Dataset distillation for audio-visual datasets

Saksham Singh Kushwaha[1] Siva Sai Nagender Vasireddy[1] Kai Wang[2] Yapeng Tian[1]
[1] The University of Texas at Dallas [2] National University of Singapore
[1]{sakshamsingh.kushwaha, sivasainagender.vasireddy, yapeng.tian}@utdallas.edu, [2]kai.wang@comp.nus.edu.sg

Data distillation aims to learn a condensed dataset such that it retains most of the essential information of the entire training data. Recent progress in data distillation techniques [1, 8, 9] have achieved remarkable performance on the image datasets but their potential in other domains remains largely underexplored. With the recent advancements in audio-visual learning [5], the size of audio-visual datasets [2, 5] has significantly increased, which leads to the heavy storage and computational cost of training on these datasets. In this work, we investigate the extension of data distillation to the audio-visual domain. Unlike image distillation [1, 8], audio-visual distillation presents unique challenges: preserving complex cross-modal correlations and addressing unique complexities due to high-resolution images and the additional audio modality.

To explore this new problem, we first extend the distribution matching [8] approach to separately train visual-only and audio-only distilled data. This helps us to demonstrate the advantage of audio-visual integration with distilled data and analyze the impact of different multimodal fusion methods on audio-visual event recognition performance. We then introduce two novel matching losses: joint matching loss and modality gap matching loss to distill the alignment between the condensed (synthetic) data from the original data. Additionally, we enhance the initialization process and storage for the synthetic data. Comprehensive experiments on recognition and cross-modal retrieval tasks demonstrate the representativeness and audio-visual alignment of our distilled data.

1. Method

Problem Formulation. Let x_i^a and x_i^v denote the audio waveform and video frame of the i-th sample, respectively, with $x_i^{av} = (x_i^a, x_i^v)$ and y_i as the corresponding ground truth category label. Given a large audio-visual training set $\mathcal{T} = \{x_i^{av}, y_i\}_{i=1}^{|\mathcal{T}|}$, our audio-visual data distillation task aims to learn a smaller, yet representative synthetic set $\mathcal{S} = \{s_i^{av}, y_i\}_{i=1}^{|\mathcal{S}|}$, where $s_i^{av} = (s_i^a, s_i^v)$. This dataset $\mathcal{S}$, with significantly fewer samples $|\mathcal{S}| \ll |\mathcal{T}|$, should encapsulate the essential information contained in $\mathcal{T}$. The goal is for models trained on each $\mathcal{T}$ and $\mathcal{S}$ to perform similarly on unseen test data:

$$\mathbb{E}_{(x^{av},y)\sim\mathcal{D}}\left[\ell(\phi_{\theta_\mathcal{T}}(x^{av}), y)\right] \simeq \mathbb{E}_{(x^{av},y)\sim\mathcal{D}}\left[\ell(\phi_{\theta_\mathcal{S}}(x^{av}), y)\right],$$

where $\mathcal{D}$ is the real test data, ℓ is the loss function (i.e. cross-entropy), ϕ_θ is a neural network parameterized by θ, and $\phi_{\theta_\mathcal{T}}$ and $\phi_{\theta_\mathcal{S}}$ are networks trained on $\mathcal{T}$ and $\mathcal{S}$ respectively.

1.1. Audio-Visual Data Distillation

Our overall approach is illustrated in Fig. 1.

Vanilla Audio-Visual Distribution Matching. Distribution matching (DM) [8] generates synthetic data by minimizing the feature distance between the distributions of real and synthetic samples. Specifically, DM minimises the visual-only loss $\mathcal{L}_{base}^v$,

$$\mathcal{L}_{base}^v = ||\frac{1}{|\mathcal{T}_v|}\sum_{i=1}^{|\mathcal{T}_v|}\psi_{\theta_v}(\mathcal{A}_\omega(\mathbf{x}_i^v)) - \frac{1}{|\mathcal{S}_v|}\sum_{j=1}^{|\mathcal{S}_v|}\psi_{\theta_v}(\mathcal{A}_\omega(\mathbf{s}_j^v))||^2,$$

where ψ_{θ_v} denotes randomly initialized visual network and $\mathcal{A}_\omega(\cdot)$ is differential siamese augmentation. The vanilla approach to audio-visual distribution matching extends this visual-only loss to optimize the following objective:

$$\mathcal{L}_{base}^{av} = \mathcal{L}_{base}^a + \mathcal{L}_{base}^v. \tag{1}$$

Where, $\mathcal{L}_{base}^a$ is the simple adaption of DM to audio-modality. A simple extension of vanilla DM for cross-modal alignment is restricted due to the modality gap created by the use of randomly initialized feature networks $\psi_{\theta_v}, \psi_{\theta_a}$. Hence we propose joint matching and modality gap matching losses.

