2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

A Case Study on Visual-Audio-Tactile Cross-Modal Retrieval

Jagoda Wojcik*, Jiaqi Jiang*, Jiacheng Wu and Shan Luo

***Abstract*— Cross-Modal Retrieval (CMR), which retrieves relevant items from one modality (e.g., audio) given a query in another modality (e.g., visual), has undergone significant advancements in recent years. This capability is crucial for robots to integrate and interpret information across diverse sensory inputs. However, the retrieval space in existing robotic CMR approaches often consists of only one modality, which limits the performance of the robot. In this paper, we propose a novel CMR model that incorporates three different modalities, i.e., visual, audio, and tactile, for enhanced multi-modal object retrieval, referred to as *VAT-CMR*. In this model, multi-modal representations are first fused to provide a holistic view of object features. Then, to mitigate the semantic gaps between representations of different modalities, a dominant modality is selected during the classification training phase to improve the distinctiveness of the representations and enhance the retrieval performance. To evaluate our proposed approach, we conducted a case study and the results demonstrate that our VAT-CMR model surpasses competing approaches. Further, our proposed dominant modality selection significantly enhances cross-retrieval accuracy.**

I. INTRODUCTION

In recent years, Cross-Modal Retrieval (CMR), the process of querying data in one modality (e.g., audio) to retrieve relevant items from another modality (e.g., vision) has made significant strides [1]–[3]. This progress has been driven by the exponential growth of multi-modal data, in various forms of images, texts and audio, available on the internet and in our daily lives. CMR holds great promise for applications such as healthcare, where it could align medical imaging with related patient profiles, thereby improving diagnostic accuracy. In the context of robotics, CMR enables the processing and interpretation of information across diverse sensory inputs, such as vision and touch, empowering robots to adapt and interact more effectively with their environment [4]–[6].

Many existing CMR approaches within the field of robotics are limited by their focus on single-modality retrieval and reliance on bi-modal CMR networks [7], [8]. While relying on a single-modality may compromise the retrieval accuracy due to restricted information scope, the use of multiple bi-modal CMR models to handle more than two modalities [9] increases the computational complexity and reduces the overall efficiency.

In contrast, human perception seamlessly integrates information across multiple modalities, such as vision, sound, and touch, to form a cohesive understanding of the environment. This natural ability to cross-reference sensory information leads to more robust and accurate information retrieval. Research on human sensory perception [10], [11] has demonstrated that the integration of multiple modalities can facilitate the development of more effective neural representations, thereby enhancing cognitive performance.

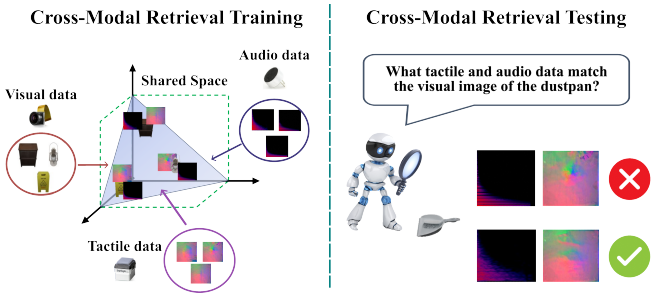

Fig. 1. Illustration of visual-audio-tactile cross-modal retrieval. Left: The visual, audio, and tactile representations of the same object converge within a shared space. Right: The robot retrieves the corresponding audio and tactile data when provided with a visual image of the dustpan.

In this work, we propose a novel CMR model that incorporates three distinct modalities for object retrieval, named *VAT-CMR*, as illustrated in Fig. 1. In the model, multi-modal representations are first learned to offer a holistic representation of the object features. This approach enhances the disambiguation of latent spaces, which might otherwise remain ambiguous when confined to a single modality. To improve the alignment among the diverse modalities, we use an attention mechanism during the multi-modal feature fusion stage. Additionally, we introduce the concept of *dominant modality selection* during the classification training phase. This approach differs from conventional methods, which often optimise based on a concatenated representation of features from multiple modalities. The emphasis on a dominant modality aims to effectively mitigate the semantic gap between modalities, thereby boosting the performance of our retrieval model.

To evaluate our proposed VAT-CMR approach, we conducted a case study and collected a synthetic dataset featuring 20 objects with data from three modalities, i.e., vision, audio, and touch. Our experiments demonstrate that VAT-CMR outperforms the state-of-the-art CMR methods, showcasing a noteworthy improvement in Mean Average Precision (MAP) when the query modality is either vision or touch. Additionally, through an ablation study, we found that both the attention feature modules and the incorporation of dominant modality selection contribute to an enhanced

This work was supported by the EPSRC project "ViTac: Visual-Tactile Synergy for Handling Flexible Materials" (EP/T033517/2).

