

D-CAT: Decoupled Cross-Attention Transfer between Sensor Modalities for Unimodal Inference

Leen Daher, Zhaobo Wang, and Malcolm Mielle[✉]

Abstract—Cross-modal transfer learning is used to improve multi-modal classification models (e.g., for human activity recognition in human-robot collaboration). However, existing methods require paired sensor data at both training and inference, limiting deployment in resource-constrained environments where full sensor suites are not economically and technically usable. To address this, we propose Decoupled Cross-Attention Transfer (D-CAT), a framework that aligns modality-specific representations without requiring joint sensor modality during inference. Our approach combines a self-attention module for feature extraction with a novel cross-attention alignment loss, which enforces the alignment of sensors’ feature spaces without requiring the coupling of the classification pipelines of both modalities. We evaluate D-CAT on three multi-modal human activity datasets (IMU, video, and audio) under both in-distribution and out-of-distribution scenarios, comparing against uni-modal models. Results show that in in-distribution scenarios, transferring from high-performing modalities (e.g., video to IMU) yields up to +10% F1-score gains over uni-modal training. In out-of-distribution scenarios, even weaker source modalities (e.g., IMU to video) improve target performance, as long as the target model isn’t overfitted on the training data. By enabling single-sensor inference with cross-modal knowledge, D-CAT reduces hardware redundancy for perception systems while maintaining accuracy, which is critical for cost-sensitive or adaptive deployments (e.g., assistive robots in homes with variable sensor availability). Code is available at <https://github.com/Schindler-EPFL-Lab/D-CAT>.

I. INTRODUCTION

The rapid miniaturization and cost reduction of sensors have allowed significant progress in robotics, particularly in human-robot interaction, collaborative systems, and imitation learning. By generating richer, more informative data, significant progress has been achieved in how robots can perceive, interpret, and respond to human behavior.

Yet, a critical challenge persists: while multi-sensor fusion enhances accuracy, deploying such systems at scale remains impractical. In real-world applications—especially in industrial environments—the feasibility of equipping every operational unit or worker with extensive multi-sensor arrays (e.g., cameras, microphones, and IMUs) is limited by cost, complexity, and scalability. Although time-limited data collection campaigns can leverage multiple sensors to curate rich training datasets, reproducing this level of sensory richness in real-world deployments is often economically or logistically infeasible.

To solve this situation, one possible solution is to develop a system that exploits multi-modal data during training but operates with only a subset of sensors at inference time. In such a framework, each modality’s classification strategy must remain independent, while knowledge is transferred

between all modalities to maximize individual performance through collaborative training. Yet, current multi-modal fusion methodologies assume the presence of all sensor modalities at both training and inference; a coupling that renders them incompatible with real-world constraints. This disparity between what is achievable at small and large scales poses the question of how to design systems that retain the benefits of multi-modal learning, while avoiding the practical limitations of real-world robotics?

In this work, we introduce Decoupled Cross-Attention Knowledge Transfer (D-CAT), a novel framework for decoupled knowledge transfer between arbitrary sensor modalities. Our approach employs a modality-specific encoder together with a self-attention mechanism to capture modality-specific features. To enable knowledge transfer between two sensor modalities, we propose a novel cross-attention loss function that optimizes the target modality’s embeddings as a linear combination of the source modality’s embeddings. This ensures knowledge transfer from source to target while preserving the target modality’s accuracy.

We validate our method through extensive experiments on three datasets, evaluating transfer performance across IMU, video, and audio modalities. Our framework enables scalable sensor-fusion by leveraging multi-modal data during training while only requiring a subset of sensors at inference.

II. RELATED WORK

A. Human Action Recognition

Human action recognition (HAR) has been extensively studied [1], [2], [3], [4] across diverse sensor modalities, driven by applications in human-robot interactions [4], video retrieval [5], or safety [6]. While RGB video or image-based approaches are the most popular due to the ubiquity of cameras [2], researchers have explored alternatives—including depth sensors [7], audio sensors [8], or infrared thermal images [9] for low-light robustness—to improve accuracy, computational cost, and privacy.

Multi-modal HAR methods frequently pairs video with depth [7], [10], [11] (e.g., via a Kinectsensor) or audio [12], [13], leveraging complementary strengths in both sensor modalities. However, visual data introduces privacy risks and processing overhead, while audio—though lightweight—may lack discriminative power for fine-grained actions. IMUs offer a low-cost, privacy-preserving alternative for HAR [14], [15], [16] that is already present in many common devices (e.g., smartphones and smartwatches). However, IMU data suffers from noise and ambiguity for similar gestures. Furthermore, a common limitation of all the previous works us-

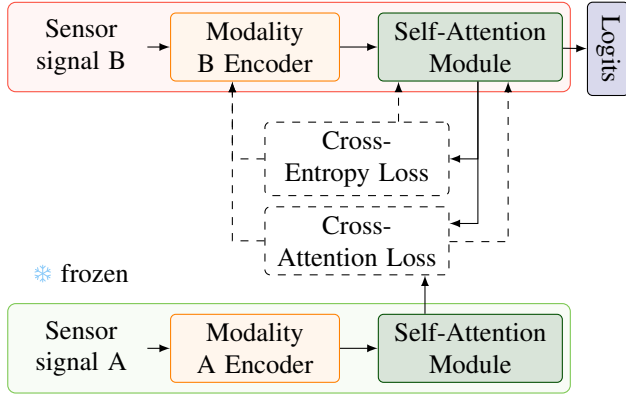


Fig. 1: **Cross-modal transfer architecture.** Sensor modality A (source) and B (target) classification networks each consist of an encoder and a self-attention module. Modality A network is pretrained and frozen during the training of modality B’s classification network. The classification network of modality B is then trained against a standard cross-entropy loss as well as a novel cross-attention loss that aims to align the two modalities’ query and value embeddings, thus enabling knowledge transfer from A to B. The dashed arrows indicate the backpropagation paths from the losses to the trainable components of the network.

ing multi-modal sensor data is the reliance on coupled classification methodologies; it assumed that all sensor modalities are always available, at training and during use.

