VIDA: 행동 인식을 위한 시각 기반 심층 오디오 피처 생성

김지은, 윤서아, 장은성, 이형준
이화여자대학교 엘텍공과대학 컴퓨터공학과
{silver397, yoonseoah, heyoom}@ewhain.net, hyungjune.lee@ewha.ac.kr

VIDA: Vision-Informed Deep Audio Feature Generation for Action Recognition

Jieun Kim, Seoah Yoon, Eunsung Jang, Hyungjune Lee
Dept. of Compter Science and Engineering, Ewha Womans University

요 약

본 연구는 오디오 모달리티가 결손된 비디오 기반 행동 인식 환경에서, 시각 정보를 활용하여 의미적으로 정합한 오디오 피처를 생성하는 새로운 멀티모달 학습 프레임워크인 VIDA (Vision-Informed Deep Audio Feature Generation for Action Recognition)를 제안한다. 제안하는 프레임워크는 SVDMap 과 CAFMap 의 두 아키텍처로 구성되며, 모두 Transformer 기반 시각 피처 추출기와 LSTM 기반 오디오 생성기를 중심으로 의미 정합성을 학습 과정에 통합하여 멀티모달 표현의 신뢰성과 활용도를 높인다. Moments-in-Time 데이터셋을 기반으로 한 실험 결과, 실제 오디오가 없는 상황에서도 생성된 오디오 피처를 활용했을 때 SVDMap 은 약 35%, CAFMap 은 약 31%의 정확도를 기록하였다. 이는 기존 오디오 미포함 모델 대비 안정적인 성능 향상을 보인 것으로, MiT 데이터셋의 SOTA[8] 정확도인 53%에는 도달하지 못했으나, 오디오 결손 환경에서도 유의미한 행동 인식 성능을 확보할 수 있음을 시사한다

1. 서론

멀티모달 학습은 인간과 유사한 인지 능력(Artificial General Intelligence, AGI) 구현을 위한 핵심 기술로 주목받고 있으며, 특히 시각과 청각 정보를 통합적으로 활용하는 비디오 기반 행동 인식 분야에서 그 중요성이 부각되고 있다. 오디오는 시각 정보만으로는 인지하기 어려운 행동의 맥락을 제공함으로써, 시각 정보의 한계를 보완하는 중요한 역할을 한다. 그러나 실제 환경에서는 오디오가 누락되거나 품질이 저하되는 경우가 빈번하다. 이러한 모달리티 결손 문제는 멀티모달 모델의 성능과 일반화 능력을 저해한다.

이를 해결하기 위한 한 가지 접근은 시각 정보를 기반으로 오디오 피처를 생성하는 것이다. 일부 기존 연구들은 비디오의 시공간적 특징을 활용하여 오디오 피처를 재구성하는 방식을 시도하였으나, 이는 시각 정보와 오디오 정보 간의 의미적 정합성을 충분히 고려하지 못했다는 한계를 가진다[1]. 의미 정합성을 고려한 연구도 존재하지만, 주로 오디오 모달리티가 존재할 때 의미적 불일치를 제거하는 오디오 필터링 (Dropout) 방식에 머물렀고, 오디오가 완전히 결손된 환경은 여전히 보완하지 못했다[2].

따라서 본 논문에서는 오디오 모달리티가 누락된 상황에서 비디오와 의미적으로 정합한 오디오 피처를 생성하여, 멀티모달 행동 인식의 신뢰성과 성능을 동시에 향상시키고자 한다. 이를 위해 우리는 두 가지 구조를 제안한다:

1.Semantic Validation Dictionary Mapping (SVD Map): 사전 구축된 시맨틱 딕셔너리를 활용해 LSTM 기반 생성 오디오 피처의 정합성을 사후 검증하는 방식이다.

2. Caption to Audio Feature Mapping (CAFMap): 대표 프레임에서 생성한 시각 캡션을 통해 오디오 라벨을 예측하고, 이를 조건으로 오디오 피처를 생성하는 방식이다.

두 구조는 모두 Transformer 기반 피처 추출기와 LSTM 생성기를 중심으로 구성되며, 의미 정합성 확보를 핵심 원칙으로 삼는다. 실험 결과, 오디오가 없는 상황에서도 행동 인식 정확도를 향상시켰으며, 특히 시맨틱 검증 및 캡션 기반 조건 주입이 오디오 피처 생성의 품질을 실질적으로 개선함을 확인하였다.

