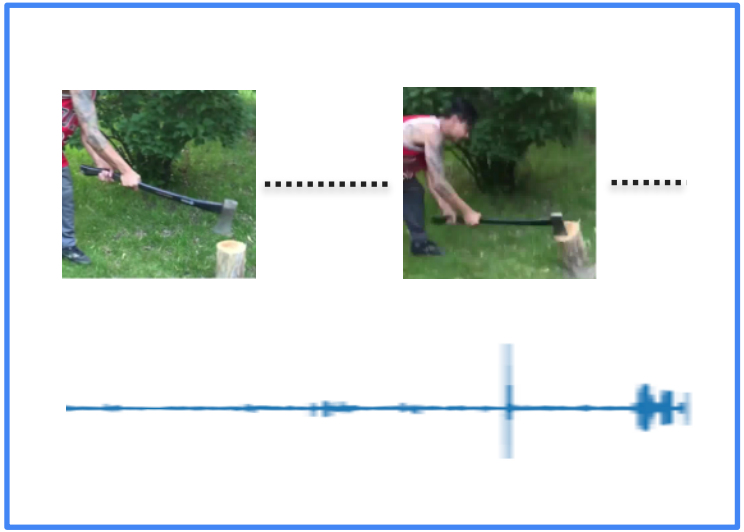
Multi-Modal Learning

Saksham Singh Kushwaha
The University of Texas at Dallas

(DL4MIR'24 Workshop)

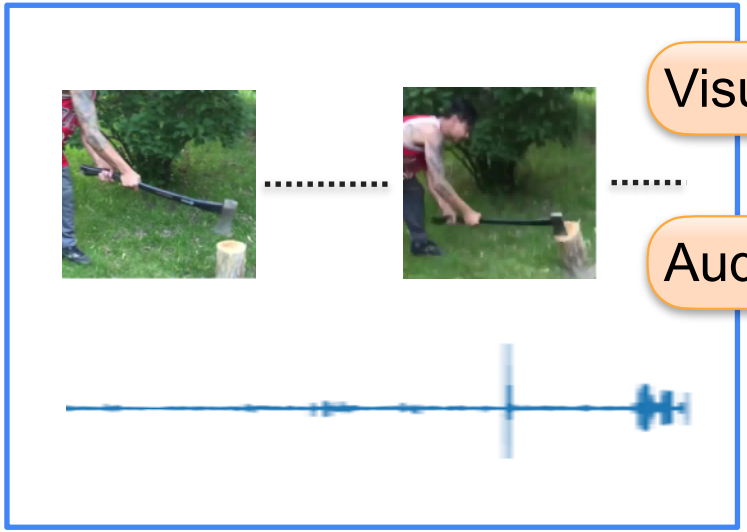
Research Interests



Audio-Visual Integration

[1] Chen et. al, "Vggsound: A Large-Scale Audio-Visual Dataset", ICASSP, 2020.

Research Interests



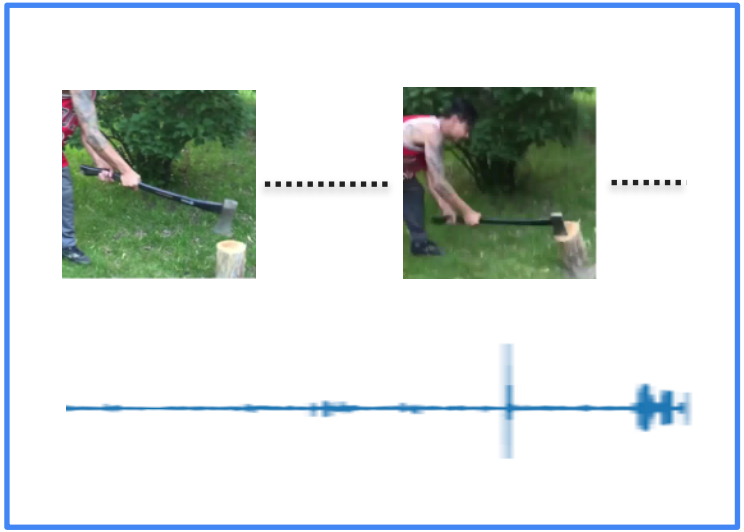
Audio-Visual Integration

Visual guided audio generation

Audio-visual data compression

[1] Chen et. al, "Vggsound: A Large-Scale Audio-Visual Dataset", ICASSP, 2020.

Research Interests

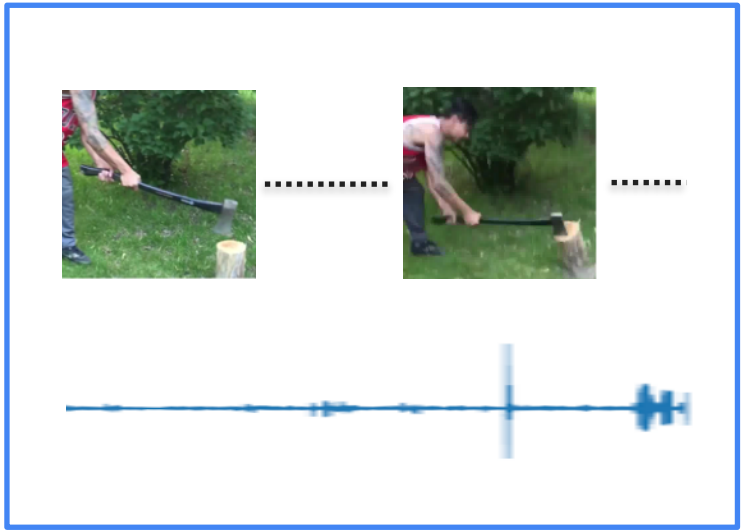


Audio-Visual Integration

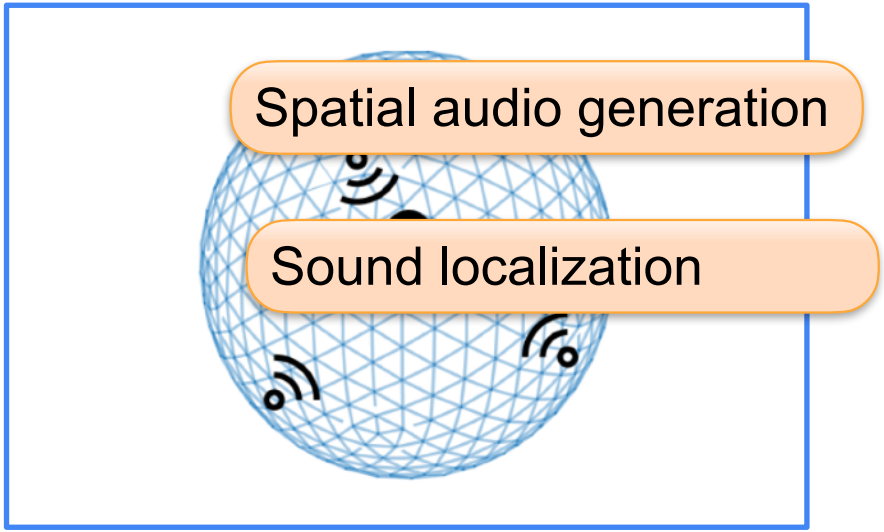
Spatial Audio

[1] Chen et. al, "Vggsound: A Large-Scale Audio-Visual Dataset", ICASSP, 2020.
[2] Dagli et. Al, “SEE-2-SOUND: Zero-Shot Spatial Environment-to-Spatial Sound”, ArXiv, 2024

Research Interests



Audio-Visual Integration



Spatial Audio

[1] Chen et. al, "Vggsound: A Large-Scale Audio-Visual Dataset", ICASSP, 2020.
[2] Dagli et. Al, “SEE-2-SOUND: Zero-Shot Spatial Environment-to-Spatial Sound”, ArXiv, 2024

Dataset Distillation for Audio-Visual Datasets

(CVPR Sight and Sound Workshop, 2024)

Saksham Singh Kushwaha[1] , Siva Sai Nagender Vasireddy[1], Kai Wang[2],Yapeng Tian[1]

[1]University of Texas at Dallas, USA, [2]National University of Singapore, Singapore

[1]{sakshamsingh.kushwaha, sivasainagender.vasireddy, yapeng.tian}@utdallas.edu, [2]kai.wang@nus.edu.sg

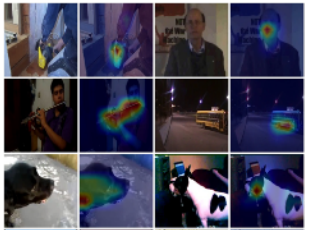
Saksham Singh Kushwaha
Ph.D. Candidate

Siva Sai Nagender Vasireddy
Ph.D. Candidate

Kai Wang
Ph.D. Candidate

Yapeng Tian
Assistant Professor

Increase in dataset sizes



AVE (2018)
(~4k)

Pandas-70M (2024)

Panda-70M

COCO (2014)
(~120k)

COCO
Common Objects in Context

VGGSound (2020)
(~200k)

[1] Chen et. al, "Vggsound: A Large-Scale Audio-Visual Dataset", ICASSP, 2020.
[2] Tian et. al, “Audio-visual event localization in unconstrained videos”, ECCV, 2018
[3] Yin et. al, “Microsoft **coco**: Common objects in context”, ECCV, 2014
[4] Chen et. al, “**Panda-70m**: Captioning **70m** videos with multiple cross-modality teachers”, CVPR, 2024

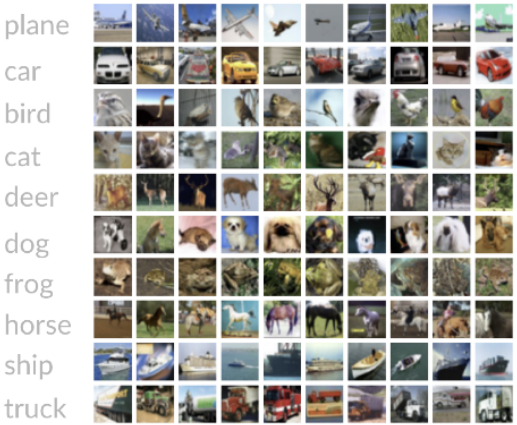
Goal: Dataset distillation/compression

Large training data

Representative data

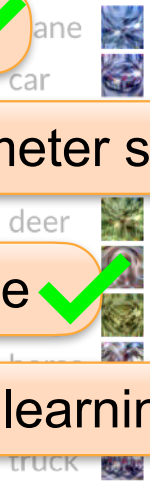
[1] Wang et. al, "Dataset Distillation", arXiv preprint, 2018.