B. Cross-Modal Transfer Learning

Cross-modal transfer learning focuses on transferring knowledge between sensor modalities. Given a source and target modality, some work [17], [18], [19] first transform the source modality data to a feature set compatible with the target modality. For example, Ma et al. [18] train a modality-specific encoder that outputs features that are directly used to pre-train a backbone model used on the target modality. Then they refine both the encoder and the backbone model using data from the target modality, leading to higher accuracy in the predictions. On the other hand, Kwon et al. [19] generate virtual accelerometry data from 3D human poses extracted from videos. The virtual IMU data is used to pre-train an HAR model that is finetuned on real IMU data, yielding a 10% improvement in accuracy.

Other methods focus on aligning the feature representations of different models [20], [21], [22], [23], [24]. Deldari et al. [22] proposes a simple yet effective cross-modal contrastive loss that learns joint representations by pulling together temporally aligned samples from different sensors (positives) and pushing apart temporally distant or unrelated samples (negatives). Wei et al. [23] propose a method that transfers knowledge between image and text models. Their method leverages both intra-and inter-modal information is implemented through self-attention and cross-attention modules, respectively.

Similarly to Section II-A, a common limitation of the

previous works is the assumption that all sensor modalities are present at both training and inference times; it is not possible to use them for applications where only a subset of sensors would be available at inference time. Unlike prior efforts focused on coupled modality pairs, we propose to transfer knowledge between arbitrary sensors at training time, while maintaining the ability to use a single modality at inference.

III. PRELIMINARIES

In this section, we present the concepts of self-attention and cross-attention, since those are instrumental to our method.

A. Self-attention

Self-attention [25] assigns scores that measure the importance of each element with respect to the other elements in the same sequence. These scores are calculated by mapping the embeddings of a model into queries, keys, and values, which are then used to generate new weighted embeddings.

The projection into the key, query, and value spaces is implemented using linear layers as in Eq. 1:

$$\begin{aligned} K &= EW_K, & W_K &\in \mathbb{R}^{d_{model} \times d_k}, & K &\in \mathbb{R}^{SL \times d_k} \\ Q &= EW_Q, & W_Q &\in \mathbb{R}^{d_{model} \times d_k}, & Q &\in \mathbb{R}^{SL \times d_k} \\ V &= EW_V, & W_V &\in \mathbb{R}^{d_{model} \times d_v}, & V &\in \mathbb{R}^{SL \times d_v} \end{aligned} \quad (1)$$

Here, SL is the sequence’s length and $E \in \mathbb{R}^{SL \times d_{model}}$ is the embedding matrix for the input sequence of the model. The matrices W_K , W_Q , and W_V are learnable parameters that project the embeddings into the key, query, and value spaces, respectively. d_{model} is the dimension of the input embeddings and may differ across modalities, while d_k and d_v are the projection dimensions, referred to as d_{out}

The self-attention output is computed as:

$$\text{Self-Attention}(K, Q, V) = \text{softmax} \left(\frac{QK^\top}{\sqrt{d_{out}}} \right) V \quad (2)$$

This operation produces a new sequence of the same length, where each element is a weighted sum of all elements in the original sequence.

B. Cross-Attention

On the other hand, cross-attention relates elements from two different sequences; the query is taken from one sequence and the keys and values from another.

Cross-attention from modality A (source) to modality B (target) is expressed as:

$$\text{Cross-Attention}(K, Q, V) = \text{softmax} \left(\frac{Q_B K_A^\top}{\sqrt{d_{out}}} \right) V_A. \quad (3)$$

where Q_B is the target model’s query matrix, and K_A and V_A are the source model’s key and value matrices, respectively.

IV. METHOD

D-CAT leverages modality-specific encoders with self-attention to extract discriminative features from individual sensor inputs. Then, to align the target and source feature spaces, we propose a novel cross-attention loss that enforces correspondence between modalities, while preserving architectural decoupling in their downstream classification networks. During training, the source encoder and self-attention modules remain frozen, enabling transfer to the target domain without changing the source’s pretrained feature space.

While we will demonstrate our framework on three common sensor modalities in robotic tasks—inertial measurement unit (IMU) signals, video sequences, and audio signals—in Section VI, it should be noted that, given a modality encoder, D-CAT’s design generalizes to arbitrary sensor types, offering a flexible solution for modality-agnostic knowledge transfer. The overall framework is presented in Figure 1.

A. Self-Attention Module

The sequences of embeddings extracted from the modality-specific encoders are used as input to a self-attention module. The self-attention module captures dependencies within each modality as defined in Section III-A, following Eq. (2). Specifically, the self-attention module provides the key, query, and value independently for each modality. Those keys, queries, and values will be used by the proposed cross-attention loss of this paper.

