Multimodal Clustering Networks for Self-supervised Learning from Unlabeled Videos

Brian Chen [1] Andrew Rouditchenko[2] Kevin Duarte [3] Hilde Kuehne[4,6] Samuel Thomas[5,6]
Angie Boggust[2] Rameswar Panda[5,6] Brian Kingsbury[5,6] Rogerio Feris[5,6]
David Harwath[7] James Glass[2] Michael Picheny[8] Shih-Fu Chang[1]

[1]Columbia University, [2]MIT CSAIL, [3]University of Central Florida, [4]Goethe University Frankfurt,
[5]IBM Research AI, [6]MIT-IBM Watson AI Lab, [7]UT Austin, [8]NYU-Courant CS & CDS,
{bc2754,sc250}@columbia.edu,{roudi,aboggust,glass}@mit.edu,kevin_duarte@knights.ucf.edu
{kuehne,rpanda}@ibm.com,{sthomas,rsferis,bedk}@us.ibm.com, harwath@cs.utexas.edu, map22@nyu.edu

Abstract

Multimodal self-supervised learning is getting more and more attention as it allows not only to train large networks without human supervision but also to search and retrieve data across various modalities. In this context, this paper proposes a framework that, starting from a pre-trained backbone, learns a common multimodal embedding space that, in addition to sharing representations across different modalities, enforces a grouping of semantically similar instances. To this end, we extend the concept of instance-level contrastive learning with a multimodal clustering step in the training pipeline to capture semantic similarities across modalities. The resulting embedding space enables retrieval of samples across all modalities, even from unseen datasets and different domains. To evaluate our approach, we train our model on the HowTo100M dataset and evaluate its zero-shot retrieval capabilities in two challenging domains, namely text-to-video retrieval, and temporal action localization, showing state-of-the-art results on four different datasets.

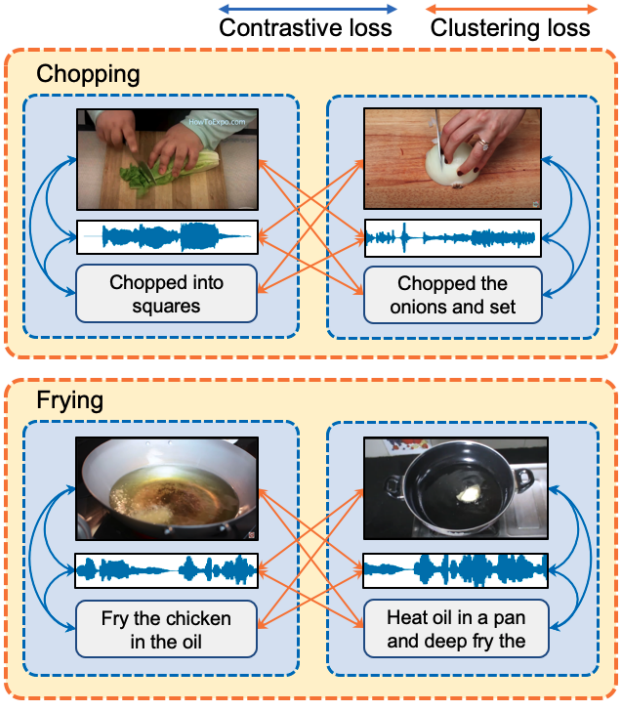


Figure 1: The Multimodal Clustering Network (MCN) combines a contrastive loss that learns feature representations to be close across different modalities such as video, audio, and text (blue box), with a clustering loss that draws instances that are semantically related together, e.g., scenes depicting the same semantic concept (e.g., chopping or frying) from different videos or different clips. (yellow box).

1. Introduction

To robustly learn visual events and concepts, humans seldom rely on visual inputs alone. Instead, a rich multimodal environment is utilized for understanding by combining multiple sensory signals along with various language representations. Many recent techniques have attempted to mimic this paradigm to train efficient computer vision models, especially those that learn from videos where multiple modalities are naturally present [1, 2, 36].

