KIPS Trans. Softw. and Data Eng.
Vol.12, No.5 pp.217~228
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)

https://doi.org/10.3745/KTSDE.2023.12.5.217

Context-Dependent Video Data Augmentation for Human Instance Segmentation

HyunJin Chun† · JongHun Lee†† · InCheol Kim†††

ABSTRACT

Video instance segmentation is an intelligent visual task with high complexity because it not only requires object instance segmentation for each image frame constituting a video, but also requires accurate tracking of instances throughout the frame sequence of the video. In special, human instance segmentation in drama videos has an unique characteristic that requires accurate tracking of several main characters interacting in various places and times. Also, it is also characterized by a kind of the class imbalance problem because there is a significant difference between the frequency of main characters and that of supporting or auxiliary characters in drama videos. In this paper, we introduce a new human instance datatset called MHIS, which is built upon drama videos, *Miseang*, and then propose a novel video data augmentation method, CDVA, in order to overcome the data imbalance problem between character classes. Different from the previous video data augmentation methods, the proposed CDVA generates more realistic augmented videos by deciding the optimal location within the background clip for a target human instance to be inserted with taking rich spatio-temporal context embedded in videos into account. Therefore, the proposed augmentation method, CDVA, can improve the performance of a deep neural network model for video instance segmentation. Conducting both quantitative and qualitative experiments using the MHIS dataset, we prove the usefulness and effectiveness of the proposed video data augmentation method.

Keywords : Drama Video, Human Instance Segmentation, Class Imbalance, Video Data Augmentation, Spatio-Temporal Context

인물 개체 분할을 위한 맥락-의존적 비디오 데이터 보강

전 현 진† · 이 종 훈†† · 김 인 철†††

요 약

비디오 개체 분할은 비디오를 구성하는 영상 프레임 각각에 대해 관심 개체 분할을 수행해야 할 뿐만 아니라, 해당 비디오를 구성하는 프레임 시퀀스 전체에 걸쳐 개체들에 대한 정확한 트래킹을 요구하기 때문에 난이도가 높은 기술이다. 특히 드라마 비디오에서 인물 개체 분할은 다양한 장소와 시간대에서 상호 작용하는 복수의 주요 등장인물들에 대한 정확한 트래킹을 요구하는 특징을 가지고 있다. 또한, 드라마 비디오 인물 개체 분할은 주연 인물들과 조연 혹은 보조 출연 인물들 간의 등장 빈도에 상당한 차이가 있어 일종의 클래스 불균형 문제도 있다. 본 논문에서는 미생 드라마 비디오들을 토대로 구축한 인물 개체 분할 데이터 집합인 MHIS를 소개하고, 등장인물 클래스 간의 심각한 데이터 불균형 문제를 효과적으로 해결하기 위한 새로운 비디오 데이터 보강 기법인 CDVA를 제안한다. 기존의 비디오 데이터 보강 기법들과는 달리, 새로운 CDVA 보강 기법은 비디오들의 시-공간적 맥락을 충분히 고려해서 목표 인물이 삽입되어야 할 배경 클립 내의 위치를 결정함으로써, 보다 더 현실적인 보강 비디오들을 생성한다. 따라서 본 논문에서 제안하는 새로운 비디오 데이터 보강 기법인 CDVA는 비디오 개체 분할을 위한 심층 신경망 모델의 성능을 효과적으로 향상시킬 수 있다. 본 논문에서는 MHIS 데이터 집합을 이용한 다양한 정량 및 정성 실험들을 통해, 제안 비디오 데이터 보강 기법의 유용성과 효과를 입증한다.

키워드 : 드라마 비디오, 인물 개체 분할, 범주 간 데이터 불균형, 비디오 데이터 보강, 시-공간 맥락

1. 서 론

최근에 활발히 연구되고 있는 비디오 개체 분할(Video Instance Segmentation, VIS)[1]은 비디오를 구성하는 연속된 영상 프레임들 안에서 관심 개체 영역들에 대해 탐지(detection), 분류(classification), 분할(segmen-tation), 트래킹(tracking) 작업을 동시에 수행하는 컴퓨터 비전 기술이다. 단 한 장의 영상을 대상으로 하는 영상 개체 분할(Image Instance Segmentation)과는 달리, 비디오 개체 분할은 비디오를 구성하는 영상 프레임 각각에 대해 관심 개체 분할을

※ 본 연구는 정보통신기획평가원의 재원으로 정보통신방송 기술개발사업의 지원을 받아 수행한 연구과제(No. 2020-0-00096. 클라우드에 연결된 개별 로봇 및 로봇그룹의 작업 계획 기술 개발)입니다. 또한, 2022년도 정부(산업통상자원부)의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임(P0008691, 2022년 산업혁신인재성장지원사업).
※ 이 논문은 2022년 한국정보처리학회 ACK 2022의 "비디오 데이터 보강을 이용한 인물 개체 분할"의 제목으로 발표된 논문을 확장한 것임.
† 준 회 원 : 경기대학교 컴퓨터과학과 석사과정
†† 비 회 원 : 경기대학교 컴퓨터과학과 석사과정
††† 종신회원 : 경기대학교 컴퓨터과학과 교수
Manuscript Received : December 13, 2022
First Revision : January 13, 2023
Accepted : January 30, 2023
* Corresponding Author : Incheol Kim(kic@kyonggi.ac.kr)

수행해야 할 뿐만 아니라 동시에 프레임 시퀀스 전체에 걸쳐 개체들에 대한 정확한 트래킹을 요구하기 때문에 난이도가 더 높은 기술이다. 이와 같은 비디오 개체 분할 기술은 자율주행, 증강 현실, 보안과 안전, 스포츠 비디오 분석, 비디오 콘퍼런스 등 다양한 분야들에 폭넓게 활용될 수 있어 많은 관심을 받고 있다.

본 연구에서는 비디오 개체 분할의 한 특수 유형으로서, 드라마 비디오에 등장하는 인물 개체들을 분할하는 비디오 인물 개체 분할(Video Human Instance Segmentation) 작업에 초점을 맞추고 있다. 사람의 경우 주로 특정 인물 한 명의 지속적인 트래킹을 요구하는 기존의 비디오 개체 분할과 비교하면, 드라마 비디오 인물 개체 분할은 다양한 장소와 시간대에서 상호 작용하는 복수의 주요 등장인물들에 대한 정확한 트래킹을 요구하는 특징을 가지고 있다. 또한, 드라마 비디오는 주연 인물들의 등장 빈도와 조연이나 보조 출연 인물들의 등장 빈도 간에는 상당한 차이가 있다는 특징도 있다. 또 현재 일반적인 비디오 개체 분할을 위한 Youtube-VIS[1]와 같은 벤치마크 데이터 집합들은 일부 존재하지만, 아직 드라마 비디오 인물 개체 분할을 위한 공개 데이터 집합은 알려진 것이 없다.

본 논문에서는 미생 드라마 비디오들을 토대로 구축한 새로운 드라마 비디오 인물 개체 분할 데이터 집합인 MHIS (Miseang Human Instance Segmentation Dataset)을 소개하고, 등장인물 클래스 간의 심각한 데이터 불균형(class imbalance) 문제를 효과적으로 해결하기 위한 새로운 비디오 데이터 보강(Video Data Augmentation) 기법 CDVA를 제안한다. 그동안 클래스 간 데이터 불균형 문제를 극복하기 위한 영상 데이터 보강(Image Data Augmentation)에 관한 연구들은 활발히 진행되어 온 것에 반해, 현재 비디오 데이터 보강에 관한 연구는 최근의 비디오 행동 인식(Video Action Recognition)을 위한 VideoMix[2]와 ObjectMix[3], 비디오 개체 분할을 위한 B-Aug[4]를 제외하고는 거의 찾아보기 어려운 실정이다.

