특집논문 (Special Paper)
방송공학회논문지 제27권 제1호, 2022년 1월 (JBE Vol.27, No.1, January 2022)
https://doi.org/10.5909/JBE.2022.27.1.13
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

준지도 비디오 객체 분할 기술을 위한 데이터 증강 기법

김 호 진[a)], 김 동 현[a)], 김 정 훈[a)], 임 성 훈[a)‡]

Data Augmentation Scheme for Semi-Supervised Video Object Segmentation

Hojin Kim[a)], Dongheyon Kim[a)], Jeonghoon Kim[a)], and Sunghoon Im[a)‡]

요 약

동영상 객체 분할(VOS) 기술은 연속된 레이블링 데이터를 필요로 하며, 현재 공개된 데이터셋으로 훈련된 VOS방법은 그 성능이 제한된다. 이 문제를 해결하기 위해 본 논문에서는 간단하면서도 효과적인 동영상 데이터 증강 기술들을 제안한다. 첫번째 증강 기술은 영상 내에서 객체를 제외한 배경을 다른 영상의 배경으로 대체하는 기법이고, 두번째 기술은 학습될 동영상 데이터의 순서를 무작위 확률로 뒤집어 역 재생되는 영상을 학습시키는 기법이다. 두 증강 기술은 객체 분할 시 배경 정보에 강인한 추정을 가능하게 하였고, 추가 데이터 없이 기존 모델의 성능을 향상시킬 수 있음을 보였다.

Abstract

Video Object Segmentation (VOS) task requires an amount of labeled sequence data, which limits the performance of the current VOS methods trained with public datasets. In this paper, we propose two effective data augmentation schemes for VOS. The first augmentation method is to swap the background segment to the background from another image, and the other method is to play the sequence in reverse. The two augmentation schemes for VOS enable the current VOS methods to robustly predict the segmentation labels and improve the performance of VOS.

Keyword: Semi supervised video object segmentation, Data augmentation

a) 대구경북과학기술원 정보통신융합전공(DGIST, Information and Communication Engineering)
‡ Corresponding Author : 임성훈(Sunghoon Im)
E-mail: sunghoonim@dgist.ac.kr
Tel: +82-053-785-6323
ORCID:https://orcid.org/0000-0001-9776-8101

※ This work was supported by Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2014-3-00123, Development of High Performance Visual BigData Discovery Platform for Large-Scale Realtime Data Analysis).
· Manuscript received November 22, 2021; Revised December 27, 2021; Accepted December 29, 2021.

Ⅰ. 서 론

동영상 객체 분할(Video Object Segmentation, 이하 VOS)은 동영상에서 각 객체가 차지한 영역을 매 프레임마다 분할하는 기술이다. 그 중 준지도(Semi-supervised) 동영상 객체 분할은 첫 프레임에서 제공된 객체의 영역 분할 정보만으로 이후 프레임에서도 이어서 객체를 검출 및 분할하는 기술이다. 준지도 동영상 객체 분할을 학습시키기 위해 YouTube VOS와 같은 데이터셋[1]이 존재하나, 그 양은 한정적이다. 이러한 데이터 부족 문제를 위해서 데이터 증강 기술이 적극적으로 사용되고 있다[2]. 데이터 증강 기술은 데이터의 라벨을 보존하면서 데이터에 무작위 변형을 일으켜 학습 가능한 새 데이터를 생성한다. 따라서 데이터의 양이 증가한 것과 유사한 효과를 낼 수 있다.

기존 VOS 연구는 마찬가지로 데이터 증강 기법을 사용한다. 하지만 VOS 기술을 위해 설계된 비디오 데이터 증강 기법이 아닌, 주로 이미지 데이터 증강 기법에서 널리 사용되는 방법을 택했다[2]. 이에 반해 최근 연구자들은 과제의 특성에 맞춘 적절한 증강 기법을 사용한다면 모델의 성능을 더 높일 수 있다고 지적하였다[3,4]. 또한 동영상 데이터와 이미지 데이터는 시간이라는 요소의 차이가 있다. 하지만 이미지 데이터 증강 기법은 각 프레임 속 이미지만 변형시킬 수 있으므로, 생성할 수 있는 데이터의 다양성이 제한된다. 위와 같은 문제점에서 착안하여 본 연구는 동영상 데이터 및 VOS 기술에 적합한 새로운 데이터 증강 기법 두 가지를 제안하였다.