본 논문의 주요 기여는 다음과 같다:

(1) 오디오 결손 상황에서 의미 기반 오디오 피처를 생성함으로써 행동 인식 정확도를 향상시키는 시맨틱 매핑 구조를 제안하였다.

(2) 두 가지 상이한 시맨틱 정렬 전략(SVDMap, CAF Map)을 설계 및 비교함으로써 구조적 특성과 응용 가능성을 제시하였다.

(3) 실제 오디오 없이도 고정밀 행동 인식이 가능한 멀티모달 프레임워크를 제시하고, 다양한 확장 가능성을 확인하였다.

2. 관련 연구

2.1 비디오 기반 오디오 생성

시각 정보를 바탕으로 오디오 피처를 생성하려는 시도는 최근 멀티모달 학습의 확장과 함께 활발히 이루어지고 있다. 특히, 영상에서 발생하는 객체의 움직임 이나 장면 전환을 기반으로 환경음을 예측하거나 오디오 스펙트로그램을 복원하는 연구들이 주를 이룬다[3]. 이러한 방식은 주로 시공간적 영상 특징을 기반으로 하며, CNN 또는 Transformer 구조를 이용해 정적인 비디오 표현을 오디오 시퀀스로 변환한다. 하지만 대부분의 접근 은 의미적 정합성 보다는 구조적 유사성에 초점을 두어, 실제 인식이나 다운스트림 태스크에서 활용하기 어려운 경우가 많다.

2.2 멀티모달 의미 정합성 검증

멀티모달 학습에서 모달리티 간 의미 정합성을 고려하려는 시도도 존재한다[4, 5]. 예를 들어, 오디오와 영상 간 의미적으로 어긋난 데이터 샘플을 식별해 제거하거나, 의미 정합성이 낮은 샘플을 학습에서 필터링하는 방식이 제안된다[4]. 이 연구는 정합성 판단 을 통해 학습 데이터의 품질을 높이는 데 기여하지만, 근본적으로 오디오가 존재하지 않는 상황을 보완하는 데에는 한계가 있다. 또한, filtering 기준으로 사용되는 semantic alignment 방식이 대부분 사후적이거나 정적임에 따라, 동적인 시퀀스 생성 과정과의 통합이 어렵다는 단점이 존재한다.

2.3 시맨틱 매핑과 자연어 조건 생성

최근에는 시각적 요소와 자연어 사이의 의미적 연결을 통해 멀티모달 모델의 표현력을 향상시키려는 연구가 주목받고 있다. CLIP, BLIP 등 사전학습된 비전-언어 모델을 활용하여 이미지에서 캡션을 생성하거나, 의미적 유사도 기반의 텍스트-오디오 간 매핑을 시도한 사례들이 대표적이다[6, 7]. 이러한 접근은 정적인 멀티 모달 태스크에서는 의미 정합성을 크게 향상시켰지만, 시계열 데이터에서의 조건 생성 및 활용 측면에서는 여전히 연구가 미흡하다.

본 논문은 위의 연구들을 바탕으로, 오디오가 결손된 상황에서도 의미적으로 정합한 오디오 피처를 생성하고 이를 행동 인식에 직접 활용할 수 있는 새로운 구조를 제안한다. 기존 연구들과 달리, 제안하는 두 아키텍처 (SVDMap, CAFMap)는 각각 (1) 시맨틱 정합 성을 통한 학습 샘플 필터링과 (2) 시각 기반 의미 조건 생성을 통해 의미적 연결성을 강화하며, 실제 행동 인식 정확도 개선을 실험적으로 입증한다는 점에서 차별성을 갖는다.

3. 구조

본 연구는 시청각 멀티모달 환경에서 오디오 모달리티의 부재 문제를 해결하기 위해, 시각 정보로부터 의미적으로 정합한 오디오 피처를 생성하는 두 가지 아키텍처를 제안한다. 첫 번째 아키텍처는 Semantic Validation Mapping 기반으로, 비디오 라벨과 생성된 오디오 라벨 간의 의미적 유사도를 활용하여 신뢰도 높은 학습 샘플을 선별한다. 두 번째 아키텍처는 Caption-to-Audio Feature Mapping 구조로, 대표 시각 프레임에서 생성된 캡션을 오디오 라벨에 매핑한 후, 해당 피처를 조건으로 하여 시계열 오디오 피처를 생성한다. 두 방식 모두 생성된 오디오 피처를 행동 인식 모델에 입력하여, 최종 분류 성능 향상을 통해 그 실효성을 평가한다.