All the authors are with Department of Engineering, King's College London, London WC2R 2LS, U.K. Emails: {jagoda.wojcik, jiaqi.1.jiang, jiacheng.2.wu, shan.luo}@kcl.ac.uk

* represents equal contributions.

retrieval accuracy by 0.04 and 0.05, respectively.

Our contributions can be summarised as follows:

- We propose VAT-CMR, a novel CMR model that utilises multi-modal feature representations for retrieval tasks;
- We introduce the concept of dominant-modality-based training for CMR, which enhances the retrieval performance;
- The proposed VAT-CMR outperforms the state-of-the-art approaches, with code publicly available[1].

The rest of the paper is structured as follows: Section II provides an overview of related work; Section III introduces our VAT-CMR framework; Section IV details our synthetic dataset and the evaluation metrics used; Section V analyses the experimental results. Finally, Section VI presents the discussion and summarises the work.

II. Related Works

The exploration of Cross-Modal Retrieval (CMR) has gained increasing attention in recent years, due to the rapid proliferation of multi-modal data in robotics, such as images, text, audio and tactile readings. Existing CMR methods fall into two categories: traditional multi-modal representation learning methods and deep multi-modal representation learning methods.

Early techniques for CMR relied on simplistic representations, with Canonical Correlation Analysis (CCA) [12] being one of the popular methods. CCA maximises the correlation between modalities, employing a semantic space to measure similarity. Other methods, like Partial Least Squares (PLS) [13] and Bilinear Model (BLM) [14], also aimed at learning latent common spaces but often faced scalability and generalisation challenges.

Recent advancements in CMR are shaped by deep network-based representation learning [15]. Notable examples include Deep Canonical Correlation Analysis (DCCA) [16], which learns intricate nonlinear transformations to ensure a strong linear correlation between bi-modal data representations. Deep Canonically Correlated Autoencoders (DCCAE) [17] extend these concepts through reconstruction objectives. In the realm of adversarial learning, Adversarial Cross-Modal Retrieval (ACMR) [18] employs a feature projector, a modality classifier and a triplet constraint to establish an effective common subspace. Deep Supervised Cross-modal Retrieval (DSCMR) [19] focuses on minimising discrimination loss in both the label space and the common representation space to learn discriminative features.

Recently, there have been works on cross-modal retrieval among visual, audio and tactile modalities [20], [21]. Liu et al. [20] investigate active visual-tactile cross-modal matching using a dictionary learning model, while their work [22] introduces a framework for weakly paired fusion of tactile and auditory modalities and cross-modal transfer for the visual modality. Zheng et al. propose a novel Discriminant Adversarial Learning (DAL) method, addressing intra-modal discrimination and inter-modal consistency for visual-tactile cross-modal retrieval in a unified training process. Despite these advancements, there is a gap in exploring multi-modal retrieval representations for cross-modal retrieval with multiple modalities.

[1] https://github.com/jagodawojcik/VAT-CMR

III. Methodology

In this section, we first outline the problem formulation and then introduce our proposed Vision-Audio-Touch Cross-Modal Retrieval (VAT-CMR) model, with an overview of the model illustrated in Fig. 2. Finally, we provide details on the training methodology to facilitate the reproduction of this work.

A. Problem Formulation

The Cross-Modal Retrieval (CMR) problem addressed in this work considers three types of data: visual, audio, and tactile readings. Let $v \in \mathbb{R}^{d_v}$, $a \in \mathbb{R}^{d_a}$, and $t \in \mathbb{R}^{d_t}$ represent the visual features, audio features, and tactile features derived from visual images, audio data, and tactile data, respectively. Here, d_v, d_a and d_t denote the dimensions of the latent representations of the vision, audio, and touch modalities, respectively. Each sample of visual, audio, and tactile data is associated with a corresponding one-hot category label $y_i \in \mathbb{R}^C$, where C represents the number of categories in the dataset.