Joint Matching (JM). This loss function effectively aligns the joint audio-visual distributions between real ($\mathcal{D}^r$) and synthetic ($\mathcal{D}^s$) data, enabling implicit cross-modal distribution matching. The loss is formally defined for each class as follows:

$$\mathcal{D}^r = \bar{R}^a + \bar{R}^v = \left[\frac{1}{|\mathcal{T}_a|}\sum_{i=1}^{|\mathcal{T}_a|}\psi_{\theta_a}(\mathcal{A}_\omega(\mathbf{x}_i^a)) + \frac{1}{|\mathcal{T}_v|}\sum_{i=1}^{|\mathcal{T}_v|}\psi_{\theta_v}(\mathcal{A}_\omega(\mathbf{x}_i^v))\right],$$

$$\mathcal{D}^s = \bar{S}^a + \bar{S}^v = \left[\frac{1}{|\mathcal{S}_a|}\sum_{j=1}^{|\mathcal{S}_a|}\psi_{\theta_a}(\mathcal{A}_\omega(\mathbf{s}_j^a)) + \frac{1}{|\mathcal{S}_v|}\sum_{j=1}^{|\mathcal{S}_v|}\psi_{\theta_v}(\mathcal{A}_\omega(\mathbf{s}_j^v))\right],$$

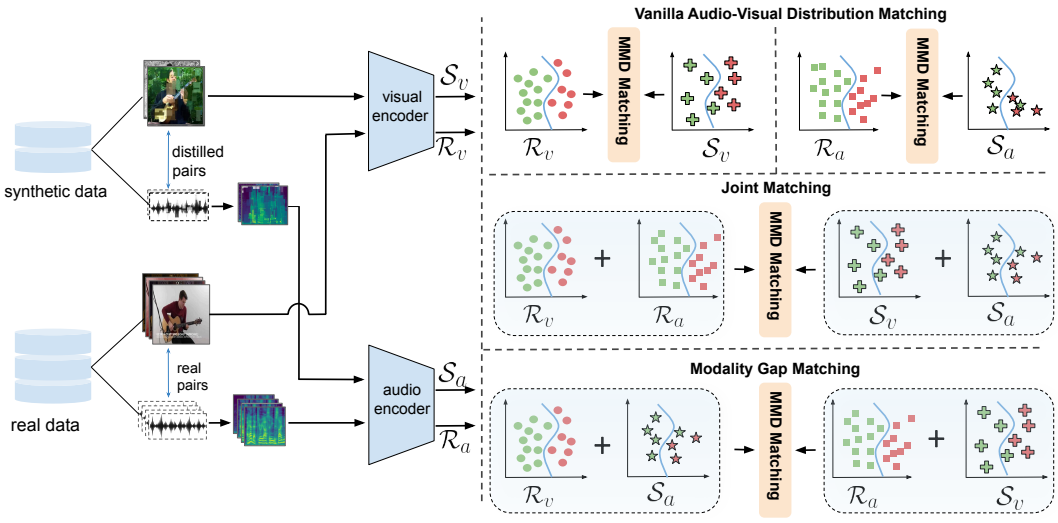

Figure 1. Overview of our audio-visual data distillation framework.

$$\mathcal{L}_{JM}^{av} = ||\mathcal{D}^r - \mathcal{D}^s||^2, \quad (2)$$

Optimizing this loss term effectively compels our model to learn synthetic audio-visual data $\{\mathcal{S}_a, \mathcal{S}_v\}$ that closely resembles and represents the real dataset $\{\mathcal{T}_a, \mathcal{T}_v\}$. Here, the loss term in Eq. 2 can be re-written as $\mathcal{L}_{JM}^{av} = ||(\bar{R}^a + \bar{R}^v) - (\bar{S}^a + \bar{S}^v)||^2 = ||(\bar{R}^a - \bar{S}^v) + (\bar{R}^v - \bar{S}^a)||^2 \leq ||(\bar{R}^a - \bar{S}^v)||^2 + ||(\bar{R}^v - \bar{S}^a)||^2$. This formulation reveals that the loss implicitly enforces cross-modal matching between real audio data and synthetic visual data, as well as between real visual data and synthetic audio data.

