RAST: Resolution-Aware Privileged Structure Transfer for Low-Resolution Audio Activity Recognition

Ji Hwan Park
The University of Texas at Austin
Austin, Texas, USA
jihwanpark98@utexas.edu

Gautham Krishna Gudur
The University of Texas at Austin
Austin, Texas, USA
gauthamkrishna@utexas.edu

Yufei Shen
The University of Texas at Austin
Austin, Texas, USA
shenyufei@utexas.edu

Dawei Liang
The University of Texas at Austin
Austin, Texas, USA
dawei.liang@utexas.edu

Edison Thomaz
The University of Texas at Austin
Austin, Texas, USA
ethomaz@utexas.edu

Abstract

Audio is increasingly used for human activity recognition (HAR) because it captures object interactions, environmental events, and contextual cues in everyday environments. High-resolution (HR) audio provides rich acoustic information for model development but incurs substantial energy and storage costs and may expose sensitive speech content. Low-resolution (LR) audio offers a more privacy-preserving and resource-efficient alternative for deployment, but reduced sampling rates can remove acoustic cues essential for activity recognition, leading to significant performance degradation. We formulate this training–deployment mismatch as *sensor-resolution privileged learning*, in which HR audio is available during training, while inference relies exclusively on LR audio. We propose *RAST*, a resolution-aware transfer framework that compresses HR teacher representations by preserving token-level information and neighborhood structure before performing localized HR–LR alignment. Experiments on the SAMoSA and AudioIMU datasets show that RAST consistently outperforms LR-only training and direct teacher-transfer baselines, improving LR-only recognition by up to $\sim 7.8\%$ while requiring only LR audio at inference.

CCS Concepts

• **Human-centered computing**; • **Computing methodologies** → **Machine learning**; **Learning paradigms**;

Keywords

human-centered sensing, audio sensing, privileged learning

HumanSys '26, Austin, TX, USA

ACM ISBN 979-8-4007-2959-1/2026/10
https://doi.org/10.1145/3842436.3843818

ACM Reference Format:
Ji Hwan Park, Gautham Krishna Gudur, Yufei Shen, Dawei Liang, and Edison Thomaz. 2026. RAST: Resolution-Aware Privileged Structure Transfer for Low-Resolution Audio Activity Recognition. In *The 4th International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems (HumanSys '26), October 26–30, 2026, Austin, TX, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3842436.3843818

1 Introduction

Audio is an attractive modality for human-centered sensing because everyday sounds encode rich contextual information about human behavior, object interactions, environmental context, and so on. These capabilities have motivated acoustic sensing for activities of daily living, classification from low-sampled microphone input, and wearable HAR [15, 16, 23]. However, continuous acoustic sensing imposes practical and social constraints, including substantial bandwidth, storage, computation, and energy costs [6], while always-on microphones raise concerns about privacy, trust, and unintended speech capture [3]. Recent work has therefore explored restricted-bandwidth and low-resolution audio as privacy-aware sensing alternatives, including analyses of speech intelligibility under bandwidth reduction [17]. Audio-based HAR thus presents a broader human-centered systems challenge: reliably recognizing meaningful behavior while limiting sensing cost, sensitive-data exposure, and resource consumption [4, 19].

Low-resolution (LR) audio is well suited for resource-constrained and privacy-sensitive deployment, but reducing the sampling rate can remove acoustic cues critical for distinguishing activities. High-resolution (HR) audio preserves high-frequency and transient structure associated with activities such as chopping, brushing, clapping, pouring, and appliance use whereas a deployed model may observe only the LR signal. Figure 1 illustrates this resolution gap: LR audio suppresses salient acoustic structure, and held-out

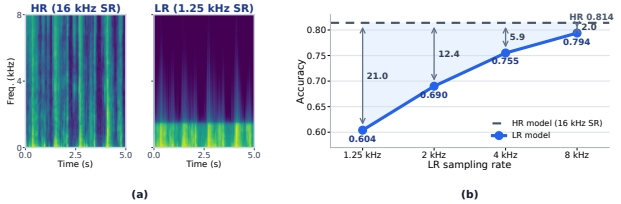

Figure 1: Resolution gap in audio-based activity recognition. (a) High-resolution (HR) and Low-resolution (LR) views of the same audio window. (b) LR accuracy decreases as the sampling rate is reduced relative to the 16-kHz HR reference.