Goal: Dataset distillation/compression



Large training data

- Reduce storage ✓
- Faster hyper-parameter search ✓
- Reduce training time ✓
- Effective Continual learning ✓



Representative data

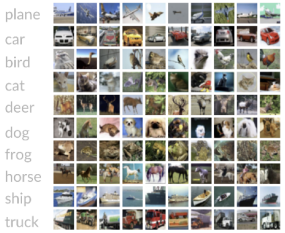
[1] Wang et. al, "Dataset Distillation", arXiv preprint, 2018.

Representativeness

- Comparable performance on unseen real test data

$T = \{(x_i, y_i)\}_{i=1}^{|T|}$

Real training data



Dataset Distillation

$S = \{(s_i, y_i)\}_{i=1}^{|S|}$

Synthetic data

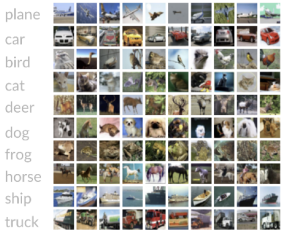
[1] Wang et. al, "Dataset Distillation", arXiv preprint, 2018.

Representativeness

- Comparable performance on unseen real test data

$T = \{(x_i, y_i)\}_{i=1}^{|T|}$

Real training data



Train

ϕ_{θ_T}

Dataset Distillation

$S = \{(s_i, y_i)\}_{i=1}^{|S|}$

Synthetic data

Train

ϕ_{θ_S}

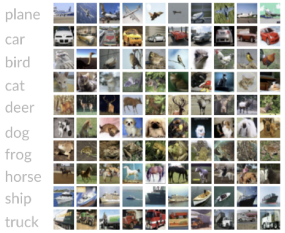
[1] Wang et. al, "Dataset Distillation", arXiv preprint, 2018.

Representativeness

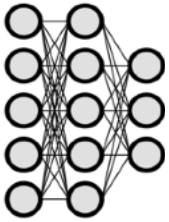
- Comparable performance on unseen real test data

$T = \{(x_i, y_i)\}_{i=1}^{|T|}$

Real training data

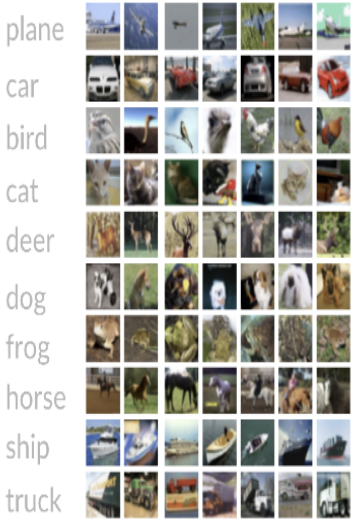


Train



ϕ_{θ_T}

Test



$\mathbb{E}_{x \sim P_D}[\ell(\phi_{\theta_T}(x), y)]$

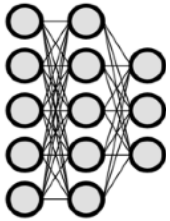
$\approx$

Dataset Distillation

$S = \{(s_i, y_i)\}_{i=1}^{|S|}$

Synthetic data

Train



ϕ_{θ_S}

Test

$\mathbb{E}_{x \sim P_D}[\ell(\phi_{\theta_S}(x), y)]$

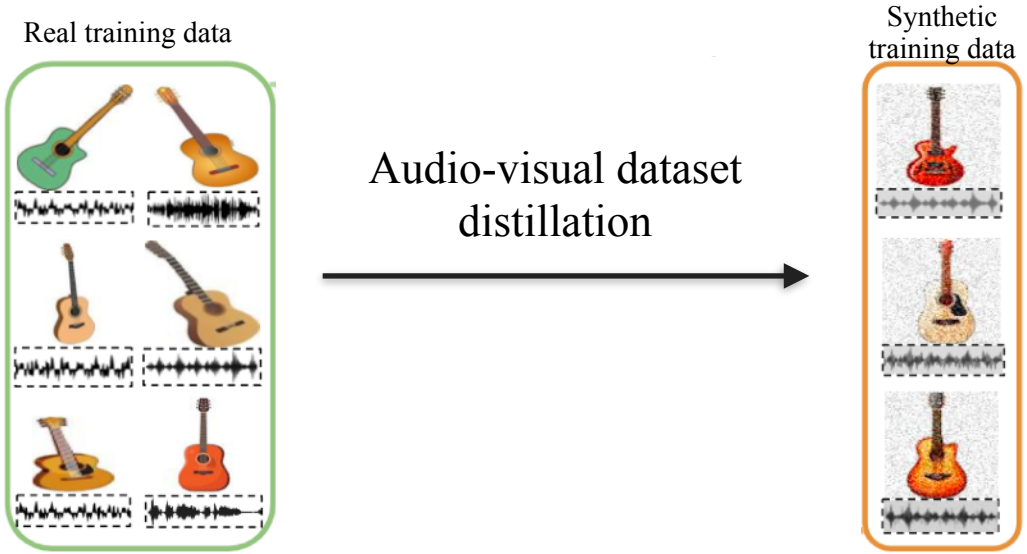
$D = \{(x_i, y_i)\}_{i=1}^{|D|}$

Real test data

[1] Wang et. al, "Dataset Distillation", arXiv preprint, 2018.

Audio-Visual dataset distillation

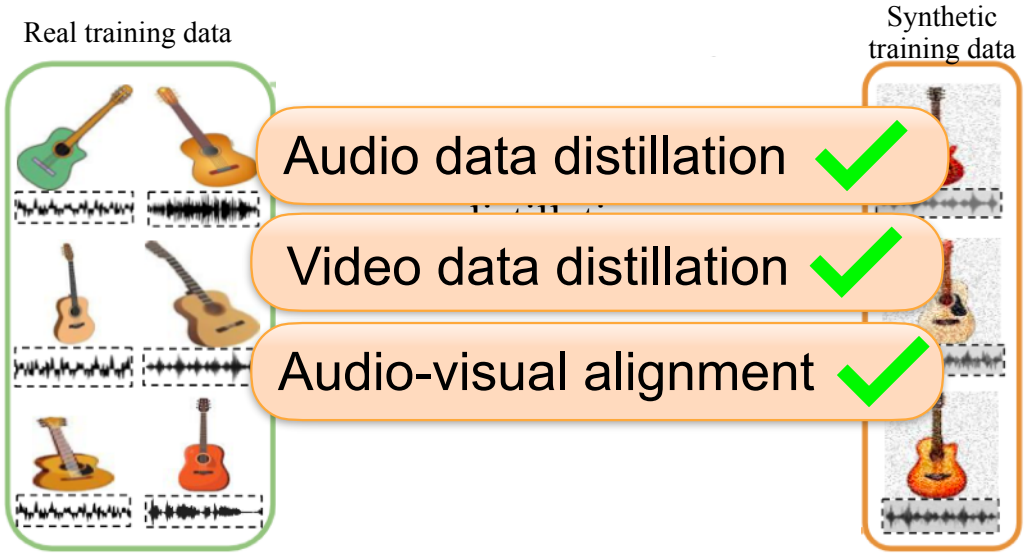
- Previous dataset distillation research focuses on image-only
- Explosion in large multimodal datasets



[1] Wang et. al, "Dataset Distillation", arXiv preprint, 2018.

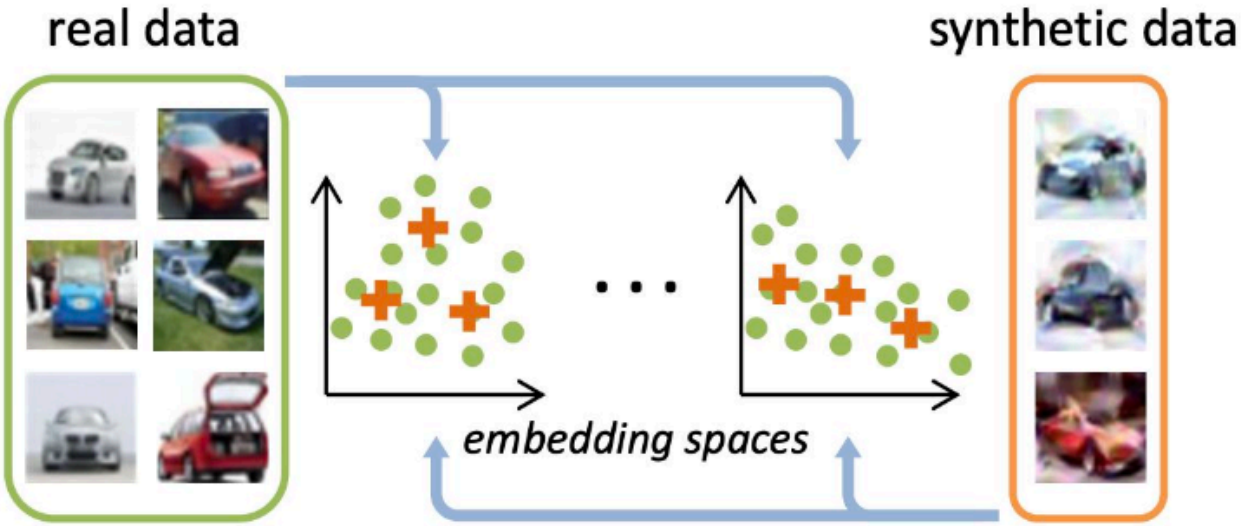
Audio-Visual dataset distillation

- Previous dataset distillation research focuses on image-only
- Explosion in large multimodal datasets



