# Dataset Distillation for Audio-Visual Datasets

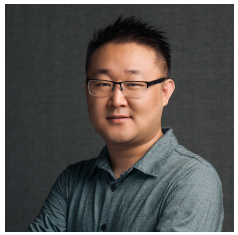
## (CVPR'24 Sight and Sound Workshop)

**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**

# What is dataset distillation?

- Synthesize tiny and high-fidelity data
- Summarises most important knowledge from target dataset



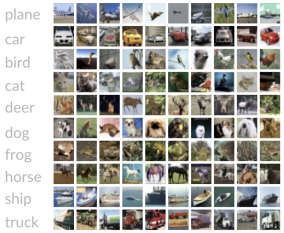
[1] Wang et. al, "Dataset Distillation", arXiv preprint, 2018.
[2] B. Zhao et. al, "Dataset Condensation with Distribution Matching", WACV, 2023.

# What is dataset distillation?

- Comparable performance on unseen real test data

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

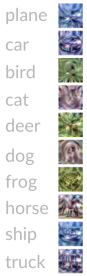
Real training data




**Dataset Distillation**

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

Synthetic data



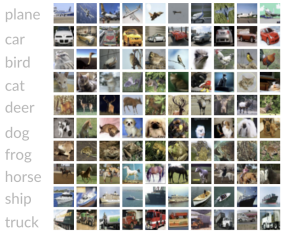

[1] Wang et. al, "Dataset Distillation", arXiv preprint, 2018.

# What is dataset distillation?

- Comparable performance on unseen real test data

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

Real training data



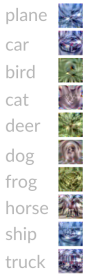

Train

$\phi_{\theta_T}$

**Dataset Distillation**

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

Synthetic data




Train

$\phi_{\theta_S}$

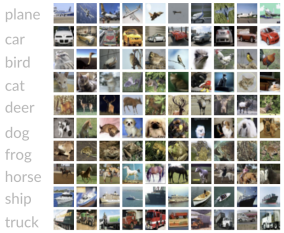
[1] Wang et. al, "Dataset Distillation", arXiv preprint, 2018.