recognition accuracy decreases as the sampling rate diverges from the 16-kHz HR reference. Prior work has shown that large-scale pretrained audio representations can improve downstream sound and activity recognition [2, 7, 12], while recent work has explored self-distillation for low-sampled microphone input [16]. Motivated by this gap, we study a *privileged-learning* setting in which HR audio is available only during training, while inference operates exclusively on LR audio. The objective is to transfer discriminative HR information without requiring HR sensing at deployment.

This setting extends beyond conventional teacher-student transfer. Classical knowledge distillation typically compresses a larger or more capable teacher into a smaller student given comparable inputs [11]. Generalized distillation and learning using privileged information allow access to additional training-only information [18, 32], while cross-modal distillation transfers supervision across paired modalities [2, 10]. In our setting, the primary challenge is instead *limited input observability*: the LR student receives a systematically degraded view of the same acoustic event. Consequently, some evidence encoded by the HR teacher may be absent from the LR signal, making direct imitation of the full HR representation both restrictive and potentially infeasible.

To address this challenge, we propose *Resolution-Aware Privileged Structure Transfer (RAST)*, a structured HR-to-LR transfer framework tailored to sensor-resolution mismatch. Rather than forcing the LR student to reproduce the complete HR embedding space, RAST compresses teacher representations into intermediate structure comprising HR-derived tokens that summarize coarse teacher states and teacher neighborhoods that encode local geometry. The LR student is first warm-started using this privileged structure and then refined through local hard-negative-mining (HNM) alignment. This design builds on prior work showing that intermediate hints, relational geometry, contrastive objectives, and information bottlenecks can transfer knowledge beyond final logits [22, 25, 29, 30]. RAST therefore focuses transfer on HR-derived structure that is both discriminative and recoverable from LR audio, avoiding imitation of fine-grained information that may be unobservable after resolution reduction.

This paper makes three contributions. **(1)** We formulate lower-resolution audio HAR as a *sensor-resolution privileged learning* problem, where an LR student must learn from an HR teacher despite limited input observability. **(2)** We design *RAST*, a resolution-aware HR-to-LR transfer framework that combines HR-derived token bottlenecks, teacher-neighborhood initialization, and local representation alignment. **(3)** We evaluate RAST on two audio-HAR datasets – SAMoSA [20] and AudioIMU [14], under subject-independent protocols. Relative to LR-only training, RAST improves balanced accuracy and macro-F1 by 6.7% and 7.5% on SAMoSA, and by 6.1% and 7.8% on AudioIMU respectively.

Together, these results demonstrate that privileged HR audio available only during training can substantially improve LR activity recognition while preserving an exclusively low-resolution sensing path at inference.

2 Related Work

Knowledge transfer with privileged information. Knowledge distillation transfers teacher knowledge through softened outputs [11], with subsequent work extending transfer to intermediate features, contrastive objectives, and relational structure [22, 25, 29]. Learning using privileged information allows additional training-only information [32], while generalized distillation integrates such information into teacher–student transfer [18]. Related paradigms include dataset distillation for compact audio-classification training sets [9], zero-shot distillation for statistical heterogeneity in privacy-preserving federated learning [8], and cross-modal distillation across paired modalities [2, 10]. HR-to-LR audio transfer introduces a distinct observability constraint because the student receives a lower-resolution view of the same event. Directly matching HR predictions or representations may therefore impose ill-suited targets that are partially unobservable or unrecoverable from LR input.

Deployment-aware activity sensing. Human-centered sensing must balance recognition accuracy with sensing cost, privacy, participant burden, and robustness. Mobile acoustic sensing has enabled context and activity recognition [15], while other prior work has characterized deployment constraints, user variability, and sensor heterogeneity [4, 27, 28, 33]. Privacy and trust concerns around always-on audio have further motivated low-resolution and bandwidth-restricted sensing [3], including self-distillation for 1–2 kHz microphone inputs [16] and discussions on privacy–utility tradeoffs in low-frequency speech [17]. RAST leverages HR audio only as privileged training supervision while requiring only LR audio at inference.

