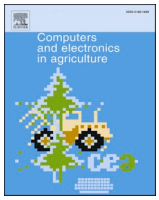
Computers and Electronics in Agriculture 239 (2025) 111132

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag



Original papers

Enhanced audio-based fish feeding intensity recognition via decomposed visually-guided cross-modality distillation

Meng Cui [a,*], Tan Wang [b], Xinhao Mei [a], Jinzheng Zhao [a], Daoliang Li [c], Wenwu Wang [a]

[a] Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford GU2 7XH, UK
[b] School of Information and Artificial Intelligence, Anhui Agricultural University, Hefei 230036, China
[c] National Innovation Center for Digital Fishery, China Agricultural University, China

ARTICLE INFO

Keywords:
Fish feeding intensity recognition
Acoustics-based methods
Cross-modal distillation
Knowledge transfer
Aquaculture management

ABSTRACT

Fish feeding intensity recognition (FFIR) is crucial for sustainable aquaculture management and production optimization. Acoustic-based methods offer non-invasive, cost-effective monitoring in turbid water conditions where visual systems fail. However, acoustic signals struggle to capture temporal behavioral dynamics (such as fish movement patterns) and spatial motion patterns (such as fish aggregation and swimming trajectories) easily detected by visual systems, limiting their discriminative capability for behavioral analysis. This limitation results in a significant performance gap between the best acoustic methods and visual approaches. To address these challenges, we propose AquaDistill, a novel cross-modal knowledge distillation framework that enhances audio-only systems by transferring knowledge from visual modalities during training while requiring only acoustic input during inference. AquaDistill incorporates a decomposed distillation strategy that separates audio features into static acoustic and dynamic behavioral branches, with hybrid distillation losses enabling effective motion knowledge transfer while avoiding feature entanglement. In addition, we introduce the cross-modal behavioral fusion (CMBF) mechanism that leverages distillation-guided knowledge to preserve temporal locality crucial for behavioral analysis through adaptive feature enhancement and cross-branch information exchange. Unlike conventional distillation methods that directly inject cross-modal knowledge, our framework maintains feature separation throughout the learning process while enabling intelligent fusion of complementary acoustic representations. Experimental results demonstrate that AquaDistill significantly improves audio-only model performance, achieving 89 % mean average precision (mAP) and 87 % accuracy, representing improvements of 7 % and 5 % respectively compared to baseline approaches, while maintaining exceptional computational efficiency with only 5.9 M parameters and 1.4 ms inference time. This effectively bridging the performance gap between acoustic and visual methods while maintaining the deployment advantages of audio-only systems. Our enhanced acoustic-based approach demonstrates significant potential for practical aquaculture monitoring applications.

1. Introduction

Fish feeding intensity recognition (FFIR) plays a pivotal role in aquaculture management, directly impacting production efficiency, feed optimization, and sustainable farming practices (Li et al., 2020; Zhao et al., 2024). Global aquaculture production has reached unprecedented levels, with worldwide output exceeding 82 million tons in 2023, making it one of the fastest-growing food production sectors (Siddik et al., 2024; Roberts et al., 2024). In practical fish farming operations, feed costs represent one of the largest expenses, often accounting for more than 50 % of total production costs (Cui et al., 2022). Accurate monitoring of feeding behaviors enables farmers to optimize feeding schedules, reduce waste, and improve fish welfare, ultimately contributing to enhanced productivity and environmental sustainability (Wang et al., 2024; Zhang et al., 2023).

The development of automated FFIR has been dominated by visual monitoring approaches, which have demonstrated remarkable progress over the past decade (Cui et al., 2025). Computer vision technology has emerged as a popular method to evaluate fish feeding intensity, leveraging the distinct visual features that fish exhibit during different feeding states. Early researchers employed background subtraction and optical flow techniques to extract target features for feeding index

* Corresponding author at: Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford GU2 7XH, UK.
E-mail address: m.cui@surrey.ac.uk (M. Cui).

https://doi.org/10.1016/j.compag.2025.111132
Received 8 August 2025; Received in revised form 15 October 2025; Accepted 17 October 2025
Available online 22 October 2025

determination (Zhao et al., 2017, Zhou et al., 2018). Although these methods could capture temporal feeding behaviors, they suffered from high computational load due to the use of complex foreground segmentation processes and performance degradation due to environmental interference such as water surface fluctuations and reflective areas (Zhou et al., 2017). Deep learning has revolutionized fish feeding analysis, with enhanced MobileNetV3 networks achieving 96.4 % accuracy and transformer-based methods like DCA-MVIT reaching 96.62 % precision (Feng et al., 2022; Zhang et al., 2023; Hu et al., 2025). Despite these achievements, vision-based approaches face fundamental limitations caused by water quality, lighting conditions, and surface reflection that restrict their widespread adoption in commercial aquaculture operations (Du et al., 2023).

Compared with vision-based methods, acoustic-based monitoring methods have emerged as an alternative solution since acoustics are unaffected by water turbidity, lighting conditions, or surface reflections that commonly degrade visual monitoring systems (Cui et al., 2024; Du et al., 2024). Additionally, acoustic signals are more compact than video data, requiring less storage and computational resources, making them particularly suitable for resource-constrained aquaculture monitoring devices (Gao et al., 2020; Lin et al., 2022). Recent agricultural AI developments have emphasized lightweight models for edge deployment (Lv et al., 2024), with knowledge distillation techniques enabling complex models to be transferred to resource limited devices (Espejo-Garcia et al., 2025; Li et al., 2025). However, existing distillation approaches focus on single-modality compression rather than cross-modal enhancement (Sai et al., 2025). Deep learning approaches using mel spectrograms have emerged as particularly effective representations for fish feeding sounds (Cui et al., 2022), as they capture the time–frequency characteristics of feeding activities while providing robustness to certain types of noise. Building upon this foundation, Du et al. (2023) introduced GhostNet, a lightweight architecture specifically tailored for acoustic FFIR that reduces computational requirements while maintaining high accuracy. Similarly, Iqbal et al. (2024) proposed an approach combining convolutional neural networks with self-attention mechanisms for *Oplegnathus punctatus* feeding intensity classification, achieving state-of-the-art performance on mel spectrograms. However, despite these compelling advantages, acoustic-based fish monitoring systems face a fundamental performance limitation: they cannot achieve the recognition accuracy levels demonstrated by advanced vision-based approaches (Cui et al., 2025). This disparity primarily stems from the limitations of single-modality audio signals, which lack the rich "physical and motion features", limiting their discriminative capability for fine-grained behavioral analysis (Cui et al., 2024; Li et al., 2024).

Recent advances in cross-modal knowledge distillation have demonstrated significant potential in bridging performance gaps between different modalities across various domains (Huo et al., 2024). Cross-modal knowledge distillation extends traditional distillation to multimodal learning, where a pretrained network from one modality provides supervision to a student network from another modality (Wang et al., 2023). Contemporary methods range from traditional response-based and feature-based approaches to advanced paradigms including self-distillation and adversarial strategies (Mansourian et al., 2025; Wang et al., 2025). Successful applications span medical imaging, computer vision, and continuous sign language recognition, where hybrid distillation losses enable effective motion knowledge transfer while avoiding feature entanglement (Gao et al., 2024; Moslemi et al., 2024; Kwak et al., 2025). However, existing cross-modal knowledge distillation methods typically employ direct feature alignment or unified representation learning, often suffering from feature entanglement when bridging significantly different modalities like audio and video. Traditional approaches force the student network to simultaneously learn both modality-specific characteristics and cross-modal knowledge, leading to conflicting optimization objectives and suboptimal performance. Moreover, the success of cross-modal distillation depends heavily on modality relationships, highlighting the need for domain-specific approaches (Hu et al., 2023). Despite these advances, cross-modal knowledge distillation remains largely unexplored in underwater acoustic monitoring.

To address these challenges, we propose AquaDistill, a novel cross-modal knowledge distillation framework that bridges the acoustic-visual performance gap through three key innovations: (1) decomposed distillation that separates static acoustic and dynamic behavioral learning, (2) cross-modal behavioral fusion for temporal locality preservation, and (3) efficient knowledge transfer while maintaining audio-only deployment advantages. Our contributions are summarized as follows:

(1) We identify and formalize the fundamental challenge of acoustic-visual performance disparity in aquaculture monitoring, providing the first systematic analysis of cross-modal knowledge transfer requirements in underwater behavioral recognition.
(2) We design a dual-branch framework that explicitly separates static acoustic and dynamic behavioral feature learning, with specialized hybrid distillation losses that prevent information interference while maximizing knowledge transfer effectiveness.
(3) We develop cross-modal behavioral fusion (CMBF) that leverages distillation-guided knowledge to preserve temporal locality crucial for behavioral analysis through adaptive cross-branch enhancement and intelligent fusion weighting, avoiding the limitations of traditional fusion approaches.
(4) We demonstrate significant performance improvements (89 % mAP and 87 % accuracy, representing 7 % and 5 % improvements over baselines) with robust cross-species generalizability from *Oplegnathus punctatus* to *Lotus carp*. Extensive experimental analysis across different architectures establishes guidelines for practical aquaculture deployment across diverse species and environmental conditions.

This paper is structured as follows. Section 2 introduces the proposed AquaDistill framework and its key components. Section 3 introduces the dataset and data preprocessing. Section 4 describes the experimental setup and implementation details. Section 5 presents the results and provides a comprehensive discussion. Section 6 concludes the study and offers perspectives for future research.

2. Methods

2.1. Problem Formulation

Let us denote the input video as $X^v \in R^{T\_v \times H \times W \times 3}$ and the corresponding audio signal converted to mel spectrogram as $X^a \in R^{T\_a \times F}$, where T_v and T_a represent the temporal dimensions for video and audio respectively, H and W are the spatial dimensions of video frames, 3 represents the number of RGB channels and F is the mel-frequency dimension. During training, we have access to both modalities with corresponding feeding intensity labels $y \in \{0, 1, 2, 3\}$ representing None, Strong, Medium, and Weak feeding intensities respectively. However, during inference, we aim to perform fish feeding intensity recognition using only acoustic input X^a.

The objective is to train an acoustic-only student model that can achieve performance comparable to a vision-based teacher model. Traditional cross-modal distillation directly transfers knowledge from the teacher to the student model, often leading to feature entanglement and suboptimal performance due to the significant modality gap between visual and acoustic representations.

2.2. Framework overview

To address the limitations of conventional cross-modal distillation,

we propose AquaDistill, a decomposed knowledge distillation framework that enhances audio-only systems through knowledge transfer from video. As illustrated in Fig. 1, our approach consists of three key components: (1) a decomposed distillation strategy that separates audio features into complementary static acoustic and dynamic behavioral branches, (2) the CMBF mechanism for effective cross-branch feature integration, and (3) hybrid distillation losses that enable systematic knowledge transfer while preserving modality-specific information and preventing feature entanglement.

The overall architecture operates in two distinct phases: during training, both log-mel spectrograms and video modalities are utilized to learn decomposed representations through teacher-guided cross-modal knowledge transfer, where the visual teacher network provides rich spatiotemporal supervision to enhance the acoustic student branches; during inference, only the acoustic input is required as the trained model predicts feeding intensity based on the internalized cross-modal knowledge. This asymmetric training-inference paradigm enables practical deployment advantages of audio-only systems while leveraging the rich supervisory signals from visual modalities during the learning process, effectively bridging the performance gap between acoustic and visual approaches.