# What is dataset distillation?

- Comparable performance on unseen real test data

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

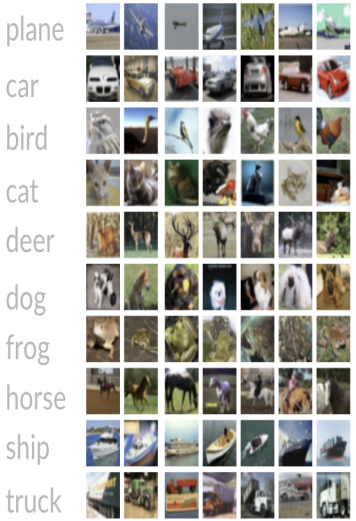
Real training data




Train

$\phi_{\theta_T}$

Test



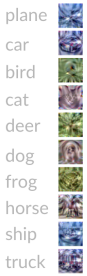

$\mathbb{E}_{x \sim P_D}[\ell(\phi_{\theta_T}(x), y)]$

$\approx$

**Dataset Distillation**

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

Synthetic data




Train

$\phi_{\theta_S}$

Test

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

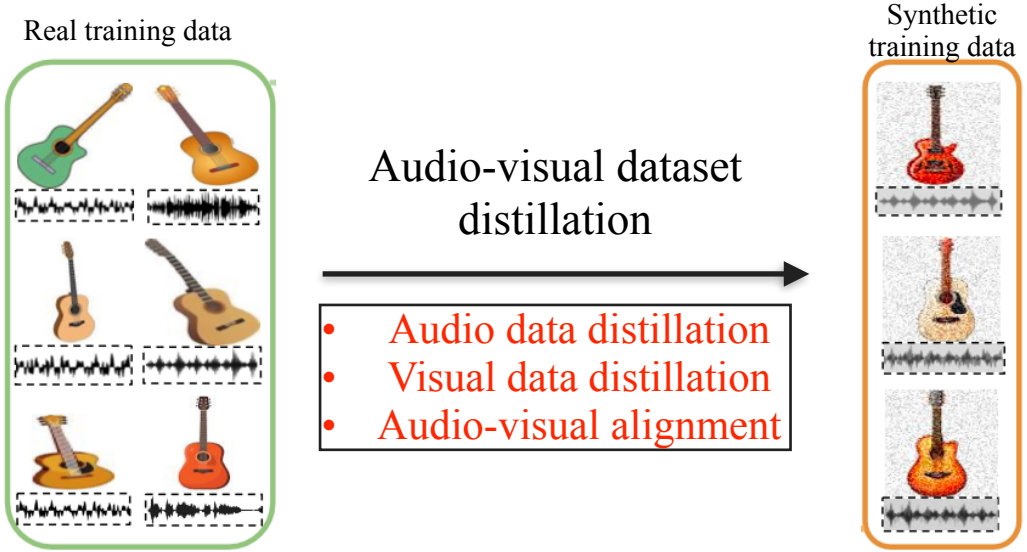
$D = \{(x_i, y_i)\}_{i=1}^{|D|}$

Real test data

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

# Audio-Visual dataset distillation

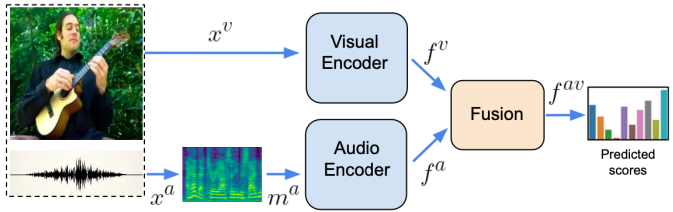
- Previous dataset distillation research focuses on image-only
- Explosion in large multimodal datasets




[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
- Experimental Setup:
  - Task: AV event recognition
  - Datasets: VGGSound-subset[1] & AVE[2]
  - Samples: 1-sec video clips

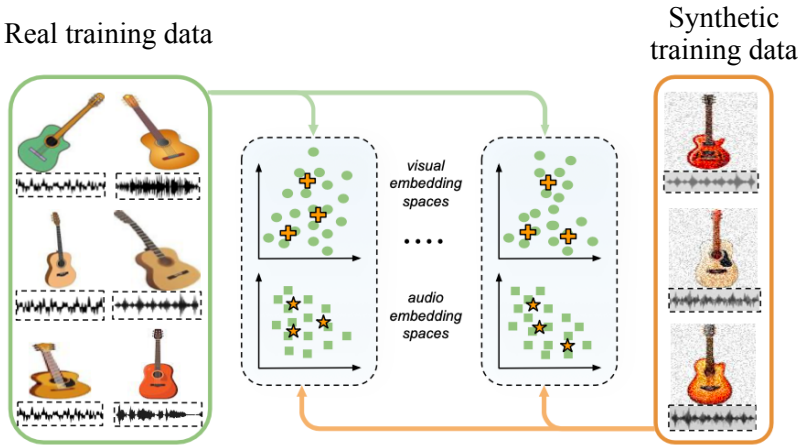



Audio Visual Event recognition

[1] H. Chen et. Al, “Vggsound: A large-scale audio-visual dataset.”, ICCASP 2020.
[2] Y. Tian et. al, "Audio-visual event localization in unconstrained videos.", ECCV 2018.

# Does AV integration hold for distilled data?

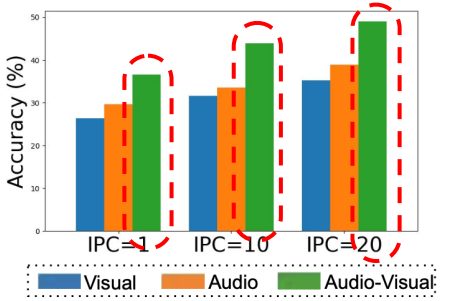
- Vanilla Audio-visual data distillation
  - Naive extension of Distribution Matching approach[1]
  - Audio-only & Visual-only data distillation




[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?