Learning on multimodal video data has both benefits and challenges. It is beneficial that each video instance has information available in multiple modalities. Textual information corresponding to the spoken narrations in the video, for example, provides a valuable language modality in addition to the visual and audio modalities [7, 21, 25]. In this work, we

focus on the problem of learning a joint embedding space across multiple modalities. Given that the features from different modalities are often not comparable, the goal is to learn the projections into a common space where features from different domains but with similar content are close to each other to allow for a direct retrieval across modalities. However, creating an effective joint multimodal embedding space is not easy. First, each of those modalities is different, *i.e*. with respect to its source, how it is sampled and processed, and its resulting feature representation. Additionally, in real-world data, the supervision available to learn these projections from each of the modalities is unfortunately weak, as *e.g*. audio sequences can be misaligned to their visual representations and corresponding narration might or might not be present in the same time interval [2, 32].

To deal with multimodal data of this nature, several recent approaches use a contrastive loss [18, 19] to learn e.g. feature representations in a joint embedding space. The goal is to bring samples drawn from the same temporal instance closer to each other while keeping samples from different times apart. Recent works [1, 32] show that such training is useful for pretraining models on large-scale data without additional supervision and that the resulting models achieve competitive performance on several tasks, *e.g*. in action classification when fine-tuned on various datasets. One problem arising from the contrastive loss is that this criterion does not consider the samples' semantic structure and similarity at different times: two samples are treated as a negative pair as long as they occur at different times regardless of their semantic similarity. This can have a considerable adverse impact on the learned representation. In a different formulation for learning representations, instead of comparing individual instances, clusters of instances are first created using a certain clustering algorithm [2, 5, 11, 29]. This approach encourages samples semantically similar to each other (namely, samples in the same cluster) to be close in the embedding space. However, if we cluster features from multi-modalities, those clusters would likely emerge only within the modalities separately, clustering audio instances with audio instances, visuals to visuals *etc*. Therefore, a mechanism that pulls the instances from different modalities together is crucial to cluster features from different modalities in a joint space. This leads to our proposed method that treats these two approaches as reciprocal information.

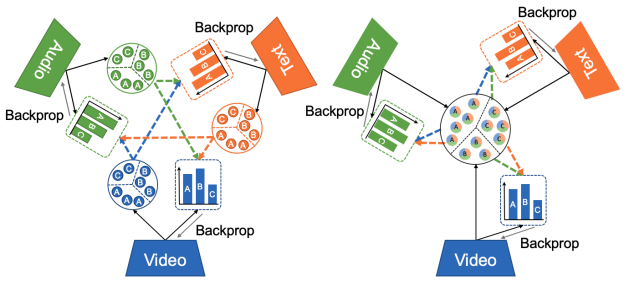
We present a multimodal learning framework that learns joint representations by training cross-modal projection heads from the visual, audio, and language modalities and accounts for the semantic similarity of embedding using a large corpus of naturally narrated videos. The proposed *Multimodal Clustering Network* (MCN) adopts a novel architecture to combine promising ideas from both representation learning paradigms described earlier: learning via the contrastive loss at the instance level and the semantic consistency at the cluster level. As another novel feature of our approach, we explore joint clusters using multimodal representations instead of clusters using separate modalities. The result features allow us to do retrieval across different modalities in linear time. Figure 1 provides a high-level overview of our approach.

To evaluate our proposed method, we address the challenging problem of zero-shot learning in two contexts: multimodal video retrieval and multimodal temporal action localization. We train our system on the HowTo100M dataset [33] and evaluate its retrieval capabilities on the YouCook2 [44] and MSR-VTT [42] dataset and its temporal action localization on the task of action detection on the CrossTask [46] dataset and on the task of temporal action segmentation on the Mining YouTube [26] dataset. Using only features from pretrained backbones, MCN significantly outperforms the best text-to-video retrieval baseline over absolute 3% in recall and outperforms the temporal action localization baseline over 3.1% in recall, both in zero-shot settings.

Contributions. The contributions of this work are threefold: (i) We propose a novel method by combining the benefits of contrastive loss and clustering loss for multimodal joint space learning. Unlike prior works that create clusters using separate modalities, our method shows the important benefits of using multimodal joint clusters. (ii) We show that the proposed model can learn across three modalities (video, audio, text) in a joint space. (iii) We demonstrate significant performance gains on multiple downstream tasks in the zero-shot setting. These results show that the learned common space representations can improve state-of-the-art results without any additional training on the target datasets.