2.3. Teacher network training

To provide rich supervisory signals for cross-modal knowledge distillation, we employ a pre-trained S3D (Separable 3D CNN) (Xie et al., 2018) model as our visual teacher network. S3D is a variant of 3D CNNs that factorizes standard 3D convolutions into separate spatial and temporal convolutions, significantly reducing computational complexity while maintaining strong performance in video understanding tasks. This architecture effectively captures spatiotemporal patterns in video data, making it particularly suitable for understanding dynamic fish feeding behaviors. We selected S3D as our teacher model for several compelling reasons: first, it demonstrates excellent performance in video classification tasks while maintaining a relatively small parameter footprint compared to other 3D CNN architectures; second, Cui et al. (2024) and Cui et al. (2025) have demonstrated in two recent studies that S3D achieves accuracy exceeding 90 % in fish feeding intensity classification tasks, establishing its effectiveness in aquaculture video analysis. We fine-tune a pretrained S3D model from Kinetics-400 on our fish feeding video dataset with four intensity categories (as discussed in Section 4.2.1). The fine-tuned visual teacher model then distills spatiotemporal knowledge to the acoustic student network.

2.4. Decomposed cross-modal distillation

Given the audio features $h^a \in R^D$ extracted by the MobileNetV2 (Kong et al., 2020) backbone from mel spectrograms, our decomposed distillation strategy separates these features into two complementary branches to avoid feature entanglement during cross-modal knowledge transfer. The core idea is to explicitly separate the learning of stable acoustic patterns from dynamic behavioral patterns, enabling more effective cross-modal knowledge transfer. The representation h^a is projected into two different feature spaces through separate projection

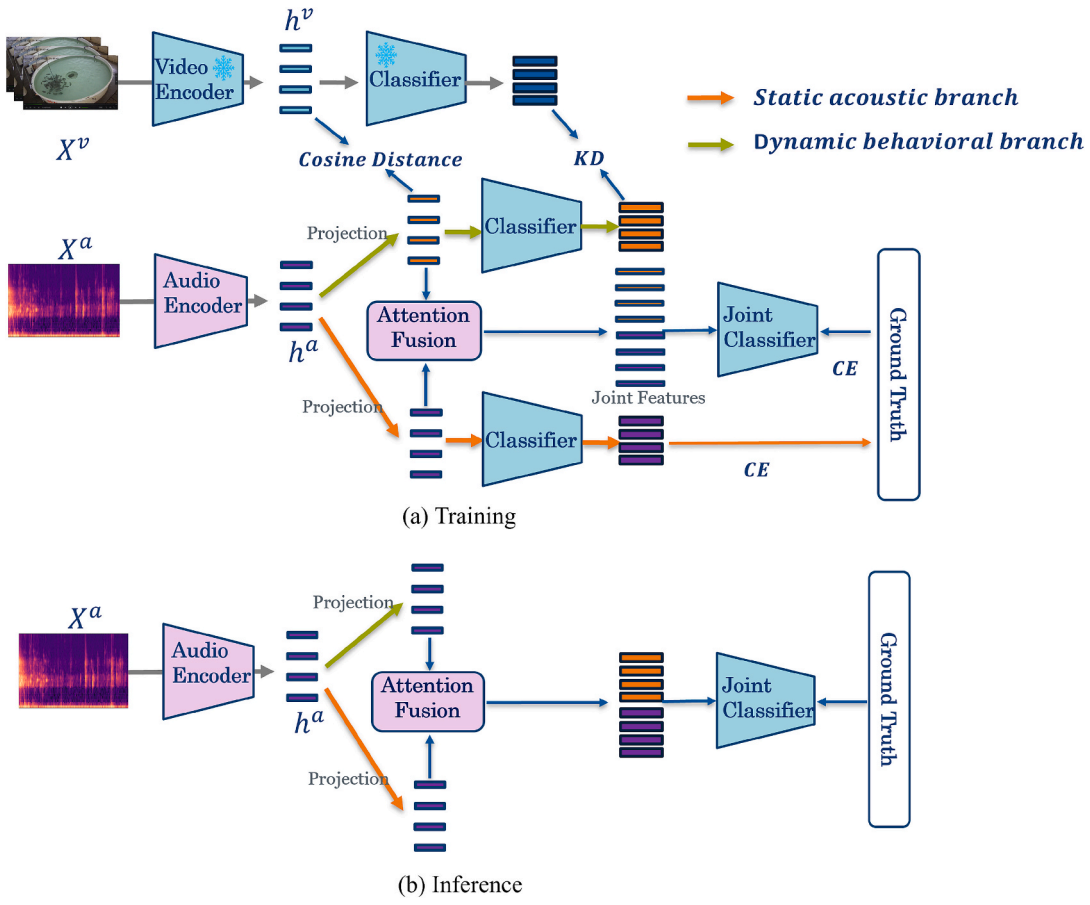


Fig. 1. Overall workflow of the proposed AquaDistill framework. (a) During training, our model performs decomposed cross-modal distillation by explicitly separating mel spectrogram inputs into static acoustic (green) and dynamic behavioral branches (orange), with each branch learning complementary representations under visual teacher guidance through hybrid distillation losses, followed by CMBF for integrated prediction. (b) At inference, our model requires only mel spectrogram input and accurately predicts fish feeding intensity through the learned cross-modal knowledge embedded in the dual-branch architecture and CMBF mechanism.

networks, as shown in Equation (1) and Equation (2):

$$f^{static} = \phi_{static}(h^a) \in R^D \tag{1}$$

$$f^{dynamic} = \phi_{dynamic}(h^a) \in R^D \tag{2}$$

where $\phi_{static}(\cdot)$ and $\phi_{dynamic}(\cdot)$ are implemented as identical two-layer linear networks with ReLU activations followed by L2 normalization for regularization. These projection layers serve as domain adaptation components that transform the shared acoustic features into modality-specific representations suitable for different learning objectives.

The static acoustic branch is designed to learn stable spectral patterns and frequency characteristics that are inherent to acoustic feeding signals, such as consistent frequency signatures of feeding sounds and background aquaculture environment noise patterns. This branch operates independently from the visual teacher to preserve audio-specific information that might be lost during cross-modal transfer. The static features $f^{static} \in R^D$ capture the intrinsic acoustic properties that are consistent across different feeding intensities but vary in their spectral characteristics.

The static features are processed through a linear classification layer to produce predictions as follows:

$$\hat{p}^{static} = \text{Linear}_{cls}\left(f^{static}\right) \tag{3}$$

The static branch is trained using standard cross-entropy loss with ground truth labels:

$$\mathscr{L}_{static} = \ell_{CE}\left(\hat{p}^{static}, y\right) \tag{4}$$

where y is the one-hot encoded ground truth label.

The dynamic behavioral branch learns temporal dynamics and motion patterns by distilling knowledge from the pre-trained visual teacher model. Unlike the static branch, this branch focuses on capturing temporal evolution and intensity variations in feeding behaviors that are more readily observable in visual data. To address the significant modality gap between acoustic and visual representations, we employ specialized normalization and similarity-based distillation losses. The visual teacher features $h^v \in R^{T\times D}$ are first temporally averaged to obtain global representations and then L2 normalized to ensure consistent feature scales, denoted as z^v_{global}.

To handle the distribution differences between modalities, we use the cosine similarity loss instead of the L2 distance-based loss, which is more robust to scale variations and modality gaps, as shown below:

$$\mathscr{L}_{feature} = 1 - CosineSimilarity\left(f^{dynamic}, z^v_{global}\right) \tag{5}$$

Where $f^{dynamic}$ is the L2 normalized dynamic features.

The dynamic features are processed through the same linear classification layer to produce predictions:

$$\hat{p}^{dynamic} = \text{Linear}_{cls}\left(f^{dynamic}\right) \tag{6}$$

Rather than using traditional knowledge distillation (KD) with soft targets, we employ the teacher's predictions as pseudo ground truth labels, which provides more direct supervision for bridging the modality gap as follows:

$$y\_pseudo = argmax\left(\hat{p}^{teacher}\right) \tag{7}$$

$$\mathscr{L}_{pred} = \ell_{CE}\left(\hat{p}^{dynamic}, y\_pseudo\right) \tag{8}$$

where y_pseudo represents the hard pseudo label obtained by selecting the class with highest probability from the teacher's soft predictions $\hat{p}^{teacher}$. ℓ_{CE} represents the cross-entropy loss using teacher's hard predictions as pseudo ground truth supervision. We employ hard pseudo labels rather than traditional soft knowledge distillation due to the significant modality gap between audio and visual features. Soft probability distributions are less reliable for cross-modal transfer, as teacher confidence scores may not translate meaningfully across modality boundaries. Hard pseudo labels provide more decisive supervision for categorical boundary learning. Our comparison shows that soft distillation ($\tau = 4$, KL divergence loss) achieves only 87.2 % mAP versus our 89.0 % mAP, while requiring 8 % longer training time and 5 % higher memory usage. The computational overhead makes hard labels more suitable for resource-constrained aquaculture applications.

The total loss for the dynamic branch combines both distillation objectives, as follows:

$$\mathscr{L}_{dynamic} = \mathscr{L}_{pred} + \mathscr{L}_{feature} \tag{9}$$

The decomposed distillation strategy described above enables effective knowledge transfer while maintaining feature separation. However, to fully leverage these complementary representations, an intelligent fusion mechanism is required that can adaptively combine static acoustic and dynamic behavioral features based on their contextual relevance. The following section introduces our cross-modal behavioral fusion (CMBF) mechanism that addresses this challenge through adaptive weighting and cross-branch information exchange.

2.5. Cross-modal behavioral fusion (CMBF)

After the decomposed cross-modal distillation process described in Section 2.4, we obtain enhanced static features f^{static} and dynamic features $f^{dynamic}$ from the two separate branches. We then propose CMBF to effectively combine these complementary representations and produce the final feeding intensity prediction. This fusion mechanism is essential because simple concatenation or averaging would ignore the varying importance of static and dynamic information across different feeding scenarios. Traditional fusion approaches either use simple concatenation that ignores feature interactions or complex attention mechanisms that suffer from computational overhead. CMBF leverages the cross-modal knowledge learned through teacher distillation to achieve effective feature integration with linearly scaled computational complexity.

To preserve modality-specific characteristics while enabling effective fusion, we project the static and dynamic features into a shared representation space through separate projection networks, as follows.:

$$P^{static} = LayerNorm\left(ReLU\left(Linear_{static}\left(f^{static}\right)\right)\right) \tag{10}$$

$$P^{dynamic} = LayerNorm\left(ReLU\left(Linear_{dynamic}\left(f^{dynamic}\right)\right)\right) \tag{11}$$

where independent projection layers maintain branch-specific information while enabling subsequent fusion operations. We employ *LayerNorm* and *ReLU* as standard choices for normalization and activation in CMBF, which are commonly used in efficient neural network designs. Our comparative evaluation confirmed that *LayerNorm* and *ReLU* provide optimal performance for our cross-modal fusion task while maintaining computational efficiency suitable for agricultural edge deployment.

Building upon the cross-modal knowledge learned through teacher distillation, we apply frequency-domain weighting to emphasize important spectral components in each branch. The formula is shown in below:

$$w^{static} = \sigma\left(Linear\left(P^{static}\right)\right) \tag{12}$$

$$w^{dynamic} = \sigma\left(Linear\left(P^{dynamic}\right)\right) \tag{13}$$

where $\sigma(\cdot)$ denotes the sigmoid activation function. The learned weights identify the most discriminative frequency components for each modality.

