 정답, it은 현재 훈련 반복 횟수를 나타낸다. s는 부트스트래핑을 적용하기 시작하는 훈련 반복 횟수을 나타내며, p는 어려운 예제에 더 가중치를 두기 위한 상수, raw loss는 개별 데이터 포인트의 크로스 엔트로피 손실 값을 나타낸다. 따라서, 이 손실 함수는 초기 훈련 단계에서는 모든 예제에 대해 일반적인 크로스 엔트로피를 계산하며, 일정 훈련 단계 이후에는 하드 예제에 더 가중치를 두어 훈련의 안정성과 성능을 향상시키는 역할을 수행하여 모델이 어려운 예제에 더 강력하게 적응하도록 돕는 역할을 한다.

Ⅳ. 실험 결과

준지도 비디오 객체 분할의 목표는 첫번째 영상의 객체 분할 정보만 주어지며, 해당 분할 정보를 통해 이후 영상들에 대하여 목표 객체를 분할하는 것을 목표로 한다. 본 논문은 STCN에서 제공하는 대용량의 BL30K[9] 및 정지 영상 기반으로 사전 학습된 모델 파라미터에 기반하여 기전 연구의 실험 프로토콜에 맞춰서 DAVIS 2017 학습셋 및 YouTubeVOS 2019[3] 학습셋을 통해 모델 학습을 진행한다. 또한, 사전 학습된 모델 파라미터를 사용하지 않는 상황에서 DAVIS 2017 학습셋과 YouTubeVOS 2019 학습셋만으로 학습을 진행하여 외부 메모리 어텐션 기법의 효과를 알아본다. 성능 평가는 DAVIS 2017 검증셋에서 진행되었

표 1. 학습 및 평가에 사용된 데이터셋 규모
Table 1. Dataset size used for model training and evaluation

	DAVIS 2017			YouTubeVOS 2019		
	학습	검증	총	학습	검증	총
비디오 개수	60	30	90	3,471	507	3,978
영상 개수	4,219	2,023	6,242	-	-	-
객체 개수	138	59	197	6,459	Seen: 1,063 Unseen: 26	7,548

표 2. DAVIS 2017 데이터셋에서의 어텐션 기법 적용 결과 비교
Table 2. Comparison of attention methods on DAVIS 2017

Method	J&F-Mean (↑)	J-Mean (↑)	F-Mean (↑)	FPS (↑)
Default	84.8	81.5	88.0	24.5
Self-Attention	85.3	81.9	88.8	24.3
External Memory Attention (Ours)	85.6	82.6	88.6	24.0

표 3. DAVIS 2017 데이터셋에서 어텐션 기법을 모델의 각 인코더에 적용시킨 결과 비교
Table 3. Comparison results of applying the attention methods to each encoder of the model on DAVIS 2017

Method	Encoder	J&F-Mean (↑)	J-Mean (↑)	F-Mean (↑)
Self-attention	Key Encoder	85.2	82.0	88.5
	Value Encoder	84.7	81.6	87.8
	Key & Value Encoder	85.3	81.9	88.8
External Memory Attention (Ours)	Key Encoder	85.1	81.8	88.4
	Value Encoder	85.2	82.1	88.4
	Key & Value Encoder	85.6	82.6	88.6

다. 표 1은 각 데이터셋에 대한 주요 정보 요약을 보여준다.

표 2는 베이스라인 모델, 셀프 어텐션을 적용한 모델, 외부 메모리 어텐션을 활용한 제안 기법간의 정량적 성능 비교를 보여준다. 셀프 어텐션 및 외부 메모리 어텐션은 모델의 키 인코더 및 밸류 인코더에 적용되었으며 정량적 평가에는 영역(J) 및 경계(F) 측정 메트릭이 사용되었다. 셀프 어텐션을 적용한 모델은 외부 메모리 어텐션을 적용한 모델보다 성능이 낮으며, 이는 단일 영상 내의 연관성에 집중하는 셀프 어텐션의 한계점을 보여준다. 외부 메모리 어텐션을 적용한 모델은 독립적인 선형 레이어가 메모리 역할을 하기 때문에 다양한 영상 프레임에 대한 정보를 기억할 수 있게 되어 모델의 전체적인 성능이 향상된 것으로 파악된다.

