Audio Focused Multimodal Representation Learning

Guided by
Prof. Preethi Jyothi
Prof. Ganesh Ramakrishnan

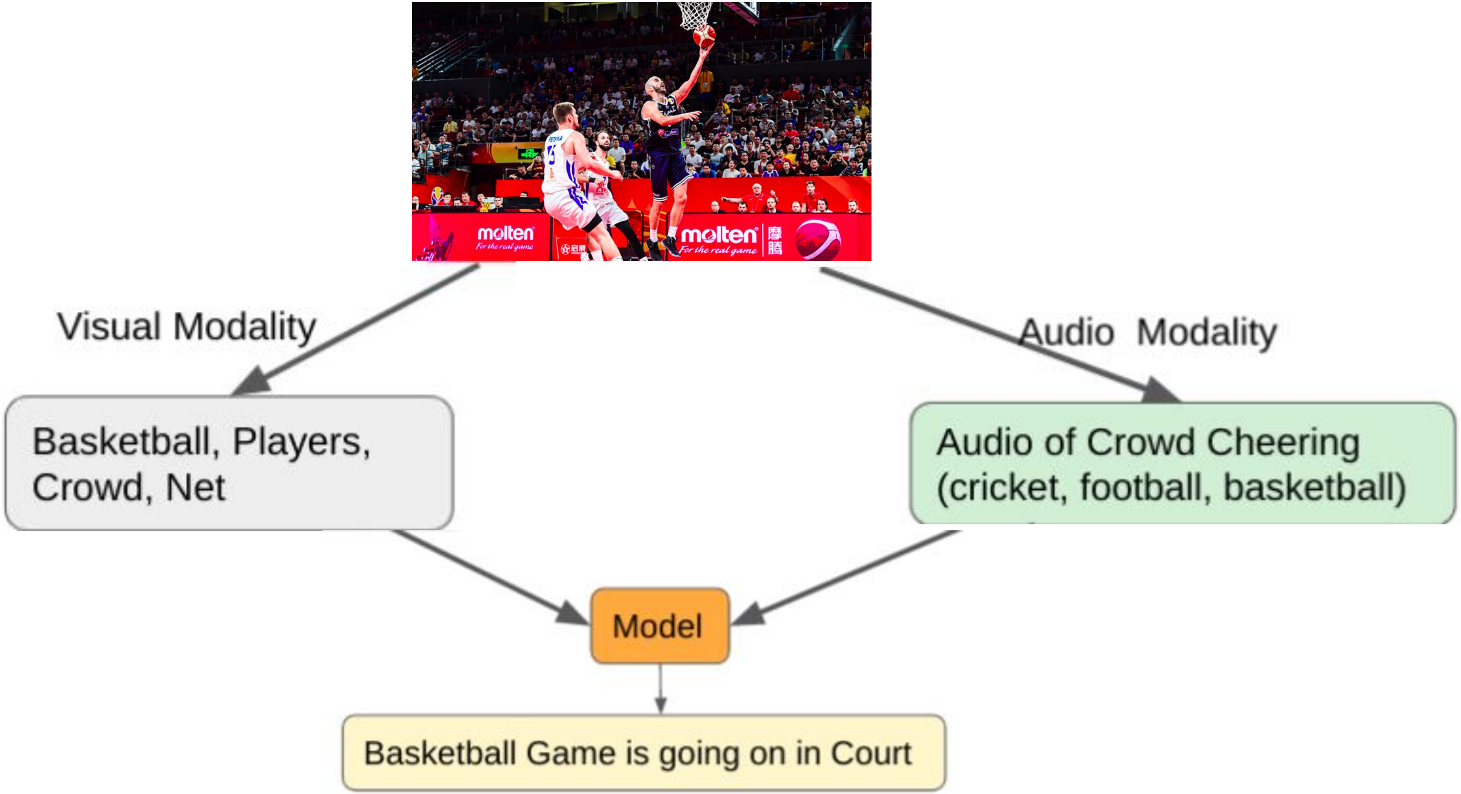
Presented by
Shubham Nemani
203050011
MTech -1 CSE

Content

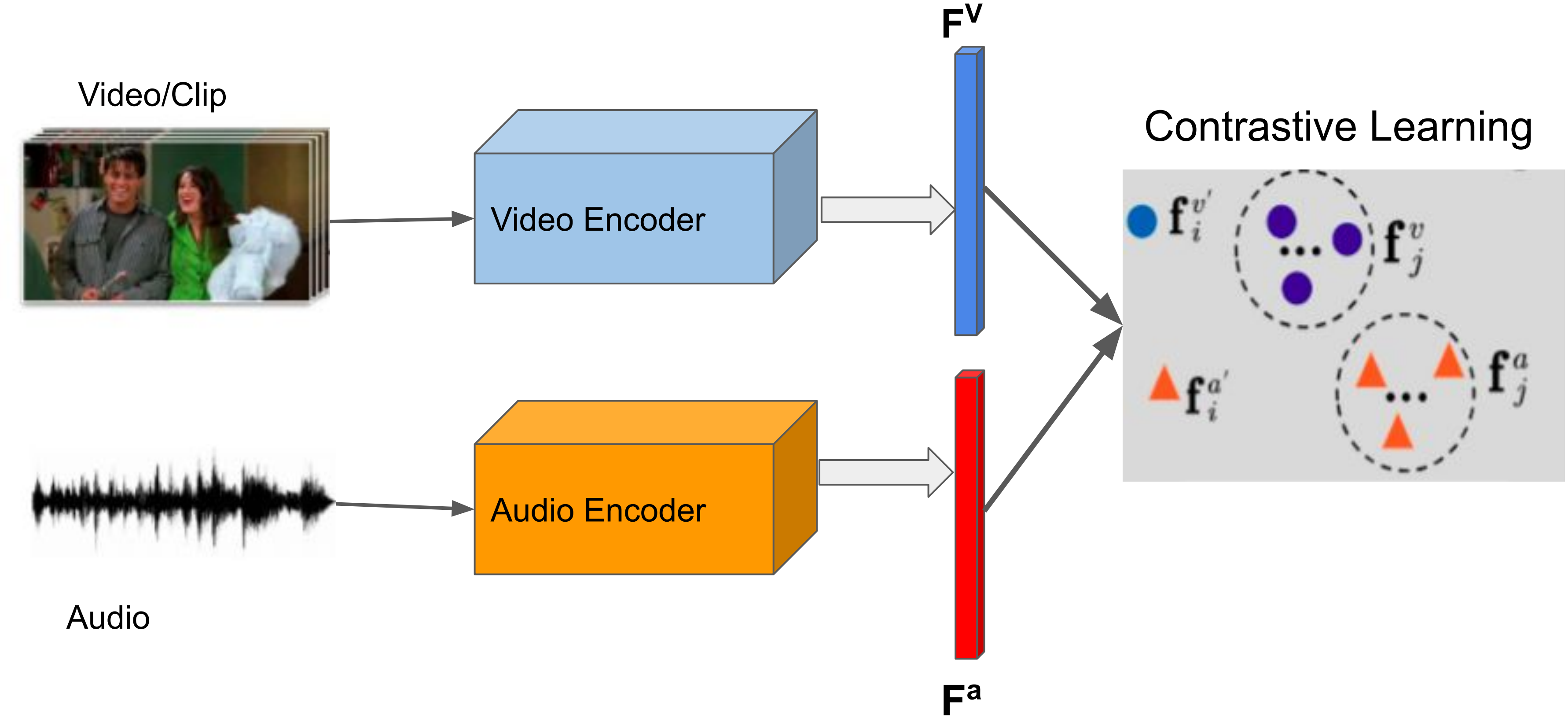
- Motivation
- Basic Architecture
- Knowledge distillation
- Listen to Look
- Contrastive learning
- Curriculum Learning
- Compositional Contrastive Learning
- Conclusion

