EMMA: Extended Multimodal Alignment for Robust Object Retrieval

Rahul Agarwal
IBM
New York, USA
e-mail: rahul.agarwal@ibm.com

***Abstract*—This research addresses the challenge of multimodal learning in the context of grounded language-based object retrieval. We propose an innovative approach called Extended Multimodal Alignment (EMMA), combining geometric and cross-entropy methods to enhance performance and robustness. Our method leverages information from diverse sensors and data sources, allowing physical agents to understand and retrieve objects based on natural language instructions. Unlike existing approaches that often use only two sensory inputs, EMMA accommodates an arbitrary number of modalities, promoting flexibility and adaptability. On the GoLD benchmark EMMA reaches 0.93 mean-reciprocal rank and 78.2% top-1 recall, outperforming the strongest baseline by +7.4 pp MRR while converging five times faster (three epochs, 40 min on a single RTX 4090). When any single modality is withheld at test time, EMMA retains 88% of its full-modality accuracy, whereas competing methods drop below 65%. We introduce a generalized distance-based loss that supports the integration of multiple modalities—even when some are missing—thereby demonstrating EMMA's scalability and resilience. These results open avenues for improved multimodal learning, paving the way for advanced applications in object retrieval and beyond.**

Keywords-Multimodal learning; Object retrieval; Sensor fusion; Contrastive loss; Grounded language

I. INTRODUCTION

Inspired by the multimodal nature of human interaction with the world, it is intuitive that agents learning about the world, upon encountering new concepts and new objects, should form a model that incorporates information from all available sensors and data sources. The benefits of integrating multiple modalities are twofold: first, complementary information can be extracted from different modalities that can help with understanding the world, and second, additional modalities can help in the cases when one or more sources of data about the world become unavailable. Grounded language understanding, in which natural language is used as a query against objects in a physical environment, allows a real-world, intuitive mechanism by which users can instruct physical agents to engage in tasks such as object retrieval. Visuolinguistic approaches to such object inference tasks typically involve training on large pools of image/text pairs and then using language to subselect elements of the sensed environment [1][2].

Although physical agents, such as robots typically have access to sensory and interactive modalities beyond vision, and learning from multiple modalities can improve performance on downstream tasks, most approaches use at most two sensory inputs (e.g., visual data such as RGB plus depth images) with single labels, such as those provided by textual natural language. Simultaneously using additional inputs from different modalities is an underexplored area, in part due to the domain-specific nature of such n-ary learning approaches. With the modern proliferation of audio and text-based communication and home agents (e.g., Alexa/Google Home), there is a growing need to handle more modalities and simultaneously their potential failures.

One difficulty with working with complex multimodal data is the increased likelihood that one or more modalities may have missing information. Hardware can become damaged or defective, sensors can get blocked or obstructed, and various adverse but not uncommon conditions can remove a modality from use. Current multimodal approaches are typically not robust to the loss of one or more modalities at test time, as may happen if, for example, a physical agent fails to retrieve data from a particular sensor. In order to fully leverage multimodal training data while being robust to missing information, we propose a generalized distance-based loss function that can be extended to learn retrieval models that incorporate an arbitrary number of modalities.

We consider the domain of grounded language-based object retrieval [3][4], in which objects in an environment must be identified based on linguistic instructions. This can be considered a special case of image retrieval [5]–[8] in which objects are identified using visual inputs in combination with other sensor modalities. Approaches to acquiring grounded language have explored various combinations of sensor inputs such as depth and RGB with labels provided by textual language or speech [9]. However, despite object retrieval's multisensory nature, much of the existing work has not previously been extended to include an arbitrary number of modalities.

To this end we introduce *Extended Multimodal Alignment* (EMMA), a retrieval framework that fuses a geometric distance objective with a cross-entropy based supervised contrastive loss function [10]. EMMA (i) accommodates an *arbitrary* number of sensory and linguistic modalities, (ii) converges approximately five times faster than strong SupCon [11] and SimCLR [10] baselines while matching or exceeding their accuracy, and (iii) remains robust when one or more modalities are ablated at test time-achieving a mean-recall improvement of 7.4 pp on the GoLD benchmark. Treating speech and text as first-class input modalities further demonstrates that label information can be leveraged even when explicit annotations are sparse.

Paper organization. Section II reviews related multimodal and contrastive learning work. Section III formalizes the EMMA objective. Section IV details the end-to-end retrieval pipeline, and Section V describes datasets, implementation,

and training protocol. Experimental results and ablations are presented in Section VI, followed by a discussion of limitations and broader impact. Section VII concludes the paper and outlines future directions.

II. Related Work

A. Image–Text Retrieval

Retrieval systems that align free-form language with images range from fashion matching [12][13], sketch search [5], and large-scale photo datasets [1][6][7], to compositional language-vision models [8]. Extensions that ground queries in external knowledge [14][15] remain limited to two modalities, motivating our focus on *robust* multimodal grounding.

B. Multimodal Datasets and Fusion

New corpora highlight the need for techniques that cope with more than vision & text. CMU-MOSEI combines video, speech, and text for sentiment analysis [16]; GoLD pairs household objects with RGB, depth, spoken and written language [17][18]. Baltrušaitis *et al.* catalogue five core challenges (*representation, translation, alignment, fusion, co-learning*) in multimodal learning [19]; our work tackles the **alignment** problem when any subset of sensors may be absent.

C. Instance- vs Class-Level Retrieval

Multi-modal retrievers such as [20][21] treat objects with the same class label as interchangeable. For grounded robotics we instead require *instance* discrimination: the agent must find *that* red mug, not *any* red mug. We therefore adopt an instance-level objective and explicitly test robustness to missing modalities, an aspect ignored in prior class-level systems.

D. Alignment Losses and Robustness

Contrastive learning methods cluster into two families: classification/cross-entropy objectives [10][11] and geometric/metric losses [22]–[24]. Hybrid approaches are rare. Alayrac *et al.* align video, audio, and text with a dual-space loss [25], and Nguyen *et al.* use cosine similarity for image-language retrieval [26]; neither scales beyond three modalities nor handles sensor drop-out. Triplet-based works [4][27] often rely on costly hard-negative mining [28]–[35], which we avoid.