2. Related Work

Learning from Multimodal Data. Instead of collecting new annotated datasets [12, 38] for building various state-of-the-art visual recognition models, current approaches leverage large amounts of videos available on multiple social media platforms. When specific language resources like automatically generated speech recognition captions are available in narrated video datasets such as How2 [39] or HowTo100M [33], an appropriate proxy task that leverages these resources is instead used. Such visual caption pairs have been widely used in self-supervised models in vision and language tasks recently [3, 16, 17, 28, 31, 35, 40, 45]. In other approaches like [2, 6, 8, 21, 30, 37], the need for these language transcripts is avoided by using just the corresponding raw speech signal. More recently, models that trained from scratch from the narrated video along with generated speech captions have also been successfully developed [32]. The three modalities naturally present in videos, the visual, audio, and language streams, are further integrated via a multimodal variant of this learning framework in [1]. Unlike these works, our goal in this paper is to learn a joint em-



(a) Separate clusters for each domain

(b) Joint cluster across domains

Figure 2: **Cross-domain Clustering vs. Joint Clustering.** (a) Previous methods such as XDC perform clustering at separate spaces and use pseudo-labels as supervision to other domains. (b) Our method performs clustering across features from different modalities in the joint space to learn multimodal clusters. Best viewed in color.

bedding in three modalitites for zero-shot multimodal downstream tasks where we create an embedding space which the features across different modalitites are directly comparable.

Contrastive Learning. A technique central to several state-of-the-art self-supervised representation learning approaches for images is instance-wise contrastive learning [13, 22]. In this paradigm, a model is trained to place samples extracted from the same instance, e.g., transforms or crops of an image, close to each other while pushing samples from different instances further apart. Given its similarity to noise contrastive estimation (NCE), where two samples are treated as a negative pair as long as they are drawn from different time segments, in MIL-NCE [32], the benefits of both multiple instance learning and NCE are combined. An advantage of this approach is that it now allows for compensation of misalignments inherently found in videos and corresponding text captions. One inherent drawback of the instance-wise contrastive learning described above is that it is agnostic to the inherent semantic similarity between the samples when positive and negative pairs are constructed. In our work, we alleviate this problem by relaxing the instance level similarity across modalities to semantic level similarity by introducing a clustering component that learns semantic similarity among multimodal instances within the batch.

Deep Unsupervised Clustering. Given the high cost of computing all pairwise comparisons in a large dataset, instead of applying the contrastive learning paradigm discussed above on each individual instance, a more practical solution is to discriminate between groups of instances during training. This is done by first pre-training a model to derive suitable feature representations of the data in a simple cascaded approach. Keeping the representations fixed, a clustering algorithm is then used to group instances before the weights of the model are updated using the derived class assignments as supervision [10, 43]. In contrast, instead of keeping the clustering step independent of the representation learning phase, more recent techniques jointly learn visual embeddings and cluster assignments [5, 6, 11, 41]. While both these approaches can produce interpretable clustering results that benefit downstream tasks by integrating global information across the entire dataset, running a clustering algorithm over a large data set slows down training. However, this issue can be addressed by performing the clustering in an online fashion [11]. These online models simultaneously learn to cluster and represent image data. To improve the performance of clustering, it is, however, also essential to leverage the correlated yet very complementary information available in the various modalities present in narrated videos [5]. To learn better feature extractors for audio and video, recent works, XDC [2] and SeLaVi [5] extend this clustering idea to the multimodal space. While these approaches focus on learning better feature extractors for each domain separately, our goal is to learn a joint multimodal embedding. As shown in Figure 2, these cross-domain clustering methods (left) create separate clusters and use cross-domain pseudo-labels as the supervision for each feature extractor. In contrast, our model (right) creates a common embedding space across all modalities and performs clustering jointly.