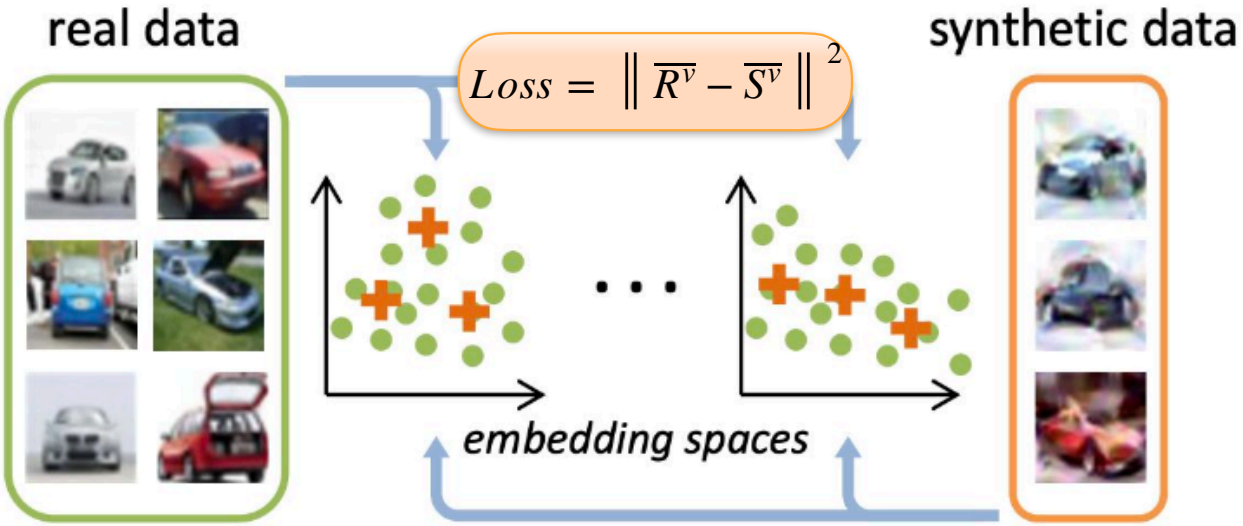
[1] Wang et. al, "Dataset Distillation", arXiv preprint, 2018.

Background: Distribution Matching



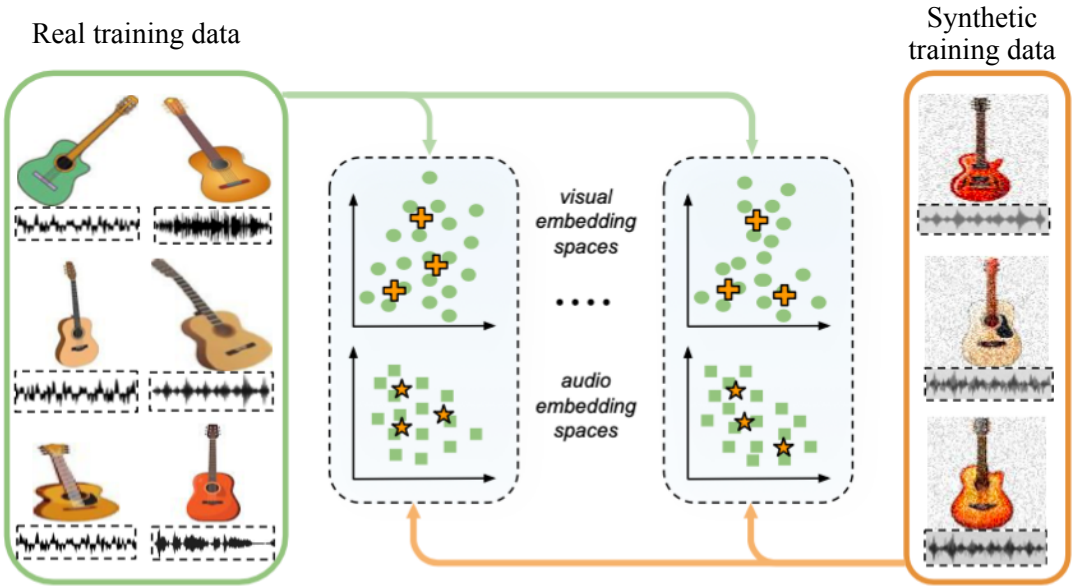
[1] B. Zhao “Data Condensation using distribution matching” WACV 2023
[2] G. Zhao “Improved Distribution Matching for Dataset condensation” CVPR 2023
[3] H. Zhang “M3D: Dataset condensation by minimizing maximum mean discrepancy” AAAI 2024

Background: Distribution Matching



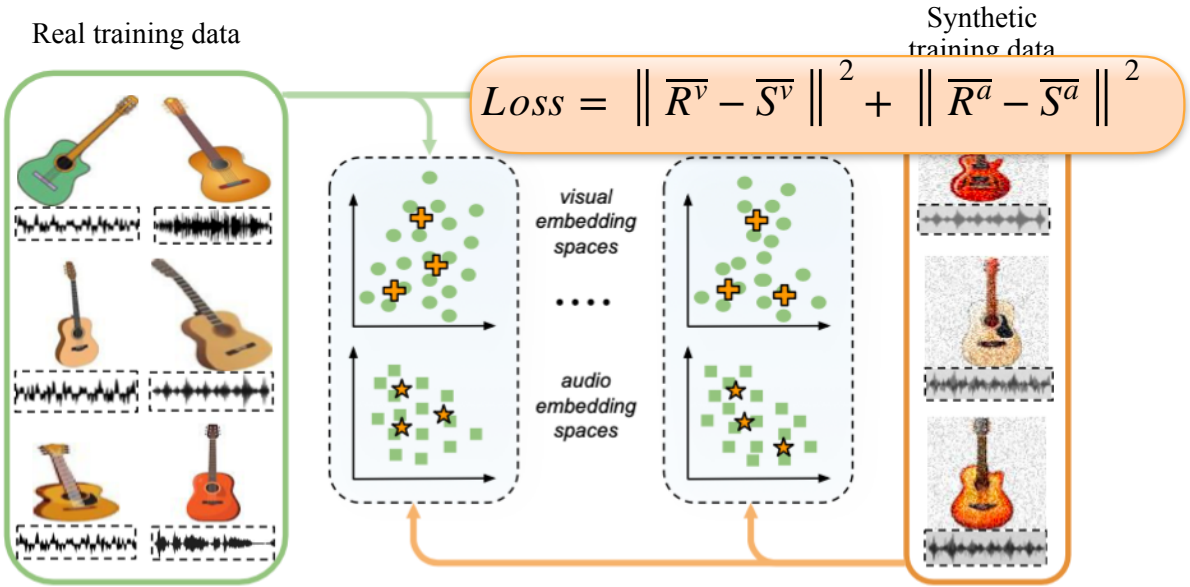
[1] B. Zhao “Data Condensation using distribution matching” WACV 2023
[2] G. Zhao “Improved Distribution Matching for Dataset condensation” CVPR 2023
[3] H. Zhang “M3D: Dataset condensation by minimizing maximum mean discrepancy” AAAI 2024

Vanilla Audio-Visual Dataset distillation



[1] B. Zhao “Data Condensation using distribution matching” WACV 2023
[2] G. Zhao “Improved Distribution Matching for Dataset condensation” CVPR 2023
[3] H. Zhang “M3D: Dataset condensation by minimizing maximum mean discrepancy” AAAI 2024

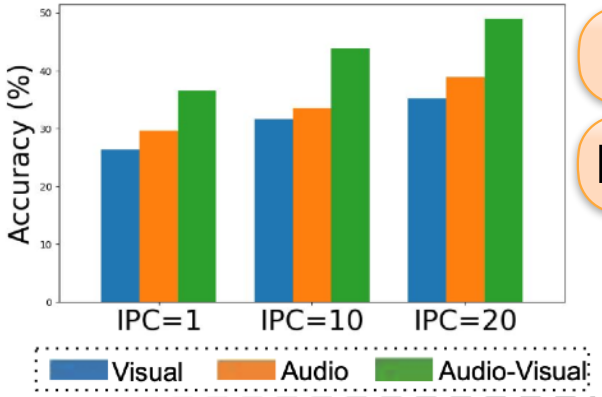
Vanilla Audio-Visual Dataset distillation



[1] B. Zhao “Data Condensation using distribution matching” WACV 2023
[2] G. Zhao “Improved Distribution Matching for Dataset condensation” CVPR 2023
[3] H. Zhang “M3D: Dataset condensation by minimizing maximum mean discrepancy” AAAI 2024

Does integration hold for AV distilled data?