B. Cross-Attention Loss

As shown in Section III, cross-attention relates the elements from two different sequences by using the query from one sequence and the key and value from the other sequence. However, cross-attention creates a strong dependency between modalities, where the two sensor data cannot be decoupled at inference. To remove this dependency, we propose a novel cross-attention loss during training to align the cross-attention embeddings (QK^TV) of each modality.

Given a pretrained source classification network from which we aim to transfer knowledge to a target classification network, we first freeze the weights of the source model. Hence, the cross-attention loss will only updates the target modality encoder and self-attention module, aligning them with the fixed source representations.

The proposed cross-attention loss aims to satisfy:

$$Q_B K_B^T V_B \approx Q_B K_A^T V_A \quad (4)$$

with Q_B , K_B , and V_B the query, key, and value projections of the target modality embeddings, and K_A , V_A the key and value projections of the source modality embeddings.

Since the aim is to align the models’ embeddings through minimizing the distance between the self-attention and cross-attention embeddings, the Frobenius norm is used to obtain the Euclidean length between each element of a matrix:

$$\|A\|_F = \sqrt{\sum_{i,j} |a_{ij}|^2}, \quad (5)$$

where a_{ij} denotes the element at the i^{th} row and j^{th} column. This yields the following cross-attention loss:

$$L_{CA} = \|\overline{Q_B K_B^T V_B} - \overline{Q_B K_A^T V_A}\|_F \quad (6)$$

where $\|\cdot\|$ denotes the normalized matrices.

The query matrices can be factored out from each side for modeling simplification, yielding a final cross-attention loss that still pushes the two attention maps to be similar:

$$L_{CA} = \|\overline{K_B^T V_B} - \overline{K_A^T V_A}\|_F \quad (7)$$

The loss aligns self-attention ($K_A^T V_A$) and cross-attention ($K_B^T V_B$) values, assuming that the softmax and scaling of Eq. (2) are dropped. Indeed, their removal does not affect the alignment quality, since, as proven in Theorem 4.1, K_B converges to a linear mapping of K_A , and similarly, V_B converges to a linear mapping of V_A . Hence, the attention terms remain comparable, ensuring that the alignment between modalities is preserved even without the softmax and scaling.

Theorem 4.1: Assume matrix $M \in \mathbb{R}^{m \times n}$ with $rank(M) = r$ where:

$$M = AB = A'B', \quad (8)$$

where $A, A' \in \mathbb{R}^{m \times k}$ and $B, B' \in \mathbb{R}^{k \times n}$. Then there exist matrices $R \in \mathbb{R}^{k \times k'}$ and $S \in \mathbb{R}^{k' \times k}$ s.t:

$$\begin{aligned} A' &= AR \\ B' &= SB \end{aligned} \quad (9)$$

Proof: Since $M = AB$, every column of M lies in the span of the columns of A . Likewise, since $M = A'B'$, every column of M lies in the span of the columns of A' . Hence, the column-space of M is contained in both the span of A and the span of A' . If that is not the case, then A' cannot produce the columns of M .

So far, we have proved that every column of A' is a linear combination of the columns of A . This can be reformulated as:

$$a'_j = r_{1j}a_1 + r_{2j}a_2 + \dots + r_{rj}a_k, \quad (10)$$

where a'_j is a column in A' , (a_1, \dots, a_k) are the columns of A , and (r_{1j}, \dots, r_{rj}) are the coefficients needed to get the value of the column in A' . These coefficients can be collected into a matrix R :

$$A' = AR \quad (11)$$

Similarly, because $M = AB = A'B'$, each row of M lies in the span of the rows of B and also in the span of the rows of B' . Therefore, every row of B' can be written as a linear combination of the rows of B . This can be written as

$$b'_i = s_{1i}b_1 + s_{2i}b_2 + \dots + s_{ri}b_k, \quad (12)$$

where b'_i is a row in B' , (b_1, \dots, b_k) are the rows of B , and (s_{1i}, \dots, s_{ri}) are the coefficients. These coefficients can be collected into a matrix S :

$$B' = SB \quad (13)$$

Finally, substituting Eq. (11) and Eq. (13) into Eq. (8) yields:

$$\begin{aligned} AB &= A'B' \\ &= (AR)(SB) \end{aligned} \quad (14)$$

Applying Theorem 4.1 to our case, where $A = K_A^T$, $A' = K_B^T$, $B = V_A$, and $B' = V_B$ yields:

$$\begin{aligned} K_B^T &\rightarrow K_A^T R \\ V_B &\rightarrow S V_A \end{aligned} \quad (15)$$

where R and S are the linear map matrices.

Thus, it is proven that the cross-attention loss enforces K_B and V_B to converge to linear mappings of K_A and V_A , respectively, ensuring that the cross-attention term $K_B^T V_B$ remains aligned with the self-attention term $K_A^T V_A$.

C. Masked Cross-Modal Alignment

Since no classification model is a perfect oracle, all models will return erroneous classifications for some input data. However, such incorrect classification should not be used for knowledge transfer since the source model has not been able to learn from the extracted features.