표 3은 셀프 어텐션과 본 연구에서 도입한 외부 메모리 어텐션을 모델의 각 인코더에 적용했을 때의 성능을 나타낸다. 셀프 어텐션은 밸류 인코더에서 성능의 하락이 있었고, 키 인코더에서는 성능의 향상이 있었다. 외부 메모리 어텐션은 전반적으로 모두 성능의 향상이 있었다. 구체적으로, 키 인코더 보다 밸류 인코더에 단일로 적용했을 때 상대적으로 높은 성능을 보이는 것을 통해 밸류 인코더에서의 특징맵이 객체 분할에 더 중요하다는 것을 파악할 수 있었다

표 4는 제안 기법과 기존 준지도 비디오 객체 모델과의 정량적 성능 비교를 보여준다. 제안 기법은 기존 기법과

표 4. DAVIS 2017 데이터셋에서의 기존 준지도 비디오 객체 분할 모델과의 성능 비교
Table 4. Performance comparison with previous semi-supervised video object segmentation models on DAVIS 2017

Model	J&F-Mean (↑)	J-Mean (↑)	F-Mean (↑)	FPS (↑)
STM*[13]	81.8	79.2	84.3	10.2
CFBI*[24]	81.9	79.1	84.6	5.9
KMN*[11]	82.8	80.0	85.6	< 8.4
LWL*[22]	81.6	79.1	84.1	< 6.0
MiVOS*[9]	84.5	81.7	87.4	11.2
STCN[15]	84.9	81.8	88.0	24.5
Ours	85.6	82.6	88.6	24.0

* : 각 논문에서 제시한 결과

비교했을 때 우수한 객체 분할 정확도를 보여주고 있으며, 실시간 적용이 가능한 수준인 20이상의 FPS 성능을 유지하고 있다.

그림 3은 기존 및 셀프 어텐션, 외부 메모리 어텐션의 적용에 따른 객체 분할 결과를 보여준다. 그림3 위의 이미지는 사람의 역동적인 움직임에 대한 분할 결과를 보여주고 있다. 그림에서 기존 및 셀프 어텐션은 사람과 배경을 잘 구분하지 못하는 모습을 보이는 반면, 외부 메모리 어텐션