The enhanced features are then computed as follows:

$$F_{enhanced}^{static} = P^{static} \odot w^{static} \tag{14}$$

$$F_{enhanced}^{dynamic} = P^{dynamic} \odot w^{dynamic} \tag{15}$$

Where $\odot$ denotes element-wise multiplication (Hadamard product), enabling selective enhancement of individual feature dimensions based on their learned importance weights. $F_{enhanced}^{static}$ represents frequency-enhanced static acoustic features that emphasize important spectral components through learned frequency weights, while $F_{enhanced}^{dynamic}$ represents behaviorally enhanced dynamic features that incorporate both frequency-domain refinement and behavioral knowledge from the visual teacher network. This dual enhancement ensures that static features capture refined acoustic patterns while dynamic features integrate cross-modal behavioral understanding.

Rather than using fixed fusion weights, CMBF computes adaptive weights based on feature similarity and interaction strength, The formula is shown below:

$$I = F_{enhanced}^{static} \odot F_{enhanced}^{dynamic} \in R^{T \times D} \tag{16}$$

$$S = \frac{\sum \left(F_{enhanced}^{static} \odot F_{enhanced}^{dynamic} \right)}{\left| \left| F_{enhanced}^{static} \right| \cdot \left| F_{enhanced}^{dynamic} \right| \right| + \varepsilon} \tag{17}$$

where S represents the normalized cosine similarity (range 0–1) measuring the complementarity between static acoustic and dynamic behavioral features. A higher similarity values ($S > 0.6$) indicate a stronger feature complementarity, where both acoustic characteristics contribute equally to feeding intensity recognition. A lower value ($S < 0.3$) suggests that one feature type is more discriminative, enabling the model to adaptively emphasize static patterns (ambient sounds) or dynamic patterns (feeding activity sounds) based on the specific feeding scenario.

$$w_{modal} = \sigma(MLP(mean(I)) + S) \tag{18}$$

where w_{modal} represents the adaptive weight that balances static and dynamic contributions based on their complementarity, and ε is a small constant (e.g., 1e-8) added for numerical stability to prevent division by zero. The adaptive weights in CMBF are computed through a learnable *MLP* consisting of a single linear layer with trainable parameters, followed by the sigmoid activation. The *MLP* parameters are initialized using Xavier uniform initialization (Ennadir et al., 2024) and optimized end-to-end with the entire framework using the Adam optimizer (Cui et al., 2024), allowing the model to learn optimal fusion strategies during training.

Finally, we enable mutual information exchange between branches before adaptive fusion as follows:

$$score = \sigma\left(\frac{\Sigma\left(F_{enhanced}^{static} \odot F_{enhanced}^{dynamic} \right)}{\sqrt{D}} \right) \tag{19}$$

$$P_{enhanced}^{static} = F_{enhanced}^{static} + score \odot F_{enhanced}^{dynamic} \tag{20}$$

$$P_{enhanced}^{dynamic} = F_{enhanced}^{dynamic} + score \odot F_{enhanced}^{static} \tag{21}$$

$$F^{fused} = w_{modal} \odot P_{enhanced}^{static} + (1 - w_{modal}) \odot P_{enhanced}^{dynamic} \tag{22}$$

where $\sqrt{D}$ is a scaling factor that normalizes the dot product to prevent saturation of the sigmoid function, ensuring meaningful gradient flow and providing discriminative attention scores across different samples. High scores enable more cross-branch information sharing, while low scores preserve branch-specific characteristics.

This allows each branch to benefit from the other's knowledge while maintaining computational efficiency with linear complexity order O(D) per sample.

3. Dataset

3.1. Data acquisition and experimental system

Our dataset was collected in a controlled aquaculture facility using *Oplegnathus punctatus* as the experimental subject. The fish were housed in a recirculating tank (3 m in diameter, 0.75 m in depth) located in Yantai, Shandong Province, China. The experimental population consisted of 40–100 fish, each weighing approximately 150 g. To capture multimodal data, we employed a high-definition digital camera (Hikvision DS-2CD2T87E(D)WD-L) with a frame rate of 25 fps (1920 × 1080) and a high-frequency hydrophone (LST-DH01) with a sampling frequency of 256 kHz. The camera was positioned on a tripod at approximately 2 m height to capture overhead video footage, while the hydrophone was submerged underwater to record acoustic data (as shown in Fig. 2). The acquisition of video and audio data was synchronized to ensure temporal alignment of multimodal information. During data collection, we adhered to feeding protocols consistent with real aquaculture production environments to ensure fish adaptation and minimize appetite loss due to environmental changes. The water conditions were maintained as follows: temperature at 26 ± 1°C, dissolved oxygen ≥ 5 mg/L, pH at 7.2 ± 0.5, nitrate ≤ 0.5 mg/L, and ammonia nitrogen ≤ 0.8 mg/L. Fish were fed twice daily at 9 AM and 4 PM. The feeding process typically lasted 3–15 min per session.

3.2. Data preprocessing and annotation

Under the guidance of aquaculture technicians, we annotated the feeding behavior based on observed feeding intensity into four categories: "Strong", "Medium", "Weak", and "None" (as shown in Fig. 3). The feeding intensity categories were defined as follows: Strong - significant water turbulence with high fish aggregation and rapid bait consumption; Medium - moderate fish movement toward bait with reduced aggregation; Weak - limited fish participation with slow feeding behavior; None - no feeding response with dispersed fish distribution. To create a fine-grained dataset suitable for cross-modal knowledge distillation, we segmented each recording session into 2-second clips, resulting in 19,021 synchronized audio–video segments. The dataset was partitioned into training (80 %), validation (10 %), and testing (10 %) sets through random selection while maintaining class balance, resulting in 13,421, 2,800, and 2,800 clips, respectively.

For acoustic data preprocessing, we converted raw audio signals into log-mel spectrograms, which have proven effective for capturing time–frequency characteristics of fish feeding sounds. The log-mel spectrogram transformation was performed using the following parameters: window size of 1024 samples, hop length of 512 samples, and 128 mel filter banks. This representation provides a compact yet informative encoding of acoustic features while maintaining robustness to environmental noise commonly present in aquaculture settings. video data was preprocessed by extracting frames at the original 25 fps and resized to 224 × 224 pixels for compatibility with standard deep learning architectures. Data augmentation techniques including horizontal flipping and random noise addition were applied during training to enhance model generalization. The final dataset distribution across feeding intensity categories is presented in Table 1, showing a comprehensive collection suitable for training robust cross-modal knowledge distillation models.

4. Experimental setup

4.1. Evaluation metrics

To comprehensively evaluate the performance of our proposed AquaDistill framework and enable fair comparison with existing

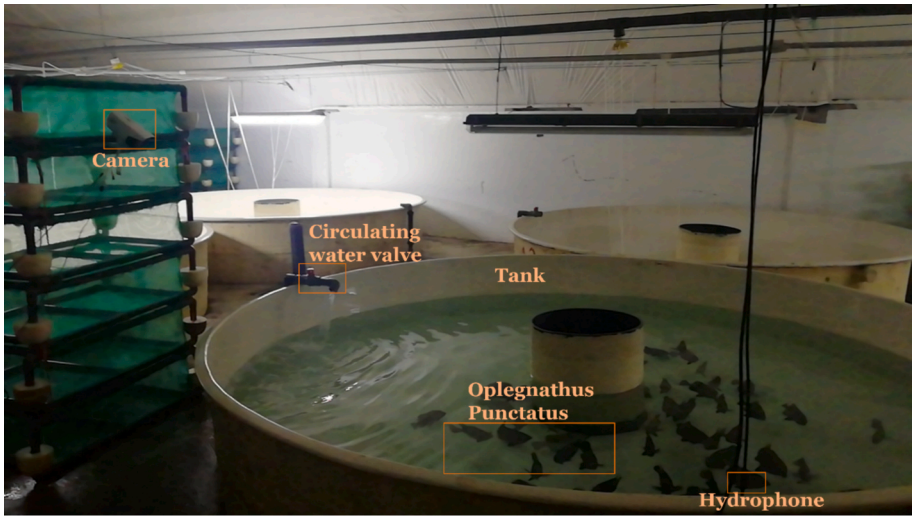


Fig. 2. Experimental systems for data collection.

methods, we employ standard classification metrics commonly used in fish feeding intensity recognition literature: Accuracy, Precision, Recall, and F1-Score and mean Average Precision (mAP). These metrics provide complementary perspectives on model performance, with Accuracy reflecting overall classification performance, Precision indicating prediction reliability, Recall measuring detection completeness, and F1-Score providing a balanced assessment particularly valuable for handling class imbalance. The metrics are computed as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{23}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{24}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{25}$$

$$\text{F1-Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{26}$$

$$\text{mAP} = \frac{1}{C}\sum_{c=1}^{C} \text{AP}_c \tag{27}$$

where *TP*, *TN*, *FP*, and *FN* represent true positives, true negatives, false positives, and false negatives, respectively. AP_c represents the average precision for class *C*, computed as the area under the precision-recall curve for that class, and *C* is the total number of classes. For multi-class classification, mAP provides a robust evaluation by considering the model's performance across all feeding intensity categories.

4.2. Model training configuration

4.2.1. Teacher model training

The visual teacher model (S3D) is initialized with pretrained weights from Kinetics-400 and fine-tuned on our fish feeding dataset. We selected S3D over the larger models such as TimeSformer (Kour et al., 2025) and videoMAE (Li et al., 2025), as they are prone to overfitting on our dataset (19,021 samples) and require excessive computational resources (150 GFLOPs vs 23.4 GFLOPs) that are impractical for agricultural applications. We use the complete training to maintain the same data split as the student model training: 13,421 samples for training (80 % of total dataset), 2,800 samples for validation (10 % of total dataset), and 2,800 samples reserved for final testing (10 % of total dataset). This ensures consistent data distribution between teacher and student training phases. During teacher model fine-tuning, we randomly sample 16 frames from each 2-second video clip in temporal order to maintain the sequential nature of feeding behaviors while reducing computational complexity. The teacher model training employs the Adam optimizer with a learning rate of 1e-3 and batch size of 32, fine-tuned for 20 epochs until convergence. The fine-tuned S3D teacher model achieves 92.8 % mAP on the validation set, demonstrating effective adaptation from the general action recognition domain (Kinetics-400) to the specific aquaculture monitoring task.

To accelerate student model training and ensure reproducible results, we pre-extract and cache all teacher model outputs after fine-tuning completion. Specifically, we extract teacher features and predictions for all 13,315 training samples and save them as tensors. This approach eliminates the computational overhead of running the teacher network during student training iterations, reducing overall training time by approximately 40 % while maintaining identical supervision quality for cross-modal knowledge distillation.

4.2.2. Student model training

All experiments were conducted on a high-performance computing platform to ensure reproducible results and efficient training. The experimental details can be seen in Table 2. We use an NVIDIA GeForce RTX 4090 chip with 24 GB of RAM as the graphics card for core computation, paired with an Intel Core i9-13900 K processor running at 3 GHz. The experimental environment utilizes CUDA 12.1 for GPU acceleration, Python 3.9.16 for implementation, and PyTorch 2.0.1 as the deep learning framework. The student model (AquaDistill) training employs the Adam optimizer with an initial learning rate of 1e-3, trained for 300 epochs with a batch size of 32. To prevent overfitting and ensure model generalization, we implement early stopping with a patience of 15 epochs and apply dropout regularization with a rate of 0.5. The training utilizes pre-cached teacher features and predictions to perform cross-modal knowledge distillation, with the decomposed distillation strategy separating acoustic features into static and dynamic branches. All random processes are controlled using a fixed seed (25) to ensure experimental reproducibility across different runs. Training requires approximately 4 h in total: including the fine-tuning of teacher model (20 epochs, ~1.5 h) and the training of the student model (300 epochs, ~2.5 h). We report mean performance across 3 independent runs with different weight initializations, following standard practice in deep learning research where multiple runs with different random seeds are the most commonly accepted method for assessing model stability and reporting variance. The results are presented as mean ± standard deviation to indicate performance consistency across different initializations.