Modality Gap Matching (MGM). Besides the JM loss, we introduce further constraints to align the distributions of ($\bar{R}_a$ and $\bar{S}_v$) and ($\bar{R}_v$ and $\bar{S}_a$), as follows:

$$\mathcal{D}^{av} = \left[\frac{1}{|\mathcal{T}_a|}\sum_{i=1}^{|\mathcal{T}_a|}\psi_{\theta_a}(\mathcal{A}_\omega(\mathbf{x}_i^a)) + \frac{1}{|\mathcal{S}_v|}\sum_{j=1}^{|\mathcal{S}_v|}\psi_{\theta_v}(\mathcal{A}_\omega(\mathbf{s}_j^v))\right]$$

$$\mathcal{D}^{va} = \left[\frac{1}{|\mathcal{T}_v|}\sum_{i=1}^{|\mathcal{T}_v|}\psi_{\theta_v}(\mathcal{A}_\omega(\mathbf{x}_i^v)) + \frac{1}{|\mathcal{S}_a|}\sum_{j=1}^{|\mathcal{S}_a|}\psi_{\theta_a}(\mathcal{A}_\omega(\mathbf{s}_j^a))\right]$$

$$\mathcal{L}_{MGM}^{av} = ||\mathcal{D}^{av} - \mathcal{D}^{va}||^2 \quad (3)$$

This additional loss ensures that the synthetic data closely represent the corresponding real data without misaligning the existing matches of $\bar{S}^a \leftrightarrow \bar{R}^a$ and $\bar{S}^v \leftrightarrow \bar{R}^v$ enforced in unimodal DM and $(\bar{S}^a + \bar{S}^v) \leftrightarrow (\bar{R}^a + \bar{R}^v)$. With a simple re-writing, we can obtain $\mathcal{L}_{MGM}^{av} = ||(\bar{R}^a + \bar{S}^v) - (\bar{R}^v + \bar{S}^a)||^2 = ||(\bar{R}^a - \bar{R}^v) - (\bar{S}^a - \bar{S}^v)||^2$. We can see that it will help to align the modality gap between real and synthetic data to strengthen the joint audio-visual distribution matching.

Final Loss. To generate synthetic audio and visual data simultaneously, we will jointly optimize the three loss terms:

$$\mathcal{L}_{final}^{av} = \mathcal{L}_{base}^{av} + \lambda_{JM} \cdot \mathcal{L}_{JM}^{av} + \lambda_{MGM} \cdot \mathcal{L}_{MGM}^{av}. \quad (4)$$

Here λ_{JM} and λ_{MGM} are the weights for joint matching and modality gap matching losses respectively. These three loss terms work collaboratively to enhance the audio-visual data distillation process. Their combined effect ensures that the synthetic data closely corresponds to the real data, aligning unimodal, cross-modal, and modality gap distributions effectively.

Improved initialization and storage. To further improve our synthetic data, we use two techniques: herding-based initialization [7] for better alignment with real data, and factor technique [9] for increasing extracted features without extra store cost.

2. Experiments

Following the evaluation protocols from previous data distillation studies [1, 8], we use audio-visual event recognition as the main proxy task to assess the classification accuracy on held-out test data of deep networks trained from scratch on our distilled audio and visual data.

2.1. Experimental Settings

Datasets. We use VGGS-10k, a randomly selected subset of 10 classes from VGGSound [2], and AVE [5]. Each data point represents a one-second video clip, comprising a center frame(of 224x224 size) and its corresponding audio.

Baselines. We compare our approach with vanilla approach-based coreset selection methods: random, herding [7] and training set synthesis methods: MTT [1] and DM [8].

