Distilling Audio-Visual Knowledge by Compositional Contrastive Learning

Yanbei Chen[1], Yongqin Xian[2], A. Sophia Koepke[1], Ying Shan[3], Zeynep Akata[1,2,4]

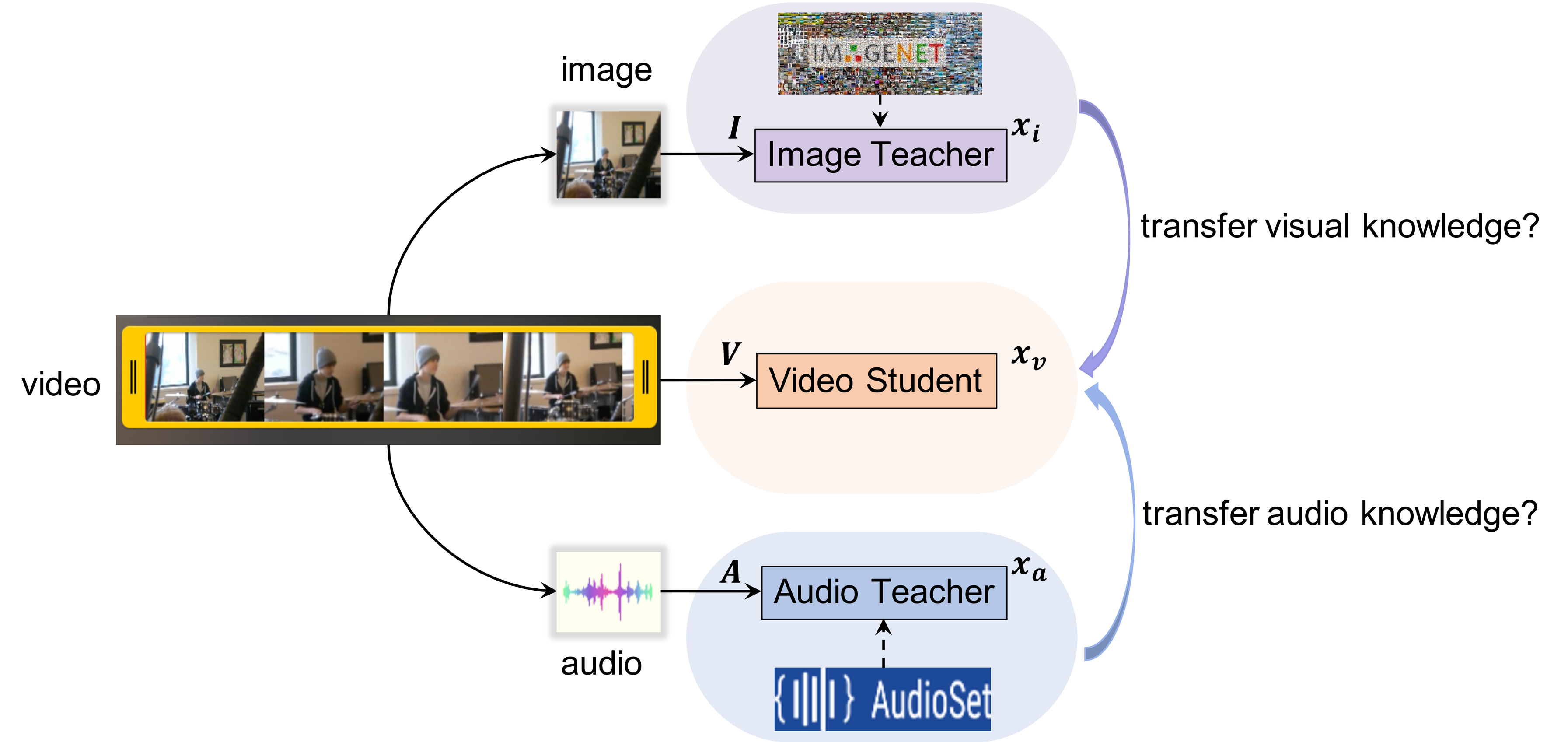
[1]University of Tübingen [2]MPI for Informatics [3]ARC, Tencent PCG [4]MPI for Intelligent Systems

CVPR VIRTUAL JUNE 19-25

Introduction

Research question

➢ How could we transfer knowledge across heterogeneous data modalities to learn more powerful representations?