To avoid aligning embeddings of wrongly classified samples from the source modality, we use an indicator function $\mathbb{1}$ with the cross-attention loss to take the cross-attention loss into account only for embeddings of correctly classified samples: This $\mathbb{1}$ is expressed as:

$$\mathbb{1}(x) = \begin{cases} 1 & \text{if } x \text{ is a correct classification,} \\ 0 & \text{if } x \text{ is an incorrect classification.} \end{cases} \quad (16)$$

Hence, the final cross-attention loss is expressed as:

$$L_{CA} = \mathbb{1}(x) \cdot \| \overline{K_B^T V_B} - \overline{K_A^T V_A} \|_F, \quad (17)$$

D. Classification Loss

Finally, the cross-entropy loss is adopted for classification:

$$L_{CE} = -\frac{1}{N} \sum_{n=1}^N \log\left(\frac{\exp(x_{n,y_n})}{\sum_{c=1}^C \exp(x_{n,c})}\right) \quad (18)$$

where N is the total number of samples, C is the total number of classes, x_{n,y_n} is the true logit of the n^{th} sample of the current batch, and $x_{n,c}$ is the model output logit for the class c .

The complete loss function is:

$$Loss = L_{CE} + \lambda L_{CA}, \quad (19)$$

where λ is a non-negative hyper-parameter that balances the contribution of the cross-attention loss L_{CA} relative to the cross-entropy loss L_{CE}

V. EXPERIMENTS

In this section, we discuss the metrics used to measure the performance of knowledge transfer from source modality A to target modality B, the datasets, and the baseline model architectures. The implementation of D-CAT and all baselines can be found online.¹

¹<https://github.com/Schindler-EPFL-Lab/D-CAT>

A. Datasets

The proposed method is evaluated on three multi-modal datasets, each containing synchronized data from two sensor modalities.

1) Video-IMU: the University of Electronic Science and Technology of China Multi-Modal Egocentric Activity Dataset for Continual Learning (UESTC-MMEA-CL) [26] contains synchronized Video-IMU data collected from 10 participants performing 32 everyday activities. Each activity lasts between 8 and 35 seconds and is recorded using wearable smart glasses equipped with an RGB camera and an IMU sensor.

To reduce training complexity and hardware requirements, we use a subset of the 8 activity classes, containing 6,522 samples. These classes—drinking, reading, floor sweeping, cutting fruits, washing hands, typing on a laptop, typing on a phone, and opening/closing a door—were selected because they exhibit a high degree of variability in motion intensity, interaction with objects, and sensor signal characteristics. For the experiment, the dataset is split into 80% training, 10% validation, and 10% test sets.

2) Audio-IMU: the Cough Audio-IMU multimodal cough dataset using wearables [27] consists of synchronized IMU-Audio data recorded from 13 participants performing 8 tasks, such as coughing, sneezing, and laughing. The dataset is imbalanced, with some classes having more samples than others. To address this, samples with the shortest IMU sequences, carrying minimal motion information, were removed. After this filtering, the dataset contains 2,576 samples, split into 65% training, 18% validation, and 17% testing.

3) Audio-Video: the Visual Geometry Group-Sound (VG-GSound) [28] dataset contains audio-video data sourced from YouTube covering 309 classes, each with a duration of 10 seconds. Due to the computational cost of processing the entire dataset, a random subset of 9 classes was selected for evaluation. This subset contains 2,886 clips, split into 70% training, 15% validation, and 15% testing.

B. Data Preprocessing and Modality-Specific Encoders

Each of the sensor pipeline starts with a modality-specific encoder to extract high-level feature embeddings specific to the sensor data. Such embeddings will be used for the self-attention module presented in Section IV-A. Each encoder should map the input data into sequences of embeddings of dimension $(T_m \times d_{model})$, where d_{model} is model dependent.

Since this paper focuses on IMU, video, and audio data, we present a data preprocessing strategy and encoder architecture for each sensor modality. Across all three modalities, the convolution block shown in the upper section of Fig. 2 serves as a fundamental building unit of the encoder backbones.

1) IMU: IMU signals are first min-max normalized to the range $[-1, 1]$. They are then processed by a 1D Convolutional Neural Network (CNN) backbone as proposed by Mollyn et al. [29], producing a sequence of embeddings $E^{imu} \in \mathbb{R}^{T_{imu} \times d_{model}}$. While the overall backbone design

TABLE I: Training Parameters

Hyper-parameter	UESTC-IMU	UESTC-Video	Cough-IMU	Cough-Audio	VGGSound-Video	VGGSound-Audio
Epochs	10	100	10	100	100	150
Learning Rate	5×10^{-4}	5×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}
Dropout	0.8	0.8	0.8	0.6	0.5	0.5
Batch Size	16	8	32	32	8	8
Weight Decay	0.005	0.005	0.001	0.0001	0.005	0.0
Window/Stride Length	70	30	60	0.5 Seconds	30	2 Seconds
Seed	11	17	7	17	17	11

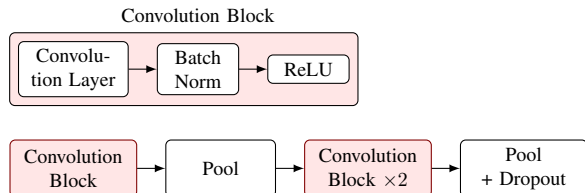


Fig. 2: **IMU encoder architecture**. The model begins with a convolution layer, batch normalization, and ReLU activation, followed by a pooling layer. Two additional convolutional blocks are applied, and the encoder concludes with a pooling and dropout stage.

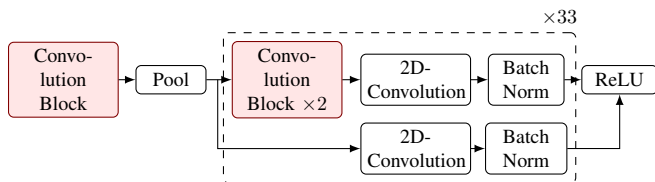


Fig. 3: **Image encoder architecture**. The model begins with an initial convolution block and pooling layer, followed by two stacked convolution blocks. A residual block, repeated 33 times, combines pairs of 2D convolutions with batch normalization and a ReLU activation.

follows the Somasa architecture, the specific parameters of the convolutional and subsequent layers (kernel sizes, strides, and number of filters) are adjusted for our datasets. The full IMU-encoder architecture can be seen in Fig. 2.