Evaluation. We report the mean recognition accuracy and standard deviation of 3 runs where the model is randomly initialized and trained for 30 epochs using the learned synthetic data. Each run consists of 5000 iterations. We evaluate over 3 images-per-class (IPC) settings and keep the training setup similar to [6] to train the audio-visual models.

Table 1. Recognition results with synthetic audio(A), visual(V), and audio-visual(AV) data on VGGS-10K. For AV we use ensemble fusion over individually learned audio and visual synthetic data.

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

Table 2. Audio-visual event recognition results for different fusion methods and images per class (IPC) on VGGS-10K. Ensemble consistently achieves the highest accuracy.

		Audio-Visual Fusion			
		Concat	Sum	Attention [5]	Ensemble
IPC	1	$33.77_{\pm 1.65}$	$34.72_{\pm 1.27}$	$9.97_{\pm 0.83}$	$\mathbf{36.54}_{\pm 2.52}$
	10	$41.71_{\pm 1.27}$	$40.49_{\pm 1.83}$	$10.11_{\pm 0.35}$	$\mathbf{43.85}_{\pm 1.75}$
	20	$46.59_{\pm 1.34}$	$46.05_{\pm 1.74}$	$11.10_{\pm 1.88}$	$\mathbf{49.01}_{\pm 2.44}$

2.2. Experimental Results

Audio-visual integration. Firstly, we compared the effect of audio-visual integration in the case of distilled data. The results, shown in Tab.1, clearly demonstrate that audio-visual integration consistently outperforms unimodal modalities in most cases.

Multimodal Fusion. Next, we investigated the effect of different audio-visual fusion strategies on the performance of models trained on distilled synthetic data. From the results in Tab. 2, we can see that the ensemble method consistently outperforms other approaches in all image-per-class settings. The comparatively low fusion results in attention fusion can be accounted for classwise alignment losses, spatial distortions (as shown in Fig. 2), and a larger number of trainable model parameters. Consequently, we employ the ensemble fusion for further experiments.

Comparison with Data Distillation Baselines. Finally, we compared the performance of audio-visual recognition with dataset distilled using ours and previous approaches. From the results in Tab. 3, we observe that our audio-visual data distillation approach consistently outperforms the other baseline approaches. A large improvement over vanilla audio-visual distillation with DM demonstrates the effectiveness of incorporating joint matching losses to strengthen cross-modal alignment. In addition, similar to previous image-only distillation methods [1], we observe diminishing returns as IPC increases. For instance, in VGGS-10k, there's a significant performance jump from 40.41% to 54.99% when moving from 1 to 10 IPC, while the improvement from 10 to 20 IPC is more modest, reaching 58.04%.

Ablation Study. To validate the contributions of two novel joint audio-visual losses and herding initialization and factor technique, we conducted an ablation study by systematic addition. The ablation study results are shown in Tab. 4, from which we can see that each of the proposed parts has a positive influence on the final result.

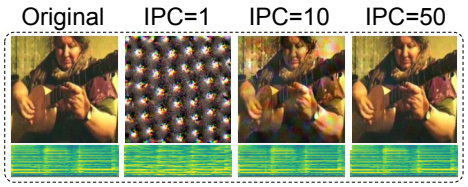

Figure 2. Visualization of distilled audio-visual data. With an IPC increase, the synthetic data gets far less away from initialization.

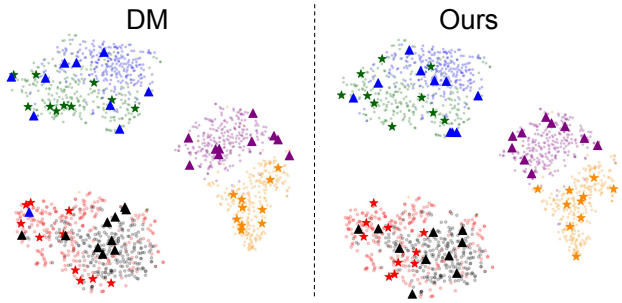

Figure 3. Distribution plot of synthetic audio-visual data (IPC=10) learned by DM, and Ours(factor=1), with same initilization. (green,blue), (red,**black**) and (purple,yellow) points are the real (audio, visual) points for the first 3 classes of VGGS-10k. The synthetic (audio, visual) data is represented by (★, ▲). We observe that Our synthetic audio and visual distributions better resemble the real distributions.