3. Learning to Cluster Multimodal Data

To effectively construct a *joint representation space* from unlabeled narrated videos, we start with n narrated video clips. Each video clip is associated with its corresponding visual representation, audio representation and text narration. Given this input, the joint embedding space is learned, where the embeddings of video clips with semantically similar visual, audio, and text content are close to each other and apart when the content is dissimilar, as illustrated in Figure 1.

Using the notation in [32], for each clip, let video $v \in \mathcal{V}$ denote its visual representation, $a \in \mathcal{A}$ represent its corresponding audio and $t \in \mathcal{T}$, its matching text narration generated using an automatic speech recognition (ASR) system. Given a set of n tuples of associated video, audio and text narrations $\{(v_i, a_i, t_i)\}_{i=1}^{n} \in (\mathcal{V} \times \mathcal{A} \times \mathcal{T})^n$, as shown in Figure 3 (a), we first construct three parametrized mappings that derive embedding representations from the original video, audio and text signals. Transform $f : \mathcal{V} \rightarrow \mathbb{R}^d$ derives a d-dimensional embedding representation $f(v) \in \mathbb{R}^d$ from a video clip v, transforms $g : \mathcal{A} \rightarrow \mathbb{R}^d$ and $h : \mathcal{T} \rightarrow \mathbb{R}^d$, produce similar d-dimensional audio and text embeddings: $g(a) = z \in \mathbb{R}^d$ and $h(t) \in \mathbb{R}^d$. In this work, f takes as input pre-extracted 2D and 3D features from a fixed-length clip, the input for g are log-mel spectrograms extracted from the audio segments, and for h, we use a sentence based neural model that transforms a set of words into a single vector. More details about model architectures are in Section 4.

Next, we introduce three loss functions to guide and properly situate these embeddings in the joint embedding space.

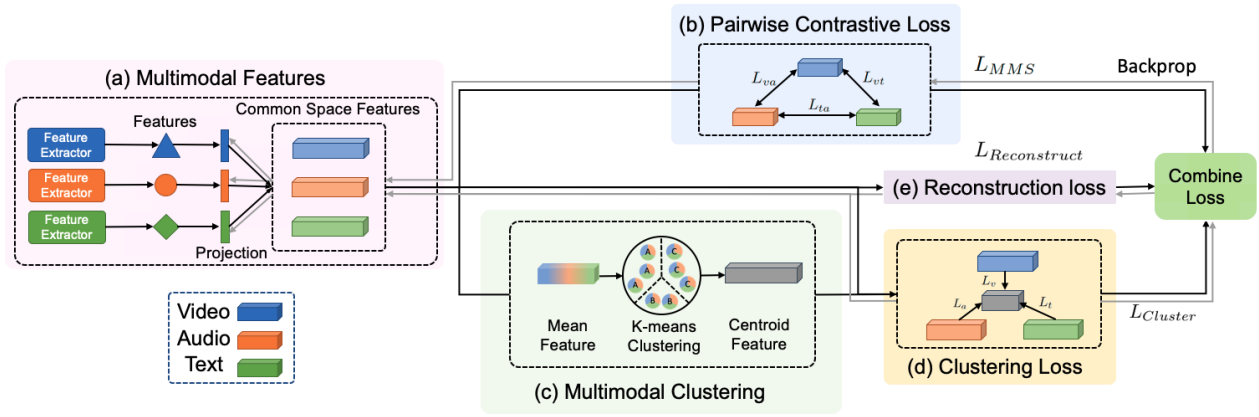


Figure 3: **Illustration of our proposed framework.** Our framework comprises four parts: (a) Extracting features from several modalities and projecting them into joint space. (b) Calculating contrastive loss pairwise to pull the features close across modalities. (c) Performing multimodal clustering across features from different domains in a batch. (d) Performing joint prediction across features to multimodal centroids to bring together semantically similar embeddings. (e) Reconstruction loss for regularization. Best viewed in color.