Given a set of training samples $V = \{v_1, v_2, \ldots, v_n\}$, $A = \{a_1, a_2, \ldots, a_n\}$ and $T = \{t_1, t_2, \ldots, t_n\}$ extracted from paired visual, audio, and tactile data, VAT-CMR first employs supervised learning to establish a shared semantic space. The primary objective is to ensure close alignment among the corresponding visual, audio, and tactile representations, facilitating instance retrieval across modalities when provided with data from a specific modality. This alignment is achieved through a function $f(\cdot)$ that maps features from all modalities into a common semantic space. The training goal is to maximise the semantic similarity between multi-modal representations of the same object. Euclidean distance is employed as a similarity metric $sim(\cdot, \cdot)$, which is used to evaluate the distance between query and retrieval feature representations:

$$f : V \times A \times T \rightarrow S \tag{1}$$

where S represents the shared semantic space.

B. The proposed VAT-CMR model

To address the visual-audio-tactile cross-modal retrieval problem, we develop a three-branch network. As shown in Fig. 2, our VAT-CMR model takes a visual RGB image, a tactile image from an optical tactile sensor [23]–[26] and an audio sample as input. Within each branch, the three distinct modalities are processed through disjoint networks during the initial layers to capture modality-specific features. An attention mechanism is used to fuse the multi-modal representations of the retrieval modalities. Finally, a cross-entropy loss based on the selected dominant modality is employed to obtain final feature representations. These are then mapped to the common latent space using triplet loss.

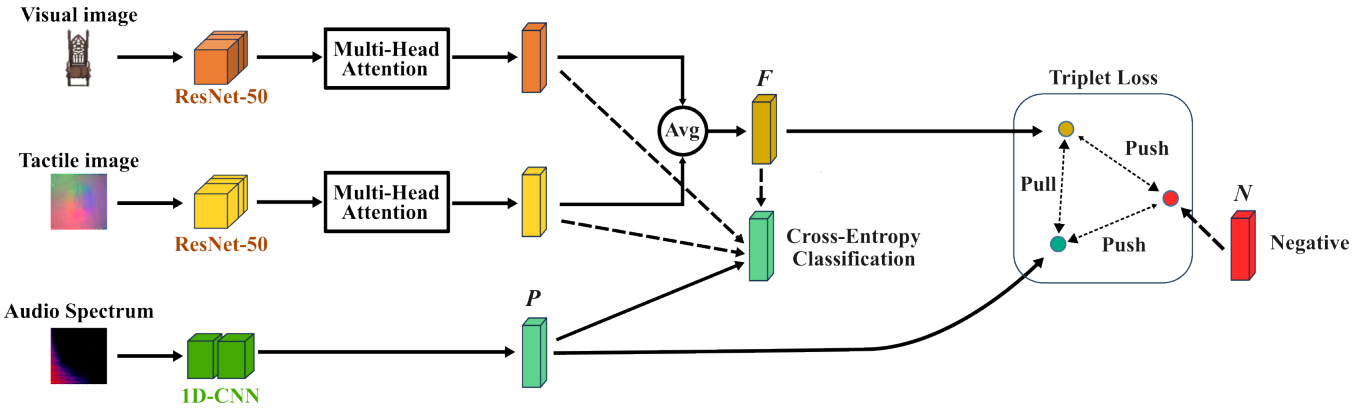

Fig. 2. **Overview of our VAT-CMR model (using audio as the query modality in this example).** From left to right: First, VAT-CMR takes a visual image, a tactile image, and an audio sample as input. These inputs are processed through three separate neural network branches. Multi-head attention modules are used to fuse the feature representations from the two retrieval modalities. With the fused feature representation F, and the positive input modality feature P, the model is trained using cross-entropy loss, where audio serves as the dominant modality and is directly linked via a solid line. The dashed lines connected to the cross-entropy module represent the cases when other modalities are selected as the dominant modality. Finally, a triplet loss function is employed to map the features extracted from the last hidden layer of each branch to a cross-sensory embedding space. N on the right-hand side represents a negative sample used in triplet loss training.

1) Disjoint neural networks architecture: Firstly, we extract the features from visual images, audio patches, and tactile images to represent the object across different modalities. Specifically, two pre-trained ResNet50 models [27] are fine-tuned using the visual images and the tactile images generated for selected objects. The visual features v and tactile features t are extracted from the last hidden layer of the fine-tuned models. To obtain audio feature embeddings, we use a 1D-CNN architecture, which comprises three convolutional layers followed by a pooling layer to reduce the spatial dimensions of the feature maps and standardise the size of the output representations. Subsequently, the pooling layer's output is fed into a fully connected layer with the intended output dimension a.