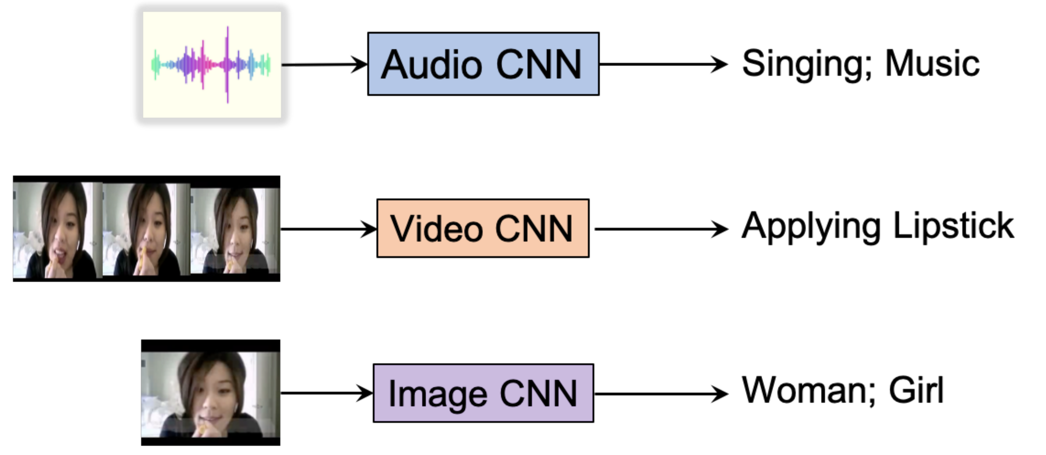
Main challenges

➢ cross-modal content may not be semantically correlated:

- e.g visual content is *applying lipstick*, while audio content is *music*.

➢ audio, image, video data exhibit heterogeneous characteristics:

- encoding temporal, spatial, and spatiotemporal information.



Main idea

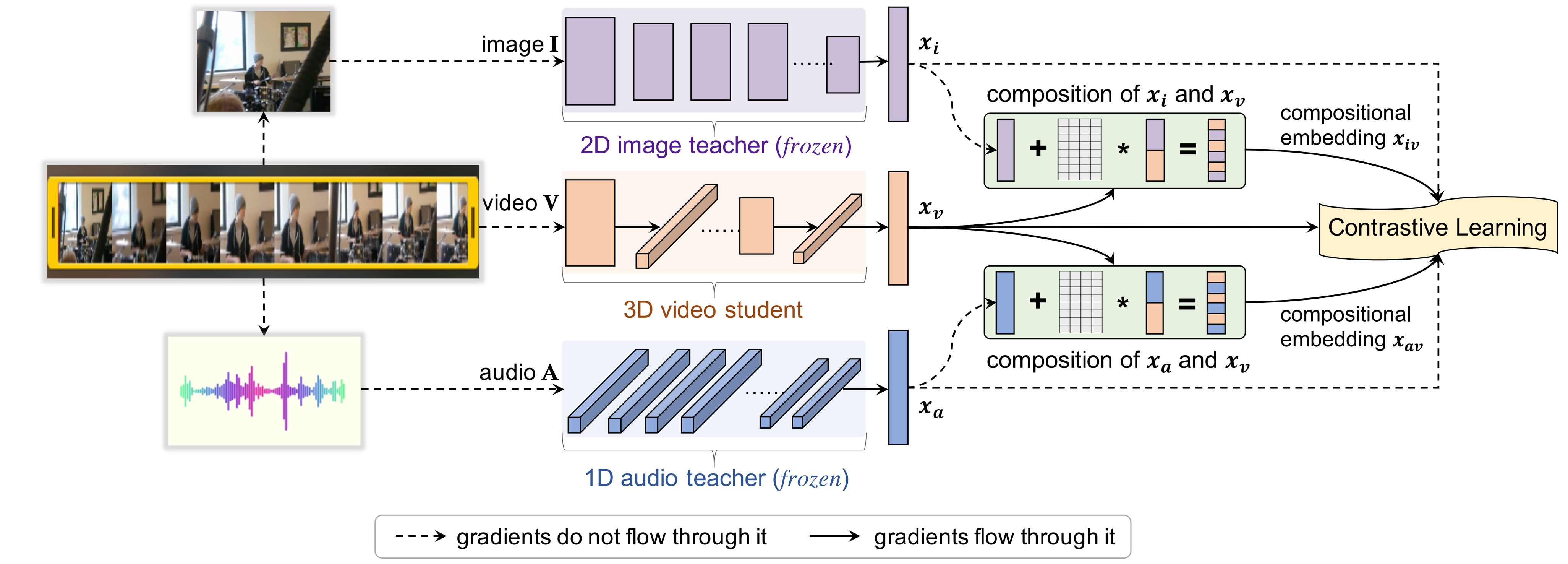
❖ **compositional contrastive learning**

