# 비디오 데이터 보강을 이용한 인물 개체 분할

전현진, 김인철
경기대학교 컴퓨터공학부
wlsrh135@kyonggi.ac.kr, kic@kyonggi.ac.kr

# Human Instance Segmentation using Video Data Augmentation

Hyun-Jin Chun, Incheol Kim
Department of Computer Science Kyonggi University

요 약

본 논문에서는 미생 드라마 비디오들을 토대로 구축한 비디오 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 비디오의 시공간적 맥락을 충분히 고려해서 부족한 인물 클래스의 훈련 비디오 데이터들을 추가 생성함으로써, 비디오 개체 분할 신경망 모델의 성능을 효과적으로 개선시킬 수 있다. 본 논문에서는 정량 및 정성 실험들을 통해, 제안 비디오 데이터 보강 기법의 우수성을 입증한다.

## 1. 서론

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(segmentation), 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 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 자율 주행, 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 분할에 비해, 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-VIS와 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

* 본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 대학ICT연구센터육성지원사업의 연구결과로 수행되었습니다. (IITP-2017-0-01642)


(그림 1) 기존의 비디오 데이터 보강 기법 간의 비교

VideoMix[2]와 ObjectMix[3], 비디오 개체 분할을 위한 B-Aug[4]를 제외하고는 거의 찾아보기 어려운 실정이다.

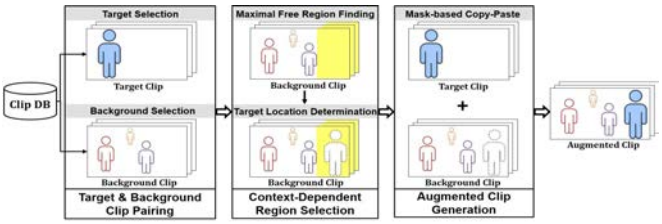
앞서 설명한 대로 드라마 비디오 인물 개체 분할에서는 각 인물이 등장하는 장소와 배경, 함께 상호작용하는 다른 등장인물 등과 같은 공간적 맥락뿐만 아니라, 이전 그리고 이후의 장면들과의 일관성을 위한 시간적 맥락이 매우 중요하다. (그림 1)은 미생 드라마의 주연인 장그래에 비해 등장 빈도가 낮은 오상식의 부족한 훈련용 비디오 데이터의 생성을 위한 서로 다른 기존의 비디오 데이터 보강 기법들의 적용 사례들을 나타낸다. VideoMix 기법은 (그림 1)의 (a)처럼 오상식이 등장하는 원본 목표 클립의 첫 번째 프레임 내 경계상자와 동일한 크기와 위치로 임의로 선정된 배경 클립의 모든 프레임에 삽입하는 방법이다. ObjectMix는 (그림 1)의 (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 (그림 1)의 (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.

(그림 1)의 (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. 이와 같이 생성된 비-현실적인 보강 데이터들은 시간적, 공간적 맥락에 민감한 비디오 개체 분할 모델의 성능을 효과적으로 향상시키기는 어렵다. 따라서 본 논문에서는 이러한 기존 비디오 데이터 보강 기법들의 한계를 극복하고자, 새로운 맥락-의존성 비디오 데이터 보강 기법 CDVA (Context-Dependent Video Data Augmentation)을 제안한

다. 본 논문에서는 MHIS 데이터 집합을 이용한 정량 및 정성 실험들을 통해, 제안 비디오 데이터 보강 기법의 우수성을 입증한다.

## 2. 비디오 데이터 보강

본 논문에서 제안하는 맥락-의존적 비디오 데이터 보강 기법 CDVA는 다양한 시공간적 맥락을 고려함으로써 부족 인물 클래스를 위한 보다 현실성 있는 보강 비디오 데이터들을 생성한다.




(그림 2) 맥락-의존적 비디오 데이터 보강

제안하는 CDVA 비디오 데이터 보강은 (그림 2)와 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.