앞서 설명한 대로 드라마 비디오 인물 개체 분할에서는 각 인물이 등장하는 장소와 배경, 함께 상호작용하는 다른 등장인물 등과 같은 공간적 맥락뿐만 아니라, 이전 그리고 이후의 장면들과의 일관성을 위한 시간적 맥락이 매우 중요하다. Fig. 1은 미생 드라마의 주연인 장그래에 비해 등장 빈도가 낮은 오상식의 부족한 훈련용 비디오 데이터의 생성을 위한 서로 다른 기존의 비디오 데이터 보강 기법들의 적용 사례들을 나타낸다. VideoMix 기법은 Fig. 1의 (a)처럼 오상식이 등장하는 원본 목표 클립의 첫 번째 프레임 내 경계 상자와 동일한 크기와 위치로 임의로 선정된 배경 클립의 모든 프레임에 삽입하는 방법이다. ObjectMix는 Fig. 1의 (b)와 같이 통합 마스크(mask)를 이용해 오상식이 등장하는 목표 클립의 각 프레임 내 모든 등장인물들을 배경 클립의 각 프레임 내 동일한 위치에 함께 삽입하는 방법이다. 이에 반해 B-Aug는 Fig. 1의 (c)와 같이 오상식을 포함한 목표 클립의 각 프레임에서 오상식만을 나타내는 단일 마스크를 이용해 배경 클립의 각 프레임으로 같은 크기 같은 위치로 복사하여 삽입하는 방법이다.

Fig. 1의 (a)~(c)의 결과에서 보듯이, 기존의 비디오 데이터 보강 기법들은 배경 비디오 클립 내의 기존 등장인물 간의 상호작용이나 장면을 고려하지 못한 채 보강 대상인 인물을 자연스럽지 못한 공간적 위치와 시간적 상황에 삽입하는 결과 비디오들을 생성하게 된다[5]. 이와 같이 생성된 비-현실적인 보강 데이터들은 시간적, 공간적 맥락에 민감한 비디오 개체 분할 모델의 성능을 효과적으로 향상하기는 어렵

Fig. 1. Previous Video Data Augmentation Methods: (a) VideoMix, (b) ObjectMix, (c) B-Aug

다. 따라서 본 논문에서는 이러한 기존 비디오 데이터 보강 기법들의 한계를 극복하고자, 새로운 맥락-의존적 비디오 데이터 보강 기법 CDVA (Context-Dependent Video Data Augmentation)을 제안한다. 본 논문에서는 MHIS 데이터 집합을 이용한 정량 및 정성 실험들을 통해, 제안 비디오 데이터 보강 기법의 유용성과 효과를 입증한다.

본 논문의 2장에서는 본 연구와 연관된 기존의 관련 연구들을 살펴보고, 3장에서는 본 논문에서 제안하는 새로운 맥락-의존적 비디오 데이터 보강 기법에 대해 설명한다. 이어서 4장에서는 드라마 미생 비디오들을 토대로 구축된 인물 개체 데이터 집합인 MHIS와 인물 개체 분할을 위한 심층 신경망 모델인 SeqFormer[6]에 대해 설명한다. 5장에서는 제안 비디오 보강 기법과 심층 신경망 모델의 구현 및 실험 환경, 그리고 이들을 이용한 정량 및 정성 실험 결과들을 소개하고, 끝으로 6장에서는 결론을 정리하고 향후 연구를 소개한다.

2. 관련 연구

2.1 비디오 개체 분할

비디오 개체 분할 연구에 앞서, 한 장의 영상 속 개체의 탐지, 분류, 분할을 동시에 수행하는 영상 개체 분할 연구가 수행되었다. 대표적인 선행 연구인 Mask RCNN 모델은 영역 제안 신경망(Region Proposal Network)으로 영상 내 개체가 있을 법한 후보 영역들을 먼저 생성한 후, 이들을 토대로 RoIAlign을 통해 개체의 공간 정보를 반영하는 특징(feature)을 추출하여 탐지, 분류, 분할을 수행한다. 한 장의 영상만을 사용하는 영상 개체 분할과는 달리, 연속된 프레임을 사용하는 비디오 개체 분할 연구는 각 프레임의 개체를 탐지, 분류, 분할할 뿐 아니라, 프레임 시퀀스 전체에 걸친 개체들을 정확히 트래킹해야 하는 어려움이 있다.

비디오 개체 분할에 관한 기존 연구들은 합성곱 신경망(Convolutional Neural Network, CNN)을 주로 사용하였지만[1,7], 최근 들어서는 토큰 간의 관계를 학습하며 광범위한 문맥을 고려하는 어텐션(attention) 기법을 적용한 트랜스포머(Transformer) 신경망 구조를 사용한 연구들이 소개되고 있다[6,8,9].

합성곱 신경망을 사용하는 비디오 개체 분할 연구 중에서 MaskTrack RCNN[1]은 영상 개체 분할을 위해 제안된 Mask RCNN에 신경망 모듈로 구성된 트래킹 분기(tracking branch)를 추가한 비디오 개체 분할 모델이다. 입력 프레임 한 장의 개체를 분할한 후, 직전 프레임에서 탐지된 개체와 현재 프레임에서 탐지된 개체가 50% 이상 겹칠 때 서로 같은 개체로 연결한다. 한편, STEM-Seg[7]은 인코더-디코더 구조에 3차원 합성곱 신경망(3D CNN)을 적용한 모델로, 여러 프레임을 3차원 특징으로 합쳐서 비디오 개체 분할을 위한 작업을 동시에 수행하는 특징이 있다. 하지만, 이처럼 합성곱 신경망을 이용한 모델들은 지역적인 정보를 취합하여 사용하기 때문에 여러 프레임의 맥락 정보를 전역적으로 추출하지 못한다. 이러한 한계점을 극복하기 위해, 합성곱 신경망 대신 어텐션 기법을 사용하는 트랜스포머 구조 기반의 모델들이 제안되고 있다.

대부분의 트랜스포머 신경망 구조 기반의 비디오 개체 분할 모델들[6,8,9]은 영상 개체 탐지 모델인 DETR[10]을 비디오 개체 분할로 확장한 구조를 가진다. 이러한 모델들의 인코더에서는 픽셀 단위에서 정보를 교환하기 위해, 백본 네트워크(backbone network)로 추출한 입력 특징맵에 어텐션 기법을 적용하는 반면, 디코더에서는 쿼리 단위에서 정보를 교환하고자, 개체 쿼리와 인코딩된 특징에 어텐션 기법을 적용하며, 개체 쿼리를 이용하여 비디오 개체 분할을 수행한다. 비디오를 구성하는 서로 다른 프레임 간 정보 교환과 트래킹 방법에 따라 모델 간 차이점이 존재한다.

최초의 트랜스포머 구조 기반의 비디오 개체 분할 모델인 VisTR[8]은 백본 네트워크로 추출한 여러 프레임의 특징 전체를 1차원으로 펼쳐서 인코더의 입력으로 사용한다. 인코더에서는 입력 전체에 어텐션을 적용하여 서로 다른 프레임 간 정보를 교환한다. 디코더에서 개체 쿼리는 인코딩 된 정보를 반영하여 각 프레임의 개체를 나타낸다. 이러한 개체 쿼리는 포착된 비디오의 개체를 순서대로 나타내기 때문에, 같은 인덱스의 쿼리를 헝가리안 알고리즘(Hungarian Algorithm)을 통해 연결함으로써 프레임 시퀀스에서 동일 개체를 트래킹한다. 하지만, 인코더에서 전체 프레임에 모두 어텐션을 적용하기 때문에, 크기가 작은 개체를 포착하기 어렵고 많은 연산량을 요구한다. 연산량을 줄이기 위해 고안된 IFC[9]는 프레임 맥락 교환 시 많은 자원이 들지 않도록 인코더 부분에 메모리 토큰 방식을 도입한 모델이다. 백본 네트워크를 통해 각 프레임의 패치(patch) 단위 특징을 추출한 후, 각 프레임의 특징을 1차원으로 펼친다. 인코더에서는 각 프레임의 특징 정보를 토큰에 임베딩(embedding)한 후, 여러 프레임의 메모리 토큰 간 교환을 통해 정보 교환을 수행한다. 디코더에서는 갱신된 프레임 정보가 메모리 토큰 정보를 개체 쿼리에 반영한다. 이러한 쿼리 정보를 토대로 개체 마스크를 탐지하고 트래킹 분기를 통해 개체들을 연결한다. 그러나, 프레임의 압축된 정보를 반영하는 메모리 토큰을 사용하기 때문에, 전체적인 시-공간적 정보를 사용하지 못하는 한계가 있다.