그 중 첫 번째는 배경 증강 기법이다. 최근 데이터 증강 연구에 영감을 받아[3], 객체와 배경을 서로 다른 동영상에서 불러오는 증강 기법을 고안하였다. VOS 모델을 시행한 결과 배경과 객체를 혼동하여 모델의 성능이 떨어지는 현상이 관찰되었다. 이것은 모델이 객체를 배경으로부터 구분하는 기준을 객체 정보가 아닌 색깔 및 위치와 같은 객체와 배경이 공유되는 특징으로 학습이 이뤄졌기 때문이다. 이에 비해서 객체를 제외한 배경을 다양한 영상으로 대체한 경우, 같은 객체여도 다른 배경을 지닌 데이터를 학습하게 된다. 생성된 데이터는 원본과 다른 데이터이지만 객체 정보 및 Ground truth는 일치하므로 같은 예측을 해야 하기에 모델은 배경과 공유되지 않는 객체 정보에 더 집중하고 배경 정보에 덜 민감하도록 학습이 이뤄지어 상기한 문제를 해결할 수 있다.

두 번째는 영상 역재생이다. 일반적인 동영상 데이터는 영상에서 이미지를 불러오되, 이미지의 순서는 바꾸지 않는 방법을 고수했다. 하지만 본 연구는 이미지의 순서를 역전시켜 시간의 흐름을 변형시킨 데이터를 학습시키는 것 또한 데이터 증강 기술로서 가치를 가진다고 제안한다. 그 이유는 VOS의 모델 구조에 있다. 모델은 이전 프레임의 정보를 활용해 현 프레임의 객체를 분할한다. 그러나 역 재생된 영상에서는 이전프레임과 현재 프레임의 관계가 역전되어 새로운 정보로 바뀐다. 따라서 역 재생된 영상은 기존 영상과 다른 영상으로 인식된다. 역 재생 영상의 이러한 특성은 데이터 증강 기법으로 활용될 수 있었다.

Ⅱ. 관련 연구

1. 준지도 동영상 객체 분할 기법

초기 준지도 VOS 모델은 Finetuning 방식을 주로 사용하였다[5,6,7]. 첫 프레임을 활용한 미세 조정 방법은 느린 학습 속도와 데이터 부족 문제에 부딪혔다. 다른 방법으로는 물체의 광학 흐름(Optical Flow)을 추정하여 다음 프레임으로 분할 정보를 전달하는 방법이 개발되었다[8,9]. 그러나 비경식(Non-rigid object) 물체에 대한 광학흐름을 추정하기 어렵다는 단점이 있었다. 최근 연구들은 STM (Space-Time Memory Networks)[10] 모델을 근간으로 첨단의 성능을 내고 있다. STM은 메모리 뱅크에 각 객체의 정보를 저장하여 다음 프레임에서 객체를 예측하기 위해 사용한다. 후속 연구들은 메모리 뱅크의 개선 혹은 손실함수의 추가 등으로 성능향상을 이끌었다. STCN (Space-Time Correspondence Networks)[11]은 STM의 구조를 기반으로 현재 가장 뛰어난 성능을 보이는 모델이다. STM 구조를 단순화하여 계산 복잡도를 낮추었다. 본 연구는 언급된 장점들로 미루어 STCN을 기반 모델로 설정하였다.

2. 데이터 증강 기법

동영상데이터 증강 기술 연구는 이미지 데이터 증강 기술로 사용된 방법을 보편적으로 따라간다[2]. 그러나 몇몇

이미지 증강 기술은 특정 과제에 적합하게 설계되었기에 동영상 데이터 증강기술 또한 과제에 맞는 적절한 선택이 필요하다. 보편적으로 쓰이는 이미지 증강 기술로는 Random Crop, color jittering 등이 있다[2].

CutMix[4]는 이미지 분류(Image Classification)를 위한 데이터 증강 기술이다. 이미지 안에 다른 이미지를 합성하여 이미지의 분류를 어렵게 만들어 성능을 향상시킨다. VideoMix[12]는 CutMix에서 영감을 받아 고안된 동영상 데이터 증강 기법이며, 동영상 동작 분류 과제를 위해 개발되었다. VideoMix 또한 영상 안에 영상을 합성하여 CutMix와 같은 효과를 내었다.