E. Higher-Order Multimodal Models

Efforts to fuse more than two modalities include three-way tensor products for images, hashtags, and users [36], quadruplet losses for sketch-image matching [37][38], and co-attention for image, sketch, and edgemap retrieval [39]. Emotion recognition combines face, speech, and text via CCA [40]; deception detection merges language, physiology, and thermal data [41]; heterogeneous transfer predicts a third modality from two inputs [42]. All scale poorly as modalities grow or assume every sensor is present. Our **EMMA** loss unifies an *arbitrary* number of modalities and demonstrates graceful degradation across nine missing-modality scenarios.

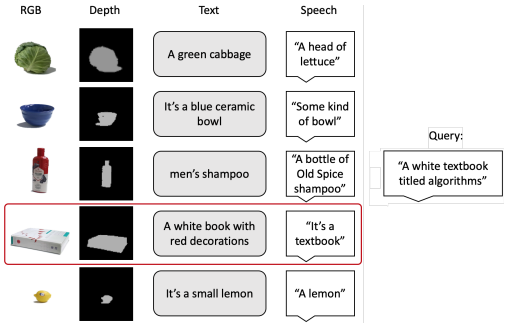


Figure 1. Multimodal object-retrieval setup (RGB, depth, speech, text).

III. Problem Description

Given a language command—either text or speech—that describes an object, we want our model to retrieve the correct object from a set of objects. This problem is an exemplar task in grounded language learning within the fields of robotics and natural language processing. Intuitively, the goal is to accept unconstrained natural-language queries and select the appropriate object by leveraging the complete set of sensor inputs available to the agent. We demonstrate a domain containing four modalities, each referring to objects in the environment: spoken language, written text, RGB images, and depth images. Figure 1 illustrates our object-retrieval task: the spoken query "A white textbook titled algorithms" is provided to our contrastive model, which identifies the item outlined in red in Figure 1 as the most likely object referred to by the query.

More formally, given a spoken-language command x_s, a textual command x_t, a set of RGB images $X_r = \{x_r^{(1..n)}\}$, and a set of depth images $X_d = \{x_d^{(1..n)}\}$, the task is to retrieve the correct object by choosing the index with the minimum distance to either language command across all modalities. Depending on which modalities are or are not ablated, we consider up to four distance vectors: *sr*, distances between x_s and all RGB images in X_r; *sd*, distances between x_s and all depth images in X_d; *tr*, distances between x_t and all RGB images in X_r; and *td*, distances between x_t and all depth images in X_d. To select the correct object, we first compute a component-wise average of the relevant modality-pair distances for the available modalities, then choose the object with the minimum of this averaged vector (i.e., we take the argmin).

Depending on which sensors are available at test time, any combination of these four distance vectors may be present. For example, if no written instructions are available—a salient setting because, although large bodies of text may exist during training, a user interacting with a physical agent might provide only spoken commands—we average sr and sd and select the object whose entry yields the lowest average distance. This method allows us to extend our model to arbitrary modality sets while remaining robust when some modalities are missing or incomplete.

IV. Approach

In keeping with previous work on the closely related problem of image retrieval, we focus on contrastive-loss approaches, where the goal is to learn an embedding in which similar samples—in our case, instances of the same object class—lie close together, while dissimilar samples are farther apart. We develop a novel geometric loss function, Geometric Alignment, that simultaneously minimizes intra-class distances and maximizes inter-class distances across every pair of modalities, yielding a model that is effective at the retrieval task defined above and robust to modality drop-outs at test time. We further combine this Geometric Alignment loss with a classification-based cross-entropy term, producing a superior model relative to either loss alone; we refer to this combination as **Extended Multimodal Alignment** (EMMA).

A. Core concepts.

The methods described in this section share terminology but differ in what they incorporate. Three terms recur: *anchor*, *positive*, and *negative*. The *anchor* is the reference data point; *positives* are samples similar to the anchor, and *negatives* are dissimilar. For example, to learn the concept "book," the anchor might be an RGB image of a book; the corresponding text description and depth image form the positive set, whereas the description and RGB image of an apple belong to the negative set. The methods below vary in how they choose these sets and in the objective functions they employ.

B. Baselines

We compare both EMMA and Geometric Alignment with the contrastive learning method of Chen *et al.* [11] and with supervised contrastive learning [10], hereafter SupCon. We treat SupCon as the principal baseline, as it generalizes several contrastive objectives, including triplet loss, the classic self-supervised contrastive loss [11], and N-pair loss [43].

1) Contrastive Loss: We re-implement the contrastive method of Chen *et al.* [11], which employs the normalized temperature-scaled cross-entropy loss (NT-Xent). Following SimCLR, we use cosine similarity; an unnormalized inner product [10] is numerically unstable because it is unbounded, but a normalized inner product is equivalent to cosine similarity. The loss is formulated in Equation (1).

$$-\sum_{i\in I}\log\frac{\exp(sim(z_i,z_{j(i)})/\tau)}{\sum_{a\in A(i)}\exp(sim(z_i,z_a)/\tau)} \tag{1}$$

where i is the index of the anchor, $j(i)$ is the index of the positive item with respect to the anchor z_i and is not the same as an anchor, $A(i)$ is the set of all negatives and the one positive indices excluding anchor, and $z = f(x)$.

We can treat different modalities of the same instance as additional input that augments the available information and consider them positive points for the anchor. Equation (1) can be rewritten with the sum over more than one positive item as formulated in Equation (2):

$$-\sum_{i\in I}\sum_{p\in P(i)}\log\frac{\exp(sim(z_i,z_p)/\tau)}{\sum_{a\in A(i)}\exp(sim(z_i,z_a)/\tau)} \tag{2}$$

where I is a batch consisting of one or more instances, each with a set of all its modalities, and $P(i)$ is the set of modalities/augmentations of the anchor i excluding itself (e.g., RGB image, depth image, speech, text) and $z = f(x)$. Therefore, if we have four modalities and the batch size is 64, the size of I is 256, the size of $P(i)$ is $M - 1 = 3$ where M is the number of modalities, and the size of $A(i)$ is $256 - 1 = 255$.

2) Supervised Contrastive Learning: [10] extend the contrastive learning method (NT-Xent) and propose a supervised way of performing contrastive learning to treat not only augmentations of the anchor but also every item that shares the same label with the anchor as positives. This loss function is shown in Equation (3).

$$\sum_{i\in I}\frac{-1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(z_i\cdot z_p/\tau)}{\sum_{a\in A(i)}\exp(z_i\cdot z_a/\tau)} \tag{3}$$

Although this loss function does not use cosine similarity, embeddings are normalized before performing the dot product, which is equivalent to cosine similarity.