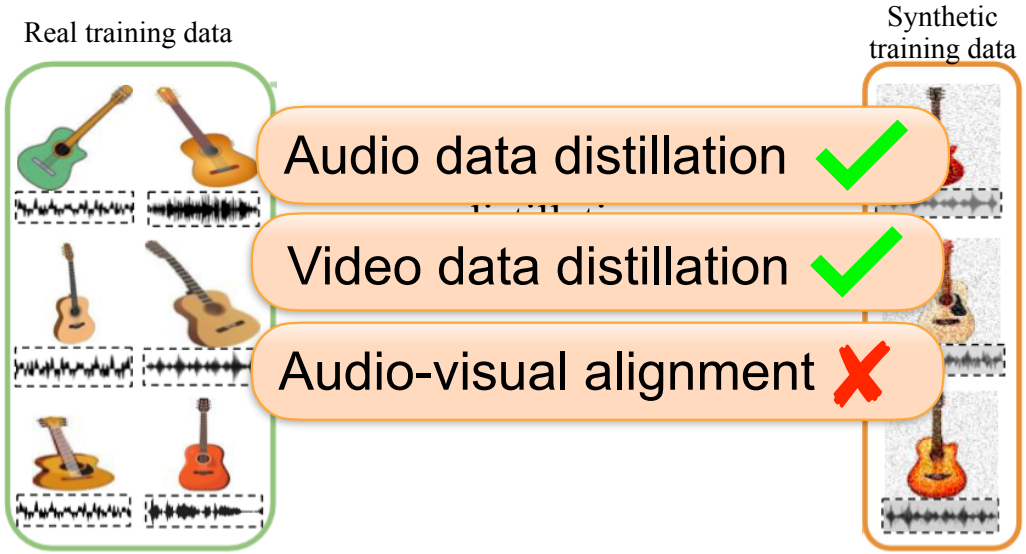
	Only-A	Only-V	Audio-Visual Fusion			
			Concat	Sum	Attention	Ensemble
1	$29.60_{\pm 2.33}$	$26.40_{\pm 1.10}$	$33.77_{\pm 1.65}$	$34.72_{\pm 1.27}$	$9.97_{\pm 0.83}$	$\mathbf{36.54}_{\pm 2.52}$
IPC 10	$33.60_{\pm 1.35}$	$31.63_{\pm 1.96}$	$41.71_{\pm 1.27}$	$40.49_{\pm 1.83}$	$10.11_{\pm 0.35}$	$\mathbf{43.85}_{\pm 1.75}$
20	$38.93_{\pm 3.52}$	$35.23_{\pm 1.16}$	$46.59_{\pm 1.34}$	$46.05_{\pm 1.74}$	$11.10_{\pm 1.88}$	$\mathbf{49.01}_{\pm 2.44}$



Audio-visual integration helps

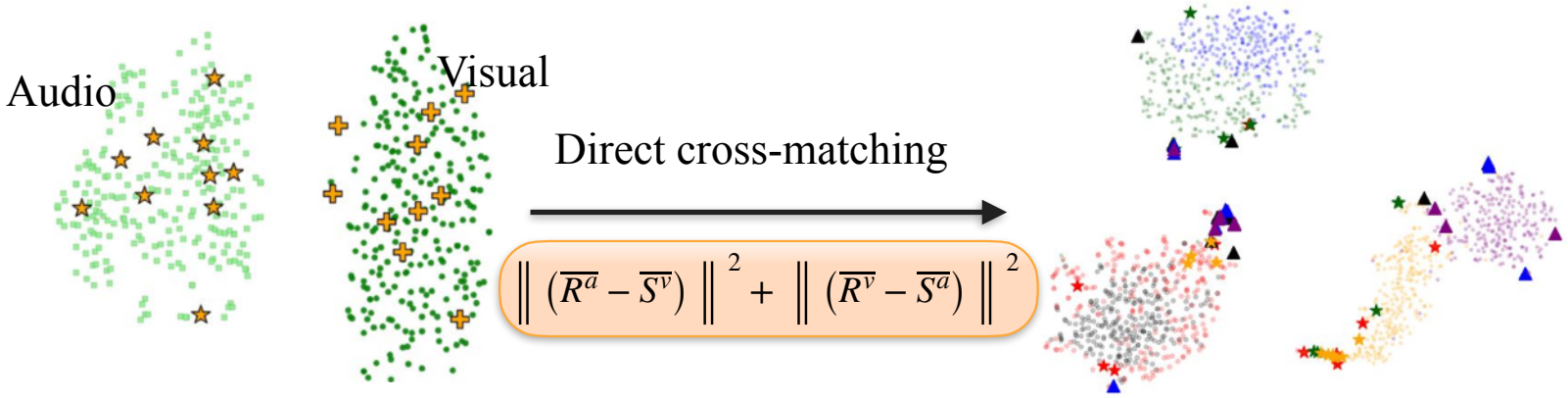
Ensemble method outperforms

Vanilla Audio-Visual Dataset distillation



[1] B. Zhao “Data Condensation using distribution matching” WACV 2023
[2] G. Zhao “Improved Distribution Matching for Dataset condensation” CVPR 2023
[3] H. Zhang “M3D: Dataset condensation by minimizing maximum mean discrepancy” AAAI 2024

Simple cross-modal alignment fails

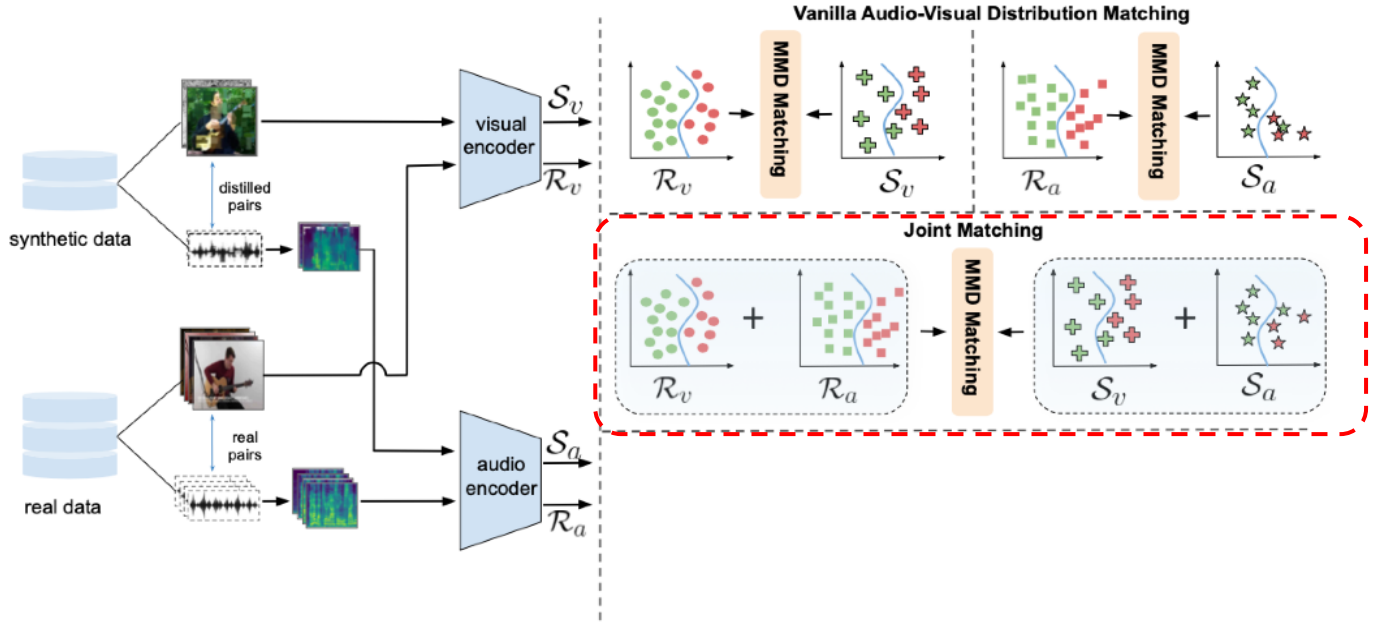


Modality Gap shown in intermediate stage of vanilla approach

Final Distribution

Joint Matching Loss

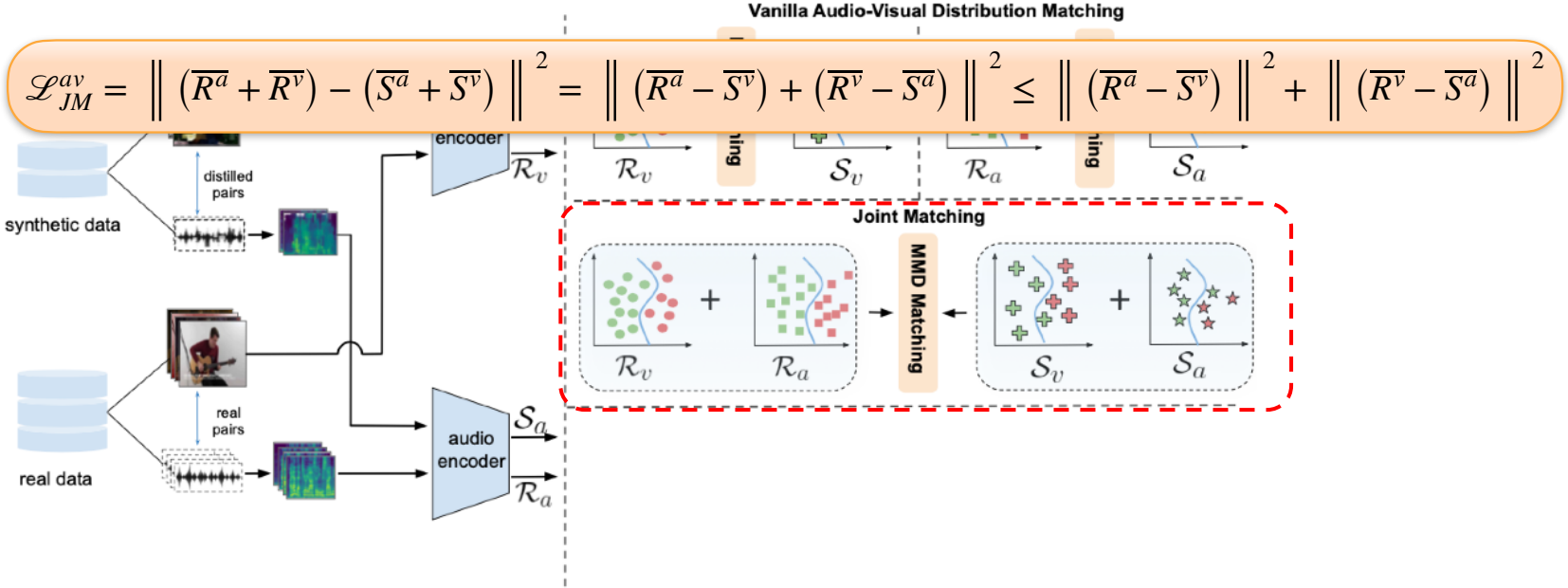
- Implicitly distills cross-modal alignment