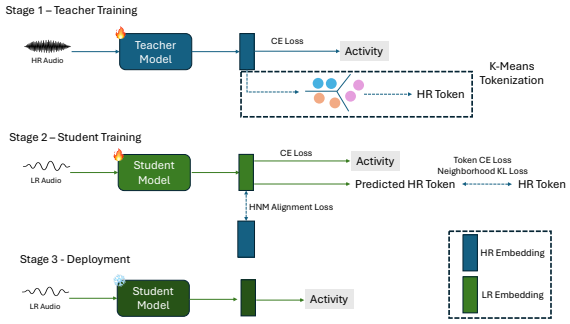

Figure 2: RAST training and deployment pipeline. HR audio provides token- and neighborhood-level privileged supervision only during training, while deployment uses only the LR student and LR audio.

Compact teacher structure. Information-bottleneck methods seek compact representations that preserve task-relevant information while suppressing unnecessary detail [1, 30], while discrete representation learning provides mechanisms for constructing compact latent codes [31]. More broadly, hint-based, relational, and contrastive distillation demonstrate that intermediate structure can convey useful teacher knowledge beyond hard labels or final logits [22, 25, 29]. RAST adapts this principle to sensor-resolution mismatch by transferring compact HR-derived tokens and local teacher-neighborhood structure instead of requiring the LR student to reproduce the full HR embedding space.

3 Methodology

3.1 Problem Formulation

Let $\mathcal{D} = \{(x_i^H, x_i^L, y_i)\}_{i=1}^N$ denote the training set, where x_i^H and x_i^L are paired high-resolution (HR) and low-resolution (LR) views of the same audio segment, respectively, and $y_i \in \{1, \ldots, C\}$ is the activity label. The HR view is available only during training. At inference time, the deployed model observes only x^L and must predict its activity label. This formulation separates the deployment constraint from the model-development setting: richer HR audio can be collected or retained for training, while the deployed classifier remains an LR-only system.

We train an HR teacher T_H and an LR student S_L, each comprising an encoder and a classifier head:

$$z_i^H = e_H(x_i^H), \quad p_i^H = g_H(z_i^H), \tag{1}$$

$$z_i^L = e_L(x_i^L), \quad p_i^L = g_L(z_i^L). \tag{2}$$

where z_i^H and z_i^L denote latent representations. The standard LR baseline minimizes the cross-entropy objective,

$$\mathcal{L}_{\text{CE}} = \frac{1}{N} \sum_i \text{CE}(p_i^L, y_i). \tag{3}$$

RAST augments this LR classification objective with compact structural supervision extracted from the HR teacher.

3.2 Framework Overview

RAST comprises three stages. In Stage 1, the HR teacher is trained for activity recognition, and its intermediate representations are distilled into two forms of compact privileged structure: discrete HR-derived tokens and local teacher neighborhoods. In Stage 2, the LR student uses these structures for initialization (warm start) and is subsequently refined through local HR–LR representation alignment. In Stage 3, the HR teacher and all auxiliary transfer targets are discarded. Deployment therefore requires only the LR student and LR audio input. Figure 2 summarizes the privileged train-time transfer path and the LR-only inference path.

RAST transfers compact, structured HR knowledge rather than requiring the LR student to reproduce the full HR embedding space. Because downsampling removes acoustic evidence available to the teacher, direct HR–LR representation matching can impose partially unrecoverable targets. RAST therefore transfers bottlenecked HR-derived structure followed by local hard-negative alignment.

3.3 Privileged Structure Extraction

HR-derived tokens. After training the HR teacher T_H, we extract representations from a fixed intermediate block and cluster the resulting feature space using K-means. Each training example receives a token label:

$$c_i = \arg \min_{k \in \{1,\ldots,K\}} \left\| \bar{z}_i^H - \mu_k \right\|_2^2, \tag{4}$$

where $\bar{z}_i^H$ denotes the normalized HR teacher representation and μ_k the kth cluster centroid. These tokens provide a compact discretization of the HR representation space, exposing coarse teacher structure to the LR student. They serve only as auxiliary training targets and are discarded at inference.