한편 SeqFormer[6]는 영상 개체 탐지 모델인 Deformable DETR을 비디오 개체 분할에 적용한 모델로, 기존 어텐션 기법 대신 변형 어텐션(deformable attention)을 사용하는 특징이 있다. 백본 네트워크로부터 추출한 프레임별 2차원 특징을 입력으로 사용한다. 인코더에서는 변형 어텐션으로 각 프레임의 공간적 맥락 추론을 수행한다. 디코더에서는 변형 어텐션으로 프레임 단위로 개체 쿼리를 갱신함으로써 프레임 간 맥락을 교환한다. 디코더의 출력인 서로 다른 프레임의 개체 쿼리 정보를 모두 합하여 프레임 시퀀스에서 개체를 트래킹한다.

2.2 비디오 데이터 보강

하나의 영상 속 개체를 복사하여 다른 영상에 붙여넣는 Copy-Paste 기법은 반전(flipping), 회전(rotation), 크기(scale) 변경 등의 기존 영상 데이터 보강 기법들보다 훨씬 다양한 영상들을 생성하며, 영상 물체 탐지 및 개체 분할 모델들을 학습하는 데 매우 효과적이다. 영상 기반의 Copy-Paste 보강에는 영상의 개체 위치를 옮기는 방법[11]과 배경 영상에 새로운 개체를 삽입하는 방법[12,13]이 있다. Insta-Boost[11]는 위치를 옮길 개체와 유사한 색상 히트맵을 이용하여 확률적으로 높은 위치에 개체를 이동시키는 방법이다. 다만, 색상 히트맵 생성, 개체 위치 이동, 기존 위치 복원 등의 부가적인 과정이 많아 상당한 연산량을 요구한다. 한편, 새로운 개체를 삽입하는 방법은 배경이 되는 영상의 맥락 고려 여부에 따라 나뉜다. Context-Aug[12]는 개체 삽입 시 신경망 모듈을 통해 배경 영상의 맥락을 파악한 후에 적절한 개체를 삽입하는 방법이다. 그러나, 맥락 파악 시 특정 구역 주변의 시각 정보만 사용하기 때문에 전체적인 맥락 정보를 파악하지 못한다. 반면, Simple Copy-Paste[13]는 새로운 개체를 본래의 크기와 위치와 동일하게 배경 영상에 삽입함으로써 개체 삽입에 대한 맥락을 고려하지 않는다. 하지만, 삽입된 개체가 배경 영상에 어울리는지를 파악하지 않기 때문에 비현실적인 데이터가 생성될 수 있다.

영상 데이터 보강 기법과는 달리, 비디오 데이터 보강 기법은 연속된 프레임으로 구성된 비디오의 전체적인 문맥을 유지하는 노력이 필요함에 따라, 비디오의 공간적 맥락과 더불어 이전 및 이후의 장면들과 일관되도록 하는 시간적 특성이 반영된 방법이 요구된다. 예컨대, 여러 프레임에 연속된 개체를 삽입하는 경우, 개체의 움직임이나 위치가 자연스럽게 이어져야 한다. 따라서 비디오의 시-공간적인 맥락을 고려하여 자연스러운 비디오를 생성해야 하는 어려움이 있다.

비디오 데이터 보강에 관한 기존 연구들은 비디오 행동 인식을 위한 기법들[2,3]과 비디오 개체 분할을 위한 기법[4]이 있다. 기존의 연구들은 각 기법에 따라 재구성한 목표 비디오를 배경 비디오에 동일한 크기와 위치로 삽입한다는 공통점이 있다. 비디오 행위 인식 작업을 위한 데이터 보강 연구 중에서 VideoMix[2]는 배경 비디오의 특정 구역에 목표 비디오의 일부를 삽입하는 방법이다. 목표 비디오의 시간적 길이를 기존보다 짧게 자르거나, 목표 비디오를 사각형으로 잘라서 공간적인 크기를 자르는 과정을 수행하여 목표 비디오를 재구성한다. 재구성한 비디오를 배경 비디오에 삽입하여 새로운 학습 데이터를 생성한다. 그러나, 목표 비디오가 사각형 형태로 삽입되어 행위의 대상이 되는 인물이 아닌 나머지 부분도 같이 삽입되는 문제점이 있으며, 목표 비디오를 재구성할 일부 영역이 무작위로 선택됨에 따라 행위를 나타내는 인물 정보가 잘릴 가능성이 있다.

이러한 불필요한 시각 정보가 삽입되는 문제점을 개선하기 위해, ObjectMix[3]는 목표 비디오의 행위를 나타내는 인물 개체들만을 배경 비디오에 삽입한다. 사전 학습된 영상 개체 분할 모델을 통해 추출한 목표 비디오 속 인물 개체들의 통합 마스크(mask) 정보를 이용하며, 목표 비디오의 시간적 길이는 전부 유지한다. VideoMix 보다 비디오 속 인물의 행위에 집중하고 목표 인물의 정보를 전체적으로 유지하는 장점이 있다.

한편, 비디오 개체 분할 작업을 위한 데이터 보강 연구로는 B-Aug[4]가 있다. 통합 마스크 정보를 사용하는 ObjectMix와는 달리, 목표 비디오의 각 프레임에서 삽입 개체의 단일 마스크 정보를 대응되는 배경 프레임에 삽입하는 방법으로, 목표 비디오의 공간적인 맥락을 유지할 수 있다. 이와 같은 선행 비디오 데이터 보강 연구들은 배경 비디오의 시-공간적인 맥락을 고려하지 않아, 삽입된 개체가 배경 개체를 과하게 가리거나 비디오의 핵심인 시각 정보를 과하게 없애는 등 문맥이 유지되지 않는 한계점이 존재한다.

3. 맥락-의존적 비디오 데이터 보강

본 논문에서 제안하는 맥락-의존적 비디오 데이터 보강 기법 CDVA(Context-Dependent Video Data Augmentation)는 다양한 시-공간적 맥락을 고려함으로써 부족 인물 클래스를 위한 보다 현실성 있는 보강 비디오 데이터들을 생성한다.

제안하는 CDVA 비디오 데이터 보강은 Fig. 2와 같이 목표 및 배경 클립 짝짓기(Target and Background Pairing), 맥락-의존적 영역 선정(Context-Dependent Region Selection), 보강 클립 생성(Augmented Clip Generation) 등 3단계로 수행된다.

3.1 목표 및 배경 클립 짝짓기

이 단계는 보강이 필요한 인물 클래스(target class)의 목표 비디오 클립(target clip)과 이 인물을 삽입할 배경 비디오 클립(background clip)의 쌍(pair)을 결정하는 과정으로서, Algorithm 1과 같이 수행된다.

Algorithm 1에서 C_i는 보강이 필요한 목표(target) 인물 클래스를, $ClipSet(C_i)$와 $numClips(C_i)$는 해당 클래스의 인물 개체를 포함한 비디오 클립들의 집합과 클립 개수를 각각 나타낸다. 또한, 집합 S_B는 보강이 필요한 목표 인물 개체를 포함하고 있지 않은 비디오 클립들의 집합을 나타내며,

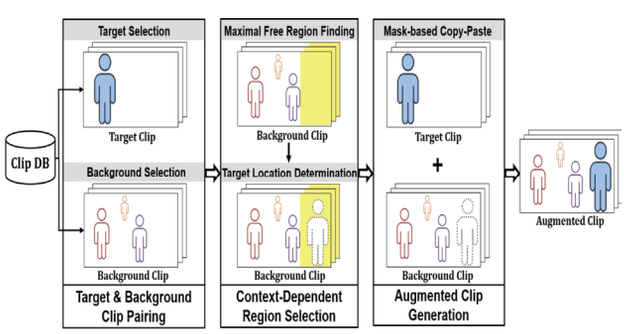


Fig. 2. Context-Dependent Video Data Augmentation

Algorithm 1: Target and Background Clips Pairing

Input: video clip set S, target classes $\{C_1, C_2, ..., C_m\}$, maximum number of clips for each class N
Output: clip pair set P

```
Initialize P = ∅
S_B = S-(ClipSet(C_1) ∪ ... ∪ ClipSet(C_m))
for i = 1, 2, ..., m do
  requiredClips = N - numClips(C_i)
  countClips = 0
  for targetClip c_t in ClipSet(C_i) do
    for backgroundClip c_b in S_B do
      if countClips < requiredClips then
        P = P ∪ {(c_t, c_b)}
        countClips++
```

P는 클립 짝짓기의 결과로서 보강 대상 인물을 포함한 목표 비디오 클립과 이것을 삽입할 배경 비디오 클립 쌍들의 집합을 나타낸다. Algorithm 1은 클래스별 최대 비디오 클립 데이터 개수인 N에서 현재 목표 클래스 C_i의 원본 비디오 클립 데이터 개수인 $numClips(C_i)$를 차감함으로써, 보강을 위해 신규 생성이 필요한 클립 개수를 구한다. 그리고 $ClipSet(C_i)$에 속한 하나의 목표 클립 c_t에 대응하는 배경 클립 c_b를 목표 인물 개체를 포함하고 있지 않은 비디오 클립들의 집합인 S_B에서 골라 비디오 데이터 보강을 위한 클립 쌍 (c_t, c_b)들을 생성하고, 이들을 집합 P에 담는다.