2) *Video*: Each frame of a video sequence is first resized to a fixed resolution (224, 224). The sequence is then passed through a 2D-CNN based backbone Image ResNet-101 that extracts features from each frame. The full architecture is shown in Fig. 3.

3) *Audio*: Audio signals are first converted to time-frequency representations using Short-Time Fourier (STFT). The resulting spectrograms are passed through a Mel filter bank to get the Mel spectrogram that represents the energy levels for each of the frequency bands. Finally, to obtain a more stable signal, the logarithm of the Mel spectrogram is calculated and flattened to be used as input data.

The processed input sequences are fed to a Pretrained audio neural network (Panns) backbone architecture from Kong et al. [30]—see Fig. 4.

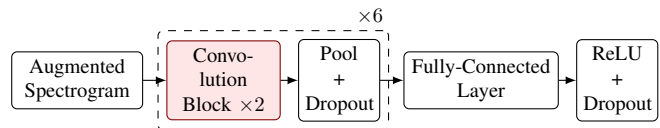


Fig. 4: **Audio encoder architecture**. The model processes augmented spectrograms through two consecutive convolutional blocks, followed by a sequence of pooling and dropout layers repeated six times, and finally a fully connected layer with ReLU and dropout.

C. Implementation Details

Unless otherwise stated, models are trained using the Adam optimizer with an attention size of 512 and ReLU activation functions. Dropout and weight decay were used for regularization and optimized using a grid search; Table I summarizes the exact value of each hyper-parameters for all experiments on a given dataset. For the UESTC dataset, the IMU model trained on an A100 GPU in 15 minutes with inference time below one minute, while the image model trained in 4 hours with 5 minutes inference time. For the Cough dataset, the IMU model trained on a T4 GPU in 15 minutes with 12 minutes inference time, and the audio model trained in 1 hour 30 minutes with 16 minutes inference time. For VGG-Sound, the image and audio models were trained on an A100 GPU in 8 and 10 hours, with inference times of 14 and 17 minutes, respectively.

D. Evaluation Metrics

To comprehensively assess the model performance, we measure 1) the *accuracy* to measure the proportion of correctly predicted samples over the total number of samples, 2) the *precision* to measure the proportion of predicted positive samples that are actually correct, reflecting the model’s ability to avoid false positives, 3) the *recall* to measure the proportion of actual positive samples that are correctly identified, indicating the model’s ability to avoid false negatives and detect only relevant samples, and 4) the *F1-score*, which is the harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives.

VI. RESULTS

We first present the overall performance across all datasets and modalities for in-distribution (ID) and out-of-distribution (OOD) scenarios. In the ID scenario, data subjects can be

present in the training, validation, and test sets. On the other hand, in OOD scenarios, the training, validation, and test sets have no overlapping subjects, making this split more challenging than ID, but also closer to usual real-world scenarios. For more details on the ID and OOD scenarios, see Al Haddad et al. [6].

To evaluate knowledge transfer, our method is compared against uni-modal baselines. We also provide ablation studies to analyze the impact of the masked cross-modal alignment of Section IV-C (denoted as MCMA in the experiments), and of the transfer knowledge weight factor (λ).

A. Comparison to Uni-modal Baseline Models

1) *In-distribution Scenario*: The results for the ID scenario are presented in Table II. As hypothesized, knowledge transfer from more informative sensor modalities (video and audio) improved the performance of the less expressive sensor modality (IMU) on all metrics—our proposed method outperforms the corresponding uni-modal IMU models in accuracy, recall, precision, and F1-score. For example, UESTC-IMU accuracy increased by 9%, its recall by 8%, its precision by 6%, and its F1-score by 8%. Similarly, Cough-IMU improves its accuracy by 4%, its recall by 3%, its precision by 6%, and its F1-score by 6%.

On the other hand, one can see that transferring knowledge from the less expressive sensor modality to the more expressive ones does not lead to improvements and can even create a slight decrease in performance. This is observed for both the image and audio models in Table II.

In conclusion, in the ID scenario, our method helps improve the general performance of IMU-based model but does not generate a positive outcome when trying to transfer knowledge from IMU to other, more expressive, sensor modalities. Hence, given two sensor modalities, our method can be used to improve the performance of lower-quality model using a better-performing model, without coupling both sensor modality for inference.

2) *Out-of-distribution Scenario*: the OOD scenario introduces a greater generalization challenge than the ID scenario since each model is tested on subjects unseen during training. Looking at the results in Table III, unlike the ID scenario, knowledge transfer in the OOD case does not follow the pattern of the stronger modality consistently improving the weaker one.

For the UESTC dataset, no gain is observed for the IMU modality, while the image modality shows a clear improvement of 8% in accuracy, 8% in recall, 9% in precision, and 9% for the F1-score. A similar result is observed for the Cough and VGGSound dataset. For the Cough and VGG Sound dataset, the IMU and image classification models only see marginal improvements from knowledge transfer from the IMU modality. On the other hand, the audio classification model performance improves thanks to knowledge transfer from the IMU modality—7% in accuracy, 1% in recall, 9% in precision, and 10% for the F1-score, on the Cough dataset, and 3% in accuracy, 5% in recall, 2% in precision, and 5% for the F1-score, on VGGSound.