Teacher neighborhoods. Although HR-derived tokens capture coarse regions of the teacher representation space, they do not preserve its local geometry. We therefore construct teacher-neighborhood distributions within each mini-batch. For samples i and j, the HR pairwise similarity is:

$$s_{ij}^H = \text{sim}(\bar{z}_i^H, \bar{z}_j^H), \tag{5}$$

where sim denotes cosine similarity. The neighborhood distribution induced by the HR teacher for anchor i is:

$$q_{ij}^H = \frac{\exp(s_{ij}^H / \tau)}{\sum_{m \neq i} \exp(s_{im}^H / \tau)}, \quad j \neq i. \tag{6}$$

where τ is the neighborhood temperature. The LR student induces an analogous distribution q_{ij}^L from z^L. We match

Table 1: Subject-held-out audio activity recognition performance. BA denotes balanced accuracy. The HR teacher is shown as the high-resolution reference, while the remaining methods use LR input at inference. Parentheses report relative gains over the LR-only baseline.

	SAMoSA		AudioIMU	
Method	BA	Macro-F1	BA	Macro-F1
HR teacher	0.8352 ± 0.0083	0.8307 ± 0.0075	0.7246 ± 0.0017	0.7066 ± 0.0032
LR-only	0.6123 ± 0.0054	0.6030 ± 0.0046	0.5157 ± 0.0033	0.4933 ± 0.0031
KD	0.6109 ± 0.0035	0.6053 ± 0.0047	0.5124 ± 0.0054	0.4843 ± 0.0064
InfoNCE	0.6137 ± 0.0041	0.6086 ± 0.0016	0.5161 ± 0.0038	0.4994 ± 0.0040
ProtoNCE	0.6112 ± 0.0065	0.6055 ± 0.0079	0.5154 ± 0.0023	0.4921 ± 0.0036
HNM	0.6218 ± 0.0082	0.6093 ± 0.0086	0.5130 ± 0.0019	0.5079 ± 0.0032
RAST	$\mathbf{0.6532 \pm 0.0072}^{(+6.7\%)}$	$\mathbf{0.6482 \pm 0.0067}^{(+7.5\%)}$	$\mathbf{0.5474 \pm 0.0042}^{(+6.1\%)}$	$\mathbf{0.5320 \pm 0.0055}^{(+7.8\%)}$

these distributions with:

$$\mathcal{L}_{\text{nbr}} = \frac{1}{B} \sum_{i=1}^{B} \text{KL}(q_i^H \| q_i^L), \tag{7}$$

where B is the mini-batch size.

3.4 Student Optimization and Alignment

The LR student is first trained with label, token, and teacher-neighborhood supervision:

$$\mathcal{L}_{\text{warm}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{tok}} \, \text{CE}(r_i^L, c_i) + \beta \mathcal{L}_{\text{nbr}}, \tag{8}$$

where r_i^L is the LR token prediction head. This warm start shapes the LR representation before direct HR–LR alignment.

We then apply local hard-negative-mining (HNM) alignment between LR and HR teacher embeddings. For anchor i, (z_i^L, z_i^H) forms the positive pair, while $\mathcal{N}_k(i)$ contains the top-k most similar non-matching HR examples selected as hard negatives in the mini-batch. We optimize:

$$\mathcal{L}_{\text{HNM}} = -\frac{1}{B} \sum_i \log \frac{\exp(\text{sim}(z_i^L, z_i^H)/\gamma)}{\sum_{j \in \mathcal{N}_k(i)} \exp(\text{sim}(z_i^L, z_j^H)/\gamma)}. \tag{9}$$

HNM focuses alignment on locally confusable examples rather than matching the full HR manifold uniformly. The structured warm start further improves alignment by first exposing recoverable HR-derived tokens and neighborhood structure to the LR student.

4 Experiments

4.1 Datasets and Sensing Setup

We evaluate our proposed framework on wearable-based HAR datasets collected in home and semi-naturalistic environments: AudioIMU [14] and SAMoSA [20]. AudioIMU contains 15 participants performing 23 activities of daily living, recorded with a Fossil Gen 4 smartwatch at 22.05 kHz audio and 50 Hz IMU. SAMoSA contains 20 participants recorded across diverse real-world environments using a Fossil Gen 5 smartwatch at 16 kHz audio and 50 Hz IMU. Both datasets provide synchronized acoustic and inertial streams; we use only audio.