Table 4. Ablation study on the different components at IPC=10.

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

Visualization. To showcase our distilled data, we plot the learned audio and visual data in Fig. 2 at different IPCs. We observe that with increased IPC, the synthetic data remains perceptually closer to the original audio-visual sample. Additionally, we compare the data distribution of the first three classes of VGGS-10k in Fig. 3. It shows how our approach better captures the underlying distribution of real data.

2.3. Audio-Visual Retrieval

We have demonstrated that audio-visual distilled data facilitates the learning of effective audio-visual representations for audio-visual event recognition. To further examine the audio-visual alignment, we investigate whether distilled data could help learn a well-coordinated audio-visual space

Table 3. Comparison with previous data distillation methods for audio-visual event recognition. Ratio(%): the ratio of condensed images to the whole training set. Whole Data: ConvNet model trained on the whole training set and is the upper bound. '-' refers to configurations for which the method couldn't scale up.

	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	$26.32_{\pm 1.01}$	$33.04_{\pm 0.38}$	-	$32.57_{\pm 0.97}$	$\mathbf{40.13}_{\pm 1.00}$	

Table 5. Audio-visual retrieval results. Whole data is the upper bound and is trained using the entire training data. We observe that our approach helps to distill better audio-visual alignment.

	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	$45.33_{\pm 5.03}$	$76.00_{\pm 2.00}$	$1.83_{\pm 0.28}$	$17.14_{\pm 0.71}$	$44.76_{\pm 1.79}$	$7.16_{\pm 0.288}$

for audio-visual retrieval.

Since our DM-based audio-visual distillation model focuses on semantic alignment rather than instance-level alignment, we evaluate audio-visual retrieval in a class-wise setting. Following [3], we create a retrieval test set by uniformly sampling a subset of five audio-visual samples per class from the original test split. We train the same recognition-based audio and visual ConvNet architecture model with a shared classifier [4] and ArcFace margin loss. The shared classifier and margin loss help to learn a joint-modal embedding space with angular margins between classes. We train the model from scratch using the distilled data of IPC=20 and the same learning setting as the classification model. We use these trained audio and visual components to get the corresponding representation of test samples and calculate the class retrieval recall at rank 1, 5, and median rank based on the cosine similarity. The results of audio-to-visual and visual-to-audio retrieval, in Tab. 5, demonstrate that our losses help distill audio-visual alignment (from real data) and hence our method outperforms DM in almost all scenarios.

3. Conclusion

In this paper, we explore a new task of multimodal distillation using audio-visual data. To evaluate the distilled audio-visual data, we use audio-visual event recognition as the proxy task. Experimental results on two audio-visual datasets show that our proposed approach outperforms other methods consistently and audio-visual integration with condensed data is still helpful. This provides a new direction in the data distillation domain.

Acknowledgments. This work was supported in part by a Cisco Faculty Research Award, an Amazon Research Award, and a research gift from Adobe. The article solely reflects the opinions and conclusions of its authors but not the funding agents.

References

[1] G. Cazenavette, T. Wang, A. Torralba, A. A. Efros, and J.-Y. Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4750–4759, 2022. 1, 2, 3, 4

[2] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 1, 2

[3] Y. Gong, A. Rouditchenko, A. H. Liu, D. Harwath, L. Karlinsky, H. Kuehne, and J. Glass. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839*, 2022. 4

[4] D. Surís, A. Duarte, A. Salvador, J. Torres, and X. Giró-i Nieto. Cross-modal embeddings for video and audio retrieval. In *Proceedings of the european conference on computer vision (eccv) workshops*, pages 0–0, 2018. 4

[5] Y. Tian, J. Shi, B. Li, Z. Duan, and C. Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018. 1, 2, 3

[6] Y. Tian and C. Xu. Can audio-visual integration strengthen robustness under multimodal attacks? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5601–5611, 2021. 2

[7] M. Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1121–1128, 2009. 2, 3, 4

[8] B. Zhao and H. Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6514–6523, 2023. 1, 2, 3, 4

[9] G. Zhao, G. Li, Y. Qin, and Y. Yu. Improved distribution matching for dataset condensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7856–7865, 2023. 1, 2