TABLE II: **In-distribution results.** Performance comparison between baseline models and the proposed transfer learning framework with and without masked alignment (positive transfer) across all datasets and modalities. Since VGGSound does not have annotations for users, we could only run the experiment on this dataset for out-of-distribution (see Table III). The baseline for IMU data correspond to the Samosa architecture [29], Resnet-101 for images, and Kong et al. [30] for audio data.

Dataset	Modality	Metric	Baseline	Ours	No MCMA
UESTC	IMU	Accuracy \uparrow	0.877	0.969	<u>0.914</u>
		Recall \uparrow	0.895	0.967	<u>0.918</u>
		Precision \uparrow	0.900	0.974	<u>0.913</u>
		F1-Score \uparrow	0.893	0.967	<u>0.914</u>
	Image	Accuracy \uparrow	0.957	0.939	<u>0.945</u>
		Recall \uparrow	0.953	0.922	<u>0.934</u>
		Precision \uparrow	0.957	0.942	<u>0.955</u>
		F1-Score \uparrow	0.954	0.925	<u>0.943</u>
Cough	IMU	Accuracy \uparrow	0.440	0.474	<u>0.463</u>
		Recall \uparrow	0.268	0.304	<u>0.297</u>
		Precision \uparrow	0.185	0.243	<u>0.238</u>
		F1-Score \uparrow	0.205	0.256	<u>0.251</u>
	Audio	Accuracy \uparrow	0.839	0.816	<u>0.825</u>
		Recall \uparrow	0.757	0.664	<u>0.683</u>
		Precision \uparrow	0.793	0.685	<u>0.694</u>
		F1-Score \uparrow	0.745	0.670	<u>0.685</u>

Through our experiments, we observed that the IMU and Image models overfitted on the training dataset. We performed a grid search of hyperparameters, including varying learning rate between $\{0.0005, 0.0001, 0.001, 0.005\}$, window size for IMU models between $\{20, 50, 70\}$, dropout rate between $\{0.5, 0.6, 0.7, 0.8\}$, and various train, test and validation splits, and were unable to not overfit the models and reach higher and similar metrics to the ones presented in Table III. While overfitting didn't impact the expected knowledge transfer in Section VI-A.1, we hypothesize that this overfitting on the training dataset led to the current results in OOD scenarios. Indeed, our hypothesis is that, while an overfitted model is still able to provide meaningful knowledge to a target model (as in the case of IMU to Image in UESTC, IMU to Audio in Cough, and Image to Audio in VGGSound) in ID scenarios where user are both in the train and test sets, the same overfitted model cannot help improving the generalization in OOD scenario since it is overfit to the specific users in the training set.

Overall, these results suggest that, in OD scenarios, even higher performing classification models benefit from knowledge transfer from lower performing models, demonstrating that cross-modal transfer can improve performances under distribution shift. While less accurate models are not able to transfer knowledge in ID scenario—since most of the information is contained in the training set—in OOD scenarios, the added knowledge provided by the weaker modality help generalize the capability of stronger classification models.

TABLE III: **Out-of-distribution results.** Performance comparison between baseline models and the proposed transfer learning framework with and without masked alignment (positive transfer) across all datasets and modalities. The baseline for IMU data correspond to the Samosa architecture [29], Resnet-101 for images, and Kong et al. [30] for audio data.

Dataset	Modality	Metric	Baseline	Ours	No MCMA
UESTC	IMU	Accuracy \uparrow	0.554	<u>0.470</u>	0.411
		Recall \uparrow	0.556	<u>0.467</u>	0.414
		Precision \uparrow	0.554	<u>0.475</u>	0.449
		F1-Score \uparrow	0.522	<u>0.450</u>	0.385
	Image	Accuracy \uparrow	0.518	0.601	<u>0.548</u>
		Recall \uparrow	0.518	0.595	<u>0.550</u>
		Precision \uparrow	0.540	0.627	<u>0.554</u>
		F1-Score \uparrow	0.508	0.597	<u>0.543</u>
Cough	IMU	Accuracy \uparrow	0.416	0.423	0.423
		Recall \uparrow	0.309	0.309	0.309
		Precision \uparrow	0.250	<u>0.259</u>	0.267
		F1-Score \uparrow	0.250	<u>0.249</u>	0.248
	Audio	Accuracy \uparrow	0.607	<u>0.680</u>	0.709
		Recall \uparrow	0.533	<u>0.618</u>	0.653
		Precision \uparrow	0.477	<u>0.572</u>	0.686
		F1-Score \uparrow	0.482	<u>0.578</u>	0.636
VGG-Sound	Image	Accuracy \uparrow	0.459	0.456	0.466
		Recall \uparrow	0.330	0.324	<u>0.325</u>
		Precision \uparrow	0.309	<u>0.285</u>	0.267
		F1-Score \uparrow	<u>0.294</u>	0.297	0.285
	Audio	Accuracy \uparrow	0.897	<u>0.926</u>	0.928
		Recall \uparrow	0.828	<u>0.883</u>	0.891
		Precision \uparrow	0.889	<u>0.909</u>	0.910
		F1-Score \uparrow	0.839	<u>0.892</u>	0.898

B. Masked Cross-Modal Alignment

Tables II and III report the results with and without masked alignment. As described in Section IV-C, our method refers to the masked alignment strategy, where knowledge transfer is restricted to samples that are correctly classified by the source modality. In contrast, no MMCA denotes the case where knowledge is transferred across all samples, regardless of correctness

In the in-distribution split, whenever knowledge transfer is effective, our method consistently outperforms no MMCA, as expected. However, in the OOD scenario, our method does not always yield better performance. For example, in the Cough-Audio dataset and in both VGGSound modalities, no MMCA slightly outperforms our method. Nevertheless, these differences are marginal; looking at the models that are improved by cross-modal alignment, it is 5% for Cough-Audio metrics, less than 1% for all metrics of VGGSound-Audio, and 5% on the metrics of UESTC-Image. Hence, in OOD scenarios, masked cross-modal alignment should be used on a per-dataset basis.