Audio is segmented into aligned 5-second windows. The HR branch uses 16 kHz audio, while the LR branch simulates constrained sensing by reducing the effective sampling rate to 1.25 kHz before computing the same log-mel representation. HR and LR inputs are paired views of the same activity window, with only LR audio available at test time.

4.2 Evaluation Protocol

We use subject-independent evaluation to measure generalization to unseen participants. SAMoSA uses leave-one-group-out validation over six participant groups, while AudioIMU uses leave-one-participant-out validation over participants 01–15. Results are reported as mean ± standard deviation over five seeds. We report balanced accuracy (BA) and macro-F1; BA accounts for uneven class frequencies, while macro-F1 measures per-class recognition without allowing frequent classes to dominate.

4.3 Methods and Implementation

All methods use the same PANNs-based audio encoder [12], pretrained on AudioSet [5], and identical subject-held-out folds. Baselines isolate different forms of HR-to-LR knowledge transfer: LR-only uses no privileged supervision; Knowledge distillation (KD) transfers teacher logits [11]; InfoNCE performs instance-level contrastive alignment [21, 29]; Prototype NCE (ProtoNCE) aligns class-level prototypes inspired by prototypical and prototype-contrastive representation learning [13, 26]; and hard-negative mining (HNM) directly aligns local HR–LR representations using hard-negative selection [24]. RAST augments HNM with the proposed token and teacher-neighborhood warm start, evaluating whether compact HR structure transfers more effectively than direct teacher imitation under resolution mismatch.

The HR token codebook uses $K = 100$ clusters extracted from an intermediate teacher block. Unless otherwise stated, RAST uses $\lambda_{\text{tok}} = 1.0$, $\beta = 0.1$, neighborhood temperature $\tau = 0.2$, and HNM temperature $\gamma = 0.2$. Experiments were conducted on a workstation with eight NVIDIA RTX A4500 GPUs, each with 20 GB of memory.

5 Results

5.1 Discussion on Overall Results

Table 1 compares RAST with LR-only training and direct HR-to-LR transfer baselines. RAST consistently achieves the best LR inference performance on both datasets. On SAMoSA, balanced accuracy improves from 0.6123 to 0.6532 and macro-F1 from 0.603 to 0.6482, corresponding to relative gains of 6.7% and 7.5%, respectively. On AudioIMU, balanced accuracy increases from 0.5157 to 0.5474 and macro-F1 from 0.4933 to 0.532, yielding relative gains of 6.1% and 7.8% respectively. Compared with KD, InfoNCE, ProtoNCE, and HNM, these gains indicate that HR supervision using teacher is more effective when first distilled into compact tokens and local neighborhood structure before localized HR–LR alignment.

5.2 Ablation Study

Table 2: Component-level performance of RAST on SAMoSA before and after HNM alignment. Balanced Accuracy (BA) and macro-F1 are reported.

	Warm start		After HNM	
Init.	BA	F1	BA	F1
CE	0.6123	0.6030	0.6218	0.6093
Token	0.6218	0.6105	0.6443	0.6240
Nbr.	0.6243	0.6131	0.6442	0.6225
Token+Nbr.	0.6308	0.6182	**0.6532**	**0.6482**

Table 2 isolates how structured warm start supervision organizes the LR representation space before subsequent HNM alignment. With CE-only initialization, HNM improves balanced accuracy from 0.6123 to 0.6218. Adding HR-token supervision raises the warm-start score to 0.6218 and the post-HNM result to 0.6443, indicating that the token bottleneck provides coarse, activity-relevant anchors that facilitate subsequent local alignment.

Teacher-neighborhood regularization further improves the alignment interface. Although neighborhood supervision alone yields a moderate warm-start score, its post-HNM performance is comparable to token supervision, suggesting that local teacher geometry helps organize nearby examples before hard-negative alignment. Combining tokens and neighborhoods yields the strongest warm start (0.6308 BA, 0.6182 macro-F1) and the final result (0.6532 BA, 0.6482 macro-F1). These results support the RAST design: tokenization supervision provides compact HR-derived learnable anchors, neighborhood regularization structures their local geometry around these anchors, and HNM then aligns the resulting LR representation more effectively.