The main difference between the contrastive loss baseline in Section IV-B1 and SupCon is that there is no notion of meaningful negative points in the contrastive loss, and everything in the batch that is not the anchor or one of the positive views is considered to be negative. In SupCon, however, all elements in the batch that have the same label as the anchor are also considered positives, in addition to different views of the same instance. While the denominators of Equations (2) and (3) stay the same, this subtle difference affects the numerator and includes more positive examples, which prevents the unintended use of actual positives as negative examples.

While this model is a strong baseline, the authors applied it to a unimodal dataset. In this paper, we extend the baseline to multimodal data and show that it learns more slowly than EMMA and performs worse when all modalities are available at test time.

Since SupCon considers all pairwise distances within a batch, with M modalities and a batch size B, each batch contains $B \times M$ items, and the computation involves $(BM)^2$ pairwise-distance terms, which depend on batch size. By contrast, the computations in our Geometric Alignment approach are agnostic to batch size, making it more scalable.

As originally proposed, the SupCon baseline was applied to unimodal datasets such as ImageNet [44], CIFAR-10 [45], and CIFAR-100 [45]. We demonstrate both that it can be used with multimodal datasets and that augmenting it with geometric components improves training speed and performance when modalities are dropped.

C. EMMA: Extended Multimodal Alignment

Our proposed multimodal method comprises two complementary parts. The first is a geometric loss based on latent-space distances; the second is a supervised contrastive loss based on cross-entropy (SUPCON). The geometric loss converges faster, whereas the cross-entropy loss aligns more closely with the downstream retrieval task. We therefore combine them to obtain **Extended Multimodal Alignment** (EMMA).

a) Geometric Alignment Loss: We define a distance-based loss function applicable to an arbitrary number of modalities. Our method is inspired by the well-known similarity-based triplet loss [4][23] and, under certain settings, resembles contrastive loss [10][11]. Triplet-loss learning forces similar concepts from different domains *together* in a shared embedding space while pushing dissimilar concepts *apart*. The name derives from the three data points it relies on: an anchor, a positive, and a negative. Standard triplet loss, however, cannot be applied to more than two modalities.

To address this limitation, we optimize pairwise distances for all data points, enabling use with an arbitrary number of modalities. In contrast, prior work that employs triplet loss [4][17] concatenates RGB and depth into a single "vision" vector, preventing robust handling of RGB or depth ablation at test time. Our method also avoids the need for hard-negative mining.

During training, we sample two object instances and gather their representations from every modality, producing a *positive* set (one object) and a *negative* set (a different object), as shown in Figure 2. Unlike some earlier triplet-loss methods [4][17], the anchor is not randomly chosen per batch. Instead, every item in the positive set becomes an anchor once; we minimize its distance to the other positive items while maximizing its distance to all negative items. Thus, our formulation is one-to-many rather than one-to-two.

To clarify our terminology:

- **Positive (Instance)** — embeddings of a single object (e.g., RGB image, depth image, text, and speech for an apple), shown in green in Figure 2.
- **Negative (Instance)** — embeddings of a different object (e.g., the same four modalities for a mug), shown in orange.
- **Anchor (Modality)** — each modality within the positive set is treated as an anchor once. In Figure 2, all four modalities serve in turn as anchors, forming the basis for distance learning.

The objective is to (i) minimize the distance between each pair of positive points from different modalities and (ii) maximize the distance between each positive and every negative point across all modalities.

We refer to this approach as GEOMETRIC ALIGNMENT, formulated in Equation (4); an illustration appears in Figure 2.

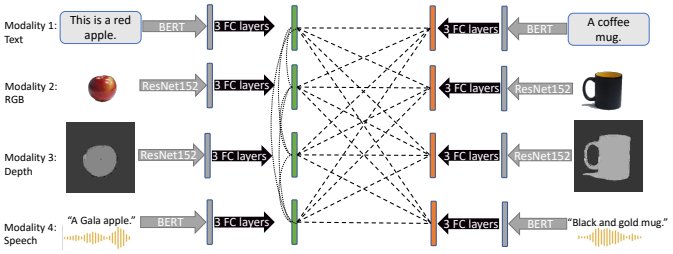


Figure 2. EMMA overview and GEOMETRIC ALIGNMENT loss (four modalities). Gray arrows = frozen encoders; black arrows = 3-layer FC + ReLU projectors. Green = positive, orange = negative embeddings; dashed lines maximize, dotted lines minimize distances.

$$L = \sum_{m_1=1}^{M} \Bigg[\sum_{m_2=1}^{M} \Big[-\max\big(dist(z^+_{m_1}, z^-_{m_2}) + \alpha, 0\big) \Big] + \sum_{m_3=m_1+1}^{M} \Big[\min\big(dist(z^+_{m_1}, z^+_{m_3}), 0\big) \Big] \Bigg] \tag{4}$$

In Equation (4), M is the number of modalities, the superscripts $+$ and $-$ represent positive and negative objects, α represents the enforced margin between each positive and negative point, which we set to 0.4 for all modalities without tuning, and z is the embedding we get by applying a mapping function f, which in our case is a neural network on our input data. In other words, $z_m = f_m(x_m)$, where each modality m has a specific model f_m that is different from the models for other modalities. These models do not share their weights.

Cosine similarity is the opposite of distance, and we need to reverse the logic for maximization and minimization. There are different options for measuring distance in embedded space. We use cosine similarity between pairs of embeddings, i.e., we measure the cosine of the angle between embeddings. Cosine similarity is a good choice for high-dimensional data as it is bounded between -1 and 1. Other distance metrics, such as Euclidean distance, grow in value with respect to their dimensionality, resulting in very large distances for data points.