3.2 맥락-의존적 영역 선정

비디오 데이터 보강 기법 CDVA에서는 배경 비디오 클립 c_b의 전체 맥락을 고려해서, 목표 클립 c_t의 인물 개체(target instance)를 삽입할 영역을 선정한다. 즉 CDVA에서는 목표 인물 개체의 크기를 재조정하지 않더라도 삽입될 배경 클립 내의 기존 인물들과 최대한 겹치지 않도록, (1) 배경 클립 속 인물들이 존재하지 않는 최대 자유 영역(maximal free region) R^*을 찾고, (2) 영역 R^* 안에 목표 인물을 실제로 삽입할 구체적인 위치(location)를 결정한다.

먼저 최대 자유 영역 R^*을 찾는 과정은 Fig. 3의 (a)와 같이, 배경 클립의 각 프레임에서 인물이 존재하지 않는 가로 범위들을 모두 탐색한다. k번째 프레임 안에서 배경 인물의 세로형 경계 상자(bbox)를 중심으로, 인물이 없는 좌측(l^k), 중간(m^k), 우측(r^k) 가로 범위들을 찾는다. 이때 m^k는 k번째 프레임의 인물이 없는 모든 중간 가로 범위들 중 최대 범위로 정한다.

드라마 비디오들은 등장인물들이 서 있거나 앉아 있는 장면이 다수여서, 세로형의 경계 상자들이 대부분을 차지한다. 따라서 이러한 데이터의 특성을 감안하여, 배경 클립 프레임의 공간적 배치를 토대로 목표 인물 삽입을 위한 각 가로 범위들을 우선 탐색하는 것이다. 배경 클립의 각 프레임에서 인물이 없는 가로 범위들을 찾고 나면, Fig. 3의 (b)와 같이 배경 비디오 클립을 구성하는 프레임 시퀀스의 시간적 맥락을 반영하기 위해, 전체 프레임들에 걸쳐 Equation (1), Equation (2)와 같이 계산한 최소 가로 범위 l^*, m^*, r^*들과 후보 영역 R_1, R_2, R_3들을 이용하여 최종적으로 목표 인물을 삽입하기 위한 맥락-의존적 영역 R^*을 결정한다.

$$l^* = \min(l^k), m^* = \min(m^k), r^* = \min(r^k), \quad k = 1,2,...,F \tag{1}$$

Equation (1)에서 F는 배경 클립의 총 프레임 수를 나타낸다. 그리고 최소 가로 범위 l^*, m^*, r^* 각각에 프레임 높이 I_h를 곱하여, 배경 클립 내 목표 인물 삽입 후보 영역들인 R_1, R_2, R_3를 Equation (2)와 같이 계산한다.

$$R_1 = l^* \times I_h,\ R_2 = m^* \times I_h,\ R_3 = r^* \times I_h \tag{2}$$

또 Equation (3)과 같이, 배경 클립 내 후보 영역들인 R_1, R_2, R_3 중 가장 넓은 영역을 최종적인 목표 인물 삽입을 위한 최대 자유 영역 R^*로 선정한다.

$$R^* = \max(area(R_i)),\ \ i = 1,2,3 \tag{3}$$

배경 클립 내의 최대 자유 영역 R^*에 목표 인물을 삽입할 구체적인 위치를 결정하는 과정은 Fig. 4의 예시와 같으며,

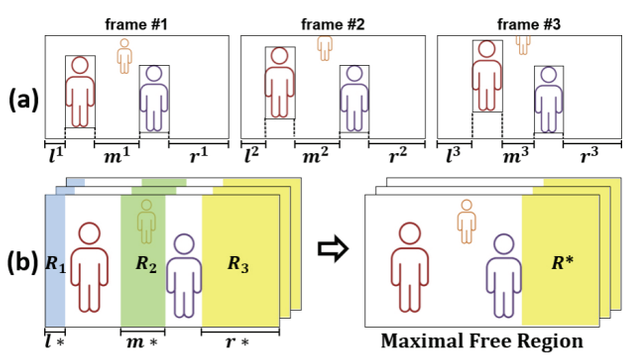


Fig. 3. Finding the Maximal Free Region

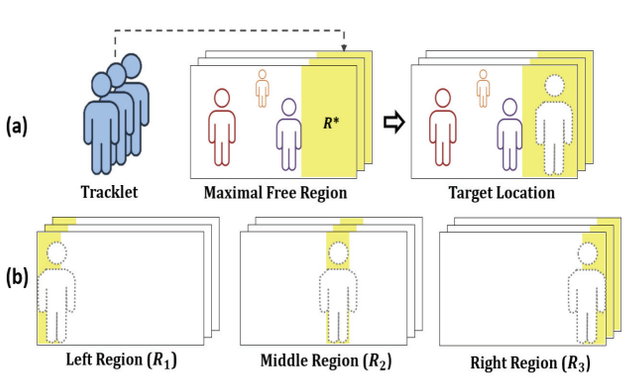


Fig. 4. Deciding the Target Instance Location

Algorithm 2: Deciding the Target Instance Location

Input: maximal free region $R^* = (R_{x_1}, R_{x_2}, R_w, R_h)$, candidate regions in the background clip R_1, R_2, R_3 size of the target instance $S_T=(T_w, T_h)$

Output: location of the target instance to be inserted $(T_{\tilde{x}},\ T_{\tilde{y}})$

1. **if** T_w 〈= R_w **then**
2. $T_{\tilde{x}}$ =*RandomSelectBetween*(R_{x1}, R_{x2}- T_w)
3. **else if** T_w 〉 R_w **then**
4. **if** R^* == R_1 **then** $T_{\tilde{x}}$ = R_{x1}
5. **else if** R^* == R_2 **then** $T_{\tilde{x}}$ = R_{x1}-(T_w-R_w)/2
6. **else if** R^* == R_3 **then** $T_{\tilde{x}}$ = R_{x2}- T_w
7. $T_{\tilde{y}}$ = *RandomSelectBetween*(0, R_h- T_h)

위치 결정의 상세 과정은 Algorithm 2와 같다.

목표 인물의 실제 삽입 위치 $(T_{\tilde{x}},\ T_{\tilde{y}})$는 배경 클립내의 최대 자유 영역 $R^*= (R_{x_1}, R_{x_2}, R_w, R_h)$, 후보 영역들 R_1, R_2, R_3, 삽입 대상 목표 인물의 크기$S_T=(T_w, T_h)$에 따라 달리 결정된다. 여기서 삽입 위치 $(T_{\tilde{x}},\ T_{\tilde{y}})$는 목표 인물의 우측 상단 모서리가 놓여질 배경 클립 내 좌표를 의미한다. 배경 클립 내 목표 인물의 실제 삽입 위치를 결정할 때 고려되어야 하는 중요한 원칙들은 다음과 같다. (1) 어떤 경우에도 보강 대상인 목표 인물의 일부가 아닌 전체가 삽입되어야 하고, (2) 피할 수 없는 경우에도 삽입되는 목표 인물이 기존의 배경 클립 내 주요 등장인물들과 겹치거나 가리는 부분을 최소화한다. (3) (1)과 (2)의 원칙이 지켜지는 한도에서 목표 인물의 삽입 상하 위치 $T_{\tilde{y}}$는 Algorithm 2의 7번째 줄과 같이 후보 영역 범위 내에서 무작위로 결정한다. 이 같은 원칙을 토대로 목표 인물의 삽입 위치는 다음과 같이 결정한다. (1) 목표 클립 내 목표 인물 마스크들의 최대 크기가 삽입 후보 영역인 R^*보다 작으면(T_w 〈= R_w), Fig. 4의 (a)와 Algorithm 2의 1-2번째 줄과 같이 목표 인물의 삽입 가로/좌우 위치 $T_{\tilde{x}}$는 영역 R^*를 벗어나지 않는 좌우 범위 내 무작위 위치로 결정한다. (2) 만약 목표 인물 마스크들의 최대 크기가 영역 R^*보다 더 크면(T_w 〉 R_w), 해당 영역에 손상 없이 목표 인물 전체를 온전히 삽입할 수 없다. 따라서 이 경우에는 Fig. 4의 (b)와 같이 영역 R^*를 벗어나 목표 인물 전체가 삽입될 수 있도록 위치를 조정하여 결정한다. (1) R^*가 좌측 영역(R_1)인 경우(R^* == R_1), Algorithm 2의 4번째 줄에 따라 목표 인물을 배경 프레임의 가장 좌측에 두는 위치로 결정된다. (2) R^*가 가운데 영역(R_2)인 경우(R^* == R_2), Algorithm 2의 5번째 줄에 따라 목표 인물이 양옆의 배경 인물들을 모두 최소로 가릴 수 있는 중간 삽입 위치로 결정된다. (3)R^*가 우측 영역(R_3)인 경우(R^* == R_3), Algorithm 2의 6번째 줄에 따라 목표 인물을 배경 프레임의 가장 우측에 두는 위치로 결정된다.