5.3 Embedding Structure Analysis

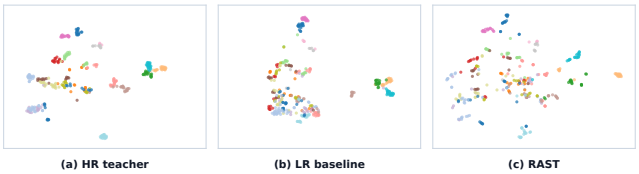

Figure 3: t-SNE visualization of intermediate embeddings. RAST transfers HR-derived structure to the LR model, producing representations that are better aligned with activity structure than the LR baseline.

Figure 3 illustrates the HR–LR representation gap and the role of RAST's structured warm start. The HR teacher preserves more coherent local neighborhood geometry, with related activities forming better-organized regions. In contrast, the LR-only model produces a more entangled representation, particularly for activities that depend on fine acoustic cues attenuated by downsampling.

RAST restructures the transfer problem before local alignment. Rather than directly matching the full HR embedding space, it first exposes the LR student to compact HR-derived tokens and teacher-neighborhood structure, providing coarse anchors and local geometric constraints for subsequent HNM. Consistent with the ablation results in Table 2, the resulting RAST embedding representation in Figure 3 shows improved separation among locally confusable activities and aligns with the overall gains in Table 1.

6 Conclusion

In this paper, we present RAST, a resolution-aware privileged structure transfer framework for low-resolution audio activity recognition. RAST uses HR audio only during training, compressing teacher representations into token- and neighborhood-level supervision before local HR–LR alignment. Across SAMoSA and AudioIMU datasets, RAST consistently outperforms LR-only training and direct teacher-student transfer under subject-independent evaluation.

The HR–LR gap motivates methods that explicitly model resolution mismatch and LR observability. Future work includes student-aware teachers, adaptive privileged bottlenecks, and alignment objectives that preserve activity-relevant structure recoverable from LR audio, with broader evaluation of privacy and real-world deployment tradeoffs.

References

[1] Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2017. Deep Variational Information Bottleneck. In *International Conference on Learning Representations*.

[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. SoundNet: Learning Sound Representations from Unlabeled Video. In *Advances in Neural Information Processing Systems*, Vol. 29.

[3] Bhawana Chhaglani, Sarmistha Sarna Gomasta, Yuvraj Agarwal, Jeremy Gummeson, and Prashant Shenoy. 2025. FeatureSense: Protecting Speaker Attributes in Always-On Audio Sensing System. *arXiv preprint arXiv:2505.24115* (2025).

[4] Sunny Consolvo, Predrag Klasnja, David W. McDonald, Daniel Avrahami, Jon Froehlich, Louis LeGrand, Ryan Libby, Keith Mosher, and James A. Landay. 2008. Activity Sensing in the Wild: A Field Trial of UbiFit Garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1797–1806.

[5] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. 776–780.

[6] Petko Georgiev, Nicholas D. Lane, Kiran K. Rachuri, and Cecilia Mascolo. 2014. DSP.Ear: leveraging co-processor support for continuous audio sensing on smartphones. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*. 295–309.

[7] Gautham Krishna Gudur, Mohit Malu, Tanmay Khandait, Reza Rahimi Azghan, Anirudh Rayas, Pavan Turaga, Joydeep Ghosh, Hassan Ghasemzadeh, Edison Thomaz, and Giulia Pedrielli. 2026. PCL: Partitioned Continual Learning via Unsupervised Latent Experts for Audio Classification. *ICML 2026 Workshop on Learning to Listen: Machine Learning for Audio* (2026).

[8] Gautham Krishna Gudur and Satheesh Kumar Perepu. 2021. Zero-Shot Federated Learning with New Classes for Audio Classification. In *Interspeech 2021*. 1579–1583.