- Vanilla Audio-visual data distillation
  - Naive extension of Distribution Matching approach[1]
  - Audio-only & Visual-only data distillation
- Distilled data: Audio-visual integration helps




| | Coreset Selection | | | | | | Training Set Synthesis | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Random | | | Herding [7] | | | MTT [1] | | | DM [8] | | | Whole data | | |
| IPC | A | V | AV | A | V | AV | A | V | AV | A | V | AV | A | V | AV |
| 1 | $14.27_{\pm 0.97}$ | $11.65_{\pm 1.45}$ | **15.44**$_{\pm 1.87}$ | **26.32**$_{\pm 1.57}$ | $14.72_{\pm 2.87}$ | $20.77_{\pm 2.77}$ | $30.99_{\pm 1.48}$ | $24.15_{\pm 2.25}$ | **34.13**$_{\pm 3.62}$ | $29.60_{\pm 2.33}$ | $26.40_{\pm 1.10}$ | **36.54**$_{\pm 2.52}$ | | | |
| 10 | $32.01_{\pm 1.64}$ | $22.71_{\pm 1.57}$ | **32.50**$_{\pm 2.03}$ | $34.58_{\pm 1.98}$ | $28.9_{\pm 1.44}$ | **39.89**$_{\pm 1.64}$ | $36.57_{\pm 2.57}$ | $25.41_{\pm 1.58}$ | **36.79**$_{\pm 1.97}$ | $33.60_{\pm 1.35}$ | $31.63_{\pm 1.96}$ | **43.85**$_{\pm 1.75}$ | $62.07_{\pm 0.54}$ | $48.19_{\pm 0.54}$ | **68.24**$_{\pm 0.75}$ |
| 20 | $36.78_{\pm 2.88}$ | $31.05_{\pm 1.17}$ | **45.10**$_{\pm 2.31}$ | $44.11_{\pm 1.47}$ | $34.58_{\pm 0.84}$ | **50.20**$_{\pm 0.74}$ | $45.73_{\pm 1.03}$ | $29.52_{\pm 1.43}$ | **51.87**$_{\pm 1.26}$ | $38.93_{\pm 3.52}$ | $35.23_{\pm 1.16}$ | **49.01**$_{\pm 2.44}$ | | | |


[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?

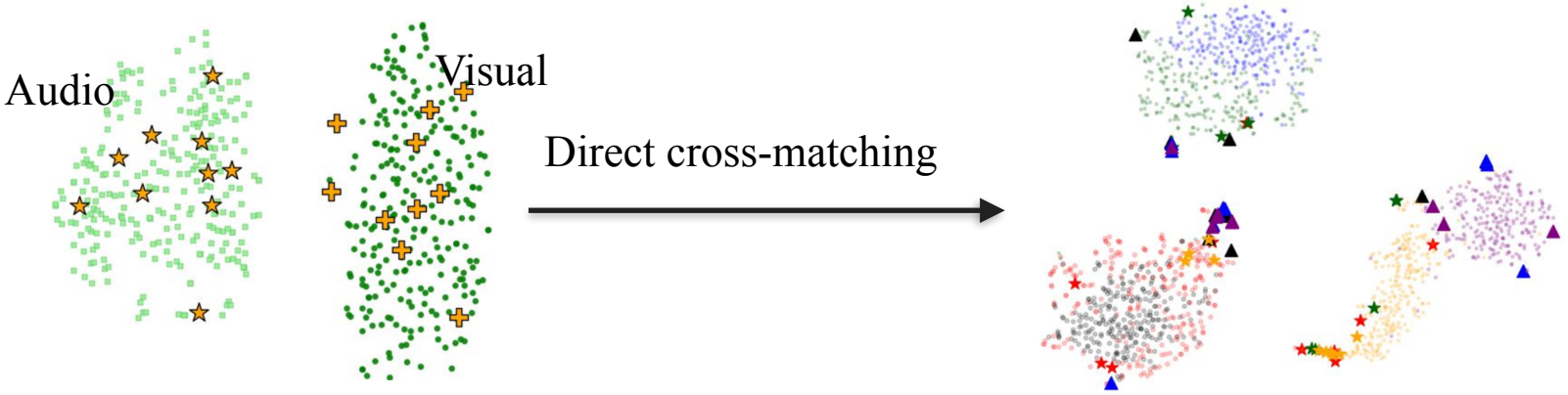
- Vanilla Audio-visual data distillation
  - Naive extension of Distribution Matching approach[1]
  - Audio-only & Visual-only data distillation
- Distilled data: Audio-visual integration helps
- Multimodal Fusion: Ensemble outperforms