A contrastive loss L_{MMS} is used to ensure that the representations from each of the three modalities are comparable. A second clustering loss $L_{Cluster}$ encourages representations from semantically similar samples across all modalities to remain close in the learned embedding space. A third reconstruction loss $L_{Reconstruct}$ regularizes the multimodal common space features for more stable clustering training. The final model is trained to minimize sum of these losses.

$$L = L_{MMS} + L_{Cluster} + L_{Reconstruct} \tag{1}$$

3.1. Contrastive Loss for Learning Joint Spaces

To learn a joint space for the three modalities, we compute a contrastive loss on all pairs of modalities, $(v,t),(t,a),(a,v)$, as shown in Figure 3 (b). This loss maximizes the similarity between representations corresponding to any two modalities from the same instance (video clip) while minimizing the similarity of imposter pairs from the two modalities from one clip of video to another. In this work, we use the Masked Margin Softmax (MMS) function [24], which defines the similarity between representations from two modalities in terms of their learned embedding vectors' dot product within a batch B. Features from each of the three modalities $\{V, A, T\}$ are assembled for each batch. The total contrastive loss L_{MMS} is the sum of pairwise losses using each of the three modalities:

$$L_{MMS} = L_{ta} + L_{vt} + L_{va} \tag{2}$$

where L_{ta}, L_{vt}, L_{va} represent the loss associated with pairwise modalities $(t,a),(v,t),(a,v)$ respectively. For a pair of modalities, for example the text and audio modalities, the individual loss L_{ta} is in turn given as:

$$L_{ta} = -\frac{1}{B}\sum_{i=1}^{B}\left[\left(\log\frac{e^{h(\mathbf{t}_i)\cdot g(\mathbf{a}_i)-\delta}}{e^{h(\mathbf{t}_i)\cdot g(\mathbf{a}_i)-\delta}+\sum\limits_{\substack{k=1\\k\neq i}}^{B} e^{h(\mathbf{t}_k^{imp})\cdot g(\mathbf{a}_i)}}\right)\right. \tag{3}$$
$$\left.+\left(\log\frac{e^{h(\mathbf{t}_i)\cdot g(\mathbf{a}_i)-\delta}}{e^{h(\mathbf{t}_i)\cdot g(\mathbf{a}_i)-\delta}+\sum\limits_{\substack{j=1\\j\neq i}}^{B} e^{h(\mathbf{t}_i)\cdot g(\mathbf{a}_j^{imp})}}\right)\right]$$

where a_j^{imp} represents imposter pairs from two modalities that are sampled from a batch but do not co-occur. As can been seen in the L_{ta} case, this loss attempts to discriminate between positive or true embedding pairs and imposter or negative pairs within each batch. Using two separate parts, the space of positive and negative samples is enumerated separately: in one case, a given text sample is paired with various negative audio samples. In the second case, an audio sample is paired with various negative text samples. (i, j, k) are various indices of video clips in a given batch. δ is a margin hyperparameter that is empirically selected. By projecting all features to the same space and ensuring that their similarities are maximized pairwise, this formulation of the pairwise contrastive loss ensures that the features across different modalities are comparable.

3.2. Clustering Multimodal Features

To ensure that representations of semantically related instances are close in the learned joint multimodal space, in addition to contrastive loss described above, a self-supervised clustering step is included as part of the training process.

Online K-means clustering. We applied standard clustering algorithm k-means that takes a set of vectors as input, in

our case, the features M produced by the fused multimodal feature:

$$M = (f(\mathbf{v}) + g(\mathbf{a}) + h(\mathbf{t}))/3 \tag{4}$$

where we take the mean over embeddings from three modalities to represent a multimodal instance. We cluster them into k distinct groups. More precisely, it outputs a $d \times k$ centroid matrix $C = \{\mu_1, .., \mu_k\}$ and the cluster assignments y_n of each multimodal instance n are defined by solving the following problem:

$$\min_{C \in \mathbb{R}^{d \times k}} \frac{1}{N} \sum_{n=1}^{N} \min_{y_n \in \{0,1\}^k} \|M_n - C y_n\|_2^2 \tag{5}$$

We then acquire a centroid matrix C^* and a set of assignments $(y_n^*)_{n \leq N}$. Unlike pseudo-labels-based methods [10] that only make use of the assignments (labels), we make use of the centroid matrix for semantic learning. To cover variant semantic information for clustering, we use features from the previous batches to gather sufficient instances for online learning.