Joint Matching Loss

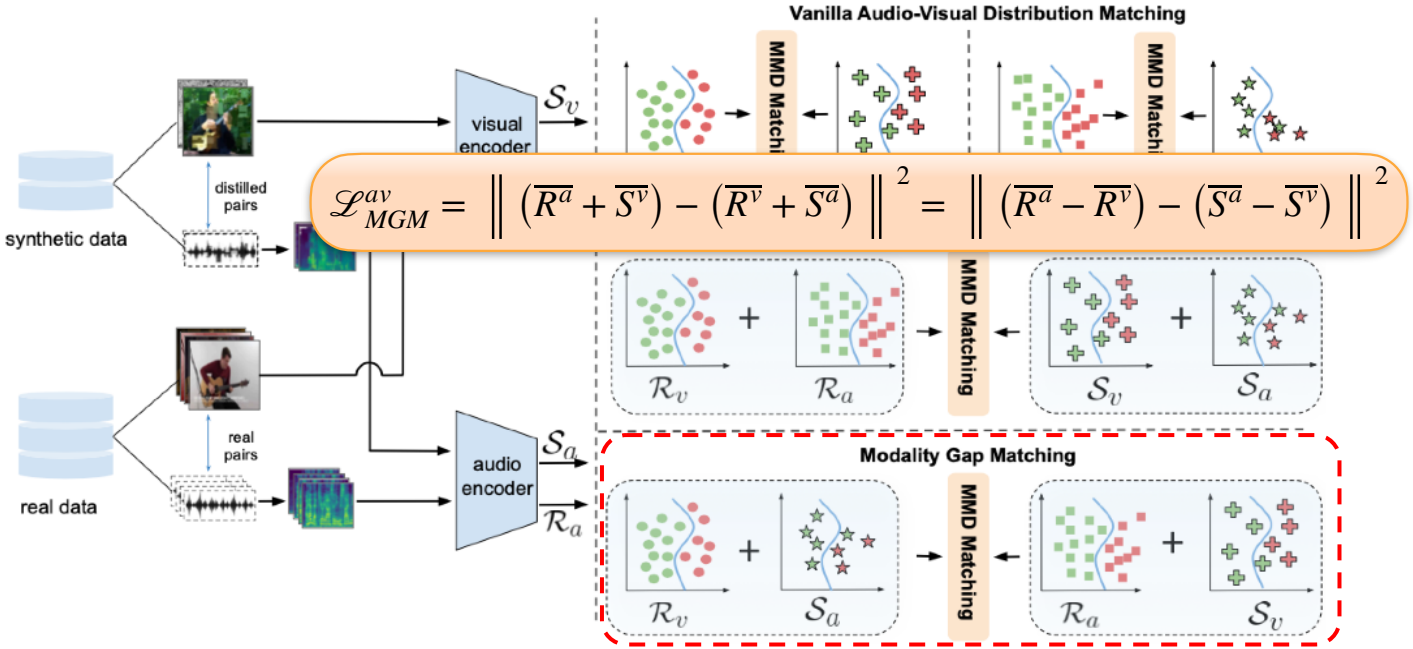
- Implicitly distills cross-modal alignment

$$\mathscr{L}_{JM}^{av} = \left\| \left(\overline{R}^a + \overline{R}^v\right) - \left(\overline{S}^a + \overline{S}^v\right) \right\|^2 = \left\| \left(\overline{R}^a - \overline{S}^v\right) + \left(\overline{R}^v - \overline{S}^a\right) \right\|^2 \leq \left\| \left(\overline{R}^a - \overline{S}^v\right) \right\|^2 + \left\| \left(\overline{R}^v - \overline{S}^a\right) \right\|^2$$

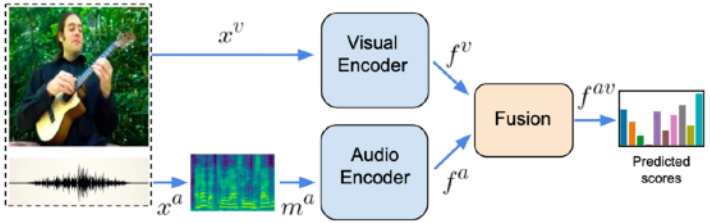


Modality Gap Matching Loss

- Implicitly aligns the audio-visual gap between real and synthetic data

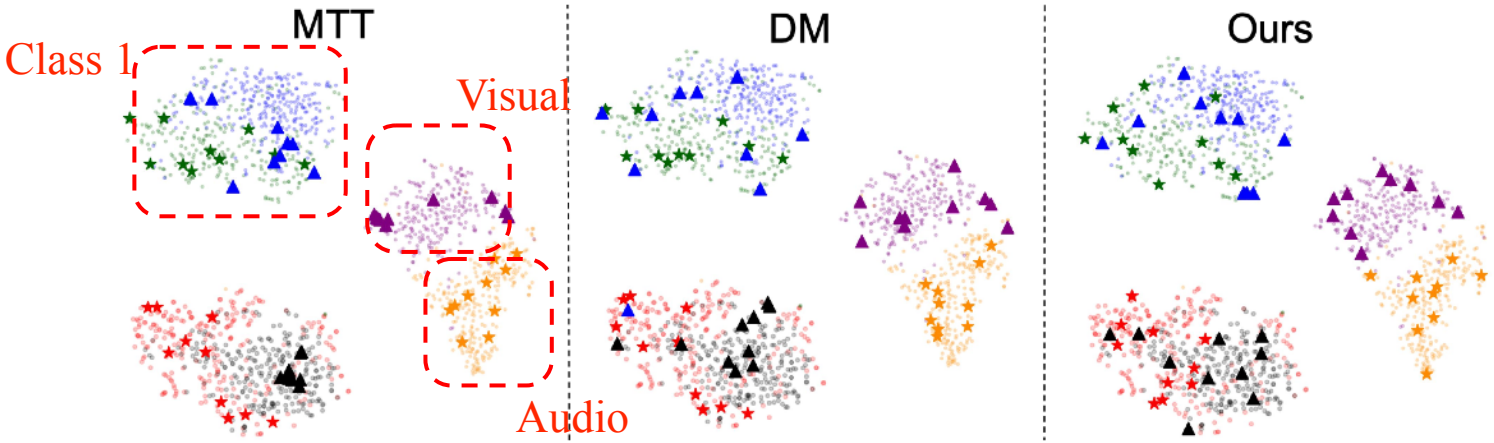


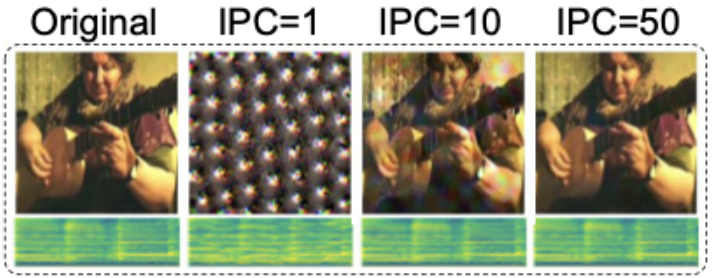
Audio-visual event recognition



	IPC	Ratio%	Coreset Selection		Training Set Synthesis				Ours	Upper Bound (Whole data)
			Random	Herding	DC	DSA	MTT	DM		
VGGS-10K	1	0.11	$15.44_{\pm 1.87}$	$20.77_{\pm 2.11}$	$18.28_{\pm 1.36}$	$19.32_{\pm 1.35}$	$34.13_{\pm 3.62}$	$36.54_{\pm 2.52}$	**40.41**$_{\pm 1.81}$	$68.24_{\pm 0.75}$
	10	1.13	$32.01_{\pm 1.64}$	$39.89_{\pm 1.64}$	$32.10_{\pm 0.84}$	$36.61_{\pm 1.04}$	$36.79_{\pm 1.97}$	$43.85_{\pm 1.75}$	**54.99**$_{\pm 1.73}$	
	20	2.26	$45.1_{\pm 2.31}$	$50.2_{\pm 0.74}$	-	-	$51.87_{\pm 1.26}$	$49.01_{\pm 2.44}$	**58.04**$_{\pm 1.68}$	
VGGSound	1	0.18	$1.38_{\pm 0.12}$	$2.14_{\pm 0.17}$	-	-	$1.55_{\pm 0.15}$	$3.08_{\pm 0.21}$	**4.97**$_{\pm 0.30}$	$25.54_{\pm 0.19}$
	10	1.87	$5.55_{\pm 0.19}$	$7.09_{\pm 0.09}$	-	-	-	$6.40_{\pm 0.28}$	**8.23**$_{\pm 0.24}$	
	20	3.74	$8.00_{\pm 0.14}$	$9.64_{\pm 0.14}$	-	-	-	$8.64_{\pm 0.14}$	**9.85**$_{\pm 0.34}$	
MUSIC-21	1	0.014	$24.12_{\pm 4.33}$	$26.15_{\pm 2.01}$	$22.60_{\pm 1.13}$	$22.98_{\pm 1.16}$	$28.71_{\pm 1.23}$	$38.26_{\pm 1.32}$	**44.02**$_{\pm 2.21}$	$85.93_{\pm 0.084}$
	10	0.14	$45.77_{\pm 1.74}$	$51.89_{\pm 1.39}$	-	-	$42.25_{\pm 1.07}$	$54.78_{\pm 1.39}$	**68.07**$_{\pm 0.98}$	
	20	0.28	$54.86_{\pm 1.85}$	$59.98_{\pm 0.85}$	-	-	-	$61.06_{\pm 1.31}$	**70.30**$_{\pm 0.69}$	
AVE	1	0.10	$10.07_{\pm 1.16}$	$11.84_{\pm 0.4}$	$10.45_{\pm 0.39}$	$10.76_{\pm 0.62}$	$12.13_{\pm 0.41}$	$21.70_{\pm 1.46}$	**23.00**$_{\pm 1.37}$	$52.20_{\pm 0.38}$
	10	1.0	$20.0_{\pm 1.45}$	$26.86_{\pm 0.52}$	-	-	$23.15_{\pm 0.95}$	$28.14_{\pm 1.80}$	**36.82**$_{\pm 0.88}$	
	20	2.0	$26.32_{\pm 1.01}$	$33.04_{\pm 0.38}$	-	-	-	$32.57_{\pm 0.97}$	**40.13**$_{\pm 1.00}$	