Motivation

Importance of modalities



Self-supervised Multimodal Representation Learning Architecture

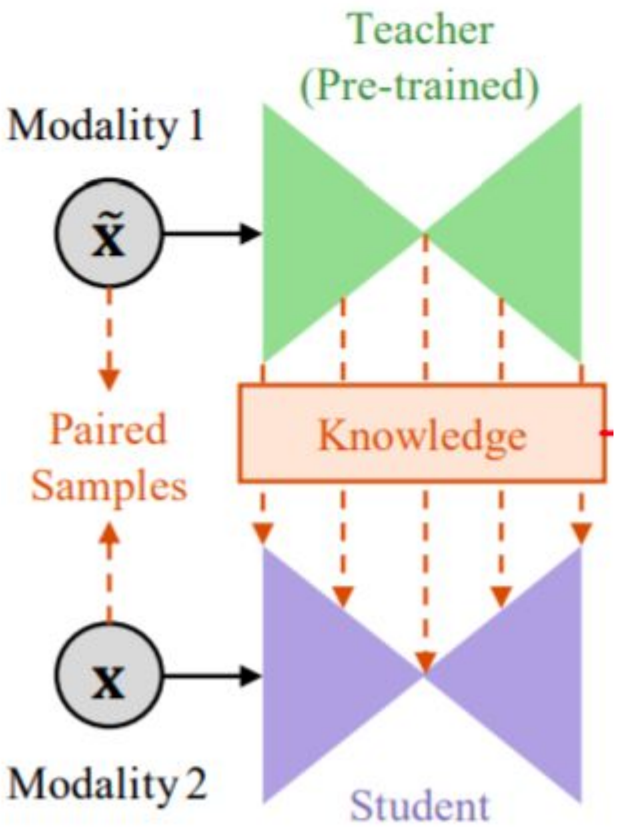


Knowledge Distillation

Cross-modal knowledge distillation deals with transferring knowledge -

from a model pre-trained with superior modality **(Teacher)**

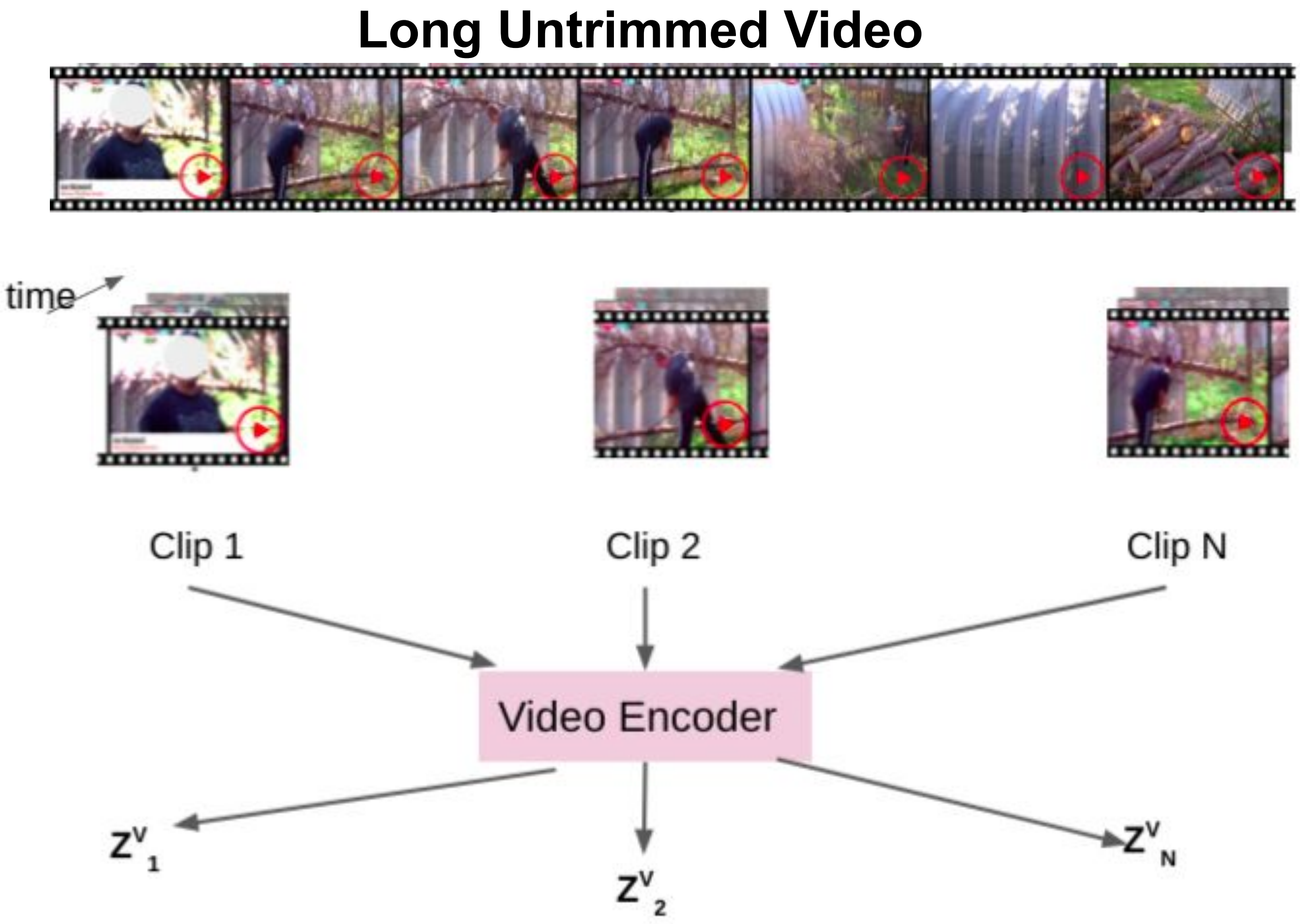
to another model training with weak modality **(Student)**