C. Transfer Knowledge Weight Factor

The transfer knowledge weight factor λ , is an essential hyper-parameter of the objective function Eq. (19). To eval-

TABLE IV: **Effect of the cross-attention weight factor λ .** Results on the OOD split for all datasets and modalities when varying the cross-attention weight $\lambda \in \{0.01, 0.1, 1.0, 10\}$

Dataset	Modality	λ	Accuracy	Recall	Precision	F1-Score
UESTC	IMU	0.01	<u>0.524</u>	<u>0.519</u>	<u>0.519</u>	<u>0.506</u>
		0.1	0.566	0.557	0.537	0.534
		1.0	0.470	0.467	0.475	0.450
		10	0.321	0.304	0.176	0.197
	Image	0.01	0.470	0.473	<u>0.536</u>	0.480
		0.1	0.427	0.427	0.472	0.416
		1.0	0.601	0.600	0.627	0.597
		10	<u>0.518</u>	<u>0.513</u>	0.531	<u>0.506</u>
Cough	IMU	0.01	<u>0.421</u>	0.303	<u>0.251</u>	<u>0.239</u>
		0.1	<u>0.421</u>	0.304	<u>0.242</u>	<u>0.239</u>
		1.0	0.423	0.309	0.259	0.249
		10	0.416	0.297	0.226	0.227
	Audio	0.01	0.686	0.653	0.728	0.608
		0.1	0.652	0.597	0.578	0.563
		1.0	<u>0.680</u>	<u>0.618</u>	<u>0.572</u>	<u>0.578</u>
		10	<u>0.611</u>	<u>0.539</u>	<u>0.598</u>	0.495
VGG-Sound	Image	0.01	0.452	0.345	<u>0.295</u>	0.315
		0.1	0.473	<u>0.337</u>	0.262	0.293
		1.0	<u>0.456</u>	0.324	0.285	<u>0.297</u>
		10	0.452	0.326	0.364	0.291
	Audio	0.01	0.937	0.910	0.938	0.916
		0.1	0.911	0.859	0.892	0.870
		1.0	<u>0.926</u>	<u>0.883</u>	<u>0.909</u>	<u>0.892</u>
		10	0.896	0.811	0.850	0.812

uate the impact of this parameter, we conduct an ablation study where λ is varied over the set $\{0.1, 1.0, 10\}$. Results for all datasets, on the more challenging OOD scenario, are shown in Table IV.

Results demonstrate that while λ has some impact on the results, especially in higher values (e.g., $\lambda = 10$), the results are best at around $\lambda = 1$ —with little variation between 0.01 and 1—apart from UESTC-IMU. Hence, in previous experiments (Table II and III) $\lambda = 1$ was used.

VII. CONCLUSION AND FUTURE WORK

While HAR often relies on multiple sensors, deploying such multi-modal sensor system in real-world settings and at scale is often infeasible due to cost or technical constraints. To address this, we present D-CAT, a decoupled cross-attention transfer learning framework that aligns modality-specific feature spaces during training while enabling single-modality inference. Unlike prior work that fuses modalities via cross-attention layers, D-CAT enforces alignment through a cross-attention loss, ensuring that each modality’s classifier remains independent and eliminates the need for paired inputs.

We evaluate D-CAT on three HAR benchmarks using IMU, video, and audio. When transferring knowledge from a high-accuracy source modality to a weaker target classifier, D-CAT improves F1 scores by up to 7% in in-distribution settings (train-test subject overlap).

Furthermore, in out-of-distribution evaluation (unseen users in test set), D-CAT even enables knowledge transfer

from low to high accuracy modalities, suggesting that its decoupled alignment may improve robustness to domain shift. These results demonstrate that D-CAT offers a scalable alternative to joint training, particularly for resource-constrained applications where sensor availability is limited.

However, D-CAT's effectiveness decreases when the target model is already overfitted, limiting its utility in some out-of-distribution scenarios. Future work will explore strategies to mitigate this limitation, as well as explore the use of more than one source modality for transfer. Overall, our findings suggest that decoupled cross-attention could unlock modular and efficient knowledge transfer for more-accurate HAR with limited sensors.