[9] Gautham Krishna Gudur and Edison Thomaz. 2024. Dataset Distillation for Audio Classification: A Data-Efficient Alternative to Active Learning. *NeurIPS 2024 Efficient Natural Language and Speech Processing (ENLSP-IV) Workshop* (2024).

[10] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. 2016. Cross Modal Distillation for Supervision Transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2827–2836.

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. In *NeurIPS Deep Learning and Representation Learning Workshop*.

[12] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. 2020. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), 2880–2894.

[13] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. 2021. Prototypical Contrastive Learning of Unsupervised Representations. In *International Conference on Learning Representations*.

[14] Dawei Liang, Guihong Li, Rebecca Adaimi, Radu Marculescu, and Edison Thomaz. 2022. AudioIMU: Enhancing Inertial Sensing-Based Activity Recognition with Acoustic Models. In *Proceedings of the 2022 ACM International Symposium on Wearable Computers*. 44–48.

[15] Dawei Liang and Edison Thomaz. 2019. Audio-Based Activities of Daily Living (ADL) Recognition with Large-Scale Acoustic Embeddings from Online Videos. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article 17 (2019), 18 pages.

[16] Dawei Liang, Alice Zhang, David Harwath, and Edison Thomaz. 2024. Improving Audio Classification with Low-Sampled Microphone Input: An Empirical Study using Model Self-Distillation. In *Proceedings of Interspeech*.

[17] Ailin Liu, Pepijn Vunderink, Jose Vargas Quiros, Chirag Raman, and Hayley Hung. 2024. How Private is Low-Frequency Speech Audio in the Wild? An Analysis of Verbal Intelligibility by Humans and Machines. In *Interspeech 2024*. 2725–2729.

[18] David Lopez-Paz, Leon Bottou, Bernhard Scholkopf, and Vladimir Vapnik. 2016. Unifying Distillation and Privileged Information. In *International Conference on Learning Representations*.

[19] Hong Lu, Wei Pan, Nicholas D. Lane, Tanzeem Choudhury, and Andrew T. Campbell. 2009. SoundSense: Scalable Sound Sensing for People-Centric Applications on Mobile Phones. In *Proceedings of the 7th International Conference on Mobile Systems, Applications, and Services*. ACM, 165–178.

[20] Vimal Mollyn, Karan Ahuja, Dhruv Verma, Chris Harrison, and Mayank Goel. 2022. SAMoSA: Sensing activities with motion and subsampled audio. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 3 (2022), 1–19.

[21] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[22] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3967–3976.

[23] Ivan Miguel Pires, Nuno M. Garcia, Nuno Pombo, and Francisco Florez-Revuelta. 2017. User Environment Detection with Acoustic Sensors Embedded on Mobile Devices for the Recognition of Activities of Daily Living. *arXiv preprint arXiv:1711.00124* (2017).

[24] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2021. Contrastive Learning with Hard Negative Samples. In *International Conference on Learning Representations*.

[25] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. FitNets: Hints for Thin Deep Nets. In *International Conference on Learning Representations*.

[26] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical Networks for Few-Shot Learning. In *Advances in Neural Information Processing Systems*.

[27] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. 2015. Smart Devices are Different: Assessing and Mitigating Mobile Sensing Heterogeneities for Activity Recognition. 127–140.

[28] Prahalathan Sundaramoorthy, Gautham Krishna Gudur, Manav Rajiv Moorthy, R. Nidhi Bhandari, and Vineeth Vijayaraghavan. 2018. HARNet: Towards On-Device Incremental Learning using Deep Ensembles on Constrained Devices. In *Proceedings of the 2nd International Workshop on Embedded and Mobile Deep Learning*. 31–36.

[29] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive Representation Distillation. In *International Conference on Learning Representations*.

[30] Naftali Tishby, Fernando C. Pereira, and William Bialek. 2000. The Information Bottleneck Method. *arXiv preprint physics/0004057* (2000).

[31] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems*.

[32] Vladimir Vapnik and Akshay Vashist. 2009. A New Learning Paradigm: Learning Using Privileged Information. *Neural Networks* 22, 5–6 (2009), 544–557.

[33] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep Learning for Sensor-Based Activity Recognition: A Survey. *Pattern Recognition Letters* 119 (2019), 3–11.