Listen to Look

Action Recognition by Previewing Audio

Kristen Grauman , Ruohan Gao (CVPR 2020)

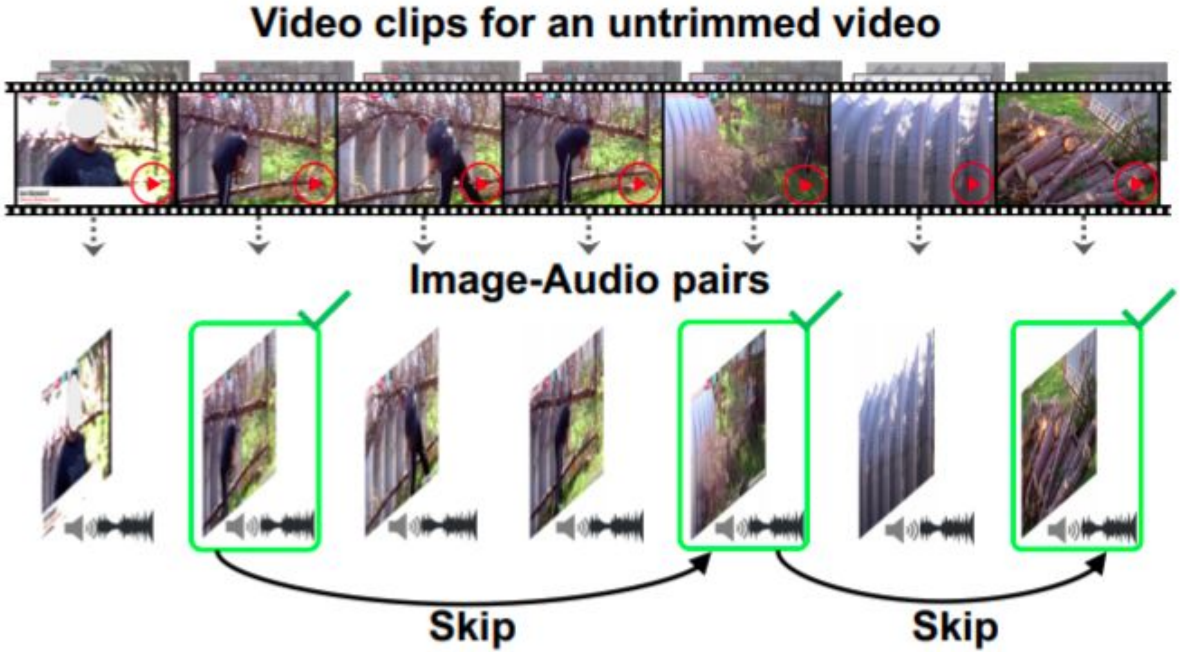


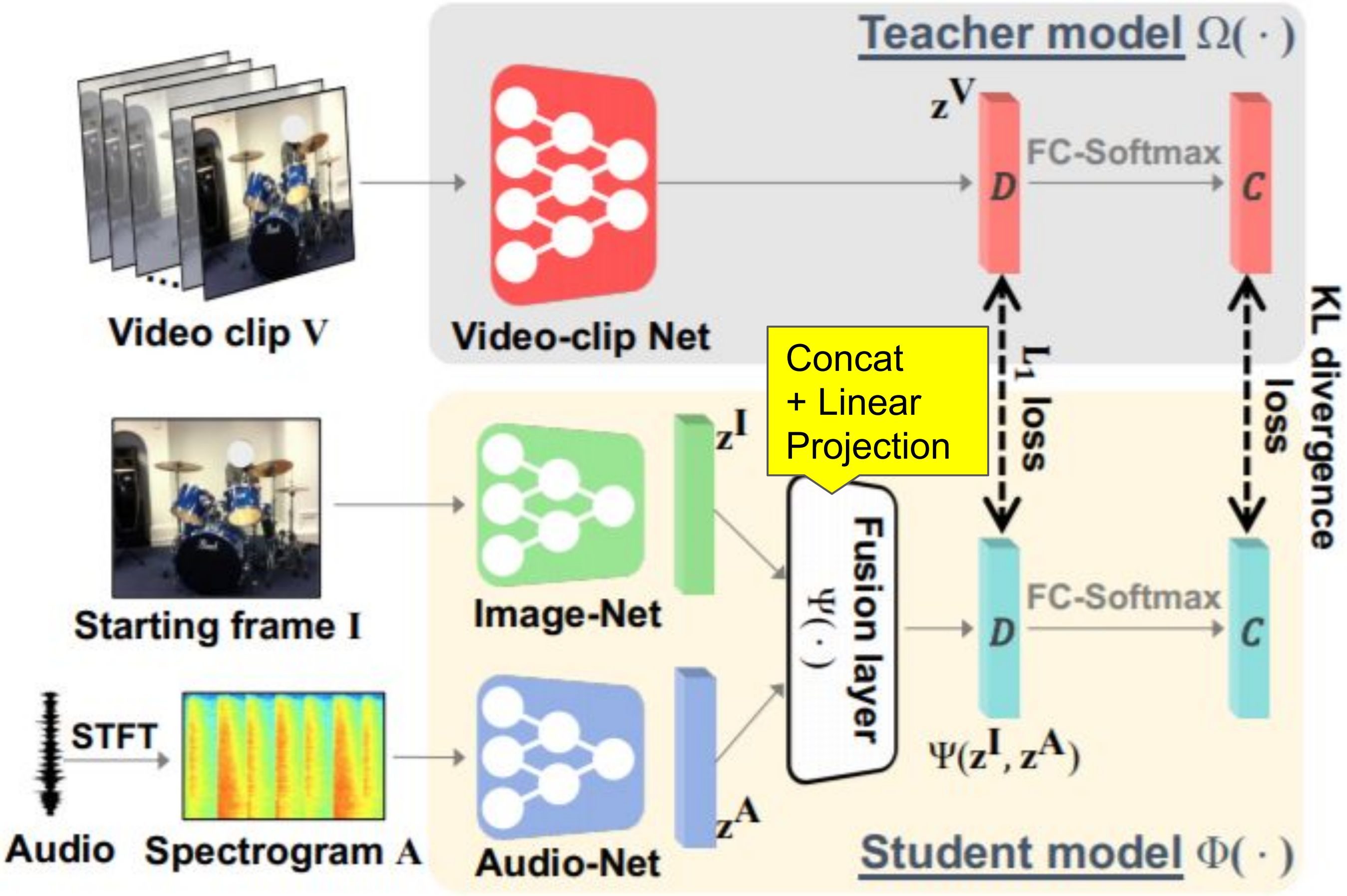
Long Untrimmed Video
time
Clip 1
Clip 2
Clip N
Video Encoder
Z^V_1
Z^V_2
Z^V_N

Redundancy in Video/Clip

1) **Clip-level** - Within each short clip , temporally close frames are visually similar,

2) **Video-level** - across all the clips in V, often only a few clips contain key moments.