2) Attention mechanism: The multi-head attention mechanism enables the model to concurrently process information from different representation subspaces at various positions. Following the approach introduced in Vaswani et al. [28], we utilise a multi-head attention mechanism to highlight shared features across two different modalities. In contrast to a single attention head, where critical information could be diluted through averaging, the multi-head mechanism maintains the integrity of these features. The fused representation F that combines two streams of modalities can be given by:

$$F = \frac{1}{2}\left(\text{MultiHead}(Q_t, K_v, V_v) + \text{MultiHead}(Q_v, K_t, V_t)\right), \tag{2}$$

where $\text{MultiHead}(\cdot,\cdot,\cdot)$ is the multi-head attention mechanism [28]; v and t in (Q_t, K_v, V_v) and (Q_v, K_t, V_t) signify the visual and tactile modalities, respectively. Here, we present the equation with audio as the query modility for example, and with visual and tactile modalities fused. However, within the VAT-CMR framework, any combination of these three modalities can be selected for integration and we conducted experiments across all possible combinations of these modalities.

3) Cross-entropy loss for dominant modality optimisation: In previous studies, a prevailing approach in guiding model training across multiple modalities has been to employ a common feature vector, as highlighted by [9], [29]. However, handling multiple modalities simultaneously can increase the complexity of the learning task, potentially leading to degraded integrated representations due to noisy data or less discriminative features in certain modalities, ultimately resulting in suboptimal performance. Our series of experiments revealed that a more effective strategy for multi-modal representation learning is to focus the learning process on a selected modality. By empirically evaluating the contributions of each modality to the learning process, we identified the most beneficial modality, which is taken as the dominant modality and used as the guiding force for optimisation. By focusing on a single modality, we reduce the complexity of the learning process, making it easier for the model to learn meaningful representations. Instead of utilising the combined representations of multiple modalities, we use the last hidden layer of the disjoint neural network pathway for the dominant modality to compute the cross-entropy loss. As a result, the cross-entropy loss is obtained by:

$$L_{dominant}(p, q) = -\sum_{i=1}^{K} p(x_i) \log(q_{dominant}(x_i)) \tag{3}$$

where $p(x_i)$ represents the true probability distribution for the target class x_i, while $q_{dominant}(x_i)$ denotes the predicted probability distribution from the selected dominant modality for class x_i. The loss calculation is conducted at the last hidden layer of disjoint pathways, and encompasses all K object classes. This ensures a thorough evaluation of the model's classification performance, focusing on the modality deemed most informative.

Focusing on a single modality during training offers significant advantages for classification tasks, as it enhances

Fig. 3. **Objects used in our experiments.** In total, there are 20 objects in our experiments, taken from the ObjectFolder dataset [30], each with unique geometric characteristics and made from materials with distinct properties.

the model's ability to discern relevant features. Importantly, this approach also yields long-term benefits for cross-modal retrieval processes. By emphasising one modality, the model becomes more adept at extracting information, leading to improved retrieval accuracy across modalities. This versatility enhances the model's utility and effectiveness in various tasks, making it a valuable tool in multi-modal applications.

4) Cross-modal correlation learning: To achieve the goal of enhancing the similarity between multi-modal representations of the same object, in this work, we use the triplet loss, as shown in Fig. 2. The Euclidean distance, also known as the L2-norm, is used to calculate the similarity between different representations. To this end, the triplet loss function is computed as follows:

$$L(F, P, N) = max(||F - P||^2 - ||F - N||^2 + \alpha, 0) \quad (4)$$

where F, P and N represent the fused representations based on the attention module, the positive input modality of the same class, and negative input which in this context denotes a foreign object, respectively. The parameter α is a safety margin, ensuring that the model does not trivialise the equation to zero by equalising the three embeddings.

C. Training Details

To enable easy replication of our work, we outline the training details in this subsection. We employed a batch size of 5 for all cross-entropy training tasks due to the extensive memory requirements associated with larger batch sizes. All cross-entropy training tasks span 50 epochs. Adam was chosen as the gradient optimisation algorithm, with learning rates set at 0.001 and 0.0001 for the cross-entropy and triplet loss training stages, respectively. One critical hyperparameter, the triplet loss margin, was empirically determined to be 0.5 after extensive testing.

IV. Experimental Setup

In this section, we first introduce the process of synthetic dataset generation. Then, we introduce the evaluation metrics for cross-model retrieval.

For the purpose of this study, we generate a total of 34,500 samples of data representing 20 randomly selected objects from the ObjectFolder 2.0 [30] dataset. The objects' visual representations are demonstrated in Fig. 3. The dataset is split into training, validation, and testing subsets, each consisting of 25,500, 4,500, and 4,500 samples, respectively. Data comprises three modalities: vision, audio, and touch, for each object, as demonstrated in Fig. 4. To render the data, a set of arguments is required, which allows us to produce a diverse range of samples for each object.