Here, the generic *dist* function is replaced with the specific $\cos(\cdot)$, and we omit the max notation for clarity by defining Equation (5):

$$\begin{aligned} g(x, y) &= \max(\cos(x, y) - 1 + \alpha, 0) \\ h(x, y) &= \min(1 - \cos(x, y), 0). \end{aligned} \tag{5}$$

The first portion of the following equation maximizes all unique pairwise distances between modalities of positive and negative instances. The second portion minimizes the unique pairwise distances among the modalities of positive cases.

$$\mathcal{L} = \sum_{m_1=1}^{M} \underbrace{\sum_{m_2=1}^{M} g(z^+_{m_1}, z^-_{m_2})}_{\text{push negatives away}} + \sum_{m_1=1}^{M} \underbrace{\sum_{m_3=m_1+1}^{M} h(z^+_{m_1}, z^+_{m_3})}_{\text{pull positives together}} \tag{6}$$

Our proposed GEOMETRIC ALIGNMENT loss function in Equation (6) can be rewritten as shown in Equation (7) by fully specifying the summations to understand better how our objective function can be reduced to well-known losses such as triplet loss and pairwise loss.

$$\begin{aligned}\mathcal{L} = &\sum_{i=1}^{M-1}\sum_{j=i+1}^{M} h(z_i^+, z_j^+) \\ &+ g(z_i^+, z_j^-) + g(z_i^-, z_j^+) \\ &+ \sum_{i=1}^{M} g(z_i^+, z_i^-).\end{aligned} \tag{7}$$

If $M = 2$, which means the number of modalities is 2, and we ignore the last two terms in the derived objective function, it results in the triplet loss method. If $M = 2$, then our objective function reduces to the quadruplet loss method [37][38] if we multiply the first term by 2, ignore the third term, and change the last summation to be up to $M - 1$ (which results in a single term). If $M = 1$, only the last term remains in the loss function, which is exactly the pairwise distance-based loss function. This loss function can be seen as a contrastive loss usually used in the domain of self-supervised learning [11]. However, our proposed loss function has two advantages over the traditional contrastive loss expressed in Equation (1). The first advantage is that our loss function does not loop over multiple positives and negatives in a large batch. Instead, we sample only two objects (positive and negative), each of which has M modalities, which gives us $2M$ datapoints (or embeddings). Hence, our model can be trained using smaller batch sizes, which reduces the number of negative samples we need. The second advantage is that this loss function can be used in a multimodal setting with an arbitrary number of modalities and is not limited to a single data type (e.g., RGB images), which is the most common usage of contrastive loss. Although our GEOMETRIC ALIGNMENT is technically quadratic in terms of a number of modalities, we observe that experimentally, training time increases only by 10 about minutes with each additional modality.

Altogether, our proposed GEOMETRIC ALIGNMENT function contains $3M^2 - M/2$ terms: $M(M-1)/2$ anchor-to-positive distance minimizations and M^2 anchor-to-negative distance maximizations. It is noteworthy that our training procedure does not perform any stochastic dropout of modalities to obtain test-time robustness to missing modalities. Moreover, our approach does not need to compute the distance between all items in the batch, as opposed to SUPCON.

b) Combining Geometric Loss and Cross-Entropy-Based SUPCON *Loss:* The main difference between GEOMETRIC ALIGNMENT and SUPCON is that GEOMETRIC ALIGNMENT focuses on a geometric notion of similarity using cosine distance, whereas SUPCON employs cosine distance inside a classification objective akin to cross-entropy. Each method has advantages the other lacks. GEOMETRIC ALIGNMENT offers an intuitive distance-based objective, interpretable embeddings, and faster convergence. SUPCON benefits from a classification loss naturally aligned with the downstream task.

Let $A(i, m)$ denote all items in the batch except $z_{i,m}$ itself, and let $P(i, m)$ include all modalities of all instances with the same label as instance i, excluding $z_{i,m}$. Formally,

$$P(i,m) = \Big\{\bigcup\nolimits_{r\neq m} z_{i,r}\Big\} \cup \Big\{\bigcup\nolimits_{l\neq i,\ y_l=y_i} \bigcup\nolimits_{r=1}^{M} z_{l,r}\Big\}.$$

Both the geometric and cross-entropy components of EMMA avoid anchoring on a specific modality; instead, they consider all available modalities. This contrasts with earlier triplet-loss approaches. For example, in Figure 2, treating the apple (left) as instance I, the dotted lines between the apple's modalities minimize intra-instance distances via $h(x, y)$, whereas the dashed lines to the mug maximize inter-instance distances via $g(x, y)$—all possible pairs are considered.

Although SUPCON and GEOMETRIC ALIGNMENT both bring similar objects together and push dissimilar ones apart, SUPCON imposes a normalized ranking, whereas GEOMETRIC ALIGNMENT allows distances to vary arbitrarily. Furthermore, SUPCON typically treats different *augmentations* of an RGB image as positives, all drawn from the same distribution. By contrast, we treat distinct modalities—drawn from different distributions—as positives. To our knowledge, this is the first use of supervised contrastive learning in such a multimodal setting, where additional language and sensor inputs provide richer supervision than single-sensor augmentations.

Combining the two losses accelerates convergence and yields slightly higher performance when all modalities are present, while preserving the gains GEOMETRIC ALIGNMENT provides when modalities are missing. Detailed results appear in Section VI.

D. Network Architecture

Transformers have become the *de facto* architecture in natural language processing and have achieved strong performance across numerous tasks. Following [17], we use BERT embeddings from the FLAIR library [47][48] to featurize textual input and wav2vec2 [49] to extract audio embeddings from speech. Both encoders output a 3 072-dimensional vector obtained by concatenating the last four hidden layers of each network. FLAIR has been applied to tasks such as named entity recognition (NER) and part-of-speech (PoS) tagging, while wav2vec2 supports various audio-processing tasks, most notably automatic speech recognition. Both BERT [46] and wav2vec2 [49] are self-supervised transformer models [50].

For images, we use ResNet-152 [51] for both RGB and depth inputs, producing 2048-dimensional embeddings; depth images are colorized before being passed to the network.

Each modality's embedding is then projected into a shared 1024-dimensional space by a dedicated multi-layer perceptron (MLP) comprising three fully connected layers with ReLU activations [52]. These MLPs are modality-specific and do not share weights.

V. EXPERIMENTS

In this section, we evaluate the quality of object retrieval models learned using the EMMA loss function. We first describe the dataset we use, then define the metrics by which we assess performance, the setup of the experiments, and the baselines against which we compare. We end by presenting and analyzing the results.