Teacher model $\Omega(\cdot)$
Video clip V
Video-clip Net
z^V
FC-Softmax
D
C
L_1 loss
KL divergence loss
Concat
+ Linear
Projection
Starting frame I
Image-Net
z^I
Fusion layer
$\Psi(\cdot)$
FC-Softmax
D
C
$\Psi(z^I, z^A)$
STFT
Audio
Spectrogram A
Audio-Net
z^A
Student model $\Phi(\cdot)$

Listen to Look Results

Method	Backbone	ActivityNet	UCF-101
ListenToLook	ResNet-152	35.5	73.5
ListenToLook	R(2+1)D-152	47.0	82.5

ActivityNet - 200 Classes

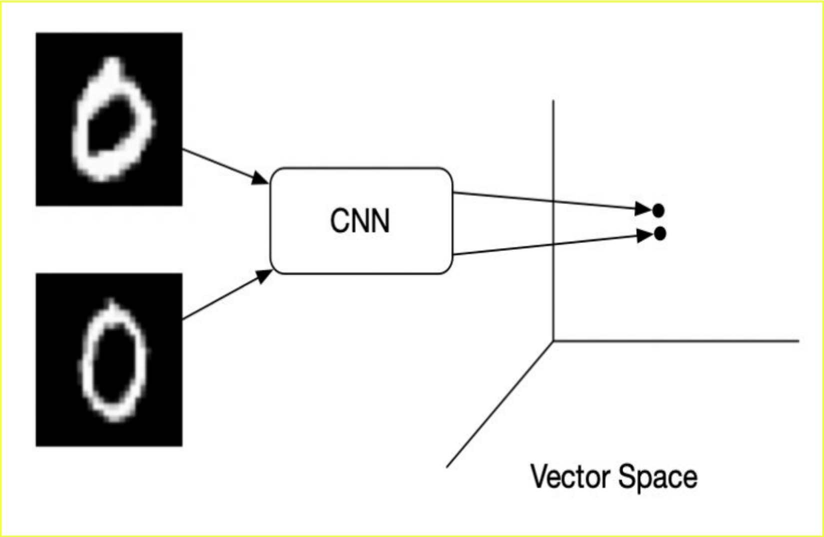
MiniSports1M - 437 Classes

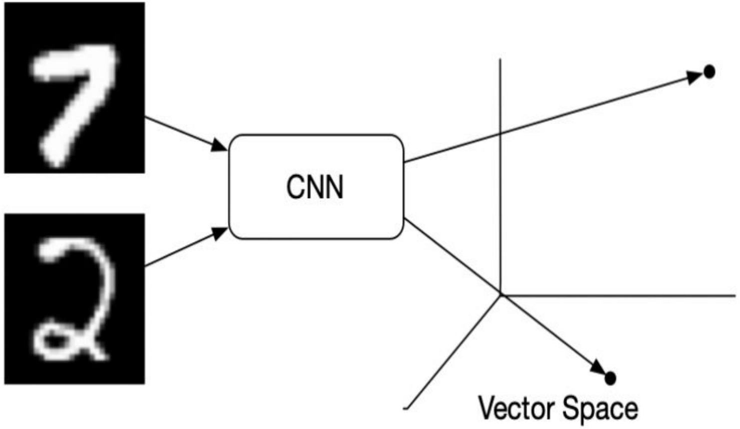
Contrastive Learning

- **End to End Contrastive Learning**
- **Memory Bank Approach**
- **Unimodal NCE**
- **Multimodal NCE**

Contrastive Learning

Contrastive learning aims to group similar samples closer and diverse samples far from each other.



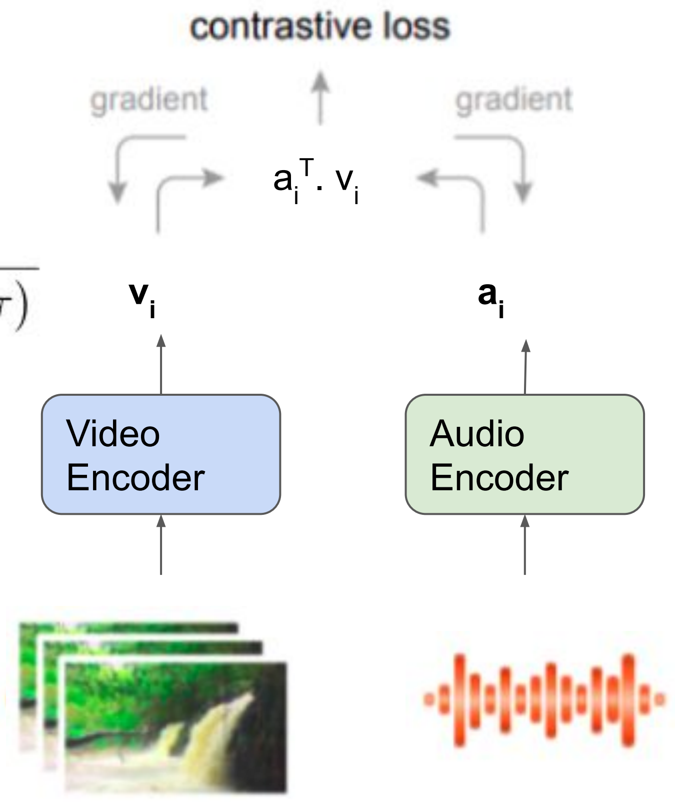


End to End Contrastive Learning -

Positive

$$\mathcal{L}_q = -\log \frac{\exp\left(a_i^T . v_i / \tau\right)}{\sum_{j=1}^{K} \exp\left(a_i^T . v_j / \tau\right)} - \log \frac{\exp\left(v_i^T . a_i / \tau\right)}{\sum_{j=1}^{K} \exp\left(v_i^T . a_j / \tau\right)}$$