Semantic centroid learning. To learn the features closer to its multimodal semantic centroids. We proposed to use the centroid as a contrastive loss reference target. This target pulls the features from three modalities closer to the centroid that is close to their multimodal instance feature M_n and pushes the features far away from the other centroid. For each modality, for example, the text modalities, the individual loss L_t is in turn given as:

$$L_t = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{e^{h(\mathbf{t}_i) \cdot \mu' - \delta}}{\sum_{k=1}^{K} e^{h(\mathbf{t}_i) \cdot \mu_k}} \tag{6}$$

where μ' is the nearest centroid for the multimodal instance feature M_i and μ'. We later sum over the loss from three modalities:

$$L_{Cluster} = L_v + L_a + L_t \tag{7}$$

In the end, the projected features learn to be closer to its centroid feature among the three and also learns to be closer in similar semantics.

Multimodal features reconstruction. Reconstruction can help in capturing features that are suppressed by contrastive learning/clustering [14]. In a video of *chopping onions*, with both the sound of chopping in the background as well as the speech/text with the word *onion* in the foreground, it is possible that contrastive learning/clustering will focus more on associating the video with either the sound (background) or the speech (foreground), but not both. We hypothesize that the reconstruction loss will force the capture of features from both background and foreground, which is important for retrieval/other downstream tasks. Reconstruction is also an auxiliary task that helps regularize training and improve generalization [27]. We performed a reconstruction loss on top of the common space features from three modalities to stabilize the feature training during clustering. For each modality, for example, the visual modalities, the individual loss $L_{v'}$ is in turn given as:

$$L_{v'} = -\frac{1}{B} \sum_{i=1}^{B} \|f'(\mathbf{v}) - f(\mathbf{v})\|^2 \tag{8}$$

where $f'(\mathbf{v})$ represented the reconstructed features by feeding $\mathbf{v}$ into two linear layers as encoder and decoder. We then sum the loss over each modality:

$$L_{Reconstruct} = L_{v'} + L_{a'} + L_{t'} \tag{9}$$

4. Experiments

4.1. Implementation details

For the visual branch of the proposed MCN model we follow [33] and use pre-trained 2D features from a ResNet-152 model [23] trained on ImageNet [15] to extract features at the rate of one frame per second, along with pre-trained 3D features from a ResNeXt-101 model [20] trained on Kinetics [12] to obtain 1.5 features per second. The video clip features were computed by concatenating the 2D and 3D features into a 4096 dimension vector and max-pooling the features over time. For the audio branch of the network, we compute log-mel spectrograms and use a pre-trained DAVEnet model [21] to extract audio features. For the textual branch, the feature extraction process proposed in [33] is adopted to extract text representations: a GoogleNews pre-trained Word2vec model [34] provides word embeddings, followed by a max-pooling over words in a given sentence to extract a sentence embedding. Note that all backbones are fixed, and they are not fine-tuned during training. Each feature extraction branch is followed by a separate fully-connected layer and a gated unit for projecting the features in a common embedding space. To allow for pairwise comparisons, features from each of the different modalities are set to be 4096-dimensional vectors. More details can be found in the supplement.

4.2. Datasets

Training Dataset. Our models are trained on the HowTo100M [33] instructional video dataset, which contains 1.2M videos along with their corresponding audio that consists of speech and environmental sound and automatically generated speech transcriptions.

Downstream Datasets. The **YouCook2** [44] dataset contains 3.5K cooking instruction video clips with text descriptions collected from YouTube. Unlike Howto100m dataset, text descriptions in YouCook2 are human-annotated. The **MSR-VTT** [42] dataset contains 200K human annotated video clip-caption pairs on various topics. We use the same