| | | | Audio-Visual Fusion | | | |
|---|---|---|---|---|---|---|
| | Only-A | Only-V | 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 | $38.93_{\pm 3.52}$ | $35.23_{\pm 1.16}$ | $46.59_{\pm 1.34}$ | $46.05_{\pm 1.74}$ | $11.10_{\pm 1.88}$ | $\mathbf{49.01}_{\pm 2.44}$ |

[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

# Distilling cross-modal alignment.

- Vanilla approach separately distills audio-only and visual-only data
- Simple/direct cross-matching results in unstable learning
  - Modality Gap due to random feature extractors in DM method




Modality Gap shown in intermediate stage of vanilla approach

Final Distribution

# Distilling cross-modal alignment.

- Vanilla approach separately distills audio-only and visual-only data
- Simple/direct cross-matching results in unstable learning
  - Modality Gap due to random feature extractors in DM method
- Novel cross-matching losses
  - Joint Matching loss
  - Modality Gap Matching loss

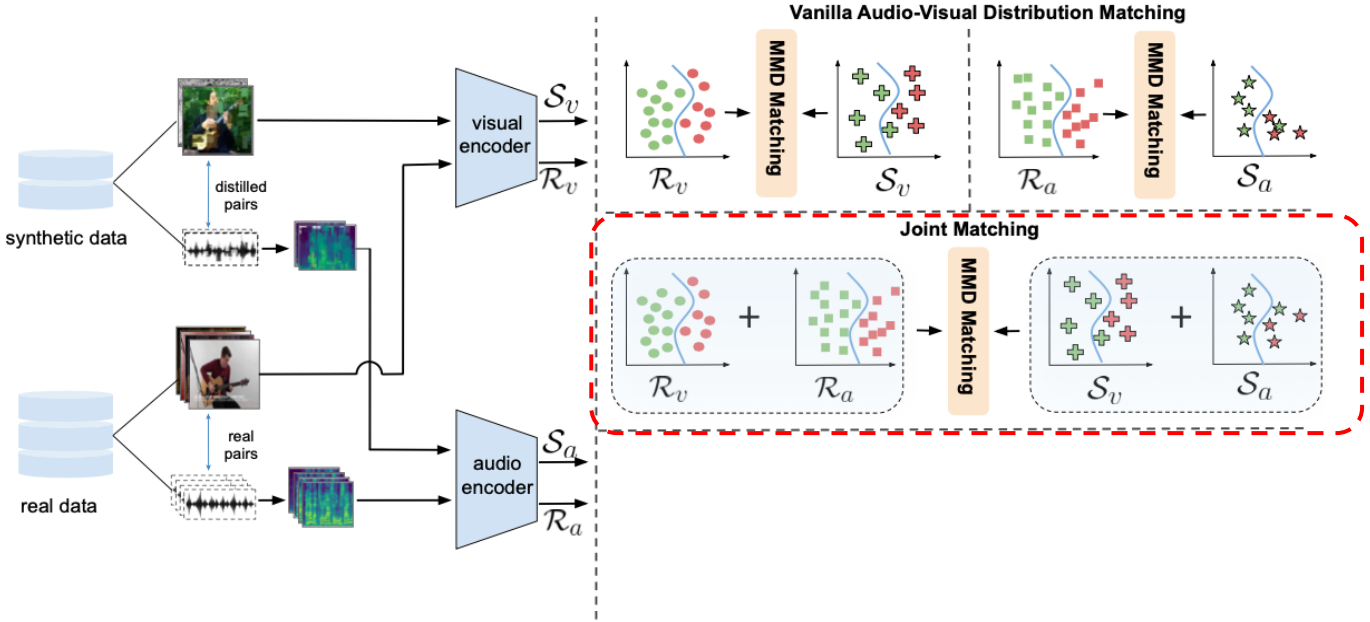
# Distilling cross-modal alignment.