이미지 객체 분할(Image Instance Segmentation)을 위한 데이터 증강 기술 중에는 Copy-Paste[3]가 있다. 이미지에서 객체를 제외한 배경을 다른 이미지로 대체하여 객체 분할 과정에서 배경의 영향을 줄일 수 있도록 설계되었다. VOS는 사용자가 선택한 객체를 분할하는 과제이므로 이미지 객체 분할을 위해 개발된 Copy-Paste 기법 또한 동영상 데이터에서 효과적일 수 있을 거라 판단하여 배경 증강 기법을 제안하였다.

3. 시간-일관성

동영상 데이터의 중요한 축인 시간대에 관련된 연구는 다양하게 진행되었으나 데이터 증강기법으로 활용되지는 않았다. 뒤섞인 시간대를 제 시간대로 재배열하는 자가지도 학습이나[13], 정 재생된 동영상에서 검출한 객체와 역 재생된 동영상에서 검출한 객체는 서로 같아야 한다는 자가지도를 이용하는 학습 방법[14]이 연구되었다. 그러나 언급된 방법은 선행학습 기법으로 별도의 학습이 필요하다는 단점이 있다. 본 연구에서는 시간 축을 변형시키는 아이디어를 데이터 증강 기법에 적용하여 더 쉽고 효과적으로 성능을 향상시킬 수 있는 방법을 제안하였다.

Ⅲ. 실험 방법

1. 배경 증강 기법

선택된 객체의 분할 영역을 다른 영상에 붙여 넣어 새로운 영상을 생성하는 데이터 증강 기법 (Background Augmentation, B-Aug)를 구상하였다. 두 영상은 기존 학습에 쓰이던 데이터베이스에서 랜덤 샘플링 하였다. 그 중 객체 정보를 가져올 영상을 F, 배경 정보를 가져올 영상을 B라고 하였다.

다음은 선택된 두 이미지를 합치는 과정을 설명한다. 먼저 동영상의 길이만큼 각 영상의 길이를 조절하였다. 완성된 영상의 길이를 k라고 명시하였다. 이 때 새 영상의 길이는 두 영상의 길이보다 짧아야 한다. 각 영상에서 잘라낼 영상의 시간대는 랜덤으로 설정하였다. 길이를 맞춘 두 동영상은 프레임 단위로 나누어 합성하였으며 그 수식은 아래와 같다.

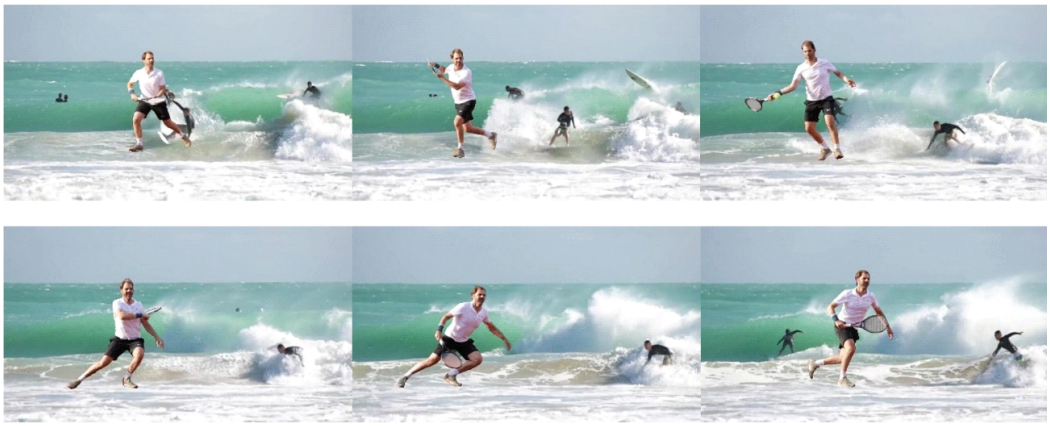

그림 1. Background augmentation으로 생성된 이미지 예시
Fig. 1. Examples of Image created by Background augmentation

$$X_i = M_i F_i + (1 - M_i) B_i,\ i = 1,2,...k$$

i는 1부터 생성될 영상의 최대 길이 k까지 각 프레임을 나타내는 번호이다. X_i는 생성된 동영상의 i번째 프레임에 해당하는 이미지이다. F_i는 객체가 담긴 동영상의 i번째 프레임 이미지이며 B_i는 배경으로 쓰일 영상의 i번째 프레임에 해당하는 이미지이다. M_i는 F_i에서 객체에 해당하는 영역을 1, 배경을 0으로 두는 이진 분할 지도이다. 이진 분할 지도는 영상 F Ground truth 라벨에서 추출하였다. 추적할 객체가 복수일 경우 객체 영역들의 합집합으로 이진 분할 지도를 구성하였다.