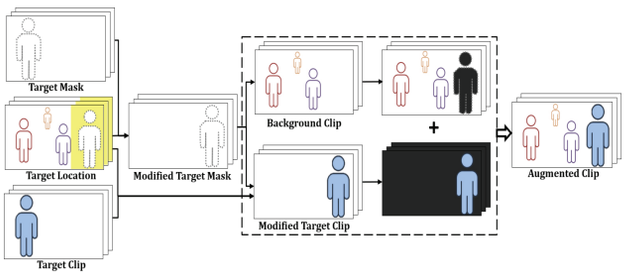


Fig. 5. Mask-based Augmented Video Clip Generation

3.3 보강 클립 생성

제안 기법의 마지막 단계는 Fig. 5와 같이 마스크를 이용해 목표 인물을 배경 클립에 삽입함으로써, 새로운 보강 클립들을 생성하는 과정이다.

먼저 새로운 위치로 목표 인물을 미리 옮기기 위한 변경된 목표 마스크들(modified target masks) M'^k_t과 변경된 목표 클립(modified target clip) c'_t을 다음과 같이 각각 생성한다. (1) 기존의 목표 인물 마스크들을 새로운 기준 위치로 옮기고 나머지 부분을 0으로 패딩(padding)하여, 변경된 목표 인물 마스크 M'^k_t들을 생성한다. (2) 기존 및 변경된 목표 인물 마스크들을 이용하여, 기존의 목표 클립 c_t로부터 목표 인물이 새로운 위치로 옮겨진 변경된 목표 클립 c'_t을 생성한다. 다음은 Fig. 5와 같이 변경된 목표 인물 마스크M'^k_t들, 변경된 목표 클립 c'_t, 그리고 배경 클립 c_b를 토대로, 마스크 기반의 Copy-Paste를 수행하여 새로운 보강 클립 c_a을 생성한다. 이때 보강 클립 c_a을 구성하는 각 프레임 c^k_a은 Equation (4)와 같이 계산한다.

$$c^k_a = M'^k_t c'^k_t + (1 - M'^k_t) c^k_b\ ,\quad k = 1, 2, ..., F \tag{4}$$

즉, 목표 인물 부분은 목표 마스크 M'^k_t를 이용해서 목표 클립 프레임 c'^k_t에서, 나머지 부분은 마스크 (1-M'^k_t)를 이용해서 배경 클립 프레임 c^k_b에서 각각 가져와 결합함으로써, 보강 클립의 각 프레임 c^k_a을 구성한다.

4. 데이터 집합과 심층 신경망 모델

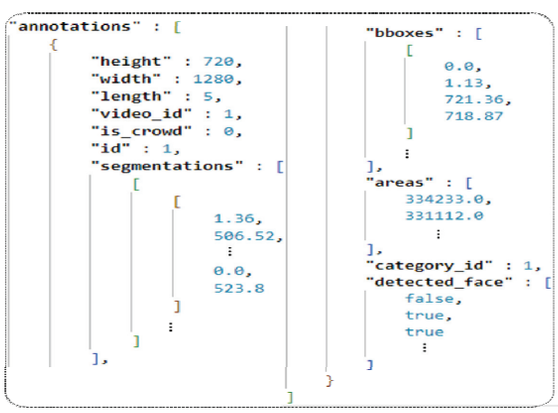
4.1 비디오 인물 개체 분할 데이터 집합

본 논문에서는 비디오 인물 개체 분할(VHIS)을 위해, 인물이 주로 등장하는 드라마 비디오들을 토대로 데이터 집합을 새롭게 구축했다. 비디오 인물 개체 분할 데이터 구축에 사용된 드라마는 tvN 드라마 “미생”이다. 미생은 등장인물들의 직장 생활을 주제로 한 드라마이다. 미생 드라마 데이터들을 토대로 MHIS 데이터 집합(Miseang Human Instance

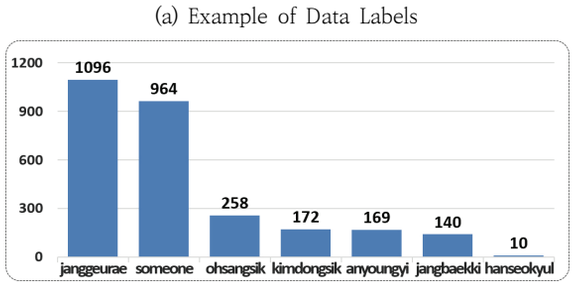
Table 1. Comparing Benchmark Datasets for Video Instance Segmentation

Statistics	YTVIS 19	YTVIS 21	OVIS	MHIS
Masks	131k	232k	296k	16k
Instances	4,883	8,171	5,223	3,490
Categories	40	40	25	7
Videos	2,883	3,859	901	1,944
Objects/frame	1.57	1.95	4.72	1.63

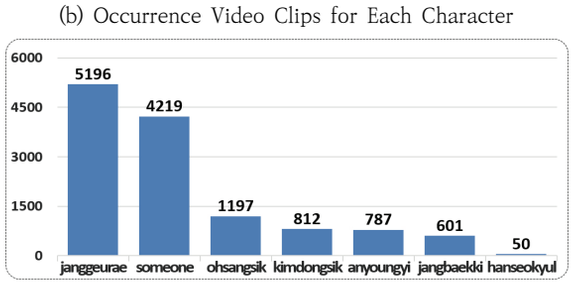
Segmentation Dataset, MHIS)을 구축하였다. 미생의 3개 화를 대상으로 Table 1과 같이 1,944개 비디오 클립에 속하는 9,623개의 프레임에 대해 데이터를 구축했다. 일반적인 비디오 개체 분할(VIS)을 위한 Youtube-VIS 2019(YTVIS 19)[1], Youtube-VIS 2021(YTVIS 21), OVIS[14] 벤치마크 데이터 집합들과 비교하였을 때, MHIS는 마스크 개수가 최소 115,500개 이상 적다.



(a) Example of Data Labels



(b) Occurrence Video Clips for Each Character



(c) Occurrence Frames for Each Character

Fig. 6. Miseang Human Instance Segmentation Dataset

데이터 구축을 위해 무료 오픈 소스 웹 기반의 컴퓨터 비전 주석 도구인 CVAT(Computer Vision Annotation Tool)를 사용하였다. 새롭게 구축한 MHIS의 레이블 스키마 구조는 기존의 비디오 개체 분할 데이터 집합인 Youtube-VIS와 유사하게 설계하였으며, Fig. 6의 (a)와 같은 JSON 형식으로 구성하였다. 비디오 인물 개체 분할 데이터 구축을 위해 비디오 클립을 대상으로 프레임마다 인물의 마스크 위치(segmentations), 경계 상자(bboxes), 클래스(category_id)를 라벨링 했다. 비디오 개체 분할 모델이 인물을 클립 단위로 탐지할 때, 인물의 움직임이나 가려짐 등에 의해 해당 인물을 특정 클래스로 식별할 수 없는 경우가 존재한다. 이러한 데이터에 의해 모델이 잘못 인식하면 전반적인 성능이 하락하므로, 인물 식별을 위한 얼굴 속성값(detected_face)을 추가했다. 비디오 클립에서 인물의 얼굴이 확인되어 인물을 식별할 수 있는 경우는 참, 아닌 경우는 거짓으로 표현하였다.