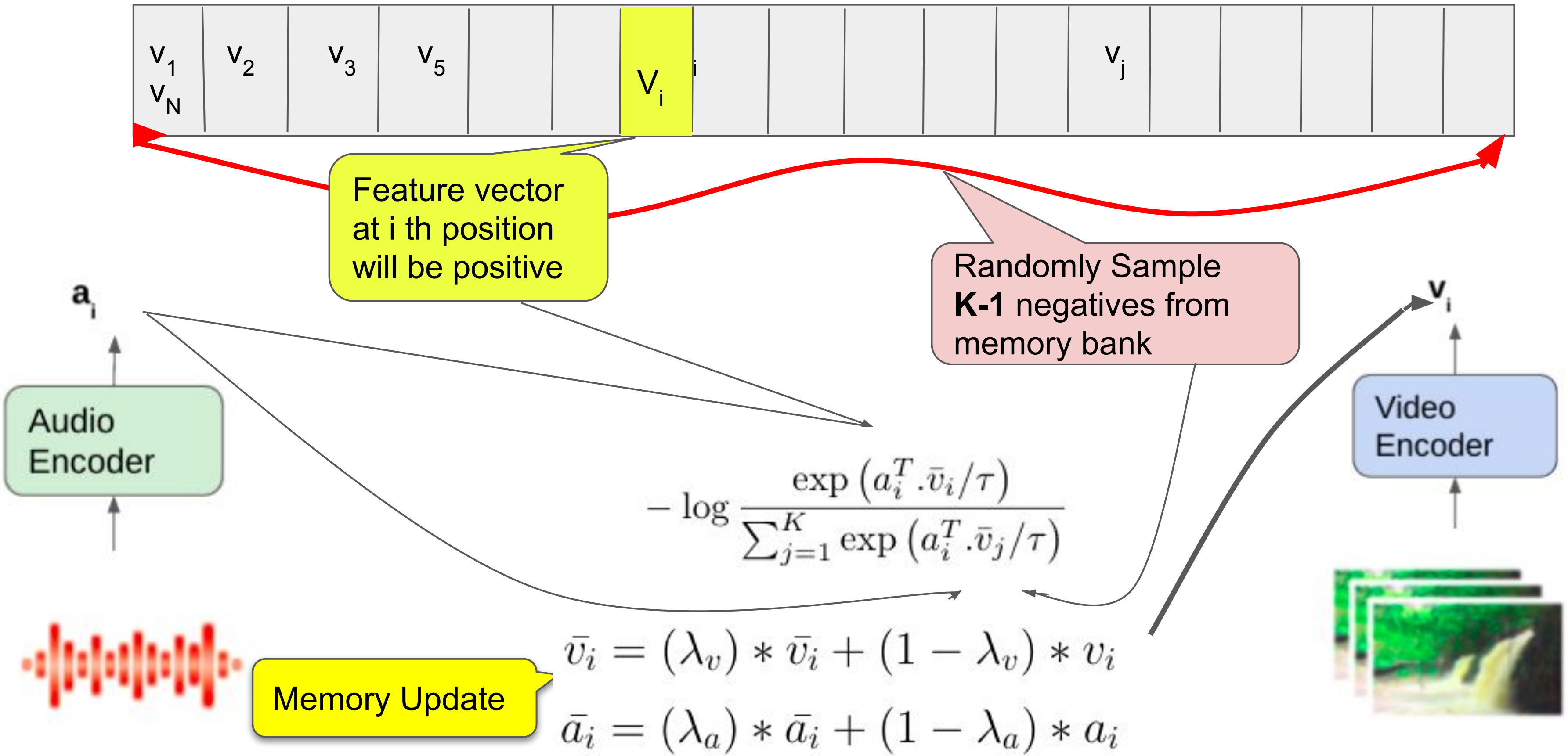
K-1 negatives



Problems -

1. Less number of negatives (batch size -1).
2. Computationally expensive .

Memory Bank Approach

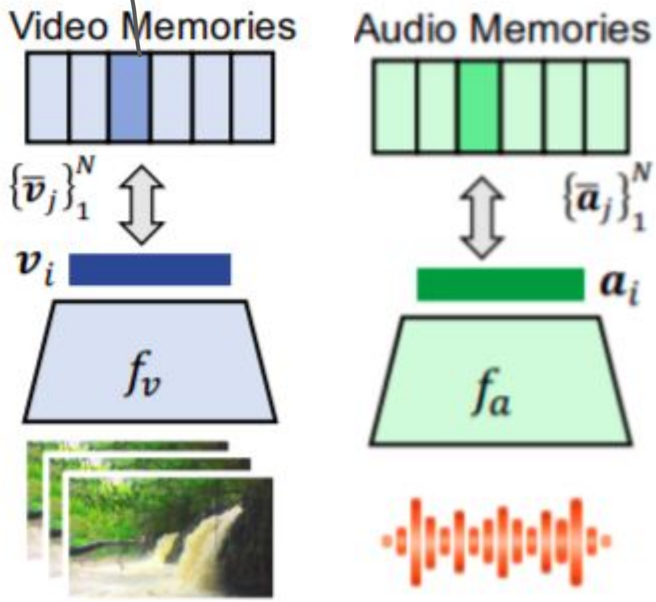


Unimodal Noise Contrastive Estimation Loss

Same index i from memory will work as positive pair

$$L_{visual}^{uni} = -\log \frac{\exp\left(sim\left(v_i, \bar{v}_i\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(sim\left(v_i, v_j\right)/\tau\right)}$$

$$L_{audio}^{uni} = -\log \frac{\exp\left(sim\left(a_i, \bar{a}_i\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(sim\left(a_i, a_j\right)/\tau\right)}$$



It can also be seen as Instance Discrimination Loss, each instance being a class.