✓ compose different modalities to close cross-modal semantic gap

✓ contrastive learning across all modalities in a shared latent space

Methodology

Proposed approach



❑ Unimodal representations of audio and vision

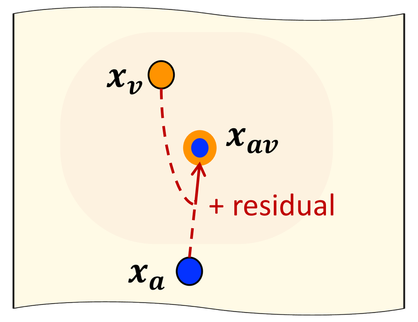
- audio recordings → 1D audio teacher network → audio embeddings
- image frames → 2D image teacher network → image embeddings
- video clips → 3D video student network → video embeddings

❑ Compositional multi-modal representation

- composition of teacher, student embeddings to *bridge the semantic gap*

$$\mathcal{F}_{av}(x_a, x_v) = x_{av} = x_a + \underline{f_{\theta_{av}}(x_a, x_v)}$$ add a learnable residual

$$\mathcal{L}_{ce}^{av}(x_{av}, k) \longleftarrow$$ constrained by task objective



❑ Compositional contrastive learning

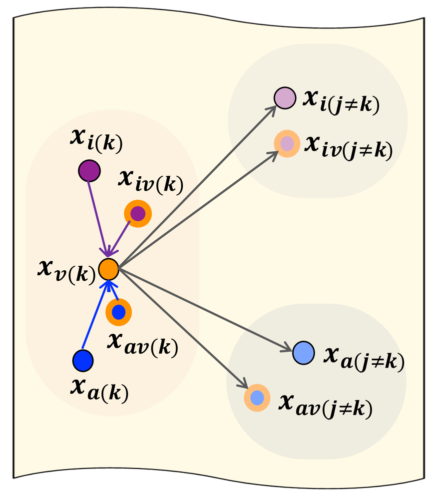
- contrastive learning to transfer multi-modal knowledge

$$-\log \frac{\exp(\Phi(x_{v(i)}, x_{a(i)})/\tau)}{\sum_{j=1}^{B} \exp(\Phi(x_{v(j)}, x_{a(j)}))/\tau)} = -\log p_{av(i)}$$ contrastive loss (NCE)

$$\mathcal{L}_{nce}(x_v, x_a) = -\frac{1}{B_p} \sum_{j=k} \log p_{av(j)} - \frac{1}{B_n} \sum_{j \neq k} \log(1 - p_{av(j)})$$ multi-class NCE

$$\mathcal{L}_a(x_v, x_a, x_{av}) = \lambda \mathcal{L}_{nce}(x_v, x_a) + (1-\lambda)\mathcal{L}_{nce}(x_v, x_{av})$$ audio distillation

$$\mathcal{L}_i(x_v, x_i, x_{iv}) = \lambda \mathcal{L}_{nce}(x_v, x_i) + (1-\lambda)\mathcal{L}_{nce}(x_v, x_{iv})$$ image distillation



Experiments

A new benchmark on multi-modal distillation

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**

Table1. Video recognition (top1 acc %)