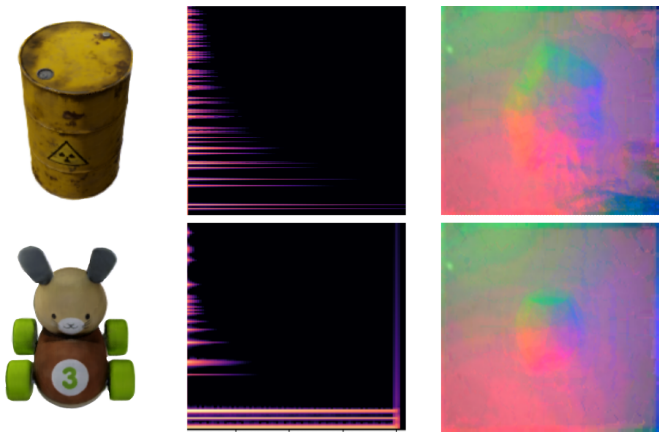

Fig. 4. Visualisation of two object examples from our generated synthetic dataset. **From left to right:** The columns represent the RGB visual images, audio spectrograms, and tactile images, respectively.

- Visual: To generate a visual image, an array of length 6 is required, where the first three coordinates specify the camera position, and the remaining three define the position of the light source. Therefore, each testing viewpoint is described by the array: ($camera_x$, $camera_y$, $camera_z$, $light_x$, $light_y$, $light_z$).
- Audio: To produce an audio sample, two sets of arrays are required. The first array takes three arguments specifying the point on the object surface (x, y, z), where force will be applied to generate an audio response. The second array, also of the same size, determines the amount of force applied to the selected coordinate point. The force is defined by providing its three directional components: (F_x, F_y, F_z).
- Tactile: Generation of tactile images also requires two sets of arrays. Similarly to the audio parameters, the first array of length 3 defines the coordinates of the point on the object's surface. The second array requires the specification of gel rotation angle and gel displacement in the form of (θ, ϕ, d).

TABLE I

EVALUATION RESULTS OF CROSS-RETRIEVAL PERFORMANCE OF OUR VAT-CMT METHOD AGAINST CCA AND OBJECTFOLDER.

Query	Retrieval	MAP Score		
		Ours	CCA [12]	ObjectFolder [9]
20 objects dataset (chance = 0.05)				
Vision	Touch	0.93	0.44	0.51
	Audio	0.90	0.54	0.51
	Touch + Audio	**0.96**	—	0.67
Touch	Vision	**0.87**	0.46	0.58
	Audio	0.84	0.33	0.73
	Vision + Audio	0.85	—	0.75
Audio	Vision	0.77	0.37	0.52
	Touch	0.77	0.51	0.63
	Vision + Touch	**0.81**	—	0.69

TABLE II

ABLATION STUDY ON VARIOUS NETWORK STRUCTURES.

Network Structure		MAP Score		
		Query Modality		
Attention	Dominant Modality	Audio	Vision	Touch
		0.69	0.67	0.75
	✓	0.77	**0.96**	0.78
✓		0.76	0.68	0.81
✓	✓	**0.81**	**0.96**	**0.85**

TABLE III

ABLATION STUDY ON THE SELECTION OF DIFFERENT DOMINANT MODALITIES.

Query	Retrieval	Dominant Modality			Joint
		Audio	Vision	Touch	
20 objects dataset (chance = 0.05)					
Audio	Vision+Touch	**0.81**	0.04	0.02	0.76
Touch	Vision+Audio	0.82	**0.85**	0.82	0.81
Vision	Touch+Audio	0.90	0.86	**0.96**	0.68
Vision+Touch	Audio	**0.94**	0.06	0.08	0.81
Vision+Audio	Touch	0.77	**0.93**	0.90	0.69
Touch+Audio	Vision	0.78	0.82	**0.85**	0.71

V. EXPERIMENT RESULTS

In this section, we conduct a series of experiments to evaluate the cross-retrieval performance of our VAT-CMR model. The goal of these experiments is three-fold: (1) To assess the performance of the VAT-CMR model in comparison to competing baseline approaches; (2) To investigate how each proposed module contributes to the cross-modal retrieval performance; (3) To analyse the impact of selecting varying dominant modalities for cross-entropy learning. In the experiment, we follow [9], [30] and employ the Mean Average Precision (MAP) [12] as the metric to evaluate the retrieval performance.