REFERENCES

- [1] J. K. Aggarwal and L. Xia, "Human activity recognition from 3d data: A review," *Pattern Recognition Letters*, vol. 48, pp. 70–80, 2014.
- [2] C. Wang and J. Yan, "A comprehensive survey of rgb-based and skeleton-based human action recognition," *IEEE Access*, vol. 11, pp. 53 880–53 898, 2023.
- [3] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods," *Frontiers in Robotics and AI*, vol. 2, p. 28, 2015.
- [4] R. Jahanmahin, S. Masoud, J. Rickli, and A. Djuric, "Human-robot interactions in manufacturing: A survey of human behavior modeling," *Robotics and Computer-Integrated Manufacturing*, vol. 78, p. 102 404, 2022, ISSN: 0736-5845.
- [5] W. Hu, D. Xie, Z. Fu, W. Zeng, and S. Maybank, "Semantic-based surveillance video retrieval," *IEEE Transactions on Image Processing*, vol. 16, no. 4, pp. 1168–1181, 2007.
- [6] F. Al Haddad, Y. Wang, and M. Mielle, "UAC: uncertainty-aware calibration of neural networks for gesture detection," *arXiv preprint arXiv:2504.02895*, 2025.
- [7] Y. Zhao, Z. Liu, L. Yang, and H. Cheng, "Combing rgb and depth map features for human activity recognition," in *Proceedings of the 2012 Asia Pacific signal and information processing association annual summit and conference*, IEEE, 2012, pp. 1–4.
- [8] D. Liang and E. Thomaz, "Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 3, no. 1, pp. 1–18, Mar. 2019, ISSN: 2474-9567.
- [9] L. Min and B. Yang, "Skeleton-based human action recognition in low-resolution infrared images," in *2023 IEEE 18th Conference on Industrial Electronics and Applications (ICIEA)*, 2023, pp. 1363–1368.
- [10] T. Singh and D. K. Vishwakarma, "A deeply coupled convnet for human activity recognition using dynamic and rgb images," *Neural Computing and Applications*, vol. 33, no. 1, pp. 469–485, 2021.
- [11] S. Mukherjee, L. Anvitha, and T. M. Lahari, "Human activity recognition in rgb-d videos by dynamic images," *Multimedia Tools and Applications*, vol. 79, no. 27, pp. 19 787–19 801, 2020.
- [12] S. Cristina, V. Despotovic, R. Prez-Rodriguez, and S. Aleksic, "Audio-and video-based human activity recognition systems in healthcare," *IEEE Access*, vol. 12, pp. 8230–8245, 2024.
- [13] M. Koch, T. Pfitzinger, F. Schlenke, F. Kohlmorgen, R. Groll, and H. Whrle, "Recognition of human activities based on ambient audio and vibration data," *IEEE Access*, vol. 12, pp. 174 399–174 412, 2024.
- [14] W. Zhuang, Y. Chen, J. Su, B. Wang, and C. Gao, "Design of human activity recognition algorithms based on a single wearable imu sensor," *International Journal of Sensor Networks*, vol. 30, no. 3, pp. 193–206, 2019.
- [15] C. Hou, "A study on imu-based human activity recognition using deep learning and traditional machine learning," in *2020 5th International Conference on Computer and Communication Systems (ICCCS)*, IEEE, 2020, pp. 225–234.
- [16] A. Ayman, O. Attalah, and H. Shaban, "An efficient human activity recognition framework based on wearable imu wrist sensors," in *2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, IEEE, 2019, pp. 1–5.
- [17] L. Zhang et al., "Multisense: Cross-labelling and learning human activities using multimodal sensing data," *ACM Transactions on Sensor Networks*, vol. 19, no. 3, pp. 1–26, Apr. 2023, ISSN: 1550-4867.
- [18] W. Ma, S. Li, L. Cai, and J. Kang, *Learning modality knowledge alignment for cross-modality transfer*, 2024. arXiv: 2406.18864 [cs.CV].
- [19] H. Kwon et al., *Imutube: Automatic extraction of virtual on-body accelerometry from video for human activity recognition*, 2020. arXiv: 2006.05675 [cs.CV].
- [20] T. Wu, Y. Liu, and S. Yongchareon, "Class-aware sample weight learning for cross-modal unsupervised domain adaptation in cross-user wearable human activity recognition," in *ECAI 2024*. IOS Press, Oct. 2024, ISBN: 9781643685489.
- [21] P. Bharti, D. De, S. Chellappan, and S. K. Das, "Human: Complex activity recognition with multi-modal multi-positional body sensing," *IEEE Transactions on Mobile Computing*, vol. 18, no. 4, pp. 857–870, Apr. 2019, ISSN: 2161-9875.
- [22] S. Deldari, H. Xue, A. Saeed, D. V. Smith, and F. D. Salim, "Cocoa: Cross modality contrastive learning for sensor data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–28, Sep. 2022, ISSN: 2474-9567.
- [23] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 938–10 947.
- [24] S. Deldari, H. Xue, A. Saeed, D. V. Smith, and F. D. Salim, "Cocoa: Cross modality contrastive learning for sensor data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–28, 2022.
- [25] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon et al., Eds., vol. 30, Curran Associates, Inc., 2017.
- [26] L. Xu et al., "Towards continual egocentric activity recognition: A multi-modal egocentric activity dataset for continual learning," *IEEE Transactions on Multimedia*, vol. 26, pp. 2430–2443, 2024, ISSN: 1941-0077.
- [27] Y. Chen et al., "Robust multimodal cough and speech detection using wearables: A preliminary analysis," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024, pp. 1–6.
- [28] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, *Vggsound: A large-scale audio-visual dataset*, 2020. arXiv: 2004.14368 [cs.CV].
- [29] V. Mollyn, K. Ahuja, D. Verma, C. Harrison, and M. Goel, "Samosa: Sensing activities with motion and subsampled audio," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 3, Sep. 2022.
- [30] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, *Panns: Large-scale pretrained audio neural networks for audio pattern recognition*, 2020. arXiv: 1912.10211 [cs.SD].