- Vanilla approach separately distills audio-only and visual-only data
- Simple/direct cross-matching results in unstable learning
  - Modality Gap due to random feature extractors in DM method
- Novel cross-matching losses
  - Joint Matching loss
  - Modality Gap Matching loss
- Additional improvements
  - Herding-based initialisation[1]
  - Factor technique[2]

[1] Max Welling, “Herding dynamical weights to learn", ICML 2009.
[2] G. Zhao et. Al., “Improved distribution matching", CVPR 2023.

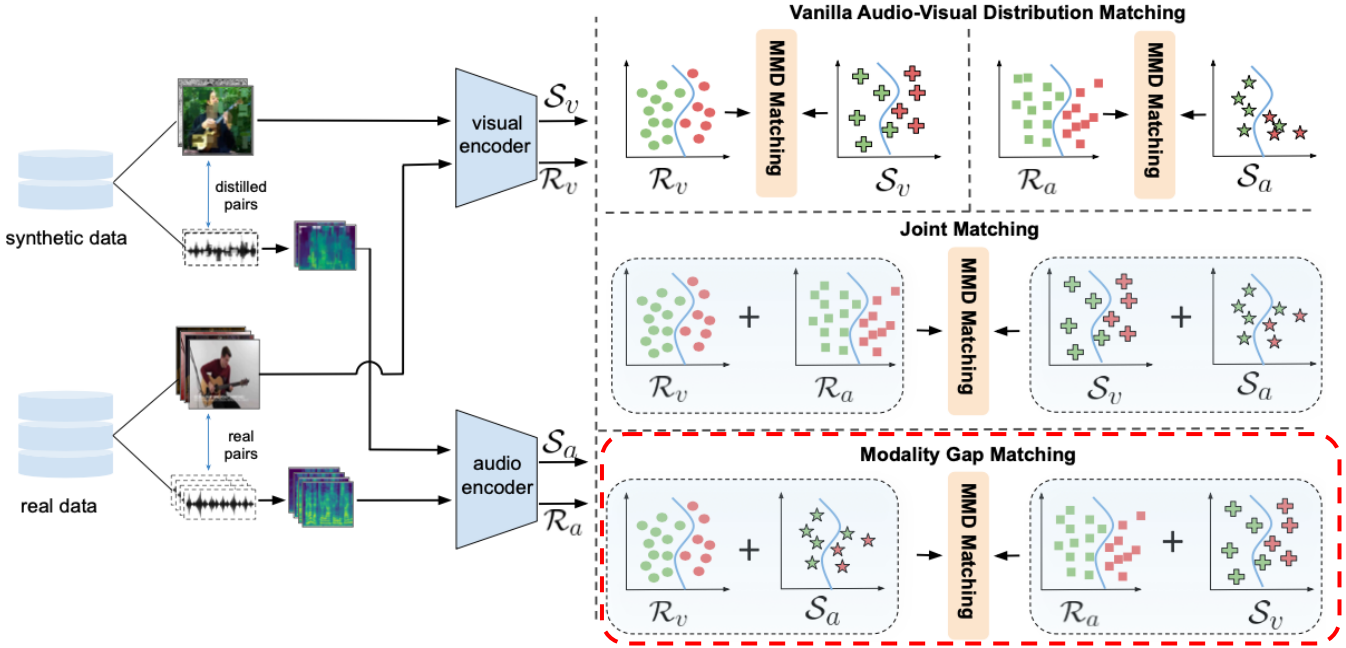
# Joint Matching Loss

- Implicitly distills cross-modal alignment



# Modality Gap Matching Loss

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



# Audio-visual event recognition

- Comparison with other baseline approaches
- Our approach performs better at different configurations
- Benefit of cross-modal alignment

| | IPC | Ratio% | Coreset Selection | | Training Set Synthesis | | | Whole data |
|---|---|---|---|---|---|---|---|---|
| | | | Random | Herding [7] | MTT [1] | DM [8] | **Ours** | |
| VGGS-10K | 1 | 0.11 | $15.44_{\pm 1.87}$ | $20.77_{\pm 2.11}$ | $34.13_{\pm 3.62}$ | $36.54_{\pm 2.52}$ | $\mathbf{40.41}_{\pm 1.81}$ | $68.24_{\pm 0.75}$ |