### 2.1 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, 알고리즘 1과 같이 수행된다.

**Algorithm 1** : Target and Background Clips Pairing

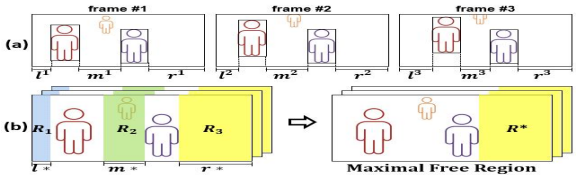
**Input:** Video Clip Set $S$, Target Classes $\{C_1, C_2, ..., C_m\}$, Maximum number of clips for each class $N$
**Output:** Clip Pair Set $P$

1. Initialize $P$ = ∅
2. $S_B$=$S$-($ClipSet(C_1) \cup ... \cup ClipSet(C_m)$)
3. **for** $i = 1, 2, ..., m$ **do**
4. requiredClips = $N$ - $numClips(C_i)$
5. countClips = 0
6. **for** targetClip $c_t$ in $ClipSet(C_i)$ **do**
7. **for** backgroundClip $c_b$ in $S_B$ **do**
8. **if** countClips < requiredClips **then**
9. $P = P \cup \{(c_t, c_b)\}$
10. countClips++

알고리즘 1에서 $C_i$는 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 $S_B$는 보강이 필요한 목표 인물 개체를 포함하고 있지 않은 비디오 클립들의 집합을 나타내며, $P$는 클립 짝짓기의 결과로서 보강 대상 인물을 포함한 목표 비디오 클립과 이것을 삽입할 배경 비디오 클립 쌍들의 집합을 나타낸다. 알고리즘 1은 클래스별 최대 비디오 클립 데이터 개수인 $N$에서 현재 목표 클래스 $C_i$의 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 $c_t$에 대응하는 배경 클립 $c_b$를 목표 인물 개체를 포함하고 있지 않은 비디오 클립들의 집합인 $S_B$에서 골라 비디오 데이터 보강을 위한 클립 쌍 $(c_t, c_b)$들을 생성하고, 이들을 집합 $P$에 담는다.

### 2.2 맥락-의존적 영역 선정

비디오 데이터 보강 기법 CDVA에서는 배경 비디오 클립(background video clip) $c_b$의 전체 맥락을 고려해서, 목표 클립 $c_t$의 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.




(그림 3) 최대 자유 영역 찾기

먼저 최대 자유 영역 $R$*을 찾는 과정은 (그림 3)의 (a)와 같이, 배경 클립의 각 프레임에서 인물이 존재하지 않는 가로 범위들을 모두 탐색한다. $k$번째 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 $m^k$는 $k$번째 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, (그림 3)의 (b)와 같이 배경 비디오 클립을 구성하는 프레임 시퀀스의 시간적 맥락을 반영하기 위해, 전체 프레임들에 걸쳐 (식 1), (식 2)와 같이 계산한 최소 가로 범위 $l^*, m^*, r^*$들과 후보 영역 $R_1, R_2, R_3$들을 이용하여 최종적으로 목표 인물을 삽입하기 위한 맥락-의존적 영역 $R$*을 결정한다.

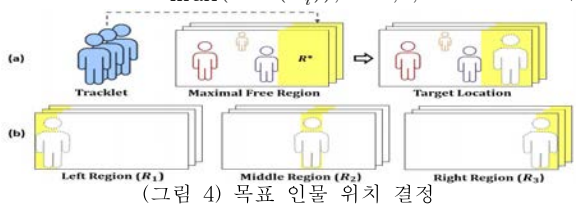
$$l^* = \min(l^k), m^* = \min(m^k), r^* = \min(r^k),\ k = 1,2,...,F \quad (1)$$

(식 1)에서 $F$은 배경 클립의 총 프레임 수를 나타낸다. 최소 가로 범위 $l^*, m^*, r^*$ 각각에 프레임 높이 $I_h$를 곱하여, 배경 클립 내 목표 인물 삽입 후보 영역들인 $R_1, R_2, R_3$를 (식 2)와 같이 계산한다.

$$R_1 = l^* \times I_h,\ R_2 = m^* \times I_h,\ R_3 = r^* \times I_h \quad (2)$$

그리고 (식 3)과 같이, 배경 클립 내의 후보 영역들인 $R_1, R_2, R_3$ 중 가장 넓은 영역을 최종적인 목표 인물 삽입을 위한 최대 자유 영역 $R$*로 선정한다.