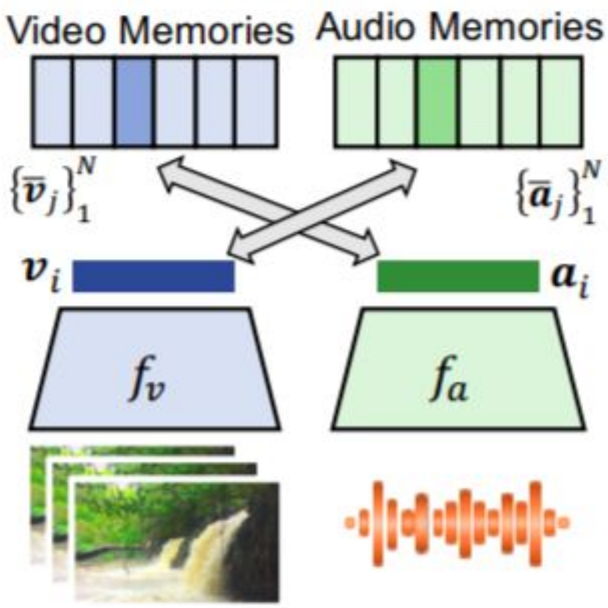
Cross Modal Noise Contrastive Estimation Loss

Visual Encoder is trained

$$L_{visual}^{cross} = -\log \frac{\exp\left(sim\left(v_i, a_i\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(sim\left(v_i, a_j\right)/\tau\right)}$$

Audio Encoder is trained

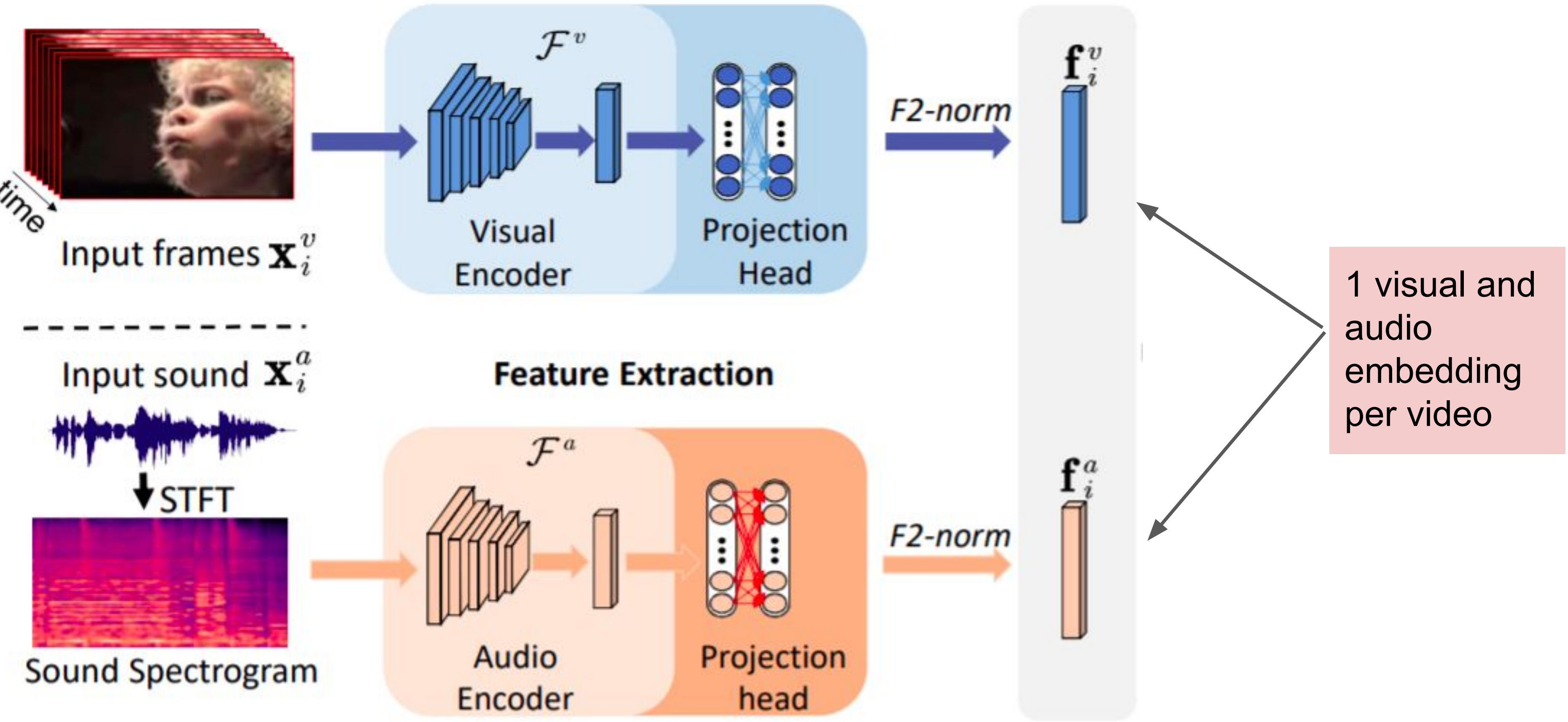
$$L_{audio}^{cross} = -\log \frac{\exp\left(sim\left(a_i, v_i\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(sim\left(a_i, v_j\right)/\tau\right)}$$



NCE loss forces audio feature vector a_i and video feature vector v_i of the i th video to come closer to each other

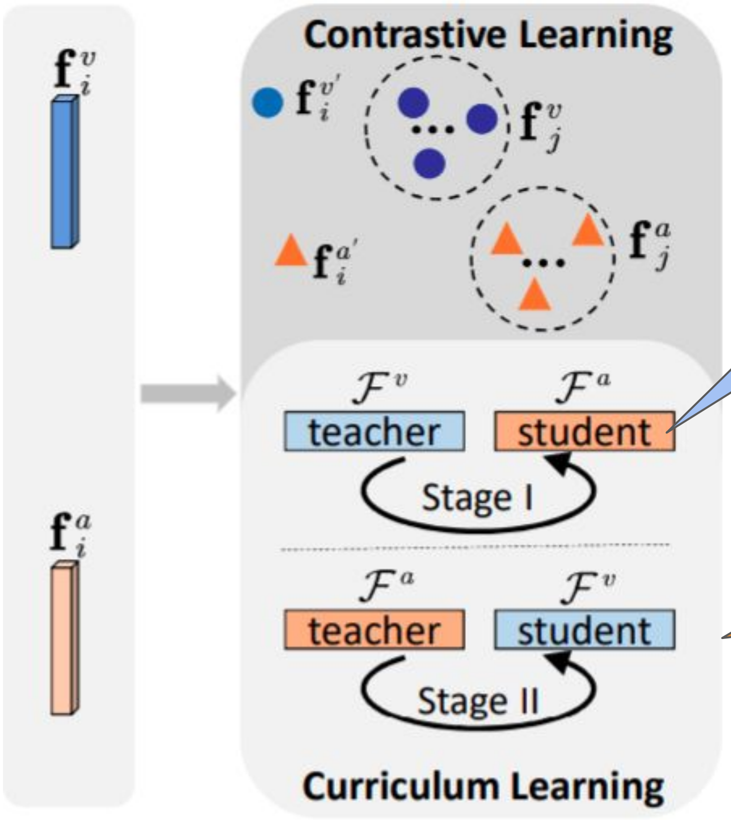
Enhancing Audio-Visual Association with Self-Supervised **Curriculum Learning**