3.1 Semantic Validation Mapping (SVDMap)

Semantic Validation Mapping (SVDMap)은 LSTM 기반으로 생성된 오디오 피처가 의미적으로 정합할 경우에만 행동 분류 학습에 활용함으로써 멀티모달 행동 인식의 성능을 향상시키는 것을 목표로 한다. 이를 위해 전체 파이프라인은 (1) 오디오 피처 생성기, (2) 의미 기반 라벨 예측기, (3) 정합성 필터링 엔진의 세 모듈로 구성되며, Semantic Dictionary 를 중심으로 오디오-비디오 라벨 간 의미 유사도를 정량화하고 이를 기반으로 학습 샘플을 선별한다.

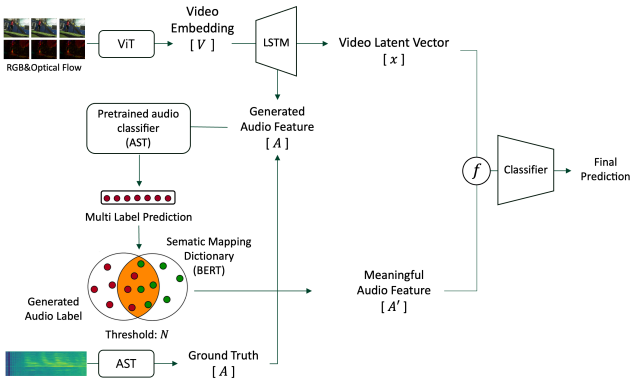

Figure 1. Semantic Validation Mapping (SVDMap)

(1) LSTM based Audio Feature Generator

입력 비디오로부터 추출한 RGB 및 Optical Flow 피처를 시간 순서대로 정렬한 시퀀스를 LSTM 에 입력하여 오디오 피처 시퀀스를 생성한다. 이때, 입력

시퀀스는 RGB 와 Flow 임베딩을 시점별로 결합하여 구성하며, 총 입력 형태는 다음과 같다:

$$X = \{x_t \in \mathbb{R}^{1536} \mid t = 1, \cdots, T\}$$

이러한 오디오 피처는 후속 라벨 예측 및 정합성 평가에 사용된다.

(2) Multi-Label Predictor

생성된 오디오 피처 $\hat{A}$는 log-Mel spectrogram 형태로 재구성된 후, AudioSet 으로 사전학습된 Audio Spectrogram Transformer(AST) 모델에 입력되어 multi-label 오디오 라벨 분류를 수행한다. AST 는 전체 시퀀스를 통합한 embedding 으로부터 해당 오디오가 포함하고 있는 라벨들의 확률 분포 $P = \{p_i\}$를 출력하며, top-k 예측 라벨을 선택하여 의미 정합성 평가에 활용한다.

(3) Semantic Validation Dictionary

Semantic Dictionary 는 비디오 데이터셋(MiT)의 액션 클래스와 AudioSet 의 오디오 라벨 간의 의미적 유사도를 BERT 임베딩을 통해 사전 구축한 딕셔너리이다. 각 비디오 클래스 $c \in C_{video}$ 에 대해 BERT 를 활용해 해당 클래스명을 문장 임베딩한 뒤, Audi Set 의 모든 라벨 임베딩과 cosine similarity 를 계산하여 top-N 유사 오디오 라벨을 선택한다:

$$sim(c, l_i) = \cos\left(BERT(c), BERT(l_i)\right)$$

이때 선택된 top-N 오디오 라벨 $L_c^{GT} = \{l_1, \cdots, l_N\}$은 해당 액션 클래스에 의미적으로 매핑된 오디오 라벨의 집합으로 간주되며, filtering 기준으로 사용된다.

실험적으로는 top-5 라벨 중 평균 3.7 개가 실제 GT 오디오 라벨과 일치하는 것으로 나타나, Semantic Dictionary 의 정밀도가 확인되었다.