A. Comparison with single modality-based retrieval methods

We compare the cross-retrieval performance between the proposed VAT-CMR model and the baseline approaches in the literature, i.e., the Canonical Correlation Analysis (CCA) method [12] and the cross-retrieval model developed in ObjectFolder [9], as presented in Table I. To ensure an equitable comparison with the model utilised in ObjectFolder [9], we have extended their model to accommodate multi-modal retrieval modalities. Our VAT-CMR model showcases a notable advancement over these baseline approaches when multiple retrieval modalities are used, i.e., with an increase of 0.29, 0.10 and 0.12 in MAP when employing vision, touch, and audio as the query modality, respectively. Further, our VAT-CMR consistently surpasses the baseline methods, also assessed against single-modal retrieval metrics. On the other hand, we observe that the fused representations retrieval space attains the highest MAP score, except when touch is utilised as the query modality. These results demonstrate the efficacy and versatility of our VAT-CMR model in facilitating cross-modal retrieval tasks, offering superior performance across a range of modalities and retrieval scenarios.

B. Ablation study

In this subsection, we analyse the effects of integrating attention-based fusion and dominant modality selection into our training methodology. Specifically, when the attention module is not applied, it will be substituted with a simple feature concatenation. Meanwhile, the dominant modality selection is replaced with joint embedding-based optimisation when not in use.

As shown in Table II, our findings reveal that the VAT-CMR model attains the highest retrieval scores when it incorporates both dominant modality selection and the attention module within its training framework. This highlights its enhanced effectiveness in object retrieval over models lacking these components, where their absence results in an average MAP score reduction of 0.17 across all test scenarios.

Furthermore, in Table III, we provide a detailed examination of the impact of dominant modality selection on cross-modal retrieval outcomes. Compared to the traditional approach, our dominant modality selection method achieves an average retrieval improvement of 0.13 in the MAP score. Moreover, we find that there is a fixed relationship between a single modality, whether used for query or retrieval, and the dominant modality. Specifically, when audio is a single modality used for query or retrieval, selecting audio as the dominant modality in the cross-entropy learning process results in the largest average scores of 0.81 and 0.94. However having touch or vision as the dominant modality results in extremely poor average scores, as low as 0.02. This poor performance likely stems from the unique nature of audio sample representations compared to those of touch and vision, which are both represented as RGB images. It is also noticed that when vision is used as a single query or retrieval modality, selecting touch as the dominant modality in the cross-entropy learning process yields the largest average score. Conversely, when touch is used as a single query or retrieval modality, selecting vision as the dominant modality will obtain the largest average score. These cases also verify that touch is very similar to vision in this proposed dataset, against audio.