A. Data

We demonstrate the effectiveness of our approach on a recent publicly available multimodal dataset called GoLD [17], which contains RGB images, depth images, written text descriptions, speech descriptions, and transcribed speech descriptions for 207 object instances across 47 object classes (see Figure 2). There are a total of 16,500 spoken and 16,500 textual descriptions. The original GoLD paper uses raw RGB and depth images in which other objects are present in the background. We use a masked version of the photos where the background is deleted (this masked version converges faster. However, masked and unmasked versions of the GoLD data converge to the same performance). Speech is converted to 16 Hz to match the wav2vec2 speech model.

B. Setup

To evaluate our model, we measure different performance metrics on a retrieval task in which the model has to select an object from a set of objects given a language description. Only one of the objects corresponds to the description, and the rest are from different object classes.

Similar to [10], we use a stochastic gradient descent (SGD) optimizer with momentum [53] with a flexible learning rate starting at 0.05.

All models are trained for 200 epochs with a batch size of 64 on a Quadro RTX 8000 GPU. We used a temperature of 0.1 for training the contrastive learning method described in Section IV-B1, and a temperature of 0.07 for training SUPCON as described in Section IV-B2.

To evaluate the performance, we compute the distance between the given natural language description and five randomly selected objects (1 of which corresponds to the description, with the others from different object classes). We compute the distance between the language embedding and all available sensory modalities of all candidates as described in Section V-D. In case we have RGB and depth, we compute the distance between language embedding and all candidate RGB embeddings, and we compute the distance between the same language embedding and all candidate depth embeddings corresponding to the RGB embeddings. We then take an average of these two distance matrices. Instead of choosing an empirical threshold beyond which objects are considered to be 'referred to,' we choose the closest image embedding (average distance of RGB and depth from language) as the prediction. In order to use cosine *distance*, we have to subtract the cosine of the *angle* between two embeddings (which represents similarity) from 1: that is, we compute $1 - \cos(e_1, e_2)$.

C. Metrics

The best metric to capture the performance in such a scenario is mean reciprocal rank (MRR, Equation (8) for Q queries). For each query, we predict the rank of all objects based on their distance from the language command, and then the inverse rank of the desired objects in all queries are averaged. For example, if the model predicts the desired object as the first rank, then MRR = $\frac{1}{1} = 1$, which means a perfect score, and if it predicts the correct object as the fourth rank among five objects, then MRR = $\frac{1}{4} = 0.25$.

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}_i} \tag{8}$$

While MRR is more meaningful when it comes to ranking in retrieval tasks, in real-world scenarios where a robot is asked to hand over an object if it fails, it does not matter whether the correct object was ranked second or last; the whole system would be considered a failure. Accuracy and micro F1 score are the same in this task, since for each prediction, we either have a true positive and no false positives and no false negatives, or we have no true positives, one false positive and one false negative. MRR is a more informative metric because it captures the idea that having the correct object as the second choice should be considered better than having it as a last choice, while in accuracy, the score is "all or nothing", either 0 or 1. Because our approach is designed to be robust to missing information across modalities, we also report MRR and accuracy for different combinations of modality dropouts.

D. Modality Ablation

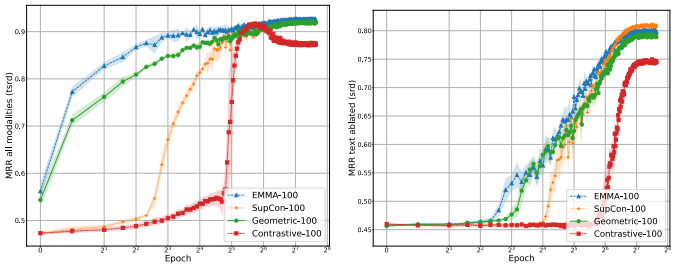
We train on four modalities—RGB, depth, speech, and written language—without altering the loss function beyond setting M in Equation (4) to the number of available modalities. Our downstream goal is non-trivial: identify the object referenced by arbitrary language given only a few examples.

When training with text, RGB, and depth, we treat written language as the query modality, compute its distances to RGB and depth, and then average those distances. Adding speech introduces a fourth sensory modality and three design choices:

1. Compute distances from both text and speech to RGB and depth (four distance matrices) and average them. 2. Treat speech like RGB and depth: compute distances from text to RGB, depth, and speech, then average the three. 3. As in option 1, but also include the distance between text and speech, averaging five matrices.

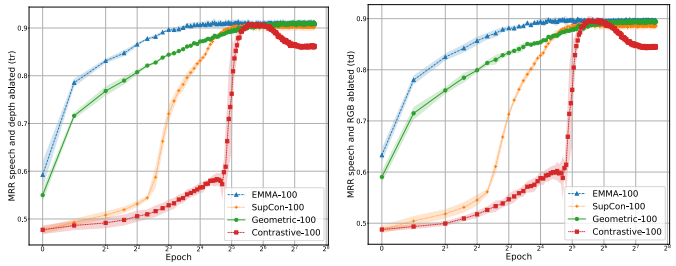
The first option is most appropriate for robust multimodal alignment. Options 2 and 3 are feasible during training, but in real-world retrieval people rarely both speak and type instructions. At test time, depending on available modalities, we use speech, text, or both to compute distances to RGB and depth and then average.

Nine dropout cases arise. Let *t* = text, *s* = speech, *r* = RGB, *d* = depth, and let K denote the final distance (a matrix for multiple queries, a vector for one).



(a) Mean Reciprocal Rank (MRR) on the held-out test set when all modalities are available.

(b) Mean Reciprocal Rank (MRR) on the held-out test set when the text modality is ablated.



(c) Mean Reciprocal Rank (MRR) on the held-out test set when speech and depth modalities are ablated.

(d) Mean Reciprocal Rank (MRR) on the held-out test set when speech and RGB modalities are dropped.

Figure 3. MRR on GoLD (avg 5 runs). Colors: red = self-sup CL, orange = sup CL, blue =GEOMETRIC ALIGNMENT, green =EMMA. Panels—(a) All inputs, (b) –Text, (c) –Speech–Depth, (d) –Speech–RGB. Higher is Better. We train all models for 200 epochs.

With two modalities we compute a single distance: K_{tr} (speech and depth missing), K_{sr} (text and depth missing), K_{td} (speech and RGB missing), K_{sd} (text and RGB missing).

With three modalities we average two distances: $K_{trd} = \frac{K_{tr}+K_{td}}{2}$ when speech is missing, $K_{srd} = \frac{K_{sr}+K_{sd}}{2}$ when text is missing, and so on.