생성된 이미지의 Ground truth는 영상 생성된 이미지의 예시는 그림 1과 같다.

2. 영상 역재생

영상을 역 재생하는 데이터 증강 기법(Temporal Augmentation, T-Aug)를 구상하였다. VOS 모델은 시간순으로 정렬된 이미지들을 학습한다. 이때 정렬된 이미지들의 순서를 바꾸는 것이 영상 역재생 데이터 증강기법이다.

동영상 데이터에 역방향 단위 행렬을 곱하여 각 프레임에 해당하는 이미지의 순서를 뒤집는 함수 $f(X)$를 구현하였다. X는 동영상을 나타내는 벡터이며, I_k는 k번째 프레임에 해당하는 이미지를 의미한다.

$$X = \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_{k-1} \\ I_k \end{bmatrix}, \quad f(X) = \begin{bmatrix} 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & & 1 & 0 \\ \vdots & & 1 & & \vdots \\ 0 & 1 & \dots & 0 & 0 \\ 1 & 0 & & 0 & 0 \end{bmatrix} \quad X = \begin{bmatrix} I_k \\ I_{k-1} \\ \vdots \\ I_2 \\ I_1 \end{bmatrix}$$

위에서 설명한 데이터 증강 기법들이 데이터에 적용될 확률은 50%로 설정하였다.

Ⅳ. 실험

1. 실험 과정

제안한 2가지 데이터 증강 기법을 STCN 모델의 학습에 추가하여 실효성을 평가하였다. 기존 모델이 이미 사용중인 기본 데이터 증강 함수에 두 가지 기능을 추가하였다.

실험은 RTX 3090 2대로 진행되었다. 성능 비교를 위해 STCN와 동일한 학습 과정을 사용했으며, Static 데이터셋에 선행학습 된 모델에 YouTube VOS 데이터셋으로 학습시켰다. 각각 300000,150000 회 반복(iteration)하여 학습시키고 성능을 측정하였다. 한 실험을 4번 진행하여 그 평균을 성능으로 기록하였다. STCN은 실험실에서 재측정한 성능을 기록하였다.

2. 실험 결과

표 1. YouTube VOS 2018 validation dataset 성능 비교
Table 1. Performance Comparison on YouTube VOS 2018 validation sets

Method	Overall	J_{seen}	F_{seen}	J_{unseen}	F_{unseen}
STM	79.4	79.7	84.2	72.8	80.9
AFB-URR	79.6	78.8	83.1	74.1	82.6
MiVOS	80.4	80.0	84.6	74.8	82.4
CFBI	81.4	81.1	85.8	75.3	83.4
KMN	81.4	81.4	85.6	75.3	83.3
RMNet	81.5	82.1	85.7	75.7	82.4
LWL	81.5	80.4	84.9	76.4	84.4
CFBI+	82.0	81.2	86.0	76.2	84.6
LCM	82.0	82.2	86.7	75.7	83.4
STCN	82.4	81.4	85.5	77.3	85.3
STCN + Ours	83.1	81.3	85.7	78.6	86.9

위 테이블에서 데이터 증강 기법을 추가한 STCN 모델과 타 모델들을 비교하였으며 다음과 같은 측정기준(metric)으로 성능을 측정하였다. J는 영역 유사도(Region Similarity)이며 F는 윤곽 정확도 (Contour Accuracy)이다. seen과 unseen은 각각 학습 데이터에 있는/없는 카테고리를 의미한다. 위 지표들의 평균을 계산하여 overall로 표기하였다. 각 항목에서 가장 높은 성능을 기록한 수치를 볼드체로 표기하였다.

STCN 모델에 T-Aug와 B-Aug 데이터 증강 기법을 적용시킬 경우 모델의 성능이 향상되었다. 또한 학습 데이터에 없던 카테고리에서 성능이 주로 향상되었다. 이것은 모델

이 다양한 데이터 형태를 학습하면서 더 보편적인 특징을 학습한 결과라 유추할 수 있다.

3. Ablation Study

적용한 두 데이터 증강 기법이 각각 모델의 성능에 어떤 영향을 미치는지를 실험하였다. 위 테이블은 각 데이터 증강기법이 추가됨에 따라 성능이 어떻게 증가하는지 그 실험 결과를 보여주고 있다. 각 데이터 증강 기술은 성능을 향상시킬 수 있었으며 두 기법을 동시에 사용한 결과 보다 높은 성능을 기록했다.