Method	Mod	Model	TR	YouCook2 R@1	YouCook2 R@5	YouCook2 R@10	MSRVTT R@1	MSRVTT R@5	MSRVTT R@10
Random		-	-	0.03	0.15	0.3	0.01	0.05	0.1
Miech [33]	VT	R152+RX101	N	6.1	17.3	24.8	7.2	19.2	28.0
MDR [3]	VT	R152+RX101	N	-	-	-	8.0	21.3	29.3
MIL-NCE* [32]	VT	R152+RX101	N	8.1	23.3	32.3	8.4	23.2	32.4
MCN (ours)	VAT	R152+RX101	N	**18.1**	35.5	45.2	**10.5**	**25.2**	**33.8**
MDR [3]	VT	R152	N	-	-	-	8.4	22.0	30.4
ActBERT [45]	VT	R101+Res3D	N	9.6	26.7	38.0	8.6	23.4	33.1
SSB [35]	VT	R(2+1)D-34+R152	N	-	-	-	8.7	23.0	31.1
MMV FAC [1]	VAT	TSM-50x2	Y	11.7	33.4	45.4	9.3	23.0	31.1
MIL-NCE [32]	VT	I3D-G	Y	11.4	30.6	42.0	9.4	22.0	30.0
MIL-NCE [32]	VT	S3D-G	Y	15.1	**38.0**	**51.2**	9.9	24.0	32.4

Table 1: Comparison of text-to-video retrieval systems. Mod indicates modality used, where V: video, A: audio, T: text. TR indicates if a trainable backbone is used or not.

Method	Mod	Model	TR	CrossTask Recall	CrossTask IOD	CrossTask IOU	MYT Recall	MYT IOD	MYT IOU
CrossTask [46]	VT	R152+I3D	N	22.4	-	-	-	-	-
CrossTask [46]	VT	R152+I3D	N	31.6	-	-	-	-	-
Mining: GRU [26]	VT	TSN	N	-	-	-	-	14.5	7.8
Mining: MLP [26]	VT	TSN	N	-	-	-	-	19.2	9.8
Miech [33]	VT	R152+RX101	N	33.6	26.6	17.5	15.0	17.2	11.4
MIL-NCE* [32]	VT	R152+RX101	N	33.2	30.2	16.3	14.9	26.4	17.8
MCN (ours)	VAT	R152+RX101	N	35.1	**33.6**	**22.2**	**18.1**	**32.0**	**23.1**
ActBERT [45]	VT	R101+Res3D	N	37.1	-	-	-	-	-
ActBERT [45]	VT	+ Faster R-CNN	N	**41.4**	-	-	-	-	-
MIL-NCE [32]	VT	I3D-G	Y	36.4	-	-	-	-	-
MIL-NCE [32]	VT	S3D-G	Y	40.5	-	-	-	-	-

Table 2: Evaluation of temporal action localization systems.

test set with 1K video clip-caption pairs constructed in [33] in our experiments. The **CrossTask** [46] dataset contains 2.7K instructional videos that cover various topics.The action steps and their order for each task were collected from *wikiHow* articles with manual annotation for each frame. The **Mining Youtube** [26] dataset focuses on YouTube videos for five simple dishes.The test set contains 250 cooking videos, 50 of each task, that are densely annotated, *i.e*. each frame is labeled with its respective action class.

4.3. Downstream Tasks

To demonstrate the effectiveness of the proposed model, we evaluate embeddings derived from the network in two downstream tasks: text-to-video retrieval and temporal action localization. We focus on the zero-shot task because we want to access the quality of the cross-modal semantic embedding that was learned during training. When performing retrieval using our model, we compare the query text features with the video and audio features by computing similarity for both and using the average. For action localization, we compute the same distance of the video-audio pair of each frame to each respective label embedding and are so able to align video frames to each of the provided action steps.

Text-to-Video Retrieval. The goal of this task is to retrieve the matching video from a pool of videos, given its ground truth text query description. The model is tested on two video description datasets and evaluated on recall metrics: R@1, R@5, R@10. These evaluations are used to demonstrate the effectiveness of the contrastive loss and learned joint embedding space across three modalities.

Text-to-Full Video Retrieval. The conventional text-to-video retrieval task attempts to match a caption (or ground-truth text query) to a single video clip. Since a single caption can refer to many individual clips within a dataset, this task is limiting. To this end, we propose the task of *text-to-full video retrieval* where the goal is to match a set of captions (or text queries) describing multiple parts of a video to an entire video. This is a more realistic task than single clip retrieval since various real-world applications require retrieving entire videos from complex textual queries. We evaluate on YouCook2 dataset with recall metrics: R@1, R@5, R@10.