Fig. 6의 (b)는 MHIS 데이터 집합에서 각 인물들이 등장하는 비디오 클립의 개수를, Fig. 6의 (c)는 각 인물들이 등장하는 프레임의 개수를 각각 나타낸다. 즉, Fig. 6의 (b)와 (c)는 MHIS 데이터 집합의 인물별 데이터 분포도를 보여준다. 미생 드라마의 주연 인물 중 한 명인 장그래가 가장 많은 등장 횟수를 보이며, 주연 이외의 인물인 someone이 두 번째로 많이 등장한다. 장그래와 someone의 프레임당 등장 빈도수를 따지면, 두 클래스가 전체 데이터의 약 73%를 차지하는 데이터 불균형(data imbalance)이 있음을 알 수 있다. 비디오 인물 개체 분할 데이터에서 인물 클래스 간의 불균형은 개체 분할 성능에 비해 클래스 분류 성능이 낮은 문제의 원인이 된다.

4.2 비디오 인물 개체 분할 심층 신경망 모델

본 논문에서는 비디오 인물 개체 분할을 위한 베이스라인 심층 신경망 모델로, 비디오 개체 분할(VIS)을 위한 트랜스포머 구조 기반의 SeqFormer[6]를 이용하였다. SeqFormer의 전체 구조는 Fig. 7과 같다.

먼저 백본 네트워크로부터 여러 크기의 특징맵을 만들어서 입력 순서를 나타내는 위치 인코딩을 적용하여 SeqFormer의 입력으로 사용한다. 인코더에서는 변형 어텐션을 통해 프레임마다 풍부한 시각적인 정보를 추출하여 입력된 특징맵

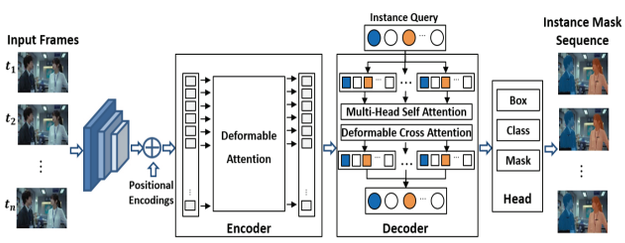


Fig. 7. The Overall Architecture of SeqFormer Model

픽셀 간의 관계를 전역적으로 학습한다. 디코더는 인코딩된 정보와 학습이 가능한 개체 쿼리(instance query)를 사용하여 변형 어텐션을 통해 서로 다른 프레임과의 맥락을 교환한다. SeqFormer에서는 개체 쿼리를 각 프레임마다의 박스 쿼리(box query)로 나눠서 어텐션 기법들을 수행하며, 여러 프레임의 박스 쿼리를 합함으로써 개체 쿼리 정보가 갱신된다. 갱신된 개체 쿼리의 정보로 경계 상자, 개체 클래스, 마스크 신경망 모듈을 통해 각 정보를 탐지한다.

비디오 개체 분할 모델인 SeqFormer 학습에 사용된 손실함수는 Equation (5)와 같다.

$$L_{match}(y,\hat{y})=\sum_{i=1}^{N}L_{class}(c_i,\widehat{c_m})+L_{bbox}(b_i,\widehat{b_m})+L_{mask}(m_i,\widehat{m_m}) \quad (5)$$

먼저, 비디오 개체 분할 모델의 탐지 결과를 대상으로 어떤 정답 값과 학습시킬지를 비교해서 정한다. 비교 과정은 헝가리안 알고리즘에 의해 수행된다. 비디오 개체 분할 모델의 출력 i는 개체 클래스 c_i, 경계 상자 b_i, 마스크 m_i이다. 손실함수는 개체 클래스에 대한 Cross-Entropy Loss, 경계 상자에 대한 GIoU Loss와 마스크에 대한 Focal Loss를 함께 사용한다.

5. 구현 및 실험

본 논문에서 제안하는 비디오 데이터 보강 기법은 Ubuntu 20.04 LTS 환경에서 PyTorch 딥러닝 라이브러리를 이용하여 구현하였으며, RTX A5000 GPU 3대가 장착된 컴퓨터를 이용하여 보강 및 학습을 진행하였다. 비디오 개체 분할 신경망 모델 SeqFormer[6]의 학습과 검증에는 본 논문에서 구축한 MHIS 데이터 집합을 이용하였다. 모델 학습을 위한 최적화 알고리즘은 AdamW를 사용하였다. 비디오 개체 분할 모델의 정량적 성능 평가 지표로 평균 정밀도(Average Precision, AP)와 평균 검출률(Average Recall, AR)을 측정하였다.

첫 번째 실험은 기존의 비디오 데이터 보강 기법들과의 성능 비교를 통해, 신규 제안 기법인 CDVA의 우수성을 입증하기 위한 실험이다. 따라서 이 실험에서는 제안 기법 CDVA를 데이터 보강 기법을 전혀 적용하지 않은 W/O, VideoMix[2], ObjectMix[3], B-Aug[4]들과 비교하였다. 이 실험에서는 비디오 인물 개체 분할을 위한 심층 신경망 모델은 모두 공통적으로 SeqFormer을 이용하고, 인물 클래스별 최대 보강 클립 수(N)는 1,800으로 설정하였다. Table 2의 실험 결과를 살펴보면, 신규 제안 기법인 CDVA를 적용한 경우가 평균 정확도인 AP 측면에서 가장 높은 성능을 보였으며, 기존의 W/O, VideoMix, ObjectMix, B-Aug을 적용한 경우들 대비 각각 14.6%. 6.28%, 7.69%, 3.70%의 성능 개선율을 보였다. 또한, 인물별 분할 정확도를 살펴보면 MHIS 데이터 집합에서 데이터가 상대적으로 부족했던 인물 클래스들인 오상식, 김동식, 장백기, 안영이, 한석율에 대한 분할 성능이 향상되었음을 확인할 수 있다. 특히 데이터 수가 현저히 적었던 한석율 클래스의 AP 성능은 비디오 데이터 보강이 이루어지지 않은 W/O 경우(AP=8.5)에 비해 CDVA 보강 기법을 적용했을 때(AP=92.5) 분할 성능이 무려 약 988%나 향상된 것을 확인할 수 있다. 이와 같은 실험 결과를 통해 부족한 인물 클래스에 대한 분할 성능 향상에 제안 보강 기법인 맥락-의

Table 2. Experimental Results with Different Video Augmentation Methods

Augmentation	AP	AP_{50}	AP_{75}	AR_1	AR_{10}	janggeurae	ohsangsik	kimdonsik	jangbaekki	anyoungyi	hanseokyul	someone
W/O	63.5	70.5	66.9	78.8	84.4	78.0	82.4	67.5	72.7	71.1	8.5	64.1
VideoMix[2]	68.5	77.0	71.7	78.8	83.0	70.4	86.9	68.0	69.4	73.1	61.1	50.7
ObjectMix[3]	67.6	75.5	70.9	78.5	82.8	65.6	84.2	70.1	64.7	73.2	64.3	51.0
B-Aug[4]	70.2	77.8	74.5	77.8	82.3	75.4	82.0	69.2	69.7	66.7	69.0	59.6
CDVA(Ours)	72.8	80.1	76.6	78.2	82.0	66.8	80.7	68.0	73.1	71.8	92.5	56.8

Table 3. Experimental Results with Different Video Instance Segmentation Models

Models	Augmentation	AP	AP_{50}	AP_{75}	AR_1	AR_{10}
MaskTrack RCNN[1]	W/O	52.7	66.7	59.1	56.5	59.7
	CDVA	58.0	70.3	64.3	59.9	62.4
IFC[9]	W/O	57.0	63.9	60.0	58.4	61.7
	CDVA	66.2	73.4	69.3	67.7	71.2
SeqFormer[6]	W/O	63.5	70.5	66.9	78.8	84.4
	CDVA	72.8	80.1	76.6	78.2	82.0

존적인 CDVA 기법의 긍정적인 효과와 우수성을 확인할 수 있었다. 하지만 데이터 보강을 통해 인위적으로 인물 클래스 간에 훈련 데이터 분포를 조정함으로써, 장그래와 같은 일부 빈발 인물 클래스에 대해서는 오히려 분할 성능이 조금 저하되는 부정적인 효과도 동시에 확인할 수 있었다. 이러한 문제점을 극복할 수 있도록 현재 CDVA의 추가적인 개선 노력도 필요할 것으로 판단한다.