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



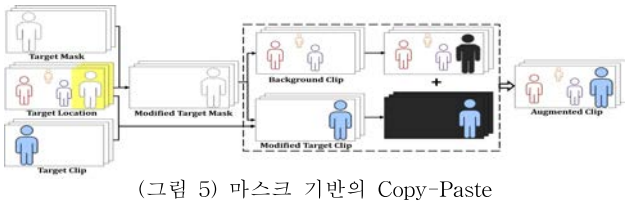

(그림 4) 목표 인물 위치 결정

배경 클립 내의 최대 자유 영역 $R$*에 목표 인물을 삽입할 구체적인 위치를 결정하는 과정은 (그림 4)와 같다. 목표 인물의 실제 삽입 위치는 후보 영역 $R$*의 크기와 삽입 대상 목표 인물의 크기에 따라 달리 결정된다. 이때 고려되어야 할 가지 원칙은 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 위치는 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, (그림 4)의 (a)와 같이 목표 인물의 삽입 위치는 영역 $R$*를 벗어나지 않는 좌우 범위 내 무작위 위치로 결정한다. (2) 만약 목표 인물 마스크들의 최대 크기가 영역 $R$*보다 더 크면, 해당 영역에 손상 없이 목표 인물 전체를 온전히 삽입할 수 없다. 따라서 이 경우에는 (그림 4)의 (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, 목표 인물을 배경 프레임의 가장 좌측에 두는 위치로 결정된다. (2) $R$*가 가운데 영역($R_2$)인 경우, 목표 인물이 양옆의 배경 인물들을 모두 최소로 가릴 수 있는 중간 삽입 위치로 결정된다. (3) $R$*가 우측 영역($R_3$)

인 경우, 목표 인물을 배경 프레임의 가장 우측에 두는 위치로 결정된다. 이처럼 결정된 목표 인물 삽입 기준 위치 $(x,y)$는 목표 클립 내 목표 인물 마스크들의 경계 상자 중 가장 좌상단의 좌표와 대응된다.

### 2.3 보강 클립 생성

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




(그림 5) 마스크 기반의 Copy-Paste

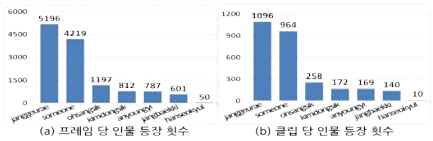
먼저 새로운 위치로 목표 인물을 미리 옮기기 위한 변경된 목표 마스크들(modified target masks) $M'^{k}_{t}$과 변경된 목표 클립(modified target clip) $c'_{t}$을 다음과 같이 각각 생성한다. (1) 기존의 목표 인물 마스크들을 새로운 기준 위치로 옮기고 나머지 부분을 0으로 패딩(padding)하여, 변경된 목표 인물 마스크 $M'^{k}_{t}$들을 생성한다. (2) 기존 및 변경된 목표 인물 마스크들을 이용하여, 기존의 목표 클립 $c_t$으로부터 목표 인물이 새로운 위치로 옮겨진 변경된 목표 클립 $c'_{t}$을 생성한다. 다음은 (그림 5)와 같이 변경된 목표 인물 마스크$M'^{k}_{t}$,들, 변경된 목표 클립 $c'_{t}$, 그리고 배경 클립 $c_b$을 토대로, 마스크 기반의 Copy-Paste를 수행하여 새로운 보강 클립 $c_a$을 생성한다. 이때 보강 클립 $c_a$을 구성하는 각 프레임 $c^{k}_{a}$은 (식 4)와 같이 계산한다.

$$c^{k}_{a} = M'^{k}_{t} c'^{k}_{t} + (1 - M'^{k}_{t}) c^{k}_{b}, k = 1, 2, ..., F \qquad (4)$$

즉, 목표 인물 부분은 목표 마스크 $M'^{k}_{t}$를 이용해서 목표 클립 프레임 $c'^{k}_{t}$에서, 나머지 부분은 마스크 $(1-M'^{k}_{t})$를 이용해서 배경 클립 프레임 $c^{k}_{b}$에서 각각 가져와 결합함으로써, 보강 클립의 각 프레임 $c^{k}_{a}$을 구성한다.

## 3. 데이터 집합과 심층 신경망 기본 모델

본 논문에서는 비디오 인물 개체 분할을 위해, 미생 드라마 비디오들을 토대로 MHIS 데이터 집합(Miseang Human Instance Segmentation Dataset)을 구축하였다. MHIS는 기존의 비디오 개체 분할 벤치마크 데이터 집합인 Youtube-VIS[1]와 유사한 레이블 스키마 구조로 설계하였다.




(그림 6) MHIS 비디오 데이터 집합 구성

인물 클래스는 드라마 주연들로 이루어진 6개 클래스와 이외의 인물들을 나타내는 someone 클래스 등 총 7개 클래스로 구성된다. (그림 6)의 (a)는 MHIS 데이터 집합에서 프레임 당 각 인물의 등장 빈도를, (b)는 클립당 각 인물의 등장 빈도를 각각 나타낸다. 이를 통해 인물 클래스 간에 상당한 수준의 데이터 불균형이 있음을 알 수 있다. 한편, 본 논문에서는 비디오 개체 분할을 위한 베이스라인 심층 신경망 모델로 Transformer 기반 SeqFormer[5]를 이용하였다.