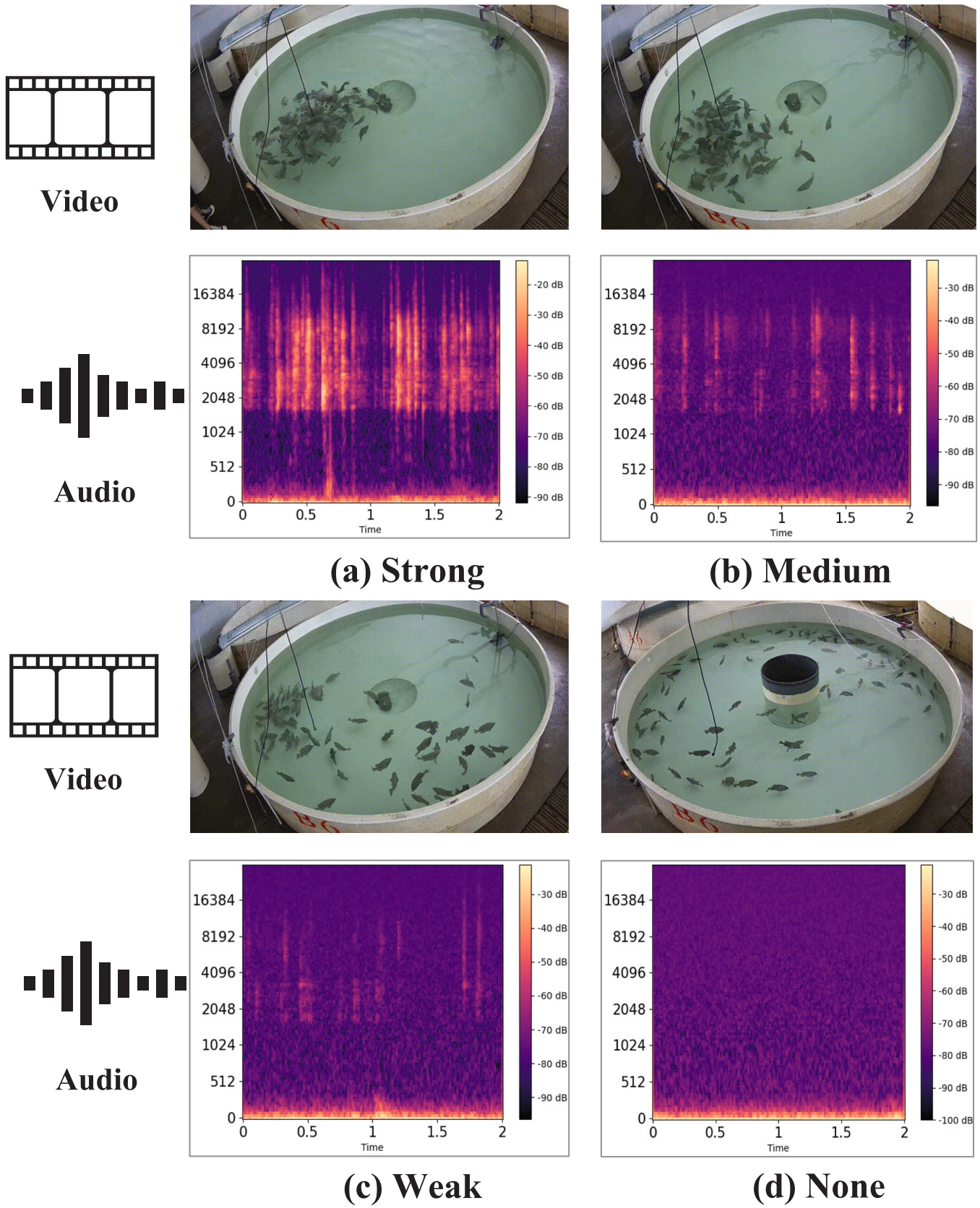


Fig. 3. Video frames and mel spectrogram visualizations of four different fish feeding intensity: *"Strong", "Medium", "Weak"* and *"None"*.

Table 1
Dataset distribution for fish feeding intensity classification.

Feeding Intensity	Training	Validation	Testing	Total
Strong	4053	869	869	5791
Medium	3801	815	815	5431
Weak	3356	719	719	4794
None	2104	603	603	3111
Total	13,421	2800	2800	19,021

Table 2
Experimental configuration and parameter settings.

Configuration	Model/Version
CPU	Intel Core i9-13900 K @ 3 GHz
GPU	NVIDIA GeForce RTX 4090 (24 GB)
Programming Language	Python 3.9.16
CUDA Version	12.1
Deep Learning Framework	PyTorch 2.0.1
Optimizer	Adam
Batch Size	128
Learning Rate	1e-3
Max Epochs	300
Early Stopping Patience	15
Dropout Rate	0.5
Random Seed	25

5. Results and discussion

5.1. Overall performance comparison

We comprehensively evaluate our AquaDistill framework against state-of-the-art methods across different modality paradigms. Our comparison includes vision-based approaches including, S3D (Xie et al., 2018), I3D (Yang et al., 2025), 3D-ResNet18 (Al-Khater and Al-Madeed, 2024) and 3D-ViT (Zhang et al., 2024), multimodal audio-visual methods including, U-FFIA (Cui et al., 2024), MFFFI (Du et al., 2023), and MMFINet (Gu et al., 2025), and audio-only baselines including, GhostNet (Du et al., 2023), Swin Transformer (Zeng et al., 2023), MobileNetV3 (Du et al., 2023), and MobileNetV2 (Kong et al., 2020). Table 3 presents detailed performance comparisons on our fish feeding intensity recognition dataset, demonstrating the effectiveness of our cross-modal knowledge distillation strategy.

Our AquaDistill achieves substantial improvements over existing audio-only approaches, establishing a new state-of-the-art for acoustic-based fish feeding intensity recognition. Compared to the best previous audio method MobileNetV2 (82.6 % mAP), our approach delivers a significant 6.4 % improvement (89.0 % mAP), representing a relatively 7.8 % improvement. This demonstrates the effectiveness of cross-modal knowledge transfer in enhancing acoustic feature discrimination capabilities. Notably, advanced architecture designed for other domains shows limitations when applied to acoustic data. The Swin Transformer, despite its remarkable success in visual tasks, achieves only 79.5 % mAP on our acoustic dataset. This suboptimal performance confirms that vision-specific inductive biases, particularly the patch-based attention mechanism optimized for spatial relationships, do not effectively transfer to the time–frequency representation of audio. In contrast, our domain-adapted approach through decomposed distillation successfully captures the temporal-spectral patterns essential for feeding behavior recognition. The complexity analysis shows AquaDistill's advantage in computational efficiency, offering a potential advantage for edge deployment. With only 1.7 GFLOPs, our method requires significantly fewer computations than visual (23.4–78.3 GFLOPs) and multimodal approaches (32.1–67.2 GFLOPs) while outperforming audio baselines. This efficiency makes AquaDistill suitable for resource-constrained aquaculture devices such as embedded controllers and low-cost terminals.

A critical contribution of our work is significantly narrowing the performance disparity between audio and visual modalities. The original gap between the best audio method (MobileNetV2: 82.6 %) and the visual teacher (S3D: 92.8 %) spans 10.2 %. Our AquaDistill reduces this gap to merely 3.8 % (89.0 % vs 92.8 %), representing a 63 % reduction in performance disparity. Fig. 4 provides detailed confusion matrix analysis revealing the classification improvements achieved by our AquaDistill framework. The original audio baseline (Fig. 4a) exhibits substantial confusion between adjacent feeding intensity levels, particularly struggling with medium-weak (classes 2–3) discrimination where 108 samples are misclassified between these categories in both directions. This bi-directional confusion indicates the inherent difficulty in distinguishing subtle intensity variations using acoustic features alone, especially between moderate feeding states. Our enhanced audio approach (Fig. 4b) demonstrates remarkable improvement in classification precision across all categories. The confusion between medium-weak feeding has been reduces significantly (75 vs 108 for medium-weak, 86 vs 98 for weak-medium), while strong feeding recognition improves dramatically with better separation from other categories. Most notably, the none-feeding category (class 0) achieves near-perfect recognition with only 20 total misclassifications versus 58 in the original model, indicating that our cross-modal distillation particularly enhances the detection of feeding absence. Compared with the visual teacher model (Fig. 4c), our enhanced audio approach achieves competitive confusion patterns despite using only acoustic input. The visual model maintains advantages in overall precision, particularly in distinguishing medium and weak categories, but our enhanced audio method successfully captures the key discriminative patterns for extreme categories (none and strong feeding).

While multimodal audio-visual methods achieve the highest absolute performance (U-FFIA: 95.1 % mAP, MMFINet: 94.6 % mAP), they require substantial computational overhead and dual-stream processing. Our audio-only AquaDistill approach (89.0 % mAP) delivers remarkable performance considering its single-modality constraint, achieving only 6 %-7% lower accuracy than the best multimodal methods while requiring significantly fewer resources. Multimodal methods (93.9–95.1 % mAP) outperform single visual modality (85.1–92.8 %

Table 3
Performance comparison with existing methods on fish feeding intensity recognition dataset.

Method	Input Modality	mAP (%)	Acc (%)	F1 (%)	Params (M)	Model Size (MB)	FLOPs (G)	Inference (ms)
S3D	Visual	**92.8 ± 0.2**	**92.3 ± 0.3**	**92.5 ± 0.2**	**7.9**	**31.7**	**23.4**	**6.4**
I3D	Visual	88.2 ± 0.4	87.5 ± 0.3	87.8 ± 0.3	12.5	50.0	42.1	15.3
3D-RestNet18	Visual	85.1 ± 0.3	84.6 ± 0.4	84.8 ± 0.3	33.5	134.2	78.3	22.8
3D-ViT	Visual	86.4 ± 0.3	85.9 ± 0.2	86.1 ± 0.3	27.8	111.2	65.7	18.7
U-FFIA	Audio + Visual	**95.1 ± 0.2**	**94.5 ± 0.2**	**95.0 ± 0.2**	21.6	86.4	67.2	28.8
MFFFI	Audio + Visual	93.9 ± 0.3	93.5 ± 0.2	93.7 ± 0.3	**10.6**	42.4	34.8	24.2
MMFINet	Audio + Visual	94.6 ± 0.2	94.1 ± 0.3	94.4 ± 0.2	10.01	**40.0**	**32.1**	**23.6**
GhostNet	Audio	81.4 ± 0.5	80.1 ± 0.4	81.7 ± 0.3	5.2	20.8	2.3	28.0
Swin Transformer	Audio	79.5 ± 0.6	76.8 ± 0.5	77.2 ± 0.4	28	112.0	8.7	26.5
MobileNetV3	Audio	80.2 ± 0.4	79.1 ± 0.3	78.7 ± 0.5	3.7	14.8	1.8	2.3
MobileNetV2	Audio	82.6 ± 0.4	80.5 ± 0.3	80.7 ± 0.3	**3.5**	**14.0**	**1.6**	**1.2**
AquaDistill (Ours)	Audio	**89.0 ± 0.3**	**87.0 ± 0.2**	**87.2 ± 0.2**	5.9	23.6	1.7	**1.4**