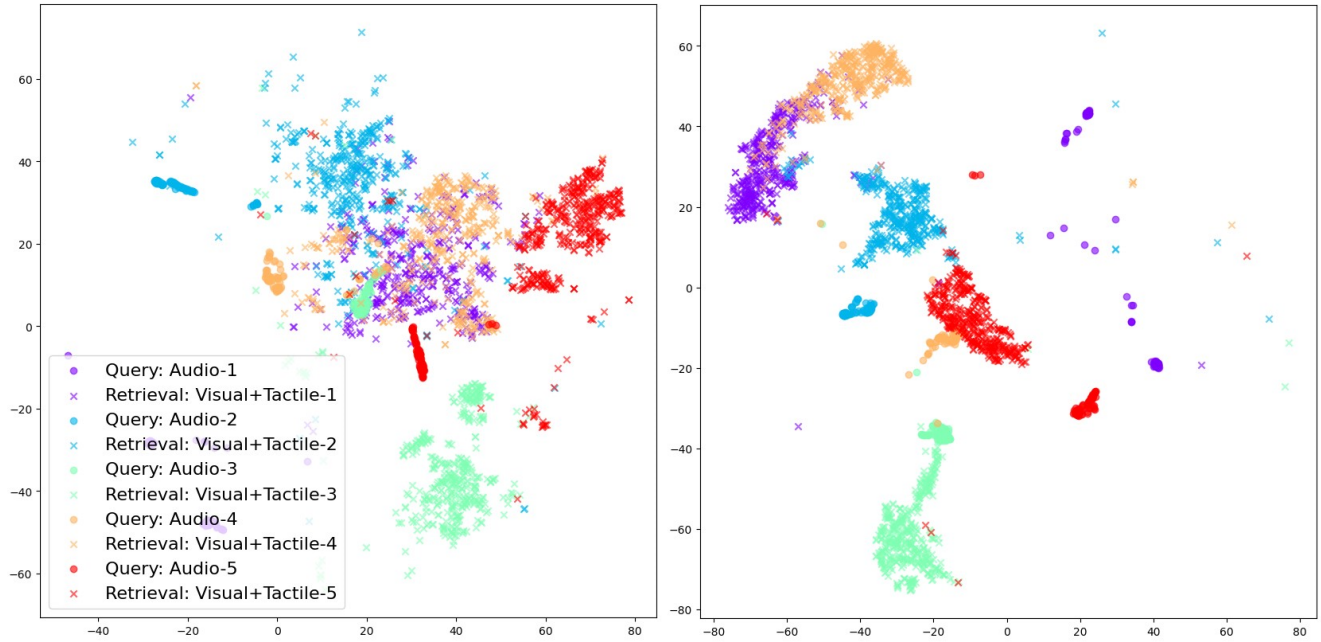

Fig. 5. 2D plots illustrating latent representations for a selected set of five classes, using Audio as the query test samples, and Visual+Tactile as the retrieval space. **Left:** Features after undergoing the cross-entropy model stage; **Right:** The same set of features after triplet loss processing.

C. Feature visualisation and learning curve

In this subsection, we visualise the distributions of latent features obtained from our VAT-CMR model using t-SNE [31]. Specifically, Fig. 5 shows the cross-retrieval latent space at different training phases, first after the cross-entropy loss stage and then after applying the triplet loss. The left-hand side of the figure clearly illustrates that when using only cross-entropy loss, features corresponding to the same object classes tend to cluster together, however still in a scattered manner. However, the introduction of the triplet loss stage substantially enhances feature organisation, both for the test set, and the retrieval space, leading to a more coherent clustering of representations. Furthermore, we present the MAP evaluation on the validation dataset during the triplet loss cross-modal correction learning in Fig. 6, further attesting to the effectiveness of the employed method.

Fig. 6. Mean Average Precision evaluated on the validation dataset during cross-modal triplet loss learning with audio as the Query modality, and vision and touch as the Fused modalities.

VI. Conclusion and Discussion

In conclusion, we have introduced VAT-CMR, a novel cross-modal retrieval model integrating vision, audio and touch modalities. When compared to previous methods, our VAT-CMR utilises a fused multi-modal retrieval representation through a multi-head attention mechanism, along with leveraging the dominant modality for cross-entropy training. Extensive experiments show that our VAT-CMR outperforms the baseline ObjectFolder method, achieving a MAP score improvement of 0.29, 0.10 and 0.12 when employing vision, touch, and audio as the query modality, respectively. A detailed ablation study further demonstrates that all the proposed modules contribute positively to cross-retrieval accuracy.

In the future work, we would like to explore a few directions based on the current findings. Firstly, in this work synthetic datasets are used, which may limit the model's generalisation to real-world scenarios, and we will design strategies to bridge the Sim2Real gap [32], [33]. Secondly, instead of using static visual and tactile data, we would like to consider active exploration strategies to augment robotic object retrieval capabilities so as to improve the real-world applicability of our proposed methods [34]. Thirdly, we will extend the study to incorporate additional perception modalities, such as force feedback, and explore zero-shot cross-modal retrieval methods [35] in the context of robotic grasping. By integrating multiple modalities, we aim to enhance each sensing modality and improve overall grasping performance.

REFERENCES

[1] P. Kaur, H. S. Pannu, and A. K. Malhi, "Comparative analysis on cross-modal information retrieval: A review," *Computer Science Review*, vol. 39, p. 100336, 2021.

[2] S. Luo, W. Yuan, E. Adelson, A. G. Cohn, and R. Fuentes, "Vitac: Feature sharing between vision and tactile sensing for cloth texture recognition," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2722–2727, 2018.

[3] S. Luo, J. Bimbo, R. Dahiya, and H. Liu, "Robotic tactile perception of object properties: A review," *Mechatronics*, vol. 48, pp. 54–67, 2017.

[4] J.-T. Lee, D. Bollegala, and S. Luo, ""Touching to see" and "seeing to feel": Robotic cross-modal sensory data generation for visual-tactile perception," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 4276–4282, IEEE, 2019.

[5] G. Cao, J. Jiang, N. Mao, D. Bollegala, M. Li, and S. Luo, "Vis2Hap: Vision-based Haptic Rendering by Cross-modal Generation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12443–12449, 2023.