## 4. 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 SeqFormer의 학습과 검증에는 본 논문에서 구축한 MHIS 데이터 집합을 이용하였다. 모델 학습을 위한 최적화 알고리즘은 AdamW를, 손실 함수는 경계 상자에 대한 GIoU Loss와 마스크에 대한 Focal Loss를 함께 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 None, VideoMix[2], ObjectMix[3], B-Aug[4]들과 비교하였다. 또 이 실험에서는 인물 클래스별 최대 보강 클립 수(N)는 1,800으로 설정하였다. <표 1>의 실험 결과를 살펴보면, CDVA를 적용한 경우가 AP 측면에서 기존의 None, VideoMix, B-Aug 대비 각각 14.6%, 6.28%, 7.69%, 3.70%의 성능 개선율을 보였다. 이를 통해 맥락 의존적인 CDVA의 우수성을 확인할 수 있었다.

<표 1> 기존 비디오 데이터 보강 기법들과의 성능 비교

| Method | AP | $AP_{50}$ | $AP_{75}$ | $AR_1$ | $AR_{10}$ |
|---|---|---|---|---|---|
| None | 63.5 | 70.8 | 66.9 | 78.8 | 84.4 |
| VideoMix[2] | 68.5 | 77.0 | 71.7 | 78.8 | 83.0 |
| ObjectMix[3] | 67.6 | 75.5 | 70.9 | 78.5 | 82.8 |
| B-Aug[4] | 70.2 | 77.8 | 74.5 | 77.8 | 82.3 |
| CDVA(Ours) | 72.8 | 80.1 | 76.6 | 78.2 | 82.0 |

두 번째 실험은 제안 기법인 CDVA의 비디오 데이터 보강 결과와 이를 이용한 비디오 인물 개체 분할 결과를 정성적으로 분석하는 실험이다.


(그림 7) 비디오 데이터 보강 결과(a)와 인물 개체 분할 결과(b)

(그림 7)의 (a)는 CDVA 기법으로 배경 클립에 오상식 인물 개체를 삽입한 보강 결과를, (b)는 보강된 비디오 데이터들을 이용해 학습한 모델이 해당 비디오에서 인물 개체를 분할한 결과를 각각 나타낸다. (a)의 결과를 통해서는 시공간 맥락에 합당한 현실적인 보강 비디오 데이터가 생성되었음을 확인할 수 있고, (b)의 결과를 통해서는 보강 비디오 데이터들을 토대로 개체 분할 신경망 모델이 등장 빈도가 높은 장그래뿐만 아니라 희소 등장인물인 오상식도 정확히 분할 가능하도록 성능 향상이 이루어졌음을 확인할 수 있다.

## 5. 결론

본 논문에서는 드라마 비디오 인물 개체 분할 작업을 위한 데이터 집합 MHIS를 구축하고, 새로운 비디오 데이터 보강 기법 CDVA를 제안하였다. 새로 제안한 기법은 시공간적 맥락을 충분히 고려해서 부족한 인물 클래스의 훈련 비디오 데이터들을 추가 생성함으로써, 비디오 개체 분할 신경망 모델의 성능을 효과적으로 개선시킬 수 있었다. 본 논문에서는 정량 및 정성 실험들을 통해, 제안 보강 기법의 우수성을 입증하였다.

## 참고문헌

[1] L. Yang, Y. Fan, and N. Xu. "Video Instance Segmentation," *Proc. of ICCV-2019,* 2019.
[2] S. Yun, S. Oh, B. Heo, D. Han, J, Kim, "VideoMix: Rethinking Data Augmentation for Video Classification," *arXiv:2020.03457,* 2020.
[3] J. Kimata, T. Nitta, T. Tamaki, "ObjectMix: Data Augmentation by Copy-Pasting Objects in Videos for Action Recognition," *arXiv:2204.00239,* 2022.
[4] H. Kim, D. Kim, et al., "Data Augmentation Scheme for Semi-Supervised Video Object Segmentation," *J. of Broadcast Engineering.* vol.27, No.1, 2022.
[5] J. Wu, Y. Jiang, et al., "SeqFormer: Frustratingly Simple Model for Video Instance Segmentation," *Proc. of ECCV-2022.* 2022.