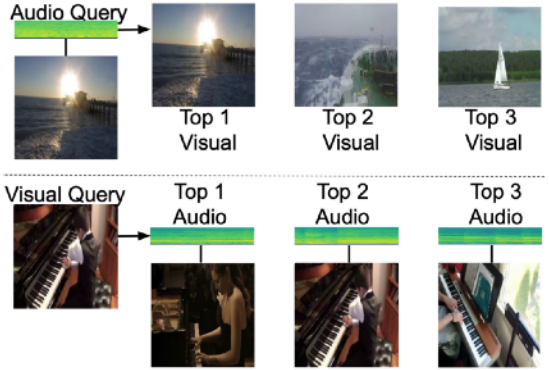
Visualization





Cross-modal retrieval

	Method	VGGS-10k test subset			AVE test subset		
		R@1↑	R@5↑	MedR↓	R@1↑	R@5↑	MedR↓
A→V	Random	$13.33_{\pm 5.03}$	$52.00_{\pm 14.00}$	$5.83_{\pm 1.75}$	$7.62_{\pm 3.21}$	$30.23_{\pm 4.06}$	$12.33_{\pm 2.30}$
	DM	$8.66_{\pm 1.15}$	$47.33_{\pm 5.77}$	$6.66_{\pm 1.52}$	$6.90_{\pm 2.29}$	$32.14_{\pm 0.71}$	$11.16_{\pm 1.75}$
	Ours	$\mathbf{19.33}_{\pm 2.30}$	$\mathbf{59.33}_{\pm 1.15}$	$\mathbf{3.66}_{\pm 0.57}$	$\mathbf{13.09}_{\pm 2.88}$	$\mathbf{35.00}_{\pm 1.88}$	$\mathbf{9.00}_{\pm 0.00}$
	Whole data	$44.00_{\pm 2.00}$	$74.00_{\pm 5.03}$	$2.00_{\pm 0.00}$	$27.61_{\pm 5.35}$	$51.66_{\pm 4.06}$	$4.66_{\pm 1.15}$
V→A	Random	$10.66_{\pm 2.30}$	$49.33_{\pm 5.77}$	$6.00_{\pm 0.86}$	$9.04_{\pm 1.48}$	$26.66_{\pm 2.29}$	$16.00_{\pm 2.00}$
	DM	$11.33_{\pm 3.05}$	$44.00_{\pm 4.00}$	$6.66_{\pm 1.15}$	$\mathbf{10.95}_{\pm 3.59}$	$29.52_{\pm 3.52}$	$14.33_{\pm 3.25}$
	Ours	$\mathbf{27.33}_{\pm 2.30}$	$\mathbf{59.33}_{\pm 7.02}$	$\mathbf{3.83}_{\pm 0.57}$	$6.43_{\pm 3.11}$	$\mathbf{34.52}_{\pm 3.30}$	$\mathbf{10.16}_{\pm 1.60}$
	Whole data	$45.33_{\pm 5.03}$	$76.00_{\pm 2.00}$	$1.83_{\pm 0.28}$	$17.14_{\pm 0.71}$	$44.76_{\pm 1.79}$	$7.16_{\pm 0.288}$



Conclusion

- New problem: Audio-Visual dataset distillation
- Audio-visual integration still hold for synthetic data
- Need carefully designed cross-modal alignment losses
- Extensive experiments on Audio-visual recognition and retrieval tasks
- Future Work and Limitation:
  - Extend to longer videos
  - Reduce gap with whole data
  - Extend to Instance-wise distillation

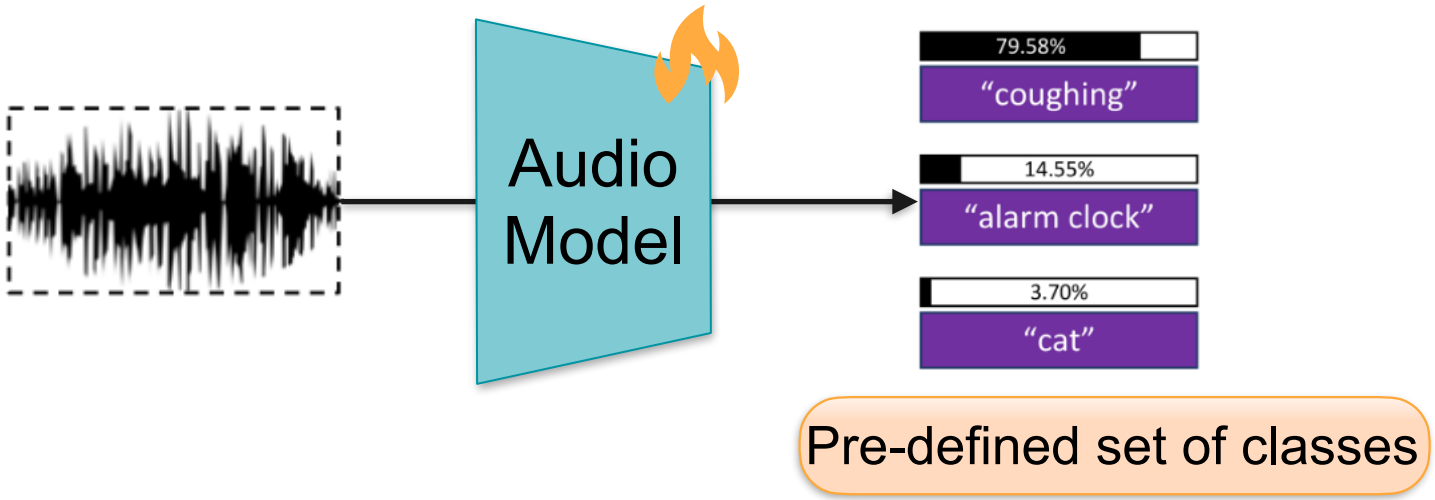
Multimodal Prototypical approach for Unsupervised sound classification