(4) IoU 기반 Feature Filter

의미 정합성 판단을 위해, AST 로부터 예측된 오디오 라벨 $\hat{L}$과 Semantic Dictionary 로부터 참조된 GT 라벨 L_c^{GT} 간의 Intersection over Union (IoU)을 계산한다:

$$IOU\left(\hat{L}, L_c^{GT}\right) = \frac{|\hat{L} \cap L_c^{GT}|}{|\hat{L} \cup L_c^{GT}|}$$

3.2 Caption-to-Audio Feature Mapping (CAFMap)

CAFMap 은 비디오 시각 정보로부터 의미적으로 정합한 오디오 피처를 생성하기 위한 프레임 중심의 조건 입력 기반 오디오 생성 구조이다. 전체 파이프라인은 (1) 대표 프레임 추출기, (2) 캡션 생성기, (3) 캡션-오디오 라벨 매핑기, (4) 조건 기반 오디오 피처 생성기의 네 모듈로 구성된다. 핵심 아이디어는 의미 있는 시각 장면이 발생한 프레임에 대응되는 오디오적 표현을 외부 데이터셋으로부터 참조하고 이를 생성 과정에 조건으로 삽입함으로써, 정합성과 시간적 일관성이 보장된 오디오 피처 시퀀스를 생성하는 것이다.

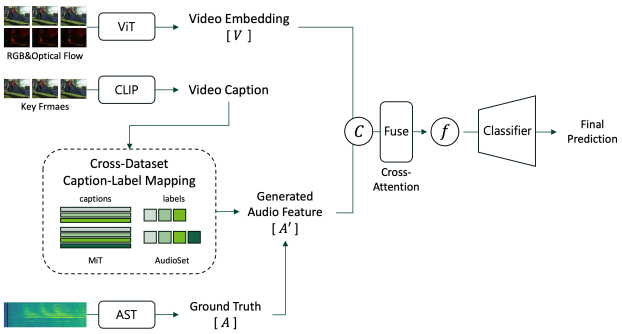

Figure 2. Caption-to-Audio Feature Mapping (CAFMap)

(1) Representative Frame Extractor

비디오 내 주요 이벤트가 발생하는 시점을 효과적으로 포착하기 위해, 모든 비디오는 초당 6 프레임의 고정 비율로 샘플링된 18 개의 RGB 프레임 시퀀스로 전처리된다. 이후 인접한 프레임 간 Optical Flow 기반의 픽셀 이동량 m_i를 계산하고, 전체 n개의 프레임에 대해 평균 이동량 $\bar{m}$ 을 산출한다. 이동량이 $t \cdot \bar{m}$ 보다 큰 프레임을 대표 프레임으로 선택하며, 본 연구에서는 사전 실험을 통해 $t = 1.2$ 가 가장 적절한 임계값으로 도출되었다:

$$\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$$

이러한 대표 프레임 인덱스는 이후 생성기 입력 시, 조건 피처가 삽입될 시점으로 고정되어 활용된다.

(2) Caption Generator

선정된 대표 프레임 이미지에 대해 자연어 기술을 생성하기 위해, 사전학습된 BLIP 기반 비전-언어 모델 [6] 을 활용한다. 본 연구에서는 추가적인 파인튜닝 없이, LAION-400M 등 대규모 이미지-텍스트 페어 데이터로 사전학습된 모델을 직접 사용하였다. 각 프레임에 대해 생성된 캡션은 해당 프레임 인덱스와 함께 JSON 포맷으로 저장되며, 이후 조건 생성 과정에서 참조된다.

(3) Caption-to-Label Mapper

생성된 캡션은 BERT 기반 문장 임베딩 모델을 통해 의미 공간 상에서 AudioSet 라벨들과 비교된다. AudioSet 의 각 라벨도 사전 계산된 BERT 임베딩을 보유하고 있으며, 각 캡션 문장과의 cosine similarity 를

계산하여 상위 k개의 라벨을 정렬한다. 단순히 Top-k 를 선택하는 대신, Top-5 의 유사도 평균을 기준으로 해당 값보다 높은 유사도를 가지는 라벨만을 최종 매핑 라벨로 선택함으로써, 의미적으로 정합한 오디오 표현만을 조건 입력으로 사용할 수 있도록 한다.

$$\sin(caption, label) = \cos\left(BERT(caption), BERT(label)\right)$$

선택된 라벨 각각은 AudioSet 에서 사전 수집된 log-Mel 기반 평균 오디오 피처를 보유하고 있으며, 이들 피처의 평균값을 대표 프레임 조건 입력으로 사용한다.