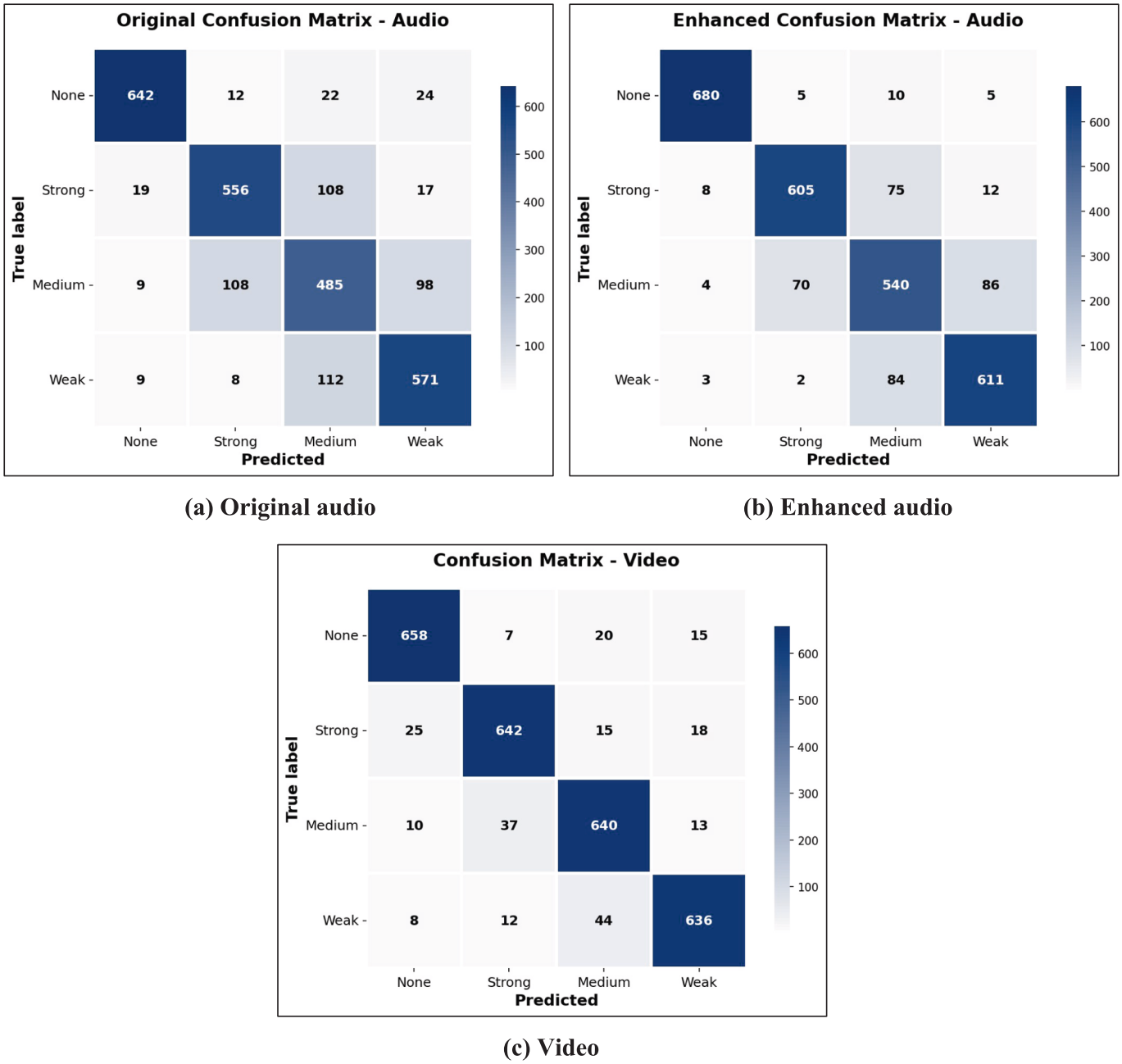


Fig. 4. Confusion matrices comparison for fish feeding intensity classification. (a) Original audio baseline using MobileNetV2 showing significant confusion between adjacent intensity levels, particularly between classes 2–3 (medium-weak) with 108 misclassifications each direction. (b) Enhanced audio performance using our AquaDistill framework demonstrating substantial reduction in inter-class confusion and improved diagonal dominance. (c) Video classification results from S3D teacher model showing superior class separation capabilities. Classes represent: 0-None, 1-Strong, 2-Medium, 3-Weak feeding intensities.

mAP), which in turn exceeds audio-only approaches (79.5–89.0 % mAP). However, the modest gains of multimodal methods (+2.3–2.8 % over best visual methods) come at substantial computational cost, requiring 17–20 × longer inference time compared to our audio-only solution. This demonstrates that while multimodal fusion provides incremental benefits, our cross-modal distillation approach offers a compelling trade-off between performance and practical deployment constraints.

5.2. Ablation study

To validate the effectiveness of our proposed AquaDistill framework, we conduct comprehensive ablation studies examining both the contribution of core components and the superiority of our fusion mechanism. All experiments are performed under identical conditions using our fish feeding intensity dataset.

5.2.1. Core component analysis

Table 4 presents the incremental contribution of each major component in our framework. To validate the necessity of our dual-branch design, we first examine single-branch performance. The static branch only (identical to baseline at 82.6 % mAP) represents training without any cross-modal distillation, preserving only inherent acoustic patterns. The dynamic branch only achieves 84.3 % mAP through cross-modal distillation from visual teacher, demonstrating the value of behavioral knowledge transfer from the visual modality. However, the

Table 4
Core component ablation study.

Components	mAP (%)	Acc (%)	F1 (%)	Params (M)
Baseline (MobileNetV2)	82.6 ± 0.4	80.5 ± 0.3	80.7 ± 0.3	3.5
+ Static Branch Only (No Distillation)	82.6 ± 0.4	80.5 ± 0.3	80.7 ± 0.3	3.5
+ Dynamic Branch Only (Video Distillation)	84.3 ± 0.3	82.4 ± 0.2	82.6 ± 0.3	4.0
+ Decomposed Distillation (Both Branches)	85.2 ± 0.3	83.1 ± 0.2	83.4 ± 0.2	5.1
+ CMBF Fusion	87.8 ± 0.2	85.7 ± 0.2	86.1 ± 0.2	5.9
+ Full Framework	89.0 ± 0.3	87.0 ± 0.2	87.2 ± 0.2	5.9

combined decomposed cross-modal distillation (85.2 % mAP) provides the most substantial gain (+2.6 % over baseline), outperforming either single branch. This significant improvement demonstrates that separating static acoustic and dynamic behavioral learning prevents feature entanglement that commonly occurs in conventional distillation approaches. By allowing the static branch to preserve audio-specific spectral characteristics while the dynamic branch focuses on motion patterns learned from visual teacher, our decomposed strategy maximizes knowledge transfer effectiveness while avoiding representational conflicts.

Unlike direct distillation methods that force audio features to simultaneously learn both acoustic properties and visual motion patterns, leading to conflicting learning objectives and suboptimal feature representations, our proposed decomposition-based strategy maintains separate learning pathways to avoid representational conflicts. As illustrated in Fig. 5, conventional single-branch distillation results in overlapping, poorly-separated clusters due to feature entanglement, while our dual-branch approach produces well-defined class boundaries with clear cluster separation. By allowing the static branch to preserve audio-specific spectral characteristics while the dynamic branch focuses on motion patterns learned from visual teacher, our proposed strategy enables effective knowledge transfer across the modalities. The addition of CMBF fusion contributes an additional 2.6 % (85.2 % – 87.8 %), indicating that simple feature combination is insufficient for optimal performance. The CMBF mechanism's adaptive weighting based on feature similarity and interaction enables sample-specific fusion strategies, allowing the model to emphasize static features for ambient-dominated samples while prioritizing dynamic features for movement-intensive feeding behaviors. The final framework optimization (+1.2 %) incorporates adaptive learning rate scheduling, label smoothing ($\alpha = 0.1$), and optimized loss weighting, demonstrating that systematic optimization of the complete pipeline yields additional gains beyond the contributions by the individual component.

5.2.2. *Fusion method comparison*

Table 5 reveals critical insights into multimodal feature fusion effectiveness. Simple concatenation (85.2 % mAP) and element-wise addition (85.5 % mAP) show limited improvement because they treat all features equally without considering their contextual importance. These naive approaches cannot adapt to the varying relevance of static vs. dynamic information across different feeding scenarios. Self-attention (86.1 % mAP) and cross-attention (86.4 % mAP) demonstrate improved performance by learning feature relationships but suffer from two critical limitations: quadratic computational complexity O (T^2D) that significantly increases inference time (1.9–2.1 ms vs 1.4 ms),

Table 5
Fusion method comparison.

Fusion Method	mAP (%)	Acc (%)	F1 (%)	Params (M)	Inference (ms)	Complexity
Concatenation	85.2 ± 0.3	83.1 ± 0.2	83.4 ± 0.2	6.1	1.5	O (D)
Element-wise Addition	85.5 ± 0.4	83.4 ± 0.3	83.7 ± 0.2	5.9	1.4	O (D)
Self-attention	86.1 ± 0.2	84.2 ± 0.2	84.5 ± 0.2	6.5	1.9	O (T^2D)
Cross-attention	86.4 ± 0.3	84.5 ± 0.2	84.8 ± 0.3	6.8	2.1	O (T^2D)
CMBF(Ours)	**87.8 ± 0.2**	**85.7 ± 0.2**	**86.1 ± 0.3**	**5.9**	**1.4**	**O (D)**

Note: T: Temporal dimension, representing the number of time frames in the processed feature; D: Feature dimension, representing the dimensionality of features used in fusion operations.

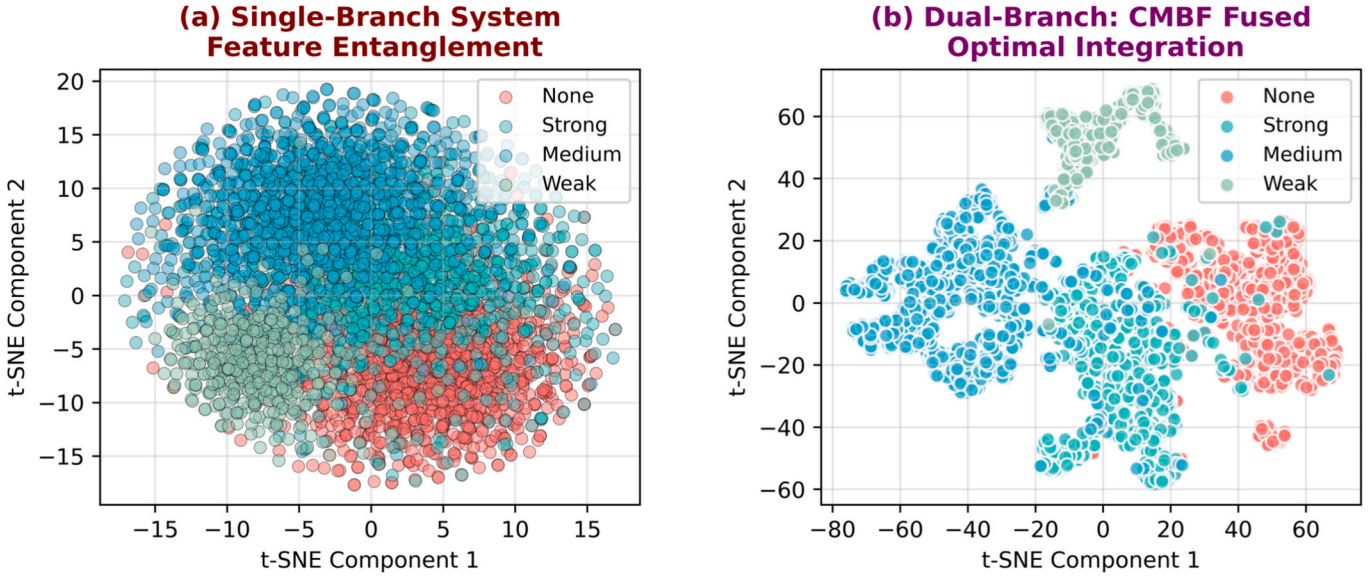


Fig. 5. T-sne visualization demonstrating feature entanglement and disentanglement. (a) conventional single-branch distillation exhibits feature entanglement with overlapping, poorly-separated clusters due to conflicting learning objectives. (b) our aquadistill approach achieves clear feature disentanglement with well-defined class boundaries, validating the effectiveness of decomposed cross-modal learning in preventing representational conflicts between static acoustic and dynamic behavioral features.