두 번째 실험은 서로 다른 비디오 개체 분할 신경망 모델들에서도 제안 비디오 보강 기법인 CDVA가 동일하게 긍정적인 효과를 볼 수 있는지 확인하는 CDVA의 일반성 테스트 실험이다. 따라서 이 실험에서는 비디오 데이터 보강 기법을 전혀 적용하지 않은 W/O 경우와 제안 보강 기법인 CDVA를 적용한 경우 각각에 대해 베이스 모델인 SeqFormer[6]를 비롯하여 MaskTrack RCNN[1], IFC[9] 등 다양한 비디오 개체 분할 모델들의 분할 성능을 서로 비교하였다. 실험 결과를 나타내는 Table 3을 살펴보면, 데이터 보강을 수행하지 않은 W/O의 경우에 비해 제안 보강 기법인 CDVA를 적용한 경우에 모든 비디오 개체 분할 모델의 분할 성능이 향상되었음을 확인할 수 있다. 즉 MaskTrack RCNN 모델은 AP가 52.7에서 58.0으로, IFC 모델은 57.0에서 66.2로, 베이스 모델인 SeqFormer 모델은 63.5에서 72.8로 각각 성능이 향상되었다.

또한, 3가지 서로 다른 비디오 개체 분할 모델들 간의 성능을 비교해보면, 본 연구에서 베이스 모델로 채택한 SeqFormer가 가장 높은 성능을 보였고, 그 다음으로는 또 다른 Transformer 기반의 비디오 개체 분할 모델인 IFC와 합성곱 신경망(CNN) 기반의 MaskTrack RCNN 모델의 순서대로 낮은 성능을 보였다. 이와 같은 실험 결과를 통해, 본 논문에서 제안한 CDVA 비디오 데이터 보강 기법이 여러 다양한 비디오 개체 분할 모델들에서도 동일하게 성능 향상에 도움을 줄 수 있다는 긍정적 효과를 확인할 수 있었다.

세 번째 실험은 제안 기법인 CDVA의 비디오 보강 결과물들을 기존의 기법들과 정성적으로 비교 분석하는 실험이다. Fig. 8은 MHIS 비디오 데이터 집합을 토대로 주연 인물인 장그래가 존재하는 배경 클립에 신규 삽입 목표 인물인 오상식을 삽입하여 새롭게 생성된 보강 비디오 클립들을 보여준다. 이들 중에서 Fig. 8의 (a)는 기존의 비디오 데이터 보강 기법의 하나인 VideoMix[2]의 보강 결과물을 나타낸다. 목표 인물인 오상식의 첫 번째 프레임의 경계 상자 크기만큼의 영상들이 배경 클립의 중앙 지점에 삽입됨으로써, 경계 상자 내의 목표 인물 외 주변의 불필요한 장그래 얼굴 부분 영상도 배경 클립에 함께 삽입되어 비현실적인 비디오 클립이 생성되었다. 한편, Fig. 8의 (b)는 또 다른 기존 비디오 데이터 보강 기법인 ObjectMix[3]의 보강 결과물을 나타내며, 목표 클립 속에 존재하던 목표 인물인 오상식 이외에 다른 모든 인물들 즉, 장그래까지 포함하는 영상 부분들을 배경 클립에 삽입한 보강 비디오 결과를 보여주고 있다. 보강이 필요한 목표 인물 외의 불필요한 다른 인물까지 삽입되어, 정작 목표 인물 클래스의 데이터 부족 문제를 해결하는 데 큰 도움을 주지 못하는 결과를 보여주고 있다. 또한, 목표 클립의 여러 인물들이 배경 비디오 클립에 함께 삽입됨에 따라 배경 인물의 중요 부위들이 상당히 가려져 보이지 않는 문제점도 발생하였다. 예컨대, 본래 배경 클립에 있던 주연

Fig. 8. Example Videos Generated by Different Augmentation Methods:
(a) VideoMix, (b) ObjectMix, (c) B-Aug, (d) CDVA(Ours)

장그래의 대부분 영역이 새로 삽입된 영상 영역들에 의해 가려진 것을 확인할 수 있다.

Fig. 8의 (c)는 비디오 개체 분할을 위한 기존의 비디오 데이터 보강 기법의 하나인 B-Aug[4]의 결과물을 나타낸다. B-Aug는 목표 클립의 각 프레임 속 목표 인물만 포함하는 단일 마스크를 이용해 목표 인물 영상 영역들을 배경 클립의 동일 위치, 동일 크기로 삽입하는 기법이다. 따라서 이 기법에서는 배경 클립 내 기존 인물의 시-공간적 맥락을 제대로 반영하지 못하여 배경 인물과 거의 겹치게 되는 공간적 위치에 목표 인물이 삽입되는 결과를 생성하였다. 한편, Fig. 8의 (d)는 본 연구에서 제안한 CDVA 기법을 적용한 결과물을 나타낸다. 목표 인물인 오상식이 영상의 좌측에 삽입된 보강 결과를 보여주며, 배경 클립의 기존 인물과 겹치지 않고 적절한 시-공간 영역에 새로운 인물 오상식을 삽입함으로써, 매우 현실성 있는 보강 비디오 데이터가 생성된 것을 확인할 수 있다.

마지막 네 번째 실험은 제안 보강 기법인 CDVA를 MHIS 데이터 집합에 적용했을 때 비디오 인물 개체 분할에 어떤 효과를 미치는지 기존의 보강 기법들과 정성적으로 비교 분석하는 실험이다. Fig. 9는 서로 다른 비디오 데이터 보강 기법들을 적용했을 때 인물 개체 분할 결과들을 비교해 보여주고 있다. 이 실험에서는 앞서 소개한 SeqFormer를 비디오 인물 개체 분할을 위한 심층 신경망 구조로 이용하였다. Fig. 9의 개체 분할 결과물들을 살펴보면, 좌우에 위치하는 주요 등장 인물들인 장그래와 오상식에 대해 각기 다른 분할 예측 결과들이 생성된 것을 알 수 있다. Fig. 9의 (a)는 비디오 데이터 보강 기법을 적용하지 않은 경우(W/O)의 분할 결과물을 나타내며, 뒷모습만 보이는 장그래의 뒷면에 대해서는 개체 분할이 제대로 이루어지지 않았지만, 옆모습을 보이는 오상식에 대해서는 올바르게 분할되었다. Fig. 9의 (b)는 기존의 VideoMix[2] 보강 기법을 적용하였을 때의 분할 결과물로서, 장그래에 대한 분할은 올바르게 수행되었지만, 장면의 전환에 따라 트래킹이 올바르게 수행되지 않았다. 또한, 오상식에 대해서는 분할이 전혀 이루어지지 않았다.

Fig. 9의 (c)는 기존 비디오 데이터 보강 기법인 ObjectMix[3]을 적용하였을 때의 분할 결과물을 나타내며, Fig. 9의 (b)의 경우와는 반대로 이번에는 장그래에 대한 개체 분할이 전혀 이루어지지 않았으나, 오상식에 대해서는 분할이 올바르게 이루어진 것을 확인할 수 있다. Fig. 9의 (d)는 기존 비디오 데이터 보강 기법인 B-Aug[4]을 적용하였을 때의 분할 결과물로서, 장그래의 뒷면에 대해서는 분할 및 트래킹이 잘 되었지만, 마지막 프레임의 앞면에 대해서는 분할과 트래킹이 제대로 이루어지지 않았다. 오상식에 대해서는 전체 프레임에