(4) Conditional Audio Feature Generator

LSTM 기반의 오디오 생성기는 18 개의 시점으로 구성된 시퀀스를 입력으로 받아, 시각적으로 정렬된 오디오 피처 시퀀스를 출력한다. 입력 시퀀스는 기본적으로 각 시점마다 $\mathbb{R}^{128}$ 의 random Gaussian noise 로 구성되며, 단 대표 프레임에 해당하는 시점에는 이전 단계에서 생성된 조건 오디오 피처가 삽입된다. 즉, 전체 입력은 다음과 같은 구조로 구성된다:

$$X_t = \begin{cases} condition\ feature & if\ t \in Representative\ Frames \\ random\ noise & otherwise \end{cases}$$

LSTM 은 시퀀스 전체를 입력받아 오디오 피처 시퀀스 $\hat{A} \in \mathbb{R}^{18\times128}$를 출력하며, 학습 단계에서는 정답 오디오 피처 A와의 평균제곱오차(MSE)를 손실 함수로 사용한다:

$$\mathcal{L}MSE = \frac{1}{T}\sum t = 1^{T}\left\|\hat{A}_t - A_t\right\|_2^2$$

이러한 구조를 통해 CAFMap 은 시각적으로 유의미한 시점에 의미적으로 정합한 오디오 조건을 삽입함으로써, 시간적으로 자연스럽고 행동 인식에 유의미한 오디오 피처 시퀀스를 생성할 수 있도록 설계되었다.

4. Experiments

4.1 SVDMap

SVDMap 구조의 실험은 다음 세 가지 조건에서 수행되었다. 모든 실험은 동일한 구조의 LSTM 기반 generator 와 classifier 를 사용하며, 입력만 다르게 구성되었다.

Baseline: 오디오 없이 RGB+Flow 피처만을 입력으로 분류기 학습
Filtered Audio: 생성된 오디오 중 의미 정합성 필터링을 통과한 샘플만 사용
Ground Truth Audio: 실제 오디오 피처를 추가 입력으로 활용 (upper bound)

(1) 학습 안정성과 수렴

Validation Loss 는 epoch 1 기준 0.1646 에서 epoch 10 기준 0.1605 까지 감소하며, 전체적으로 완만한 수렴 추세를 보였다. 특히 epoch 6 이후부터는 loss 값이 0.1605~0.1615 사이에서 안정적으로 유지되며, 학습이 빠르게 안정화되었음을 보여준다.

(2) 과적합 방지 및 일반화 성능

Train 과 Validation Loss 간의 차이는 평균 약 0.0038 로 매우 작게 유지되었다. 모든 epoch 에서 validation loss 가 train loss 보다 약간 높게 형성되었으며, 이는 과적합 없이 정상적인 일반화 성능을 보여주는 지표다. 후반부 epoch 에서도 두 값 간의 차이가 일정하여 모델의 일반화 성능이 안정적으로 유지되었음을 확인할 수 있다.

(3) 행동 분류 성능 향상

분류기 실험에서는 오디오 없이 학습한 baseline 대비, 의미 정합성이 확보된 오디오 피처만을 사용하는 경우 정확도가 평균 +14%p, 최대 +17%p 향상됨을 확인하였다. 필터링을 적용하지 않고 전체 생성 피처를 사용할 경우보다도 성능이 높았으며, 이는 noise sup-pression 효과뿐 아니라 의미 기반 선택 학습이 유효했음을 시사한다.

	Accuracy
Baseline	17.7%
Filtered Audio	35.6%
Ground Truth Audio	34.3%

Table 1. SVDMap 오디오 입력 조건별 행동 분류 정확도

또한 epoch 이 진행됨에 따라 성능이 점진적으로 수렴하며 향상 폭이 +5%p 를 초과하는 구간이 다수 존재하여, semantic filtering 기반 학습이 행동 인식 성능 향상에 실질적으로 기여했음을 입증하였다.

4.2 CAFMap

CAFMap 구조의 실험은 다음 세 가지 조건에서 수행되었다. 모든 실험은 SVDMap 과 동일하게 LSTM 기반의 generator 와 classifier 를 사용하며, 입력만 다르게 구성 되었다.

Baseline: 오디오 없이 RGB+Flow 피처만을 입력으로 분류기 학습
Filtered Audio: 생성된 오디오 피처를 입력으로 추가

Ground Truth Audio: 실제 오디오 피처를 추가 입력으로 활용 (upper bound)

(1) 학습 안정성과 수렴

Validation Loss 는 초기 epoch 에서 0.1635 로 시작하여, epoch 9 기준 0.1600 까지 점진적으로 감소하였다. 전체적으로 일정한 수렴 추세를 보이며, 후반 epoch 에서는 loss 값이 0.1602–0.1600 사이에서 안정적으로 유지되었다. 이는 모델이 빠르게 안정화되며 학습이 수렴 했음을 나타낸다.