| | 10 | 1.13 | $32.01_{\pm 1.64}$ | $39.89_{\pm 1.64}$ | $36.79_{\pm 1.97}$ | $43.85_{\pm 1.75}$ | $\mathbf{54.99}_{\pm 1.73}$ | |
| | 20 | 2.26 | $45.1_{\pm 2.31}$ | $50.2_{\pm 0.74}$ | $51.87_{\pm 1.26}$ | $49.01_{\pm 2.44}$ | $\mathbf{58.04}_{\pm 1.68}$ | |
| AVE | 1 | 0.10 | $10.07_{\pm 1.16}$ | $11.84_{\pm 0.4}$ | $12.13_{\pm 0.41}$ | $21.70_{\pm 1.46}$ | $\mathbf{23.00}_{\pm 1.37}$ | $52.20_{\pm 0.38}$ |
| | 10 | 1.0 | $20.0_{\pm 1.45}$ | $26.86_{\pm 0.52}$ | $23.15_{\pm 0.95}$ | $28.14_{\pm 1.80}$ | $\mathbf{36.82}_{\pm 0.88}$ | |
| | 20 | 2.0 | $26.32_{\pm 1.01}$ | $33.04_{\pm 0.38}$ | - | $32.57_{\pm 0.97}$ | $\mathbf{40.13}_{\pm 1.00}$ | |

# Further experiments

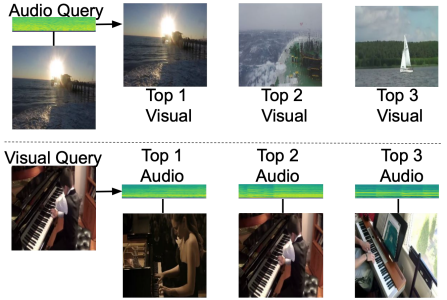
- Ablation study

| Random | Herding | Factor | Base | JM | MGM | VGGS-10k | AVE |
|---|---|---|---|---|---|---|---|
| ✓ | | | | | | $32.01_{\pm 1.64}$ | $20.00_{\pm 1.45}$ |
| | ✓ | | | | | $39.89_{\pm 1.64}$ | $26.86_{\pm 0.52}$ |
| | ✓ | ✓ | | | | $40.28_{\pm 2.34}$ | $31.80_{\pm 1.28}$ |
| | ✓ | ✓ | ✓ | | | $45.31_{\pm 2.68}$ | $34.80_{\pm 1.68}$ |
| | ✓ | ✓ | ✓ | ✓ | | $49.07_{\pm 1.97}$ | $35.13_{\pm 1.14}$ |
| | ✓ | ✓ | ✓ | ✓ | ✓ | $\mathbf{54.99}_{\pm 1.73}$ | $\mathbf{36.82}_{\pm 0.88}$ |