B-Aug와 T-Aug 모두 모델의 성능을 향상시키는 데에 도움이 되었다. 두 데이터 증강기법 모두 학습 데이터에 없는 카테고리에 해당하는 데이터에서 높은 성능 향상을 기록했다. 그 이유는 다양하 데이터를 학습하면서 모델의 보편성(Robustness)가 증가했기 때문이다. 그러나 B-Aug는 T-Aug에 비에서 소폭 낮은 성능 향상을 기록했다. 그 이유는 생성되는 데이터의 무작위성이 더 크기 때문이다. 어떤 두 이미지를 합치는지에 따라 더 쉽거나 더 어려운 데이터를 생성하게 되고, 이 비율이 무작위이기 때문에 불안정성이 동반된다. 반면 T-Aug는 각 프레임의 이미지 정보를 훼손하지 않으므로 안정적으로 새로운 데이터를 생성할 수 있었으며 성능에서 더 높은 수치를 기록했다. 두 증강 기법을 모두 사용할 경우에도 성능이 오르는 결과를 보여주었는데, 그 이유는 B-Aug의 단점이 T-Aug의 장점으로 보완되었기 때문이라고 해석된다.

표 2. YouTube VOS 2018 validation dataset 성능 비교
Table 2. Performance Comparison on YouTube VOS 2018 validation sets

Method	Overall	J_{seen}	F_{seen}	J_{unseen}	F_{unseen}
W.O Aug	82.4	81.4	85.5	77.3	85.3
+ B-Aug	82.7	81.5	85.7	77.8	86.0
+ T-Aug	82.9	81.6	85.8	78.1	86.0
+ Both	83.1	81.3	85.7	78.6	86.9

4. 시각화

아래 이미지는 YouTube VOS Validation Dataset 중에서

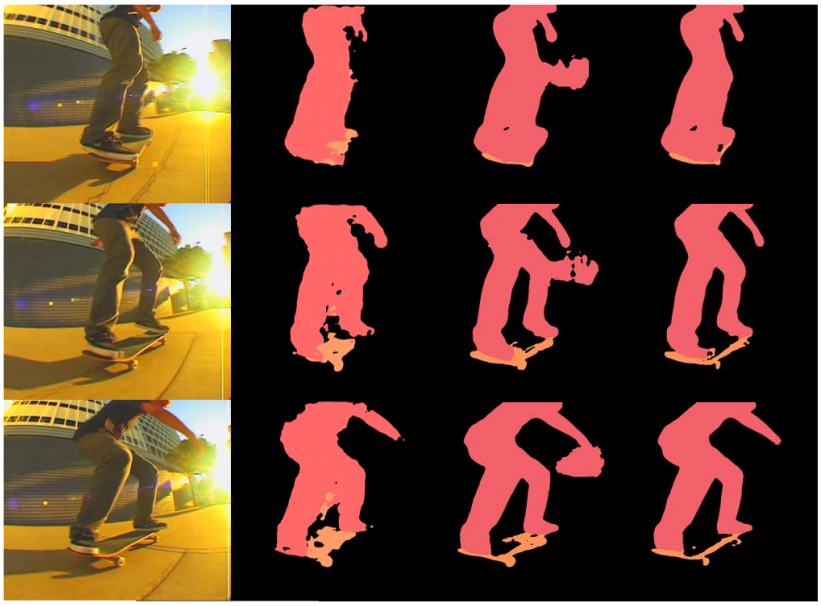

그림 2. 학습 데이터 증강 기법을 활용한 객체 분할 성능 개선 결과
Fig. 2. Result of Video Object Segmentation using Data Augmentation

VOS로 생성된 객체 분할 지도이다. STCN 모델은은 기존 모델인 CFBI에 비해 윤곽의 안정성이 크게 상승했다. 하지만 배경과 객체를 혼동하여 배경까지 객체로 인식하는 결함이 발견되었다. 위 예시에서 STCN 모델은 객체에 해당하는 바지와 배경에 해당하는 나무의 특징이 비슷하다는 이유로 배경의 나무까지 객체로 인식하였다. 그 이유는 모델이 객체의 정보를 잘 학습하였으나, 객체와 배경이 비슷한 특징을 공유하는 경우에는 이 둘을 혼동했기 때문이다. 반면, 데이터 증강 기법을 사용한 경우 이러한 문제가 개선되었다. 모델이 무작위 배경에서 객체를 분할하면서 배경과 객체의 차이점을 더 잘 학습했기 때문이다. 또한 신발과 같은 세세한 부분에서도 성능향상이 관찰되었다.