[6] S. Luo, W. Mou, K. Althoefer, and H. Liu, "Localizing the object contact through matching tactile features with visual map," in *Proc. IEEE Int. Conf. Robot. Autom.*, pp. 3903–3908, 2015.

[7] L. Pecyna, S. Dong, and S. Luo, "Visual-tactile multimodality for following deformable linear objects using reinforcement learning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3987–3994, 2022.

[8] J. Jiang, G. Cao, A. Butterworth, T.-T. Do, and S. Luo, "Where shall i touch? vision-guided tactile poking for transparent object grasping," *IEEE/ASME Transactions on Mechatronics*, vol. 28, no. 1, pp. 233–244, 2022.

[9] R. Gao, Y.-Y. Chang, S. Mall, L. Fei-Fei, and J. Wu, "Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations," in *Conference on Robot Learning*, pp. 466–476, PMLR, 2022.

[10] A. A. Ghazanfar and C. E. Schroeder, "Is neocortex essentially multisensory?," *Trends in cognitive sciences*, vol. 10, no. 6, pp. 278–285, 2006.

[11] B. E. Stein and T. R. Stanford, "Multisensory integration: current issues from the perspective of the single neuron," *Nature reviews neuroscience*, vol. 9, no. 4, pp. 255–266, 2008.

[12] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 251–260, 2010.

[13] R. Rosipal and N. Krämer, "Overview and recent advances in partial least squares," in *International Statistical and Optimization Perspectives Workshop" Subspace, Latent Structure and Feature Selection"*, pp. 34–51, Springer, 2005.

[14] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, "Generalized multiview analysis: A discriminative latent space," in *2012 IEEE conference on computer vision and pattern recognition*, pp. 2160–2167, IEEE, 2012.

[15] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," *Advances in neural information processing systems*, vol. 25, 2012.

[16] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*, pp. 1247–1255, PMLR, 2013.

[17] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *International conference on machine learning*, pp. 1083–1092, PMLR, 2015.

[18] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proceedings of the 25th ACM international conference on Multimedia*, pp. 154–162, 2017.

[19] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10394–10403, 2019.

[20] H. Liu, F. Wang, F. Sun, and X. Zhang, "Active visual-tactile cross-modal matching," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 11, no. 2, pp. 176–187, 2018.

[21] W. Zheng, H. Liu, B. Wang, and F. Sun, "Cross-modal surface material retrieval using discriminant adversarial learning," *IEEE transactions on industrial informatics*, vol. 15, no. 9, pp. 4978–4987, 2019.

[22] H. Liu, F. Wang, F. Sun, and B. Fang, "Surface material retrieval using weakly paired cross-modal learning," *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 2, pp. 781–791, 2018.

[23] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[24] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, *et al.*, "Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.

[25] D. F. Gomes, Z. Lin, and S. Luo, "Geltip: A finger-shaped optical tactile sensor for robotic manipulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pp. 9903–9909, 2020.

[26] G. Cao, J. Jiang, C. Lu, D. F. Gomes, and S. Luo, "Touchroller: A rolling optical tactile sensor for rapid assessment of textures for large surface areas," *Sensors*, vol. 23, no. 5, p. 2661, 2023.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[29] W. Zheng, H. Liu, and F. Sun, "Lifelong visual-tactile cross-modal learning for robotic material perception," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 3, pp. 1192–1203, 2020.

[30] R. Gao, Z. Si, Y.-Y. Chang, S. Clarke, J. Bohg, L. Fei-Fei, W. Yuan, and J. Wu, "Objectfolder 2.0: A multisensory object dataset for sim2real transfer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10598–10608, 2022.

[31] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[32] T. Jianu, D. F. Gomes, and S. Luo, "Reducing tactile sim2real domain gaps via deep texture generation networks," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 8305–8311, IEEE, 2022.

[33] X. Jing, K. Qian, T. Jianu, and S. Luo, "Unsupervised adversarial domain adaptation for sim-to-real transfer of tactile images," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–11, 2023.

[34] G. Cao, Y. Zhou, D. Bollegala, and S. Luo, "Spatio-temporal attention model for tactile texture recognition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pp. 9896–9902, 2020.

[35] G. Cao, J. Jiang, D. Bollegala, M. Li, and S. Luo, "Multimodal zero-shot learning for tactile texture recognition," *Robotics and Autonomous Systems*, vol. 176, p. 104688, 2024.