Jingran Zhang , Heng Tao Shen (AAAI-21)



time
Input frames $\mathbf{x}_i^v$
$\mathcal{F}^v$
Visual Encoder
Projection Head
F2-norm
$\mathbf{f}_i^v$
Input sound $\mathbf{x}_i^a$
STFT
Sound Spectrogram
Feature Extraction
$\mathcal{F}^a$
Audio Encoder
Projection head
F2-norm
$\mathbf{f}_i^a$
1 visual and audio embedding per video

2 Stage Recursive Process



$$L_{visual}^{uni} = -\log \frac{\exp\left(sim\left(v_i, \bar{v}_i\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(sim\left(v_i, v_j\right)/\tau\right)}$$

$$L_{audio}^{cross} = -\log \frac{\exp\left(sim\left(a_i, v_i\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(sim\left(a_i, v_j\right)/\tau\right)}$$

$$L_{audio}^{uni} = -\log \frac{\exp\left(sim\left(a_i, \bar{a}_i\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(sim\left(a_i, a_j\right)/\tau\right)}$$

$$L_{visual}^{cross} = -\log \frac{\exp\left(sim\left(v_i, a_i\right)/\tau\right)}{\sum_{j=1}^{K} \exp\left(sim\left(v_i, a_j\right)/\tau\right)}$$

Results of Curriculum Learning

Action Recognition

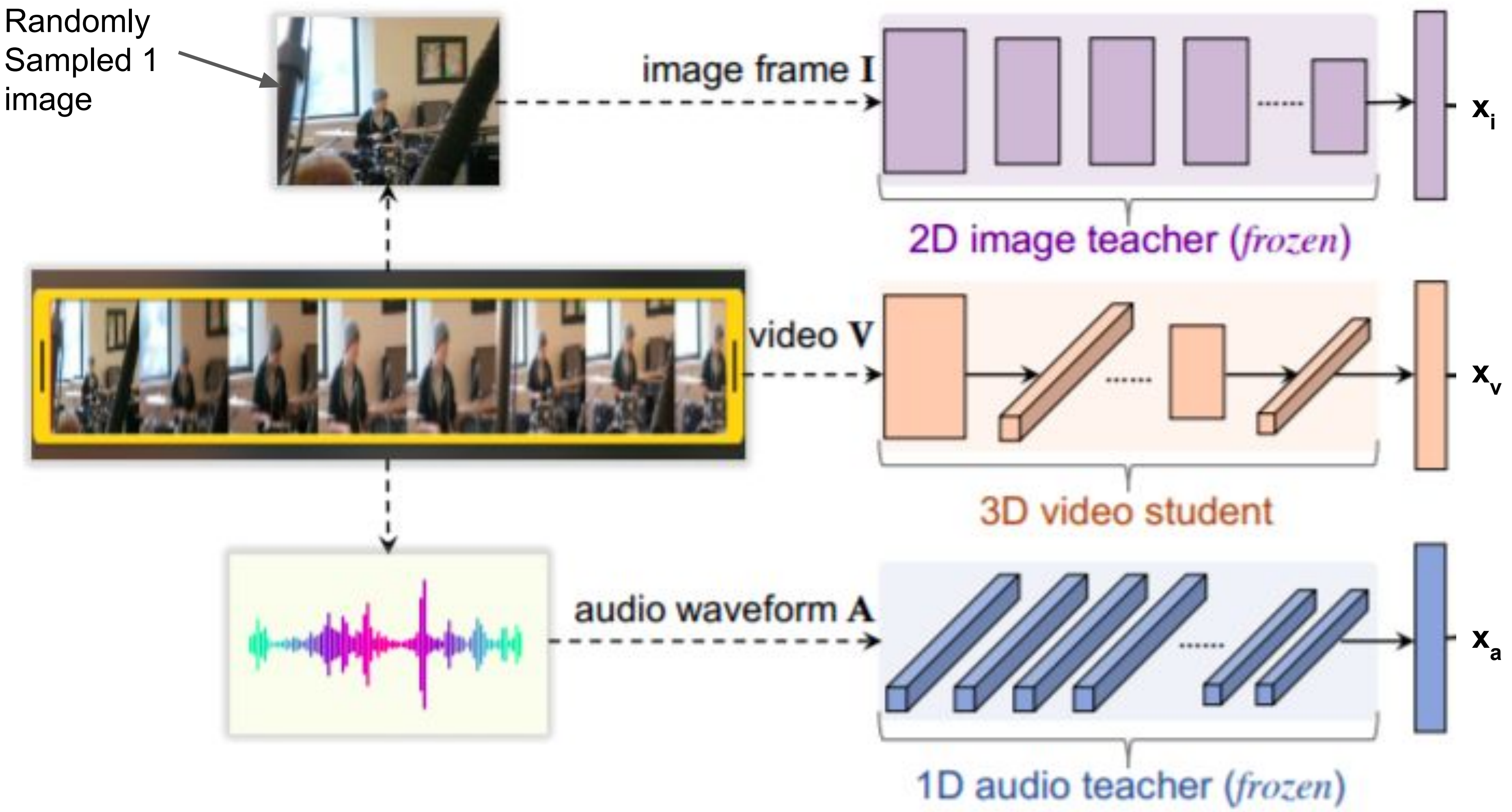
	Clip Size	UCF101	HMDB51
SSCL-stage-I	16x112x112	81.4	47.7
SSCL-stage-II	16x112x112	82.6	49.9
SSCL-stage-II	16x224x224	84.3	54.1
SSCL-stage-II	32x224x224	87.1	57.6