Ⅴ. Conclusion

본 연구에서 배경 증강 및 영상 역재생이라는 두 가지 동영상 데이터 증강 기법을 제시하였다. 두 기법은 준지도 VOS 과제의 성능을 효과적으로 높였으며 실험적으로 증명되었다. 하지만 해당 과제 외에도 비디오 데이터를 다루는 다른 과제에서 유의미한 결과를 얻을 수 있을지 검증할 필요가 있다. 제시된 기법이 다른 연구에서도 응용되어 도움이 되기를 희망한다.

참 고 문 헌 (References)

[1] Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., & Huang, T., Youtube-vos: A large-scale video object segmentation benchmark, arXiv preprint arXiv:1809.03327, 2018.

[2] Shorten, C., & Khoshgoftaar, T. M., A survey on image data augmentation for deep learning. Journal of Big Data, 6(1), 1-48, 2019.

[3] Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T. Y., Cubuk, E. D., ... & Zoph, B., Simple copy-paste is a strong data augmentation method for instance segmentation, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 2918-2928, 2021.

[4] Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y., Cutmix: Regularization strategy to train strong classifiers with localizable features, In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6023-6032), 2019.

[5] Caelles, Sergi, et al., "One-shot video object segmentation.", Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

[6] Xiao, Huaxin, et al, "Monet: Deep motion exploitation for video object segmentation.", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[7] Voigtlaender, Paul, and Bastian Leibe., "Online adaptation of convolutional neural networks for video object segmentation.", arXiv preprint arXiv:1706.09364, 2017.

[8] Perazzi, Federico, et al., "Learning video object segmentation from static images.", Proceedings of the IEEE conference on computer vision and pattern recognition, 2017.

[9] Luiten, J., Voigtlaender, P., & Leibe, B. Premvos: Proposal-generation, refinement and merging for video object segmentation. In Asian Conference on Computer Vision (pp. 565-580). Springer, Cham, December 2018.

[10] Oh, Seoung Wug, et al., "Video object segmentation using space-time memory networks." Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[11] Cheng, H. K., Tai, Y. W., & Tang, C. K., Rethinking Space-Time Networks with Improved Memory Coverage for Efficient Video Object Segmentation, arXiv preprint arXiv:2106.05210, 2021.

[12] Yun, S., Oh, S. J., Heo, B., Han, D., & Kim, J., VideoMix: Rethinking Data Augmentation for Video Classification, arXiv preprint arXiv: 2012.03457, 2020.

[13] MISRA, Ishan; ZITNICK, C. Lawrence; HEBERT, Martial., Shuffle and learn: unsupervised learning using temporal order verification. In: European Conference on Computer Vision. Springer, Cham, p. 527-544, 2016.

[14] Lai, Z., & Xie, W., Self-supervised learning for video correspondence flow, arXiv preprint arXiv:1905.00875, 2019.

저 자 소 개

김 호 진

- 2021년 2월 : 대구경북과학기술원 학사
- 2021년 3월 ~ 현재 : 대구경북과학기술원 정보통신융합전공 석사과정
- ORCID : https://0000-0002-1324-8641
- 주관심분야 : 딥러닝, 데이터증강

김 동 현

- 2021년 2월 : 대구경북과학기술원 학사
- 2021년 3월 ~ 현재 : 대구경북과학기술원 정보통신융합전공 석사과정
- ORCID : https://0000-0002-4634-3661
- 주관심분야 : 딥러닝, 비디오 객체 분할

김 정 훈

- 2019년 2월 : 대구경북과학기술원 학사
- 2019년 3월 ~ 현재 : 대구경북과학기술원 정보통신융합전공 석사 및 박사과정
- ORCID : https://0000-0002-7568-4115
- 주관심분야 : 딥러닝, 자가지도학습

임 성 훈

- 2018년 2월 ~ 2018년 8월 : Microsoft Research Asia(MSRA) 연구 인턴
- 2019년 6월 ~ 2019년 8월 : Carnegie Mellon University(CMU) 방문 연구
- 2019년 8월 : 한국과학기술원(KAIST) 박사
- 2019년 9월 ~ 현재 : 대구경북과학기술원 정보통신융합전공 조교수
- ORCID : https://0000-0001-9776-8101
- 주관심분야 : 컴퓨터비전, 머신러닝