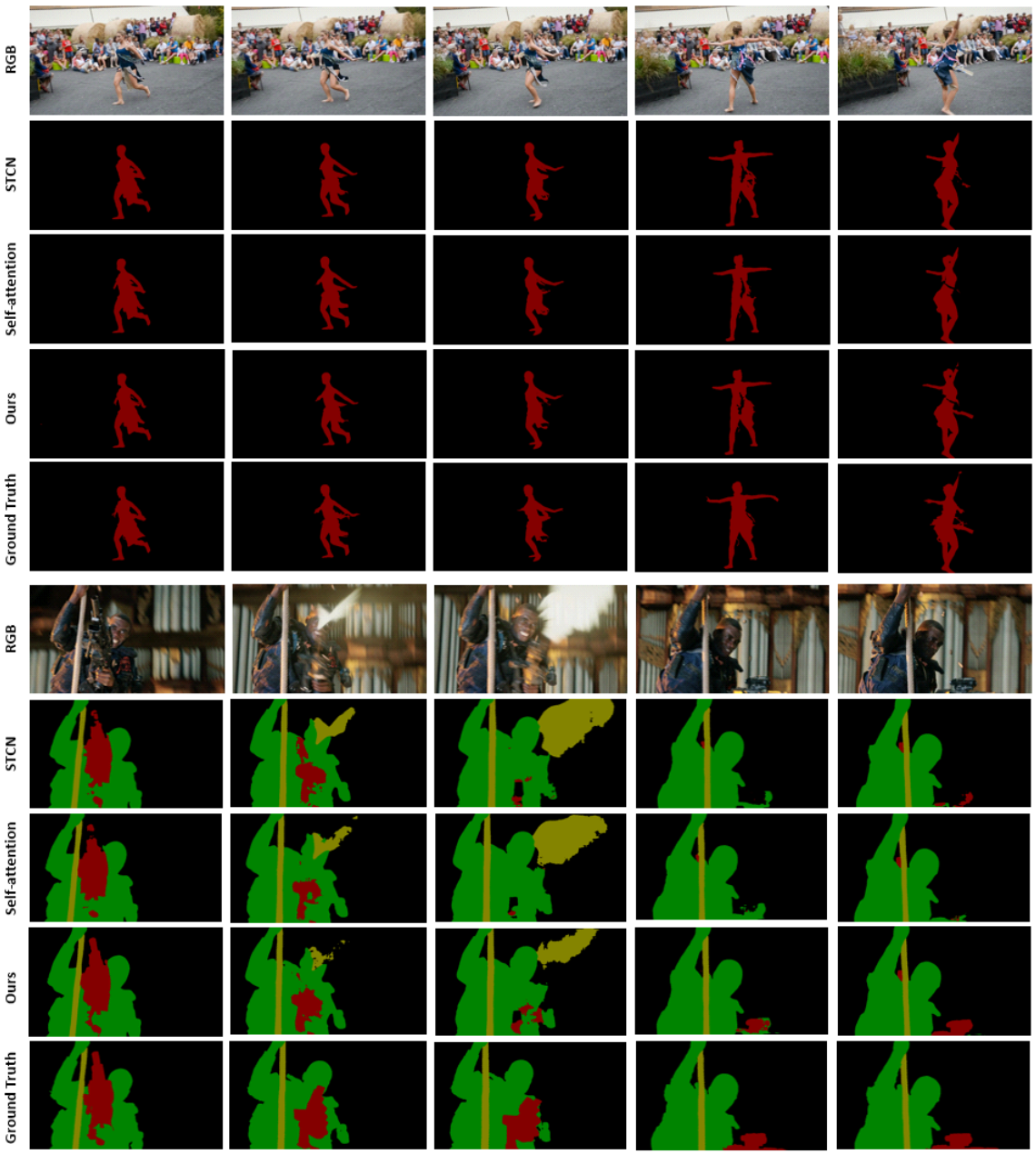

그림 3. 기존의 기법과 적용 기법 사이의 정성적 평가 결과
Fig. 3. Qualitative evaluation results between the existing methods and the proposed method

은 사람을 효과적으로 분할했음을 확인할 수 있다. 마지막 열에서 기존 및 셀프 어텐션은 역동적인 움직임에서 발생하는 옷의 움직임을 잘 분할하지 못했지만, 외부 메모리 어텐션은 분할에 성공했음을 확인하였다. 그림3의 아래의 이미지에서 주목해야할 부분은 총과 총구에서 발생되는 빛의 분할 결과이다. 기존 및 셀프 어텐션은 총이 다른 객체와 확실하게 구분되지 못하면서 총구에서 발생되는 빛을 밧줄로 잘못 분할하고 있다. 반면, 외부 메모리 어텐션은 보다 정확하게 총을 분할하며, 총구에서의 빛을 최대한 제거하려는 현상을 보이고 있다. 또한, 기존 및 셀프 어텐션은 총이 가려진 상황에서 총을 잘 인식하지 못하거나 사람으로 인식한 반면, 외부 메모리 어텐션은 명확히 사람과 분리하여 총을 분할하는 점을 확인할 수 있었다.

V. 결 론

본 논문에서는 준지도 비디오 객체 분할에서 어텐션 기법을 활용한 특징맵 추출에 대한 연구를 제안하였다. 기존의 준지도 비디오 객체 분할 모델에서는 장거리 종속성 문제와 다양한 영상 간의 연관성을 고려하는 어려움이 있었다. 따라서, 본 논문에서는 외부 메모리 어텐션 기법을 적용하여 이러한 문제를 해결하고자 했다. 실험 결과, 메모리 어텐션 기법을 적용한 모델은 기존 모델에서 성능 향상을 보여주었으며, STM과 비교해봤을 때 3.8%의 성능 향상을 이루어내었다. 이는 준지도 비디오 객체 분할 분야에서 어텐션 기법의 유용성을 보여준다. 또한, 다양한 실험 설정과 데이터셋을 활용하여 제안한 기법의 성능을 확인하였으며, 정성적 평가 결과를 통해 외부 메모리 어텐션의 역할을 확인하였다. 결론적으로, 제안한 외부 메모리 어텐션 기법이 특징맵 추출에 있어 그 유효성을 입증하였고, 성능을 향상을 보여주었다. 이러한 결과는 비디오 객체 분할 분야에서 외부 메모리 어텐션 기법의 적용 및 응용에 기여할 것으로 기대된다. 향후 연구에서는 더욱 복잡한 데이터셋과 모델을 사용하여 성능을 더욱 개선하고, 다양한 응용 분야에서의 적용 가능성을 탐색할 수 있을 것이다.

참 고 문 헌 (References)

[1] PERAZZI, Federico, et al. A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724-732. 2016.

[2] PONT-TUSET, Jordi, et al. The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675, 2017.

[3] XU, Ning, et al. Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327, 2018.

[4] YANG, Zhao, et al. Hierarchical interaction network for video object segmentation from referring expressions. In: BMVC. 2021.

[5] CAELLES, Sergi, et al. One-shot video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 221-230. 2017.

[6] MANINIS, K.-K., et al. Video object segmentation without temporal information. IEEE transactions on pattern analysis and machine intelligence, 41.6: 1515-1530, 2018.
doi: https://doi.org/10.1109/TPAMI.2018.2838670

[7] VOIGTLAENDER, Paul; LEIBE, Bastian. Online adaptation of convolutional neural networks for video object segmentation. arXiv preprint arXiv:1706.09364, 2017.

[8] XIAO, Huaxin, et al. Monet: Deep motion exploitation for video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1140-1148. 2018.