and temporal over-smoothing that diminishes the fine-grained temporal boundaries essential for accurate feeding intensity classification. The attention mechanisms aggregate information across temporal dimensions, potentially blurring the sharp transitions between different feeding states. We also evaluate alternative attention mechanisms adapted for our dual-branch fusion, including squeeze-and-excitation (SE) modules and convolutional block attention module (CBAM). When applied to our concatenated features, SE modules achieve only 84.1 % mAP, struggling with cross-branch interaction modeling due to their channel-wise focus. CBAM applied to fused features demonstrates 84.8 % mAP but introduces significant computational overhead (2.4 ms inference time vs 1.4 ms for CMBF) and fails to preserve the temporal locality crucial for feeding behavior recognition. These results highlight the importance of our specialized CMBF design for cross-branch adaptive weighting in acoustic behavioral analysis.

Our CMBF achieves the best performance (87.8 % mAP) while maintaining excellent efficiency (1.4 ms, 5.9 M parameters) due to its intelligent design that addresses the fundamental limitations of conventional fusion approaches. The adaptive sample-specific weighting mechanism enables CMBF to dynamically adjust the contribution balance between static and dynamic features based on the content characteristics of each audio sample. This contextual adaptation is particularly crucial for feeding intensity recognition, where optimal performance requires balancing static acoustic patterns and dynamic behavioral features. The CMBF's adaptive weighting mechanism based on feature similarity and interaction strength enables sample-specific fusion strategies, automatically adjusting the relative contributions of static and dynamic branches, according to the underlying acoustic characteristics of each input. Furthermore, CMBF's temporal locality preservation design maintains the discriminative temporal patterns essential for distinguishing between feeding intensity levels, avoiding the over-smoothing effects that plague traditional attention mechanisms. This preservation of fine-grained temporal boundaries is critical for detecting rapid transitions in feeding behavior, which often occur over short time scales but carry significant information about feeding intensity changes. The linear computational complexity O(D) ensures that these sophisticated fusion capabilities come without substantial computational overhead, making CMBF both effective and practical for real-time aquaculture monitoring applications.

5.3. Qualitative analysis

To gain deeper insight into the effectiveness of our decomposed cross-modal distillation, we conduct comprehensive qualitative analysis through feature visualization and training dynamics examination.

Fig. 6 demonstrates the progressive alignment between audio dynamic features and visual teacher representations throughout the training process. The cosine similarity curve reveals three distinct phases of knowledge transfer: an initial rapid alignment phase (epochs 0–50) where similarity increases from 0.55 to 0.68, indicating effective initial knowledge absorption; a gradual refinement phase (epochs 50–200) with steady improvement to 0.77, suggesting continuous feature space adaptation; and a convergence phase (epochs 200–300) where similarity stabilizes at the maximum value of 0.7693. This training trajectory demonstrates that our decomposed distillation strategy successfully guides the audio dynamic branch to learn visual behavioral patterns without forcing premature convergence, allowing for natural feature space evolution.

Fig. 7 demonstrates the progressive improvement of cross-modal alignment throughout training across three critical epochs. The t-SNE visualizations in Fig. 7(a) reveal the evolution of audio-visual feature alignment from epoch 40 to 280. At epoch 40, the overlap between audio dynamic features (red points) and visual teacher features (blue points) is limited, with distinct separation indicating initial learning stages. By epoch 150, substantial convergence begins to emerge, showing increased co-location of the two modalities. At epoch 280, the features achieve remarkable alignment with extensive overlap, confirming successful knowledge transfer from the visual teacher to the audio dynamic branch. Fig. 7(b) provides complementary evidence by visualizing feature distributions colored by class labels across the same temporal progression. The consistent class separability maintained throughout training demonstrates that the cross-modal alignment does not compromise discriminative capability. Each feeding intensity category—None (Class 0, green), Strong (Class 1, red), Medium (Class 2, orange), and Weak (Class 3, blue)—forms distinct, well-separated clusters that become increasingly coherent as training progresses. This preservation of class boundaries while achieving cross-modal alignment validates our decomposed distillation approach. The cosine similarity distributions in Fig. 7(c) provide quantitative validation of the alignment progression. The mean similarity improves substantially from

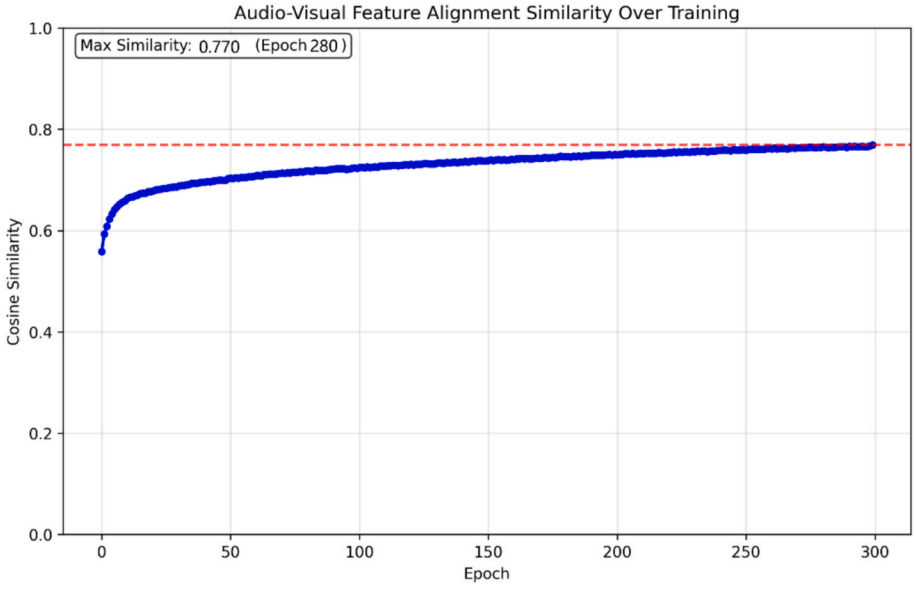


Fig. 6. Cross-modal feature alignment dynamics during training. The cosine similarity between audio dynamic features and visual teacher features progressively increases from 0.55 to 0.770 over 300 epochs, demonstrating effective knowledge transfer through our decomposed distillation approach. The red dashed line indicates the final maximum similarity achieved.

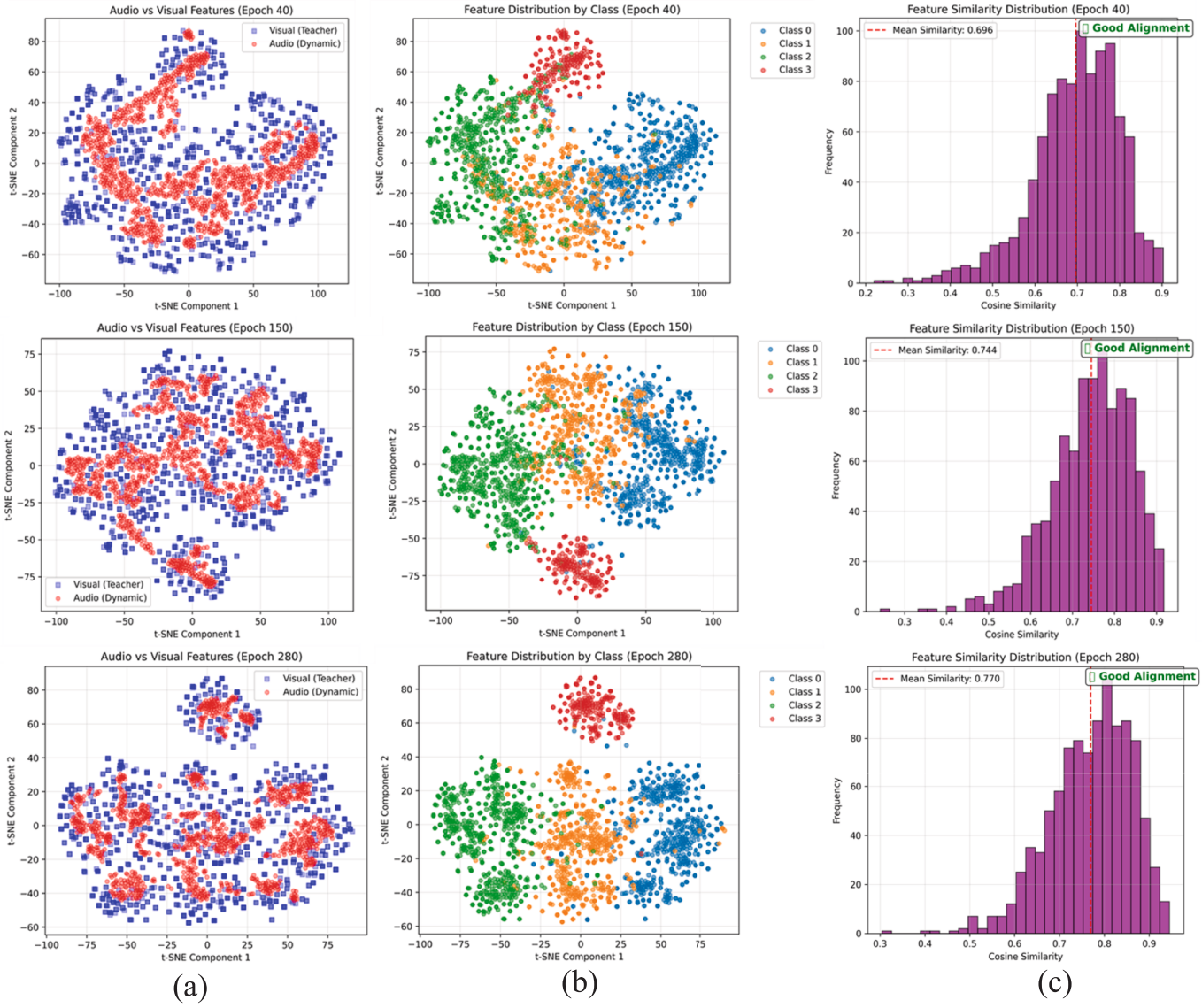


Fig. 7. Progressive cross-modal alignment analysis across training epochs. (a) t-SNE visualization showing the evolution of audio-visual feature alignment from epoch 40 to 280. Red points represent dynamic audio features and blue points represent visual teacher features. The increasing overlap demonstrates progressive knowledge transfer, with substantial alignment achieved by epoch 280. (b) Feature distribution by class labels (None: green, Strong: red, Medium: orange, Weak: blue) across the same epoch, showing maintained class separability throughout the alignment process. (c) Cosine similarity distributions between audio and visual features, with mean similarity improving from 0.696 (epoch 40) to 0.770 (epoch 280), indicating robust and consistent cross-modal knowledge transfer.