INTERSPEECH 2023

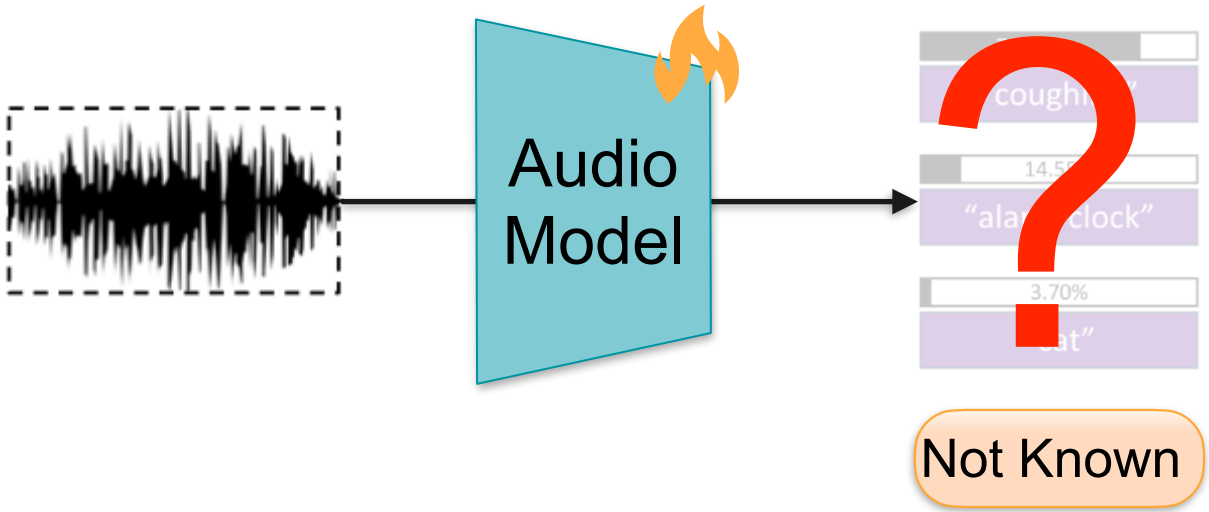
Saksham Singh Kushwaha
MSCS, NYU

Magdalena Fuentues
Assistant Prof. , NYU Tandon

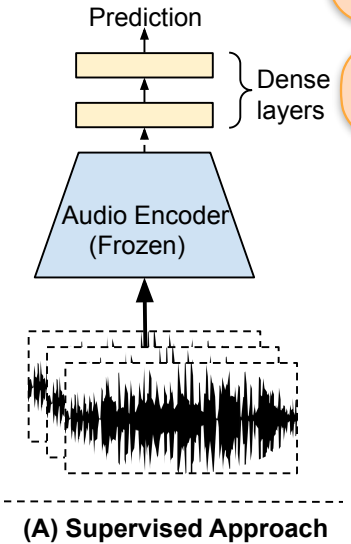
Audio classification



Audio classification



Previous approaches

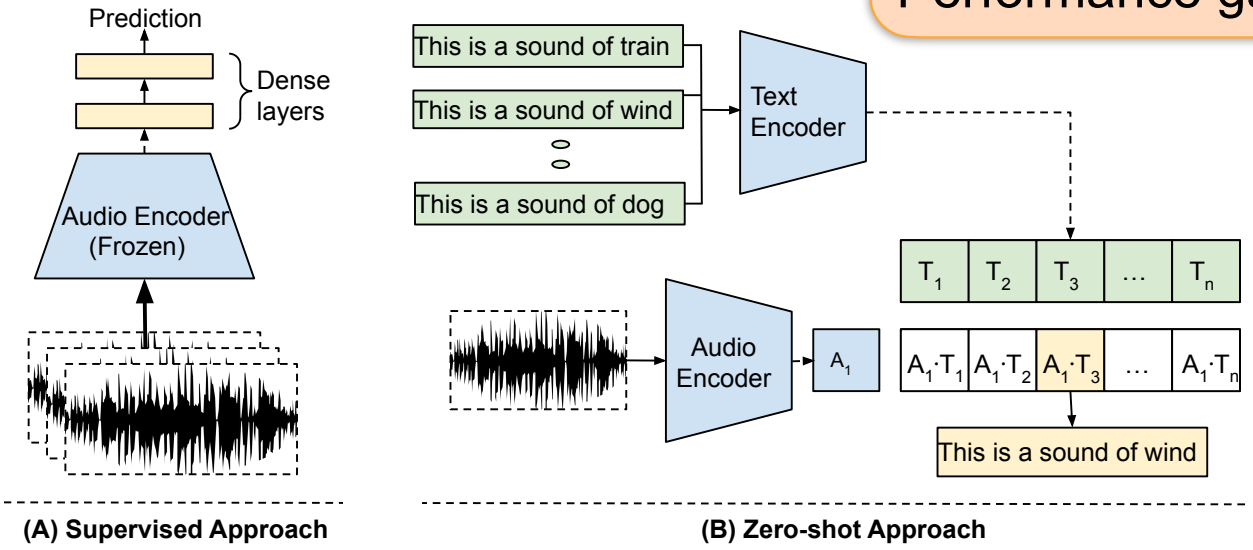


(A) Supervised Approach

Requires training ✗

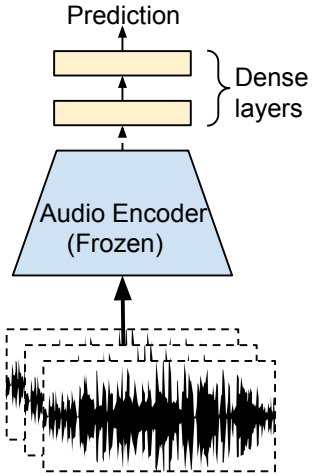
Labelled dataset ✗

Previous approaches

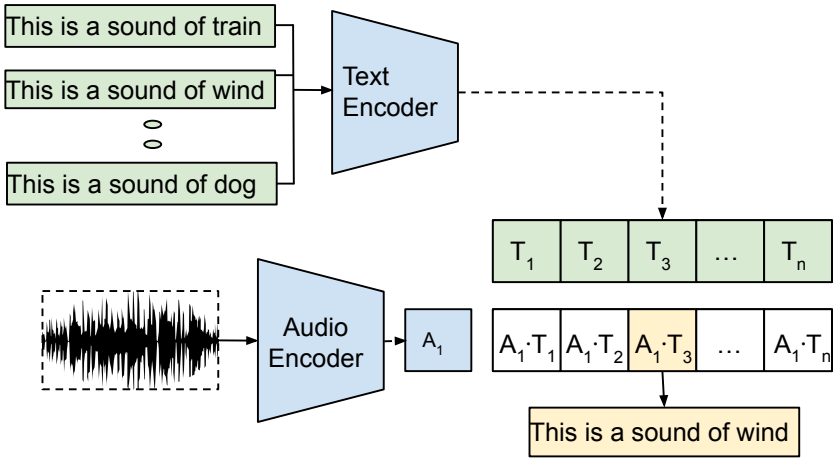


Performance gap with supervised ✗

Previous approaches

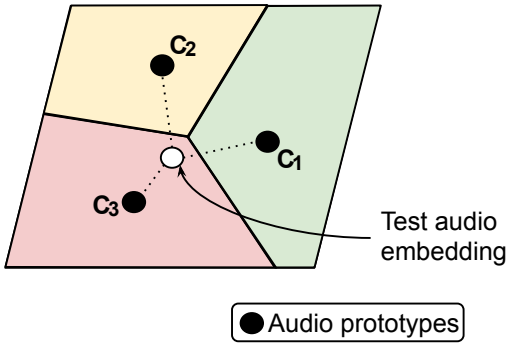


(A) Supervised Approach



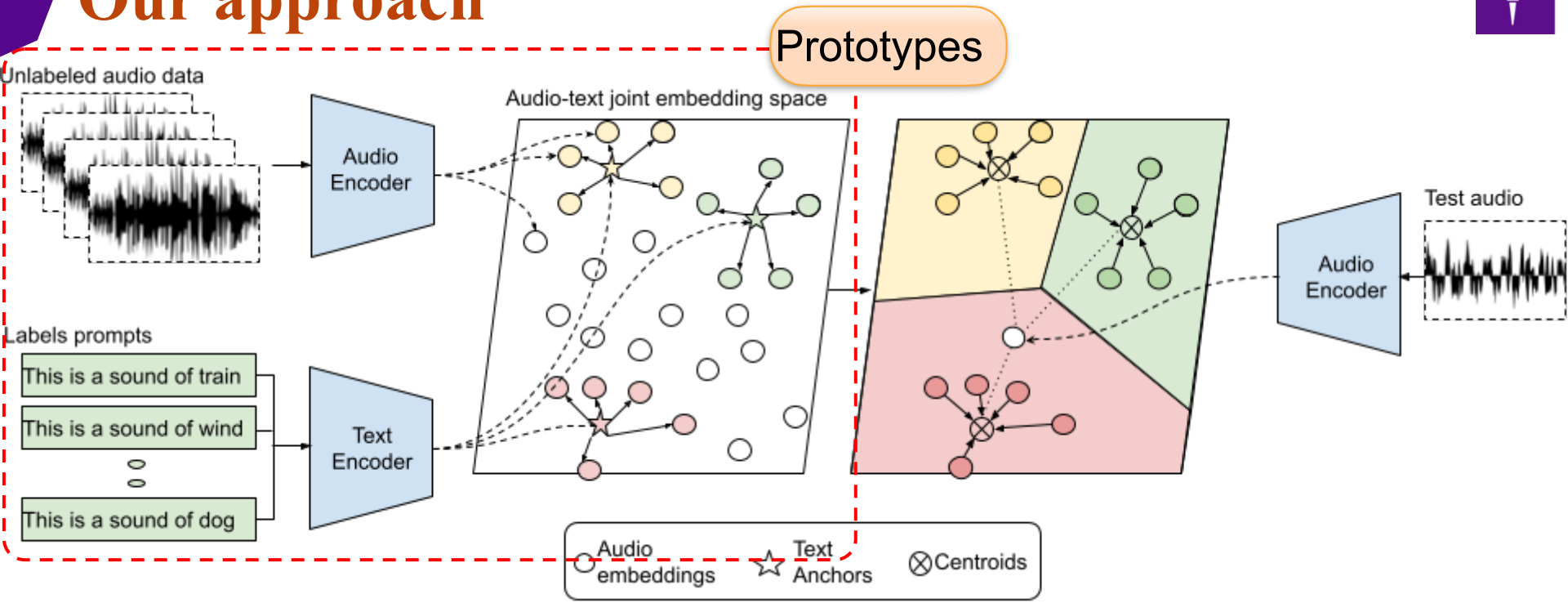
(B) Zero-shot Approach

Labelled data ✗

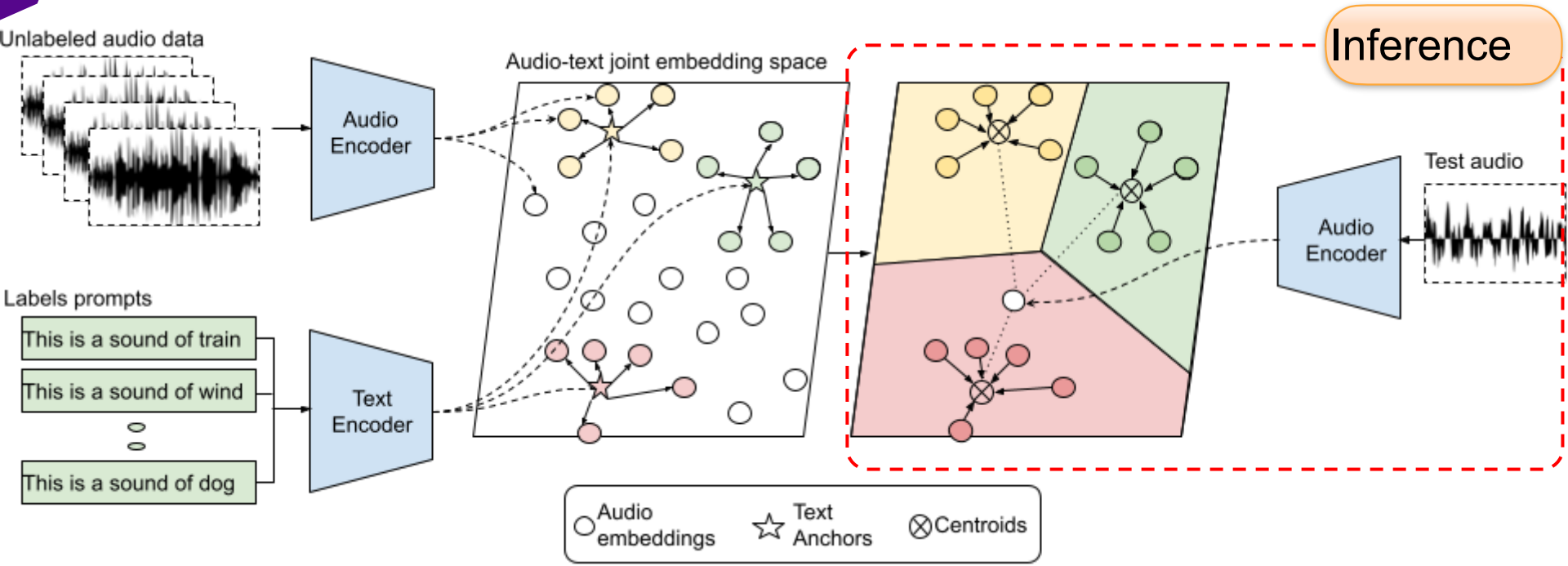


(C) Prototypical Approach

Our approach



Our approach

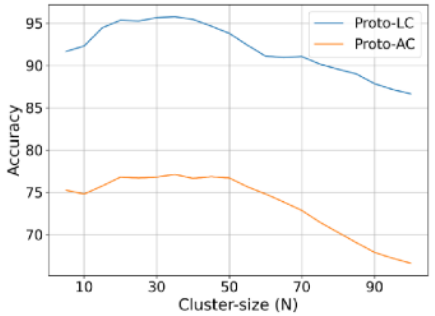


Results

	ESC-50 (acc)		US8k (acc)		FSD50K (mAP)	
	Zero Shot	Supervised	Zero Shot	Supervised	Zero Shot	Supervised
Wav2Clip	0.41	0.86	0.40	0.81	0.03	0.43
AudioClip	0.68	0.88	0.62	0.86	0.20	0.50
CLAP	0.83	**0.97**	**0.73**	0.88	0.30	0.59
LAION-CLAP	0.91	0.96	0.72	**0.89**	0.22	0.61
Proto-AC(Ours)	0.78	0.82	0.71	0.77	0.40	0.48
Proto-LC(Ours)	**0.96**	**0.97**	**0.73**	0.83	**0.52**	**0.65**

Ablation Studies

- Cluster-size vs. accuracy
- Prompt template vs. accuracy



	AudioClip	LAION-CLAP	Proto-AC	Proto-LC
Prompt 1	0.67	0.83	0.77	0.94
Prompt 2	0.68	0.86	0.72	0.92
Prompt 3	**0.69**	0.90	0.72	**0.96**
Prompt 4	0.68	0.88	**0.78**	0.95