[9] CHENG, Ho Kei; TAI, Yu-Wing; TANG, Chi-Keung. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5559-5568. 2021.

[10] HU, Li, et al. Learning position and target consistency for memory-based video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4144-4154. 2021.

[11] LIANG, Yongqing, et al. Video object segmentation with adaptive feature bank and uncertain-region refinement. Advances in Neural Information Processing Systems, 33: 3430-3441, 2020.

[12] SEONG, Hongje; HYUN, Junhyuk; KIM, Euntai. Kernelized memory network for video object segmentation. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23 – 28, 2020, Proceedings, Part XXII 16. Springer International Publishing, pp. 629-645. 2020.
doi: https://doi.org/10.1007/978-3-030-58542-6_38

[13] OH, Seoung Wug, et al. Video object segmentation using space-time memory networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9226-9235. 2019.

[14] HE, Kaiming, et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770-778. 2016.

[15] WANG, Xiaolong, et al. Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794-7803. 2018.

[16] FU, Jun, et al. Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3146-3154. 2019.
[17] ZHANG, Han, et al. Self-attention generative adversarial networks. In: International conference on machine learning. PMLR, pp. 7354-7363, 2019.
[18] YUAN, Yuhui, et al. OCNet: Object context for semantic segmentation. International Journal of Computer Vision, 129.8: 2375-2398, 2021.
doi: https://doi.org/10.1007/s11263-021-01465-9.
[19] VASWANI, Ashish, et al. Attention is all you need. Advances in neural information processing systems, 30, 2017.
[20] GUO, Meng-Hao, et al. Beyond self-attention: External attention using two linear layers for visual tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45.5: 5436-5447, 2022.
doi: https://doi.org/10.1109/TPAMI.2022.3211006
[21] ROBINSON, Andreas, et al. Learning fast and robust target models for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7406-7415. 2020.
[22] BHAT, Goutam, et al. Learning what to learn for video object segmentation. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23 – 28, 2020, Proceedings, Part II 16. Springer International Publishing, pp. 777-794, 2020.
doi: https://doi.org/10.1007/978-3-030-58536-5_46.
[23] YANG, Linjie, et al. Efficient video object segmentation via network modulation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6499-6507. 2018.
[24] YANG, Zongxin; WEI, Yunchao; YANG, Yi. Collaborative video object segmentation by foreground-background integration. In: Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23 – 28, 2020, Proceedings, Part V. Cham: Springer International Publishing, pp. 332-348, 2020.
doi: https://doi.org/10.1007/978-3-030-58558-7_20
[25] VOIGTLAENDER, Paul, et al. Feelvos: Fast end-to-end embedding learning for video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9481-9490. 2019.
[26] CHEN, Yuhua, et al. Blazingly fast video object segmentation with pixel-wise metric learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1189-1198. 2018.
[27] HU, Yuan-Ting; HUANG, Jia-Bin; SCHWING, Alexander G. Videomatch: Matching based video object segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 54-70. 2018.
[28] HUANG, Xuhua, et al. Fast video object segmentation with temporal aggregation network and dynamic template matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8879-8889. 2020.
[29] WANG, Ziqin, et al. Ranet: Ranking attention network for fast video object segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3978-3987. 2019.
[30] ZHOU, Zhishan, et al. Enhanced memory network for video segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0-0. 2019.
[31] WANG, Haochen, et al. Swiftnet: Real-time video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1296-1305. 2021.
[32] CHENG, Ho Kei; TAI, Yu-Wing; TANG, Chi-Keung. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. Advances in Neural Information Processing Systems, 34: 11781-11794, 2021.

저 자 소 개

김 지 윤

- 2019년 8월 : 공주대학교 환경공학과 학사
- 2022년 3월 ~ 현재 : 인하대학교 전기컴퓨터공학과 인공지능전공 석사과정
- ORCID : https://orcid.org/0009-0002-9089-2139
- 주관심분야 : Video Object Segmentation, Attention Mechanism, Deep Learning

저 자 소 개

홍 성 은

- 2010년 2월 : 한양대학교 컴퓨터공학과 학사
- 2012년 8월 : 카이스트 전산학과 석사
- 2018년 2월 : 카이스트 전산학과 박사
- 2018년 1월 ~ 2020년 8월 : SK telecom (T-Brain, AI Center) 연구원
- 2020년 9월 ~ 2023년 8월 : 인하대학교 정보통신공학과 조교수
- 2023년 9월 ~ 현재 : 성균관대학교 실감미디어공학과 조교수
- ORCID : https://orcid.org/0000-0003-1774-9168
- 주관심분야 : Domain Adaptation, Multimodal Learning, Face Understanding, Video Object Segmentation