0.696 at epoch 40 to 0.744 at epoch 150, and finally reaches 0.770 at epoch 280, marked as "Good Alignment." The similarity histograms show a clear shift toward higher values over time, with the distribution becoming increasingly concentrated around 0.7–0.8 by epoch 280. This progression demonstrates consistent and robust knowledge transfer most samples rather than selective alignment for easy cases.

These results collectively validate the effectiveness of our decomposed cross-modal distillation approach. First, the progressive similarity improvement curve confirms that decomposed distillation enables controlled knowledge transfer without catastrophic forgetting of audio-specific information. Second, the maintained class separability while achieving high cross-modal similarity indicates that our method successfully transfers behavioral understanding without compromising discriminative capability. Finally, the consistent alignment across diverse samples demonstrates the generalizability of our cross-modal knowledge transfer mechanism, supporting its practical applicability for real-world aquaculture monitoring scenarios.

5.4. Backbone generalizability analysis

A key strength of our AquaDistill framework lies in its architecture-agnostic design, consistently delivering performance improvements regardless of the underlying backbone network. To demonstrate the generalizability and robustness of our AquaDistill framework, we conduct comprehensive experiments across various lightweight backbone architectures commonly used in resource-constrained applications. We evaluate our framework using five representative lightweight architectures: MobileNetV2 (baseline), MobileNetV3, MobileViT, EfficientNet-B0, and ShuffleNetV2. All backbones are adapted for audio spectrogram processing with identical input resolution ($1 \times 126 \times 128$) and training configurations to ensure fair comparison. Each backbone undergoes the same decomposed distillation process with our S3D visual teacher, maintaining consistent distillation loss weights and CMBF fusion mechanisms across all experiments.

Table 6 presents comprehensive results demonstrating that AquaDistill consistently improves performance across all tested backbones.

Table 6
Backbone generalizability analysis results.

Backbone	Params (M)	Model Size (MB)	Baseline mAP (%)	AquaDistill mAP (%)	Improvement	Acc (%/M)	Inference (ms)
MobileNetV2	5.9	23.6	82.6	89.0	+6.4	15.1	1.4
MobileNetV3	6.2	24.8	80.2	85.4	+5.2	13.8	1.5
MobileViT	8.7	34.8	79.8	84.6	+4.8	9.7	1.8
EfficientNetB0	5.3	21.2	81.1	86.2	+5.1	16.3	1.6
ShuffleNetV2	4.1	16.4	78.9	84.3	+5.4	20.6	1.2

The improvements range from 4.8 % (MobileViT) to 6.4 % (MobileNetV2) in mAP, with an average improvement of 5.6 %. Notably, our framework achieves the best absolute performance with MobileNetV2 (89.0 % mAP), which also maintains the optimal parameter-efficiency ratio (15.1 Acc%/M). This superior performance stems from MobileNetV2′s depth wise separable convolutions that naturally align with the spectral decomposition patterns in audio data, making it particularly receptive to our distillation-guided feature enhancement. MobileNetV3 shows substantial improvement (+5.2 % mAP) but with slightly higher computational overhead due to its squeeze-and-excitation modules. EfficientNet-B0 demonstrates good performance gains (+5.1 % mAP) while maintaining balanced parameter count (5.3 M), though its compound scaling design introduces unnecessary complexity for our audio processing task. MobileViT, despite being a vision transformer variant, shows the smallest improvement (+4.8 % mAP), confirming our earlier observation that transformer architectures designed for spatial relationships struggle with audio spectrograms' time–frequency characteristics. ShuffleNetV2 achieves notable improvement (+5.4 % mAP) with the smallest parameter footprint (4.1 M), making it attractive for ultra-lightweight deployments despite slightly lower absolute performance. The efficiency analysis reveals crucial insights for practical deployment decisions. While MobileNetV2 achieves the best accuracy-to-parameter ratio (15.1 Acc%/M), ShuffleNetV2 provides the best accuracy-to-size ratio for memory-constrained environments. MobileViT, despite its transformer heritage, requires significantly more parameters (8.7 M) for modest performance gains, highlighting the mismatch between vision transformers and audio processing requirements. These findings suggest that depthwise separable convolution architectures (MobileNetV2/V3, ShuffleNet) are naturally better suited for our cross-modal distillation approach.

Across all architectures, we observe consistent improvement patterns: decomposed distillation contributes the largest gains (2.4–2.8 %), CMBF fusion adds significant value (2.1–2.6 %), and framework optimization provides additional refinements (0.8–1.2 %). This consistency validates that our technical innovations address fundamental challenges in audio-based feeding recognition rather than exploiting architecture-specific characteristics. All tested backbones maintain real-time processing capabilities with inference times ranging from 1.2 ms (ShuffleNetV2) to 1.8 ms (MobileViT). The minimal speed differences demonstrate that our framework's computational overhead is negligible compared to backbone computation, ensuring that architecture selection can prioritize accuracy-parameter trade-offs without sacrificing real-time performance requirements for aquaculture monitoring applications.

5.5. Real-world environment validation

To demonstrate the real-world applicability of our method, we conducted experiments with a commercial aquaculture facility at the Guangzhou Aquatic Products Promotion Station, China (as shown in Fig. 8). We collected 8,900 audio-visual samples of *Lotus Carp* fish, a common aquaculture species, from a tank (4 m × 2 m × 3 m). This real-world environment presented challenges such as environmental noise, water surface reflection, and foams, which are not present in controlled settings. The dataset maintained the same specifications as our controlled dataset, with 2-second audio-visual clips annotated by experienced technicians. We randomly split the real-world dataset into training (70 %), validation (10 %), and testing (20 %) sets, resulting in 6,230, 890, and 1,780 samples, respectively.

Due to the turbid water conditions typical of commercial aquaculture facilities, visual classification performance was significantly impacted, achieving only 84 % accuracy compared to 92.3 % in controlled settings. The challenging visual conditions resulted from multiple factors: water turbidity reduced fish visibility, surface reflections created visual artifacts, and foam formation periodically obscured the monitoring area. The audio-only baseline similarly decreased to 79.3 % accuracy due to increased environmental noise and acoustic interference, representing a smaller performance degradation than the visual modality. We fine-

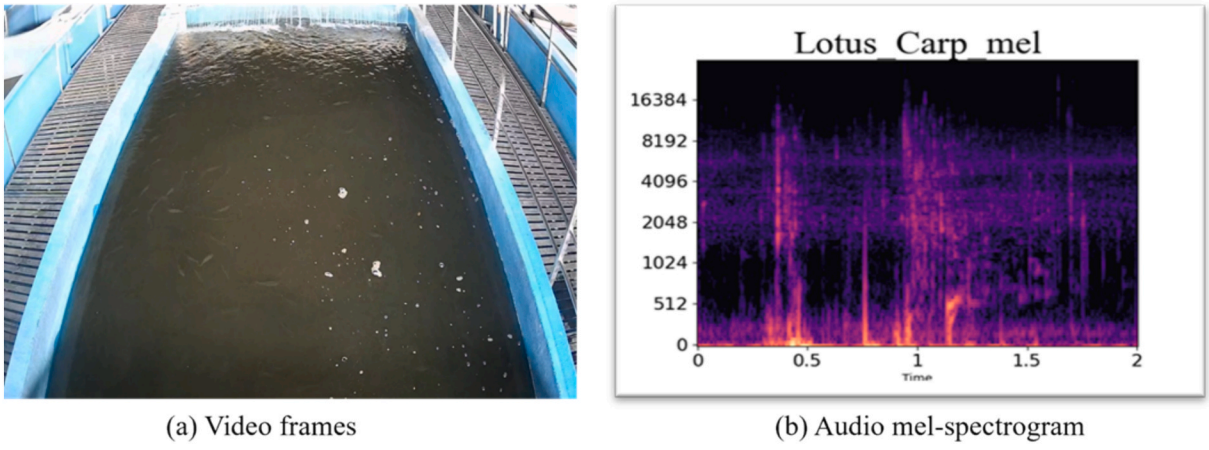


Fig. 8. Real-world aquaculture facility experimental setup and data characteristics. (a) A Commercial aquaculture tank at Guangzhou Aquatic Products Promotion Station showing turbid water conditions, surface foam, and *Lotus Carp* fish in a 4 m × 2 m × 3 m tank. (b) The corresponding mel-spectrogram of *Lotus Carp* feeding audio showing complex acoustic environment with background noise, water circulation sounds, and environmental interference patterns across the frequency spectrum from 0-16384 Hz over a 2-second duration.

tuned our pre-trained AquaDistill model on 6,230 Lotus Carp training samples using a two-stage approach to preserve the learned cross-modal alignment. Initial frozen-backbone training (20 epochs, lr = 5e-4) adapts the fusion module without disrupting audio-visual knowledge, followed by end-to-end refinement (30 epochs, lr = 1e-4) with early stopping to prevent overfitting on the smaller dataset. This fine-tuned AquaDistill achieved 83.6 % accuracy, demonstrating a 4.3 % improvement over the audio baseline and effectively narrowing the audio-visual gap to merely 0.4 % in this challenging real-world scenario.

The substantial improvement (79.3 % to 83.6 %) in real-world conditions validates several key aspects of our approach. First, our framework demonstrates excellent transferability across different species and environments, with the pre-trained knowledge successfully adapting to *Lotus Carp* from the original *Oplegnathus punctatus* dataset. Second, the reduced modality gap (0.4 % vs 5.3 % in controlled settings) indicates that acoustic signals maintain more consistent quality across environmental conditions, supporting our hypothesis that audio-based systems offer superior deployment reliability. Finally, the 5.4 % relative improvement in challenging conditions exceeds the controlled environment gains, suggesting that our method provides greater value in practical deployment scenarios where environmental factors limit visual system effectiveness.

6. Conclusion

In this paper, we have introduced AquaDistill, a novel cross-modal knowledge distillation framework designed to enhance audio-only fish feeding intensity recognition by transferring visual knowledge during training while requiring only acoustic input during inference. Our approach incorporates a decomposed distillation strategy that separates audio features into static acoustic and dynamic behavioral branches, CMBF for intelligent feature integration, and hybrid distillation losses enabling effective knowledge transfer while avoiding feature entanglement. Our framework achieves 89 % mAP and 87 % accuracy, representing improvements of 7 % and 5 % respectively over baseline approaches while maintaining exceptional computational efficiency with only 5.9 M parameters and 1.2 ms inference time. The method successfully reduces the audio-visual performance gap by 63 %, demonstrating robust generalizability across various lightweight backbones and superior performance in challenging commercial aquaculture conditions. Our results show the potential for deploying high-accuracy feeding intensity recognition systems in challenging underwater environments where visual systems may fail due to water turbidity or lighting conditions, indicating that AquaDistill may be valuable for resource-constrained edge deployments in commercial aquaculture monitoring applications. The successful cross-species adaptation from *Oplegnathus punctatus* to *Lotus Carp* further confirms the framework's practical applicability across diverse aquaculture species and environments. The cross-modal feature alignment visualization provides compelling evidence of successful knowledge transfer, validating the effectiveness of our decomposed distillation approach for optimizing feeding management in aquaculture settings. Our study has several limitations. First, validation is limited to two fish species (*Oplegnathus punctatus* and *Lotus Carp*), constraining generalizability claims across diverse aquaculture species. Second, the performance gap between controlled (89.0 % mAP) and real-world environments (83.6 % mAP) suggests scalability challenges in highly variable commercial settings with different water conditions and environmental factors. To address these limitations, future work could explore extending applicability to diverse aquaculture species and integrating additional modalities such as environmental sensors. Furthermore, investigating the application of our cross-modal distillation approach to other aquaculture monitoring tasks such as fish health assessment could establish a comprehensive framework for intelligent aquaculture management systems.