(2) 과적합 방지 및 일반화 성능

Train 과 Validation Loss 간 차이는 평균 약 0.0039 로 작게 유지되었다. 전 구간에서 Validation Loss 가 Train Loss 보다 높게 형성되었으며, 과적합 없이 일반화 성능이 안정적으로 유지됨을 확인할 수 있다. 특히 에포크 후반까지 train 과 validation 간의 손실 차이가 일정하게 유지된 점은 모델의 일반화 관점에서 긍정적이다.

(3) 행동 분류 성능 향상

CAFMap 에서 생성된 오디오 피처를 활용한 분류 정확도는 평균 약 30.7%로, 오디오가 없는 경우의 baseline 정확도 22.5% 대비 약 +8.2%p 향상되었다. 최고 성능은 epoch 6 에서 31.3%로 관측되었으며, 이는 ground truth 오디오를 활용한 upper bound 성능 (33.9%)에 근접하는 수치이다.

	Accuracy
Baseline	22.5%
Generated Audio	30.7%
Ground Truth Audio	33.9%

Table 2. CAFMap 오디오 입력 조건별 행동 분류 정확도

또한 성능 향상폭은 대부분의 epoch 에서 +5%p 를 초과하였으며, 이는 의미 정합 기반 조건 입력이 생성 오디오 품질과 분류 기여도를 동시에 개선했음을 시사한다. 비록 GT 수준에는 미치지 못하지만, 안정적인 성능 향상과 모델 일반화가 동시에 달성되었다는 점에서 CAFMap 아키텍처의 실질적 유효성을 입증하였다.

5. 결론

본 연구는 오디오 모달리티가 결손된 멀티모달 환경에서 행동 인식 정확도를 회복하기 위한 방안으로, 의미 기반 오디오 피처 생성 아키텍처 두 가지(SVDMap, CAFMap)를 제안하였다. SVDMap 은 비디오 라벨과 오디오 라벨 간 의미 정합성을 평가하여 신뢰도 높은 생성 피처만을 선별 학습에 활용하였고, CAFMap 은 시각적 대표 프레임에서 생성한 캡션을 기반으로 의미적으로 연관된 오디오 라벨 피처를 조건 입력으로 사용하여 시계열 오디오를 생성하였다. 두 아키텍처 모두 오디오가 없는 경우보다 향상된 분류 정확도를 기록하였으며, 특히 SVDMap 은 의미 기반 filtering 을 통해 실제 오디오를 사용하는 upper bound 수준의 성능에 도달하는 등 생성 피처의 유효성을 입증하였다.

한편, 본 연구는 caption-to-label 매핑의 한계, 시간적 맥락 반영 부족, 특정 클래스 성능 편중 등의 과제를 드러냈으며, 향후에는 BERT 기반의 soft matching, context-aware 주입 방식, human sound penalty 조정 등을 통해 의미 정합성과 클래스 다양성 확보를 도모할 예정이다. 또한, 하드웨어 제약으로 인해 전체 클래스와 원본 데이터를 모두 활용하지 못한 점은 한계로 작용하였으며, 향후 실험 규모 및 데이터셋 확장을 통해 제안한 아키텍처의 일반화 성능과 실용성을 보다 명확히 검증할 계획이다.

참고 문헌

[1] Lee, H.-C. et al. (2019). Audio Feature Generation for Missing Modality Problem in Video Action Recognition. Proc. IEEE ICASSP, 3550 – 3558.

[2] Alfasly, M. et al. (2022). Learnable Irrelevant Modality Dropout for Multimodal Action Recognition on Modality-Specific Features. Proc. CVPR, 19049 – 19059.

[3] Zhou, Y. et al. (2018). Visual to Sound: Generating Natural Sound for Videos in the Wild. Proc. CVPR, 3550 – 3558.

[4] Gao, R. et al. (2018). Learning to Separate Object Sounds by Watching Unlabeled Video. Proc. ECCV, 35 – 53.

[5] Miech, A. et al. (2020). End-to-End Learning of Visual Representations from Uncurated Instructional Videos. Proc. CVPR, 9879 – 9889.

[6] Li, J. et al. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. Proc. ICML, 162:12888 – 12900.

[7] Wu, Y. et al. (2022). Wav2CLIP: Learning Robust Audio Representations from CLIP. Proc. CVPR, 19166 – 19176.

[8] Srivastava, S. et al. (2024). OmniVec2: A Novel Transformer-Based Network for Large-Scale Multimodal and Multitask Learning. Proc. CVPR, 26354 – 26364.