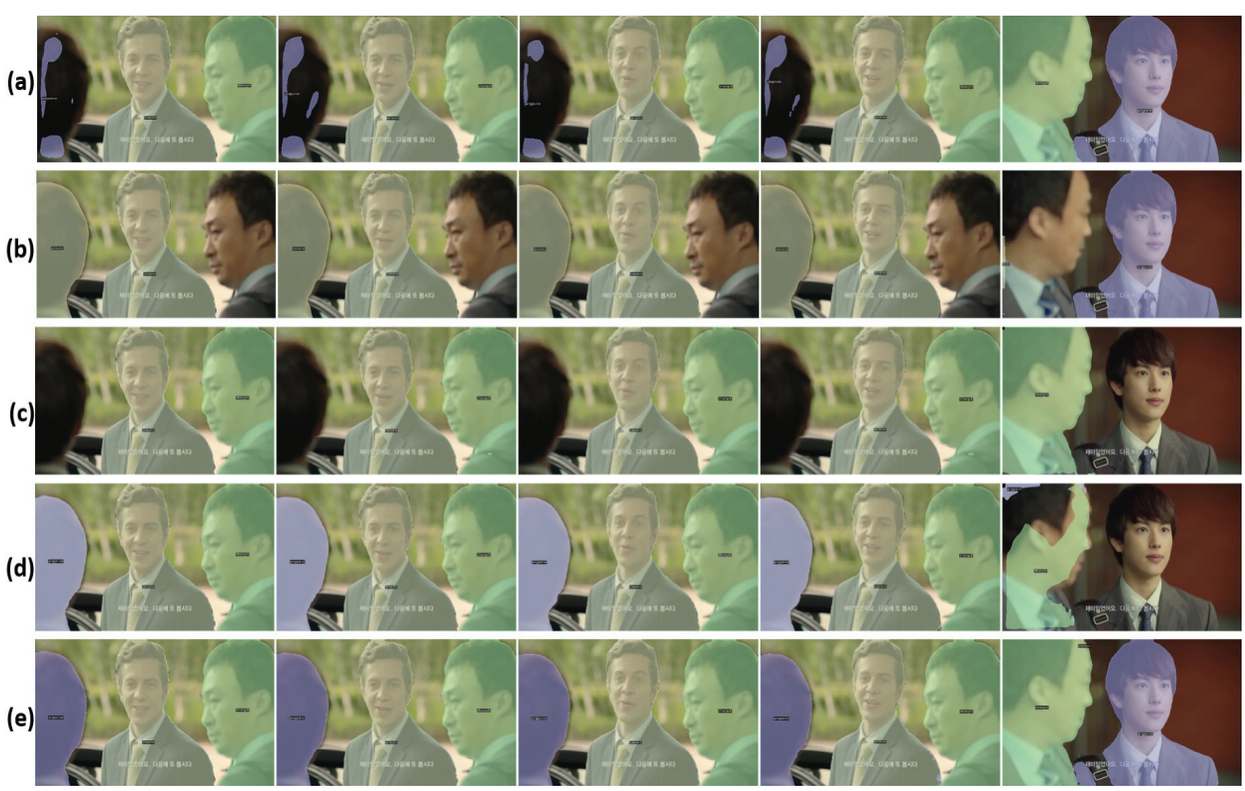


Fig. 9. Human Instance Segmentation Results with Different Augmentation Methods: (a) W/O, (b) VideoMix, (c) ObjectMix, (d) B-Aug, (e) CDVA(Ours)

대해 분할 및 트래킹이 이루어졌지만, 마지막 프레임에서는 인물 개체 분할이 제대로 이루어지지 않았다. 한편, Fig. 9의 (e)는 본 연구에서 제안한 CDVA 기법을 적용한 결과물로서, 장그래와 오상식 모두에 대해 탐지, 분류, 분할, 트래킹이 모두 올바르게 이루어진 것을 알 수 있다. 이와 같은 정성적 실험 결과들을 종합해볼 때, 기존의 비디오 데이터 보강 기법들을 사용했을 때에 비해 제안한 CDVA 기법을 사용했을 때 비디오 인물 개체 분할의 성능이 더 향상될 수 있음을 다시 확인할 수 있다.

6. 결 론

본 논문에서는 드라마 비디오 인물 개체 분할 작업을 위한 데이터 집합 MHIS를 구축하고, 새로운 비디오 데이터 보강 기법 CDVA를 제안하였다. 새로 제안한 기법은 시-공간적 맥락을 충분히 고려해서 부족한 인물 클래스의 훈련 비디오 데이터들을 추가 생성함으로써, 비디오 개체 분할 신경망 모델의 성능을 효과적으로 개선할 수 있었다. 본 논문에서는 정량 및 정성 실험들을 통해, 제안 보강 기법의 우수성을 입증하였다. 하지만 현재의 제안 기법은 시-공간 맥락에 따라 결정된 배경 클립의 삽입 영역에 맞추어 삽입 대상 목표 인물의 크기를 자동 조절하는데 부분적인 한계성이 있다. 따라서 배경 인물을 일부 가리는 현상이 발생하기도 한다. 또한, 현재의 CDVA 기법은 데이터 보강을 통해 인위적으로 인물 클래스간의 훈련 데이터 분포를 조정함으로써, 일부 빈발 인물 클래스에 대해서는 오히려 성능이 저하되는 부정적인 효과를 줄 수 있는 부분도 있다. 향후에는 이러한 문제점들을 해결하여 더 현실적인 보강 데이터를 생성하고 다양한 비디오 개체 분할 모델과 비디오 데이터 집합들에 폭넓게 활용할 수 있도록, 현재의 제안 기법을 확장하는 연구를 추가로 진행할 예정이다.

References

[1] L. Yang, Y. Fan, and N. Xu, “Video instance segmentation,” *Proceedings of IEEE/CVF International Conference on Computer Vision*, 2019.

[2] S. Yun, S. J. Oh, B. Heo, D. Han, and J. Kim, “VideoMix: Rethinking data augmentation for video classification,” *arXiv preprint arXiv: 2012.03457*, 2020.

[3] J. Kimata, T. Nitta, and T. Tamaki, “ObjectMix: Data augmentation by copy-pasting by copy-pasting objects in videos for action recognition.” *arXiv preprint arXiv: 2204.00239*, 2022.

[4] H. Kim, D. Kim, J. Kim, and S. Im, “Data augmentation scheme for semi-supervised video object segmentation,” *Journal of Broadcast Engineering*, Vol.27, No.1, 2022.

[5] H. J. Chun and I. Kim, “Human instance segmentation using video data augmentation.” *Proceedings of the Annual Conference of Korea Information Processing Society Conference (KIPS) 2022*, Vol.29, No.2, pp.532-534, 2022.

[6] J. Wu, Y. Jiang, S. Bai, W. Zhang, and X. Bai, “SeqFormer: Sequential transformer for video instance segmentation,” *European Conference on Computer Vision,* Springer, Cham, 2022.

[7] A. Athar, S. Mahadevan, A. Osep, L. Leal-Taixé, and B. Leibe, “STEm-Seg: Spatio-temporal embeddings for instance segmentation in videos.” *European Conference on Computer Vision, Springer*, Cham, 2020.

[8] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, “End-to-end video instance segmentation with transformers.” *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[9] S. Hwang, M. Heo, S. W. Oh, and S. J. Kim, “Video instance segmentation using inter-frame communication transformers.” *Advances in Neural Information Processing Systems*, Vol.34, 2021.

[10] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” *European Conference on Computer Vision*, Springerm Cham, 2020.

[11] H. S. Fang, J. Sun, R. Wang, M. Gou, Y. L. Li, and C. Lu, “InstaBoost: Boosting instance segmentation via probability map guided copy-pasting,” *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[12] N. Dvornik, J. Mairal, and C. Schmid, “Modeling visual context is key to augmenting object detection datasets,” *Proceedings of the European Conference on Computer Vision*, 2018.

[13] G. Ghiasi et al., “Simple copy-paste is a strong data augmentation method for instance segmentation,” *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[14] J. Qi et al., “Occluded video instance segmentation.” *arXiv preprint arXiv: 2102.01558*, 2021.

전 현 진

https://orcid.org/0000-0001-9793-1271
e-mail : wlsrh135@kyonggi.ac.kr
2022년 경기대학교 컴퓨터공학부(학사)
2022년 ~ 현 재 경기대학교 컴퓨터과학과 석사과정
관심분야 : 인공지능, 기계학습, 컴퓨터비전

이 종 훈

https://orcid.org/0000-0001-7436-5561
e-mail : jhlee17139@kyonggi.ac.kr
2020년 경기대학교 컴퓨터과학과(학사)
2020년 ~ 현 재 경기대학교 컴퓨터과학과 석사과정
관심분야 : 인공지능, 기계학습, 컴퓨터비전

김 인 철

https://orcid.org/0000-0002-5754-133X
e-mail : kic@kyonggi.ac.kr
1985년 서울대학교 수학과(학사)
1987년 서울대학교 전산과학과(석사)
1995년 서울대학교 전산과학과(박사)
1996년 ~ 현 재 경기대학교 컴퓨터공학부 교수
관심분야 : 인공지능, 기계학습, 로봇지능