CRediT authorship contribution statement

Meng Cui: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Tan Wang:** Writing – original draft, Data curation. **Xinhao Mei:** Visualization, Investigation. **Jinzheng Zhao:** Writing – review & editing, Formal analysis. **Daoliang Li:** Validation, Software, Project administration, Funding acquisition. **Wenwu Wang:** Writing – review & editing, Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the Key Projects of Natural Science Research in Anhui Universities (Grant No. 2024AH050462), the Key Laboratory of Agricultural Sensors, Ministry of Agriculture and Rural Affairs (Grant No. KLAS2023KF002), the National Natural Science Foundation of China "Dynamic Regulation Mechanism of Nitrogen in Industrial Aquaponics Under Asynchronous Life Cycle Conditions" (Grant No. 32373186), the China Scholarship Council Research Scholarship (No. 202006350248) and the University of Surrey (through a fee-waiver studentship).

Data availability

Data will be made available on request.

References

Li, D., Wang, Z., Wu, S., Miao, Z., Du, L., Duan, Y., 2020. Automatic recognition methods of fish feeding behavior in aquaculture: A review. Aquaculture 528, 735508.
Zhao, H., Wu, J., Liu, L., Qu, B., Yin, J., Yu, H., Zhou, C., 2024. A real-time feeding decision method based on density estimation of farmed fish. Front. Mar. Sci. 11, 1358209.
Siddik, M.A., Julien, B.B., Islam, S.M., Francis, D.S., 2024. Fermentation in aquafeed processing: Achieving sustainability in feeds for global aquaculture production. Rev. Aquac. 16 (3), 1244–1265.
Roberts, S., Jacquet, J., Majluf, P., Hayek, M.N., 2024. Feeding global aquaculture. Sci. Adv. 10 (42), eadn9698.
Cui, M., Liu, X., Zhao, J., Sun, J., Lian, G., Chen, T., Wang, W., 2022. In: August). Fish Feeding Intensity Assessment in Aquaculture: A New Audio Dataset AFFIA3K and a Deep Learning Algorithm. IEEE, pp. 1–6.
Zhang, L., Li, B., Sun, X., Hong, Q., Duan, Q., 2023. Intelligent fish feeding based on machine vision: a review. Biosyst. Eng. 231, 133–164.
Wang, Y., Xin, C., Gao, Y., Li, P., Wang, M., Wu, S., Hu, J., 2024. Advancing selective breeding in leopard coral grouper (P. leopardus) through development of a high-throughput image-based growth trait. Agric. Commun. 2 (2), 100042.
Cui, M., Liu, X., Liu, H., Zhao, J., Li, D., Wang, W., 2025. Fish tracking, counting, and behaviour analysis in digital aquaculture: a comprehensive survey. Rev. Aquac. 17 (1), e13001.
Zhao, J., Bao, W.J., Zhang, F.D., Ye, Z.Y., Liu, Y., Shen, M.W., Zhu, S.M., 2017. Assessing appetite of the swimming fish based on spontaneous collective behaviors in a recirculating aquaculture system. Aquac. Eng. 78, 196–204.
Zhou, C., Lin, K., Xu, D., Chen, L., Guo, Q., Sun, C., Yang, X., 2018. Near infrared computer vision and neuro-fuzzy model-based feeding decision system for fish in aquaculture. Comput. Electron. Agric. 146, 114–124.
Zhou, C., Zhang, B., Lin, K., Xu, D., Chen, C., Yang, X., Sun, C., 2017. Near-infrared imaging to quantify the feeding behavior of fish in aquaculture. Comput. Electron. Agric. 135, 233–241.
Feng, S., Yang, X., Liu, Y., Zhao, Z., Liu, J., Yan, Y., Zhou, C., 2022. Fish feeding intensity quantification using machine vision and a lightweight 3D ResNet-GloRe network. Aquac. Eng. 98, 102244.
Zhang, Y., Xu, C., Du, R., Kong, Q., Li, D., Liu, C., 2023. MSIF-MobileNetV3: an improved MobileNetV3 based on multi-scale information fusion for fish feeding behavior analysis. Aquac. Eng. 102, 102338.
Hu, W., Yang, X., Ma, P., Fu, T., Zhou, C., 2025. DCA-MVIT: Fused DSGated convolution and CA attention for fish feeding behavior recognition in recirculating aquaculture systems. Aquaculture 598, 742008.
Du, Z., Cui, M., Wang, Q., Liu, X., Xu, X., Bai, Z., Li, D., 2023. Feeding intensity assessment of aquaculture fish using Mel Spectrogram and deep learning algorithms. Aquac. Eng. 102, 102345.

Cui, M., Liu, X., Liu, H., Du, Z., Chen, T., Lian, G., Wang, W., 2024. Multimodal fish feeding intensity assessment in aquaculture. IEEE Trans. Autom. Sci. Eng.

Du, Z., Cui, M., Xu, X., Bai, Z., Han, J., Li, W., Li, D., 2024. Harnessing multimodal data fusion to advance accurate identification of fish feeding intensity. Biosyst. Eng. 246, 135–149.

Gao, R., Oh, T.H., Grauman, K., Torresani, L., 2020. Listen to look: Action recognition by previewing audio. In: In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10457–10467.

Lin, Y.B., Lei, J., Bansal, M., Bertasius, G., 2022. In: October). Eclipse: Efficient Long-Range Video Retrieval Using Sight and Sound. Springer Nature Switzerland, Cham, pp. 413–430.

Lv, X., Zhang, X., Gao, H., He, T., Lv, Z., Zhangzhong, L., 2024. When crops meet machine vision: a review and development framework for a low-cost nondestructive online monitoring technology in agricultural production. Agric. Commun. 2 (1), 100029.

Li, D., Sun, J., Liu, Y., 2025. Hierarchical semantic alignment heterogeneous knowledge distillation model for smart agriculture crop leaf disease recognition. Expert Syst. Appl. 129100.

Espejo-Garcia, B., Güldenring, R., Nalpantidis, L., Fountas, S., 2025. Foundation vision models in agriculture: DINOv2, LoRA and knowledge distillation for disease and weed identification. Comput. Electron. Agric. 239, 110900.

Sai, S., Kumar, S., Gaur, A., Goyal, S., Chamola, V., Hussain, A., 2025. Unleashing the power of generative AI in agriculture 4.0 for smart and sustainable farming. Cogn. Comput. 17 (1), 63.

Iqbal, U., Li, D., Du, Z., Akhter, M., Mushtaq, Z., Qureshi, M.F., Rehman, H.A.U., 2024. Augmenting aquaculture efficiency through involutional neural networks and self-attention for oplegnathus punctatus feeding intensity classification from log mel spectrograms. Animals 14 (11), 1690.

Li, D., Du, Z., Wang, Q., Wang, J., Du, L., 2024. Recent advances in acoustic technology for aquaculture: A review. Rev. Aquac. 16 (1), 357–381.

Huo, F., Xu, W., Guo, J., Wang, H., Guo, S., 2024. C2kd: Bridging the modality gap for cross-modal knowledge distillation. In: In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16006–16015.

Wang, Z., Li, D., Luo, C., Xie, C., Yang, X., 2023. Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation. In: In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8637–8646.

Mansourian, A. M., Ahmadi, R., Ghafouri, M., Babaei, A. M., Golezani, E. B., Ghamchi, Z. Y., ... & Kasaei, S. (2025). A comprehensive survey on knowledge distillation. *arXiv preprint arXiv:2503.12067*.

Wang, T., Li, F., Zhu, L., Li, J., Zhang, Z., Shen, H.T., 2025. Cross-modal retrieval: a systematic review of methods and future directions. *Proceedings of the IEEE*.

Gao, L., Shi, P., Hu, L., Feng, J., Zhu, L., Wan, L., Feng, W., 2024. Cross-modal knowledge distillation for continuous sign language recognition. Neural Netw. 179, 106587.

Kwak, M.G., Mao, L., Zheng, Z., Su, Y., Lure, F., Li, J., 2025. A cross-modal mutual knowledge distillation framework for alzheimer's disease diagnosis: addressing incomplete modalities. IEEE Trans. Autom. Sci. Eng.

Moslemi, A., Briskina, A., Dang, Z., & Li, J. (2024). A survey on knowledge distillation: recent advancements. *Machine Learning with Applications*, 100605.

Hu, C., Li, X., Liu, D., Wu, H., Chen, X., Wang, J., & Liu, X. (2023). Teacher-student architecture for knowledge distillation: A survey. *arXiv preprint arXiv:2308.04268*.

Kour, V., Kumar, S., Reddy, T.V., Poojary, K., Misra, R., Singh, T.N., 2025. Exploring spatiotemporal relational learning with timesformer for identifying the severity of the road accidents. IEEE Trans. Comput. Social Syst.

Li, B., Chen, J., Li, G., Zhang, D., Bao, X., Huang, D., 2025. Cross-modal contrastive masked autoencoder for compressed video pre-training. IEEE Trans. Image Process.

Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K., 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 305–321.

Cui, M., Yue, X., Qian, X., Zhao, J., Liu, H., Liu, X., ... & Wang, W. (2025). Audio-visual class-incremental learning for fish feeding intensity assessment in aquaculture. *arXiv preprint arXiv:2504.15171*.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., Plumbley, M.D., 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 28, 2880–2894.

Ennadir, S., Lutzeyer, J., Vazirgiannis, M., Bergou, E.H., 2024. If you want to be robust, be wary of initialization. Adv. Neural Inf. Proces. Syst. 37, 23796–23823.

Yang, J., Jia, Q., Han, S., Du, Z., Liu, J., 2025. An Efficient Multi-Scale attention two-stream inflated 3D ConvNet network for cattle behavior recognition. Comput. Electron. Agric. 232, 110101.

Al-Khater, W., Al-Madeed, S., 2024. Using 3D-VGG-16 and 3D-Resnet-18 deep learning models and FABEMD techniques in the detection of malware. Alex. Eng. J. 89, 39–52.

Zhang, D., Liang, D., Tan, Z., Ye, X., Zhang, C., Wang, J., Bai, X., 2024. In: September). Make Your Vit-Based Multi-View 3d Detectors Faster via Token Compression. Springer Nature Switzerland, Cham, pp. 56–72.

Gu, X., Zhao, S., Duan, Y., Meng, Y., Li, D., Zhao, R., 2025. MMFINet: a multimodal fusion network for accurate fish feeding intensity assessment in recirculating aquaculture systems. Comput. Electron. Agric. 232, 110138.

Du, Z., Xu, X., Bai, Z., Liu, X., Hu, Y., Li, W., Li, D., 2023. Feature fusion strategy and improved GhostNet for accurate recognition of fish feeding behavior. Comput. Electron. Agric. 214, 108310.

Zeng, Y., Yang, X., Pan, L., Zhu, W., Wang, D., Zhao, Z., Zhou, C., 2023. Fish school feeding behavior quantification using acoustic signal and improved Swin Transformer. Comput. Electron. Agric. 204, 107580.