With all four modalities we average four distances: $K_{tsrd} = \frac{K_{tr}+K_{td}+K_{sr}+K_{sd}}{4}$.

Figure 3 shows the relative performance of EMMA and GEOMETRIC ALIGNMENT against state-of-the-art methods when different modalities are ablated.

VI. RESULTS AND DISCUSSION

We evaluate on the GoLD test split using mean reciprocal rank (MRR) and top-1 accuracy (Acc).

Table I reports the *average* $\pm$ *standard deviation* over five random seeds; all models use SGD (batch 64). For five candidate objects, random guessing yields MRR = 0.33 and Acc = 0.50. EMMA matches or exceeds the strongest baseline in every modality setting.

To interpret MRR, note that a system that always ranks the correct object second would score $1/2 = 0.5$.

Figure 3a Shows that EMMA learns faster and results in a better performance compared to both SUPCON and contrastive learning [11] when trained using all modalities and with all modalities available during test. We observe that not only does contrastive loss learn more slowly, but it is prone to overfitting; while this can be addressed with careful tuning of the learning process, an approach that is innately robust to overfitting without tuning is preferable.

When we drop the text modality (Figure 3b), we can see that the performance decreases from about 0.93 to about 0.82, showing that speech cannot completely replace text. In Figure 4, the alignment of shared embeddings for a randomly sampled set of classes is visualized for all four modalities under consideration, suggesting that the speech modality is not aligned as well as the text modality. For this reason, when we drop text and use speech as the main query, the performance decreases. This supports our hypothesis that a geometric alignment of the latent space is crucial to good performance in object retrieval and multimodal understanding.

In Figure 3b, we observe that when speech is used as the query, and the text modality is ablated, the SUPCON baseline works slightly better than EMMA, although EMMA still learns faster. The reason is that SUPCON optimizes for the classification task, and since the speech modality is less well aligned, using GEOMETRIC ALIGNMENT makes the downstream task more difficult by trying to pull and push similar and dissimilar data points, respectively. Future research will consider strategies to align more chaotic modalities.

There is very little gap in performance when depth or RGB are dropped in Figures 3c and 3d compared to when we have all modalities in Figure 3a, showing that our model is robust when RGB or depth sensors fail. Also, when depth is dropped in Figure 3c, performance decreases less compared to when RGB is dropped in Figure 3d. This suggests that depth is less informative when compared to RGB, which is consistent with existing vision research results.

Our time analysis shows that EMMA takes almost 8 epochs to converge, and each epoch takes roughly 0.7 minutes, which makes it 5.6 minutes until convergence. In comparison, SUPCON takes about 36 epochs to converge, and each epoch takes 0.52 minutes, which amounts to 18.72 minutes. That is when we use all four modalities for training. When we ablate one or two modalities, the training takes less time.

Qualitative Results: In order to help visualize the performance of learned embeddings, we consider projections of a randomly selected subset of classes of the high-dimensional learned embeddings into a 3-dimensional space using t-SNE [54], a dimensionality reduction technique to visualize high-dimensional data. T-SNE creates a probability distribution over pairs of high-dimensional data where similar pairs have a higher probability, and dissimilar pairs have a lower probability. A similar probability distribution is also defined over pairs of data in the lower dimension (either 2D or 3D), and T-SNE minimizes the KL divergence between these two probability distributions.

Figure 4 shows the projection onto 3D space to give a better view of the location of embeddings. Although these projections are not perfect, combined with the quantitative results, they demonstrate that our model is learning to map instances of the same class closer to each other regardless of their modalities. Interestingly, toothbrush and toothpaste are mapped almost on top of each other in the text modality,

TABLE I. AVERAGE AND STANDARD DEVIATION OF MEAN RECIPROCAL RANK (MRR) AND ACCURACY (ACC) (HIGHER IS BETTER, BOLD = BEST).

Methods	speech/depth	speech/RGB	text/depth	text/RGB	text/speech/ depth	text/speech/ RGB	speech/RGB/ depth	text/RGB/ depth	all
Geometric	76.82±0.34	78.34±0.29	89.64±0.38	91.13±0.73	89.21±0.45	90.95±0.83	79.37±0.29	92.29±0.51	92.14±0.45
SupCon	**78.18**±0.58	**79.69**±0.54	89.04±0.88	90.56±0.74	88.75±0.66	90.5±0.69	**81.2**±0.39	91.96±0.42	92.03±0.7
EMMA	77.63±0.29	78.66±0.64	**89.87**±0.5	**91.26**±0.86	**89.66**±0.36	**90.97**±0.66	80.32±0.45	**92.71**±0.5	**92.72**±0.47
Contrastive	71.74±0.73	73.37±0.39	89.72±0.54	90.82±0.37	89.13±0.61	90.26±0.58	74.96±0.44	91.92±0.41	91.72±0.53

(a) AVERAGE AND STANDARD DEVIATION OF MRR (HIGHER IS BETTER, BOLD = BEST).

Methods	speech/depth	speech/RGB	text/depth	text/RGB	text/speech/ depth	text/speech/ RGB	speech/RGB/ depth	text/RGB/ depth	all
Geometric	61.95±0.55	64.34±0.53	82.03±0.57	84.6±1.1	81.08±0.81	84.0±1.4	65.84±0.63	86.41±0.83	85.94±0.74
SupCon	**64.17**±0.92	**66.52**±1.07	81.05±1.22	83.65±1.4	80.58±1.12	83.54±1.23	**68.7**±0.66	86.06±1.21	85.82±1.29
EMMA	63.54±0.53	65.07±1.01	82.78±0.97	**85.07**±1.42	**82.16**±0.64	**84.37**±1.23	67.69±0.81	**87.38**±0.71	**87.15**±0.72
Contrastive	54.82±1.4	57.27±0.64	**82.88**±0.88	84.35±1.01	81.55±0.93	83.26±1.02	59.38±0.6	86.31±0.67	85.75±0.87

(b) AVERAGE AND STANDARD DEVIATION OF ACC ACROSS 5 RANDOM SEEDS ON THE HELD-OUT TEST SET; COLUMN HEADERS SHOW MODALITIES PRESENT AT QUERY TIME (HIGHER IS BETTER, BOLD = BEST).