# Further experiments

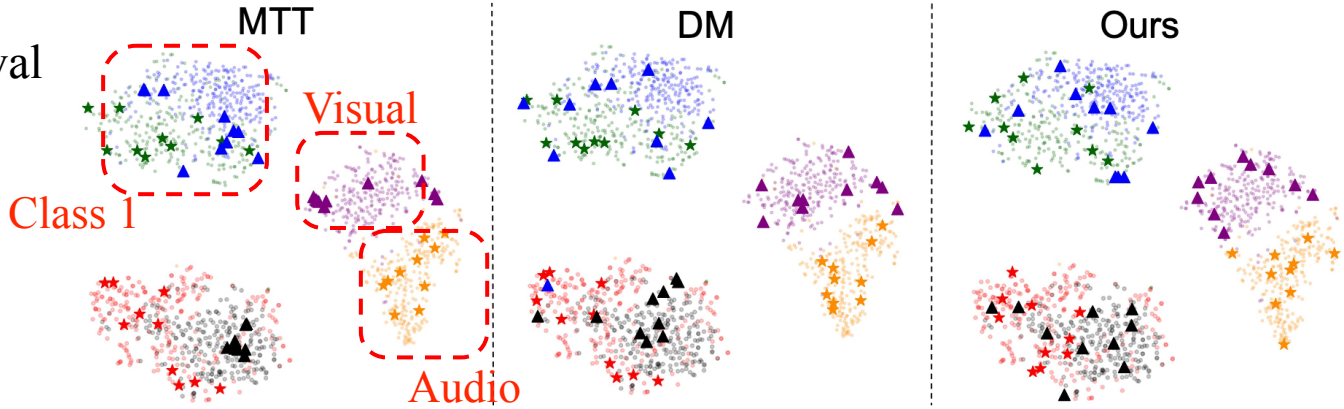
- Ablation study
- Cross-modal retrieval

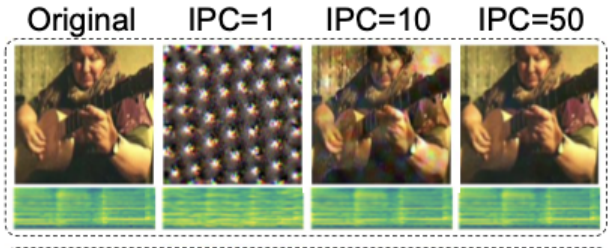
| | Method | VGGS-10k test subset | | | AVE test subset | | |
|---|---|---|---|---|---|---|---|
| | | R@1↑ | R@5↑ | MedR↓ | R@1↑ | R@5↑ | MedR↓ |
| A→V | Random | $13.33_{\pm 5.03}$ | $52.00_{\pm 14.00}$ | $5.83_{\pm 1.75}$ | $7.62_{\pm 3.21}$ | $30.23_{\pm 4.06}$ | $12.33_{\pm 2.30}$ |
| | DM[8] | $8.66_{\pm 1.15}$ | $47.33_{\pm 5.77}$ | $6.66_{\pm 1.52}$ | $6.90_{\pm 2.29}$ | $32.14_{\pm 0.71}$ | $11.16_{\pm 1.75}$ |
| | **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[8] | $11.33_{\pm 3.05}$ | $44.00_{\pm 4.00}$ | $6.66_{\pm 1.15}$ | $\mathbf{10.95}_{\pm 3.59}$ | $29.52_{\pm 3.52}$ | $14.33_{\pm 3.25}$ |
| | **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 | $45.33_{\pm 5.03}$ | $76.00_{\pm 2.00}$ | $1.83_{\pm 0.28}$ | $17.14_{\pm 0.71}$ | $44.76_{\pm 1.79}$ | $7.16_{\pm 0.288}$ |



# Further experiments

- Ablation study
- Cross-modal retrieval
- Visualisations
  - Distribution
  - Distilled data



# 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