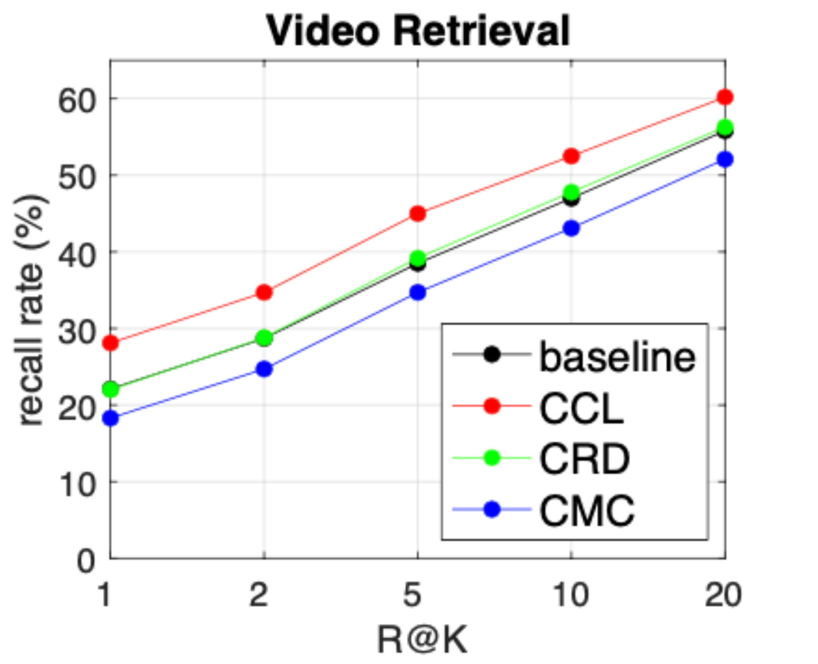


Figure 1. Retrieval on VGGSound.

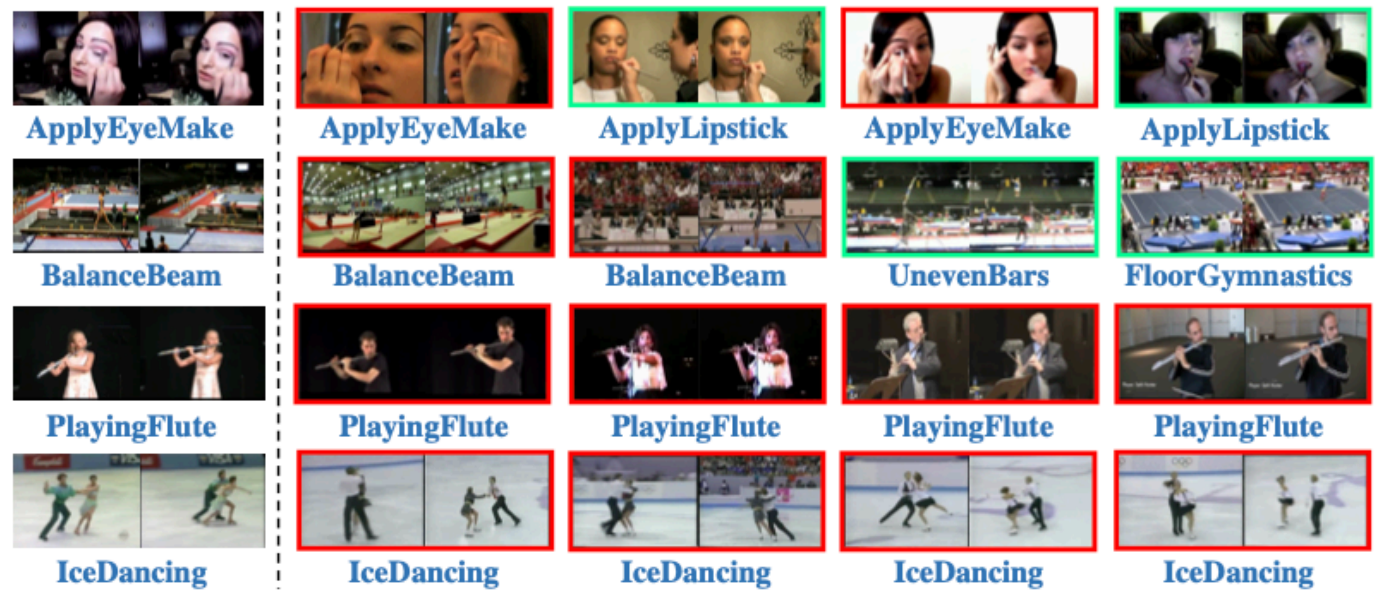


Figure 2. Qualitative results for video retrieval.

Take-home message:

- distilling audio or visual knowledge helps video recognition/retrieval
- audio and visual knowledge are complementary

Conclusion

➢ a new approach for distilling audio-visual knowledge

➢ state-of-the-art performance on multi-modal distillation benchmark

References

- Hinton et al. Distilling the Knowledge in a Neural Network. NeuRIPSW2014
- Gupta et al. Cross Modal Distillation for Supervision Transfer, CVPR2016
- Tian et al. Contrastive Representation Distillation. ICLR2020; Contrastive Multiview Coding. ECCV2020