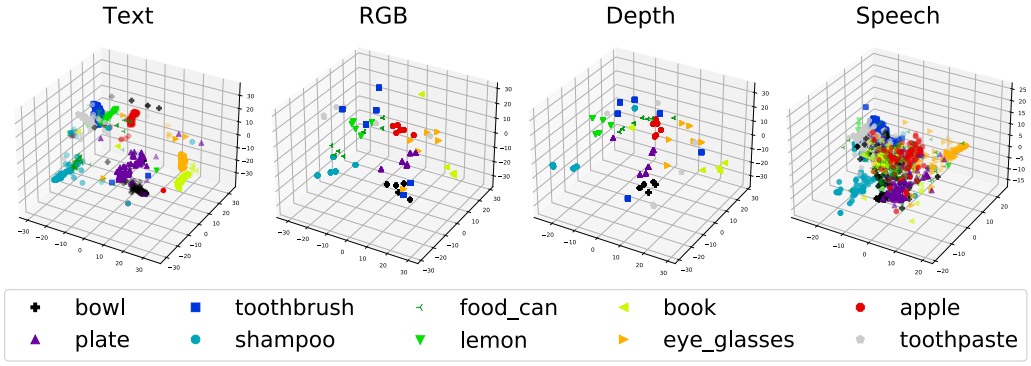


Figure 4. 3-D t-SNE of EMMA embeddings for 10 random object classes. RGB, depth, speech, and text appear as separate point clouds; dense language points reflect multiple descriptions. Tight cross-modal clusters reveal a shared manifold for reliable retrieval.

showing similar semantic and syntax. However, in the RGB and depth modality, they are close but not on top of each other since they do not look the same. Also, we can see that apple and lemon are mapped close to each other in all modalities, which suggests that our proposed EMMA learns some notion of the concept of fruits. These qualitative results show that our propose GEOMETRIC ALIGNMENT and EMMA have an interpretable latent space.

An example of the need to consider multiple modalities jointly is shown in Figure 5, showing how EMMA is able to correctly select an object instance from several similarly shaped and describable objects.

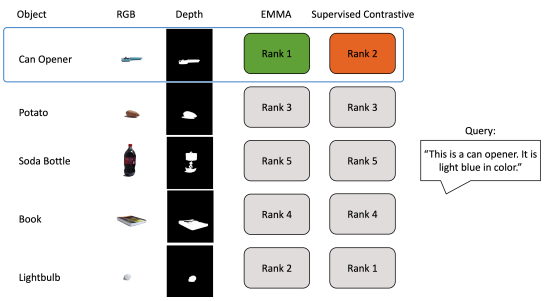


Figure 5. Qualitative retrieval: EMMA ranks the target first, whereas SUPCON mis-ranks a "light bulb" due to phrase similarity.

Our proposed model performs well and learns fast, has been demonstrated to handle four modalities of shared information effectively, and is robust to test-time situations where information from one or more modalities is missing. The bottleneck for agents in different settings may differ, and training speed may not be critical in offline learning scenarios. However, since we usually need to finetune models for other tasks when it comes to transfer learning, the training speed becomes relevant.

There remains room for improvement. Specifically, the speech modality is harder to handle. Figure 4 shows that although the relative position of instances are correct in the speech space, the distinction and clustering of different objects are not as good as the other three modalities.

The text seems to be the best-clustered modality, and that makes sense because the variation in written text is much smaller than the other three modalities. Variation in speech is higher because there are a number of factors affecting speech understanding, including different accents, native language, gender, and age [18]. Variation in RGB and depth is higher than in text due to variations in lighting conditions, an object's texture and shape, the angle of the camera, and other factors.

VII. Conclusion

In this work, we have demonstrated the effectiveness of a novel approach to learning from high-dimensional multimodal information even when one or more modalities are unavailable at test time. Our approach performs well on an object retrieval task from a testbed that contains four separate modalities, consistent with the information that might be available to a physical agent, and outperforms state-of-the-art contrastive learning approaches. Our proposed method is general enough to be applied to a variety of multimodal retrieval problems and is not limited to purely language-based image retrieval.

In the future, this work will be extended to solve less clearly delineated problems, such as differentiating among members of a class and across classes. However, this work represents a significant step towards handling such retrieval problems while not arbitrarily limiting the number of sensors and other modalities that can be incorporated.

References

[1] W. Hong *et al.*, "Gilbert: Generative vision-language pre-training for image-text retrieval," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1379–1388.

[2] M. Zhuge *et al.*, "Kaleido-bert: Vision-language pre-training on fashion domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2021, pp. 12 647–12 657.

[3] R. Hu *et al.*, "Natural language object retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4555–4564.

[4] A. T. Nguyen *et al.*, "Practical cross-modal manifold alignment for robotic grounded language learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2021, pp. 1613–1622.

[5] F. Huang, Y. Cheng, C. Jin, Y. Zhang, and T. Zhang, "Deep multimodal embedding model for fine-grained sketch-based image retrieval," in *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2017, pp. 929–932.

[6] C. Ma, C. Gu, W. Li, and S. Cui, "Large-scale image retrieval with sparse binary projections," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 1817–1820.

[7] D. Novak, M. Batko, and P. Zezula, "Large-scale image retrieval using neural net descriptors," in *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, 2015, pp. 1039–1040.

[8] N. Vo *et al.*, "Composing text and image for image retrieval-an empirical odyssey," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6439–6448.

[9] L. E. Richards, K. Darvish, and C. Matuszek, "Learning Object Attributes with Category-Free Grounded Language from Deep Featurization," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2020, pp. 8400–8407. DOI: 10.1109/IROS45743.2020.9340824.

[10] P. Khosla *et al.*, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 18 661–18 673.

[11] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[12] D. Gao *et al.*, "Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020, pp. 2251–2260.

[13] H. Wen, X. Song, X. Yang, Y. Zhan, and L. Nie, "Comprehensive linguistic-visual composition network for image retrieval," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1369–1378.

[14] W. Zheng and K. Zhou, "Enhancing conversational dialogue models with grounded knowledge," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM '19, Beijing, China: Association for Computing Machinery, 2019, pp. 709–718, ISBN: 9781450369763. DOI: 10.1145/3357384.3357889.

[15] C. Meng *et al.*, "Dukenet: A dual knowledge interaction network for knowledge-grounded conversation," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1151–1160, ISBN: 9781450380164.