Prompt 5	0.67	**0.92**	0.70	**0.96**

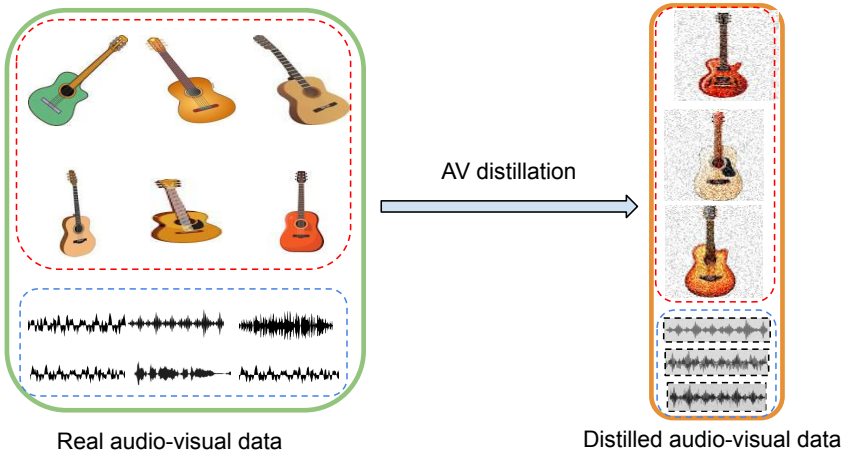
Table 2: *Accuracy on ESC-50 with different prompts. Prompt 1: '{Class label}', Prompt 2: 'I can hear {class label}', Prompt 3: 'This is an audio of {class label}', Prompt 4: 'This is {class label}', Prompt 5: 'This is a sound of {class label}'.*

Conclusion

- Personalised audio-classification without human intervention
- Outperform previous zero-shot approach
- Test on different pretrained models and single/multi-label datasets

Future research directions

- Multimodal data compression/distillation
  - Current: Class-wise distillation



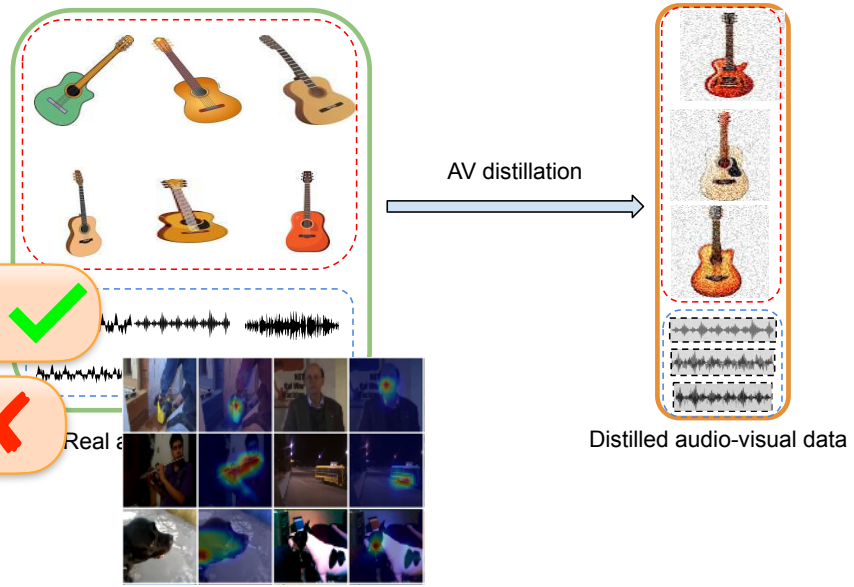
Future research directions

- Multimodal data compression/distillation
  - Current: Class-wise distillation

Classification ✓

Cross-modal retrieval ✓

Source localization ✗



Future research directions

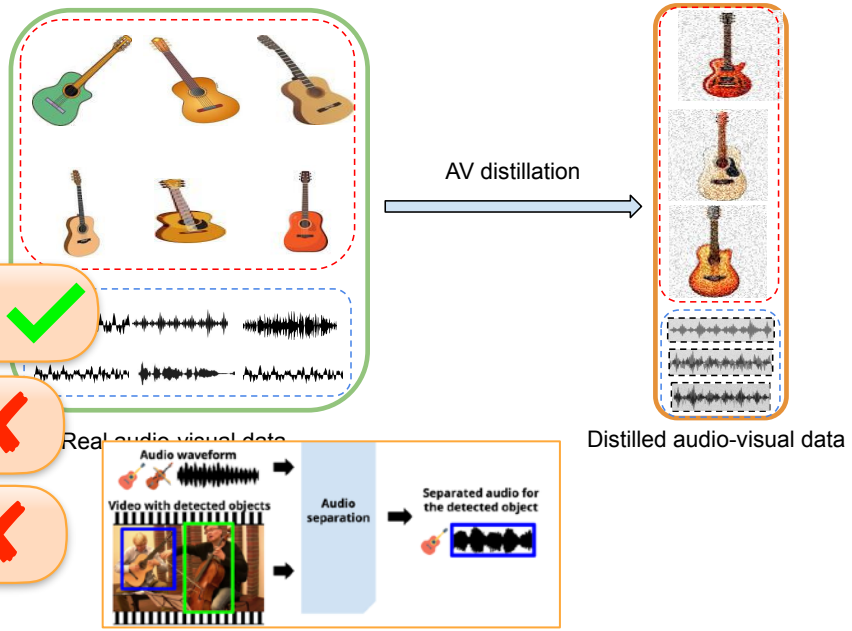
- Multimodal data compression/distillation
  - Current: Class-wise distillation

Classification ✓

Cross-modal retrieval ✓

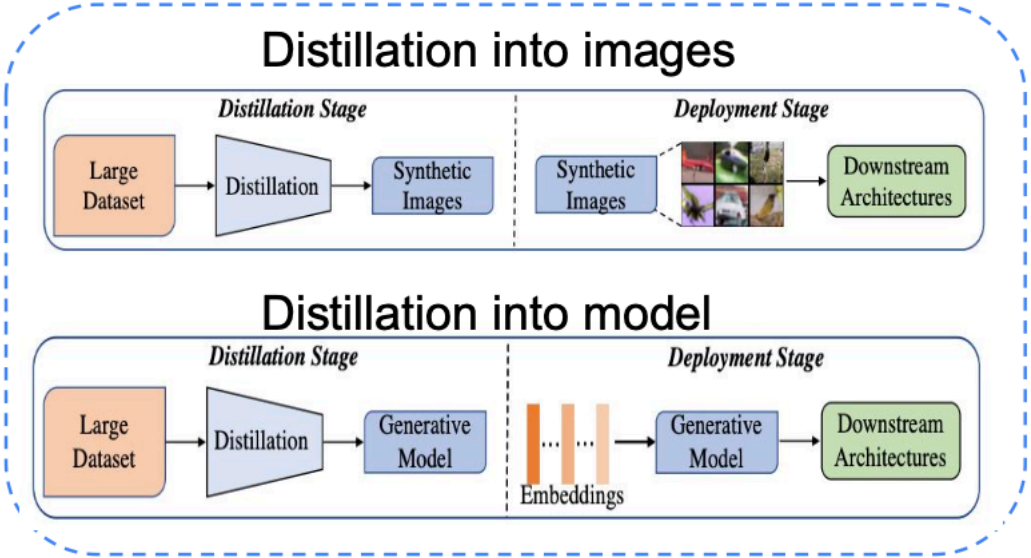
Source localization ✗

Source Separation ✗



Future research directions

- Multimodal data compression/distillation
  - Current: Class-wise distillation
  - Distillation into model



Future research directions

- Spatial audio generation
  - Text/360° videos to spatial audio generation
  - Limited data
  - Computational requirements

Thank You

- Questions?