Sound Recognition

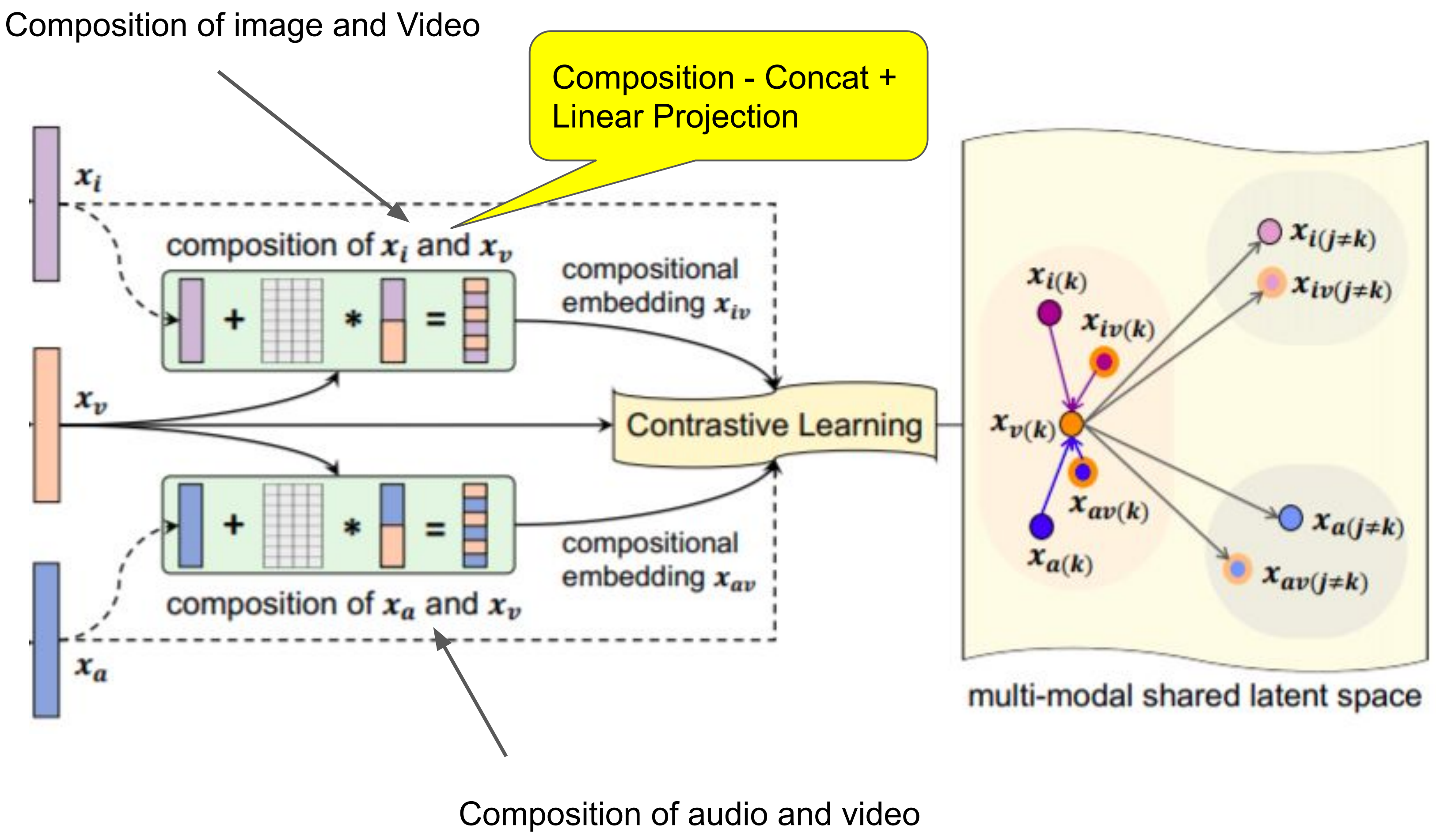
	Backbone	ESC-50	DCASE
SSCL-stage I	2D-ResNet10	85.8	91.0
SSCL-stage II	2D-ResNet10	88.3	93.0

Distilling Audio-Visual Knowledge by **Compositional Contrastive Learning**

Yanbei Chen , Yongqin Xian
(CVPR-21)



Randomly Sampled 1 image
image frame I
video V
audio waveform A
2D image teacher (frozen)
3D video student
1D audio teacher (frozen)
x_i
x_v
x_a


Composition of image and Video
Composition - Concat + Linear Projection
x_i
composition of x_i and x_v
compositional embedding x_{iv}
x_v
Contrastive Learning
compositional embedding x_{av}
composition of x_a and x_v
x_a
$x_{i(k)}$
$x_{iv(k)}$
$x_{v(k)}$
$x_{av(k)}$
$x_{a(k)}$
$x_{i(j \neq k)}$
$x_{iv(j \neq k)}$
$x_{a(j \neq k)}$
$x_{av(j \neq k)}$
multi-modal shared latent space
Composition of audio and video

Why Composition ?

1. the student and teacher embeddings may be semantically unaligned

2. an image frame may capture only partial visual cues not directly related to the video event,

3. audio of an action video may be irrelevant music or speech.

Results of Compositional Contrastive Learning

Method	UCF51			ActivityNet		
	A	**I**	**AI**	**A**	**I**	**AI**
baseline	57.5	57.5	57.5	32.6	32.6	32.6
FitNet	48.4	67.4	62.4	21.3	45.8	34.6
PKT	53.2	58.2	62.0	33.4	35.4	35.1
COR	57.7	65.5	66.3	31.4	43.1	41.7
RKD	53.0	55.4	58.2	-	34.3	-
CRD	60.3	61.4	63.2	36.4	37.3	36.6
IFD	56.3	54.2	64.2	34.6	33.8	35.4
CMC	59.2	60.4	63.1	34.4	23.7	33.9
CCL	**64.9**	**69.1**	**70.0**	**36.5**	**46.3**	**47.3**

Conclusion

- We have discussed the importance of multimodal learning.
- We have also focused on efficiency for learning video representations, that can be used in wide variety of downstream tasks.
- How cross modal knowledge distillation helps in learning better features.
- We have explored both supervised and self supervised approaches for multimodal learning.

References (1/2)

- R. Gao, T. Oh, K. Grauman, L. Torresani. "Listen to Look: Action Recognition by Previewing Audio". In CVPR, 2020.

- Yanbei Chen , Yongqin Xian , A. Sophia Koepke , Ying Shan, Zeynep Akata. “Distilling Audio-Visual Knowledge by Compositional Contrastive Learning”. In CVPR, 2021.

- Triantafyllos Afouras, Andrew Owens, Joon Son Chung, Andrew Zisserman. “Self-Supervised Learning of Audio-Visual Objects from Video”. In CVPR, 2021

- Pedro Morgado, Nuno Vasconcelos, Ishan Misra. Audio-Visual Instance Discrimination with Cross-Modal Agreement. In CVPR-2021.

Reference (2/2)

- Jingran Zhang, Xing Xu, Fumin Shen, Huimin Lu, Xin Liu, Heng Tao Shen. Enhancing Audio-Visual Association with Self-Supervised Curriculum Learning. In AAAI-21.

- Ruohan Gao, Kristen Grauman. VISUALVOICE: Audio-Visual Speech Separation with Cross-Modal Consistency. In CVPR-2021.

- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In CVPR-2020.