[16] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2236–2246. DOI: 10.18653/v1/P18-1208.

[17] G. Y. Kebe *et al.*, "A spoken language dataset of descriptions for speech-based grounded language learning," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[18] G. Y. Kebe, L. E. Richards, E. Raff, F. Ferraro, and C. Matuszek, "Bridging the gap: Using deep acoustic representations to learn grounded language from percepts and raw speech," in *Proceedings of the AAAI Conference on Artificial Intelligence*, AAAI Press, 2022.

[19] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019, ISSN: 1939-3539. DOI: 10.1109/TPAMI.2018.2798607.

[20] A. Jangra, S. Saha, A. Jatowt, and M. Hasanuzzaman, "Multimodal summary generation using multi-objective optimization," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1745–1748, ISBN: 9781450380164.

[21] P. Hu, L. Zhen, D. Peng, and P. Liu, "Scalable deep multimodal learning for cross-modal retrieval," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR'19, Paris, France: Association for Computing Machinery, 2019, pp. 635–644, ISBN: 9781450361729. DOI: 10.1145/3331184.3331213.

[22] P. Poklukar *et al.*, "Geometric multimodal contrastive representation learning," in *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri *et al.*, Eds., ser. Proceedings of Machine Learning Research, vol. 162, PMLR, 17–23 Jul 2022, pp. 17 782–17 800.

[23] M. Carvalho *et al.*, "Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings," in *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '18, Ann Arbor, MI, USA: Association for Computing Machinery, 2018, pp. 35–44, ISBN: 9781450356572. DOI: 10.1145/3209978.3210036.

[24] A. Salvador *et al.*, "Learning cross-modal embeddings for cooking recipes and food images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[25] J.-B. Alayrac *et al.*, "Self-Supervised MultiModal Versatile Networks," in *NeurIPS*, 2020.

[26] T. Nguyen *et al.*, "Robot Object Retrieval with Contextual Natural Language Queries," in *Proceedings of Robotics: Science and Systems*, Corvalis, Oregon, USA, Jul. 2020. DOI: 10.15607/RSS.2020.XVI.080.

[27] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large Scale Online Learning of Image Similarity Through Ranking," *J. Mach. Learn. Res.*, vol. 11, pp. 1109–1135, Mar. 2010, ISSN: 1532-4435.

[28] E. Hoffer and N. Ailon, "Deep Metric Learning Using Triplet Network," in *SIMBAD 2015: Similarity-Based Pattern Recognition*, A. Feragen, M. Pelillo, and M. Loog, Eds., Cham: Springer International Publishing, 2015, pp. 84–92, ISBN: 978-3-319-24261-3. DOI: 10.1007/978-3-319-24261-3_7.

[29] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2015, pp. 815–823, ISBN: 978-1-4673-6964-0. DOI: 10.1109/CVPR.2015.7298682.

[30] V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., *Computer Vision - {ECCV} 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part {IX}*, vol. 11213, Lecture Notes in Computer Science, Springer, 2018, ISBN: 978-3-030-01239-7. DOI: 10.1007/978-3-030-01240-3.

[31] Y. Zhao, Z. Jin, G.-J. Qi, H. Lu, and X.-S. Hua, "An Adversarial Approach to Hard Triplet Generation," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., ser. Lecture Notes in Computer Science, vol. 11213, Springer, 2018, pp. 508–524, ISBN: 978-3-030-01239-7. DOI: 10.1007/978-3-030-01240-3_31.

[32] Y. Zhai, X. Guo, Y. Lu, and H. Li, "In Defense of the Triplet Loss for Person Re-Identification," *ArXiv e-prints*, 2018. arXiv: 1809.05864.

[33] K. Musgrave, S. Belongie, and S.-N. Lim, "A Metric Learning Reality Check," in *ECCV*, 2020. arXiv: 2003.08505.

[34] E. Raff, "Research Reproducibility as a Survival Analysis," in *The Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021. arXiv: 2012.09932.

[35] E. Raff, "A step toward quantifying independently reproducible machine learning research," in *Advances in Neural Information Processing Systems*, H. Wallach *et al.*, Eds., vol. 32, Curran Associates, Inc., 2019, pp. 14–25.

[36] A. Veit, M. Nickel, S. Belongie, and L. van der Maaten, "Separating self-expression and visual content in hashtag supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2018.

[37] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: A deep quadruplet network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 403–412.

[38] O. Tursun, S. Denman, S. Sridharan, E. Goan, and C. Fookes, "An efficient framework for zero-shot sketch-based image retrieval," *arXiv preprint arXiv:2102.04016*, 2021.

[39] J. Lei *et al.*, "Semi-heterogeneous three-way joint embedding network for sketch-based image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 9, pp. 3226–3237, 2020. DOI: 10.1109/TCSVT.2019.2936710.

[40] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 02, pp. 1359–1367, Apr. 2020. DOI: 10.1609/aaai.v34i02.5492.

[41] M. Abouelenien, V. Pérez-Rosas, R. Mihalcea, and M. Burzo, "Deception detection using a multimodal approach," in *Proceedings of the 16th International Conference on Multimodal Interaction*, ser. ICMI '14, Istanbul, Turkey: Association for Computing Machinery, 2014, pp. 58–65, ISBN: 9781450328852. DOI: 10.1145/2663204.2663229.

[42] Z. Liu, W. Zhang, S. Lin, and T. Q. Quek, "Heterogeneous sensor data fusion by deep multimodal encoding," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 3, pp. 479–491, 2017. DOI: 10.1109/JSTSP.2017.2679538.

[43] K. Sohn, "Improved deep metric learning with multi-class n-pair loss objective," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29, Curran Associates, Inc., 2016.

[44] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[45] A. Krizhevsky, G. Hinton, *et al.*, "Learning multiple layers of features from tiny images," 2009.

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[47] A. Akbik *et al.*, "Flair: An easy-to-use framework for state-of-the-art nlp," in *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, 2019, pp. 54–59.

[48] A. Akbik, T. Bergmann, and R. Vollgraf, "Pooled contextualized embeddings for named entity recognition," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 724–728. DOI: 10.18653/v1/N19-1078.

[49] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 12 449–12 460.

[50] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 770–778, ISBN: 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.90.

[52] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010, pp. 807–814.

[53] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

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