Temporal action localization. We further evaluate our model on two temporal action localization tasks. The CrossTask [46] dataset considers the task of clip level action detection. Here, an unordered set of action labels is given for a set of clips of the same video, and clips have to be classified with the respective action labels. The performance is reported as recall and computed as a ratio of the correctly predicted clips over the total number of clips in the video as used in [46]. The MiningYoutube [26] dataset considers the task of frame-level temporal action segmentation.Here, each test video is provided together with the respective actions and their ordering, including the background. The goal is to find the correct frame-wise segmentation of the video given the action order. We follow the inference procedure outlined in [26] to compute the alignment given our similarity input matrix. The dataset employs two evaluation metrics: intersection over detection (IoD) [9], defined as $\frac{G \cap D}{D}$: the ratio between the intersection of ground-truth action G and prediction D to prediction D, and the Jaccard index, which is an intersection over union (IoU) given as $\frac{G \cap D}{G \cup D}$.

4.4. Comparison with State-of-the-art Methods

Zero-shot Video Retrieval. We first examine the results of the text-to-video retrieval task on the YouCook2 and MSR-VTT datasets (Table 1). We compare only with baseline models that were not fine-tuned on the respective dataset for a fair comparison. To allow comparability between different approaches, we use a fixed visual feature extraction backbone as described in [33] whenever possible. For the baseline MIL-NCE* [32], we apply their training strategy on the same visual feature set we use, ResNet-152 (R152) and ResNeXt-101 (RX101) [33]. On YouCook2, our model significantly outperforms prior works on the same architecture and shows even competitive results compared to models with trainable visual backbone (TR). Our method also performs better than the other baselines on MSR-VTT. The gains are,

Method	CrossTask				
	NMI ↑	**ARI** ↑	**Acc.** ↑	⟨**H**⟩ ↓	⟨$\mathbf{p}_{\max}$⟩ ↑
Random	3.2	3.2	9.4	1.30	47.5
Miech *et al.* [33]	61.8	46.1	57.0	0.39	81.5
MIL-NCE* [32]	62.0	45.6	56.7	0.37	82.4
MCN (ours)	**65.5**	**48.5**	**57.6**	**0.34**	**83.8**

Table 3: Performance on clustering metrics on the CrossTask dataset evaluated by GT text annotations on video segments.

however, not as significant as on YouCook2. We attribute this to the fact that neither the available audio nor the textual description is instructional in nature and, therefore, semantically further away from our training set.

Zero-shot Action Localization. We examine the action localization tasks on the CrossTask and the MiningYouTube dataset in Table 2. For CrossTask, given each frame in the video, we perform a zero-shot classification of the given labels and calculate the recall. In this zero-shot setting, the model computes video text similarity to localize action step labels similar to [33]. Our method outperforms state-of-the-art approaches for self-supervised learning [32, 33] and a fully supervised approach [46] especially in the IOU and IOD metrics, which also consider false-positive predictions from the background class as an action step. Approaches in [33] and MIL-NCE* [32] are directly comparable with our method since they use the same feature extractor as us. In contrast, MIL-NCE [32] uses a stronger video backbone and [45] uses additional feature modalities such as region features along with a stronger language model. We also evaluate our model on the MiningYoutube [46] temporal action localization benchmark. Our method outperforms state-of-the-art approaches for both self-supervised [32, 33] and weakly supervised [26] learning. More settings, including data and computing resources for each model, are in the supplement.

Clustering Metrics. We further evaluate our system with respect to various clustering metrics as proposed by [5]. Results are shown in Table 3. The definition of each metric is included in the supplement. It shows that our learned multimodal features are closer to the ground-truth distribution and have higher purity within the cluster.

4.5. Full Video Retrieval

To address the problem of full video retrieval from a set of captions, we divide each video into a set of clips, which are compared with the queries. We evaluate three different methods: In **majority vote over clip** predictions, we obtain the top-k predictions of each clip/caption pair as votes and select the video which has the majority of votes. For **majority vote over videos**, the maximal prediction over all the clips of a video is taken for each caption to obtain video/caption pairs. Then, the top-k of these predictions are selected as votes, and the video with the most votes is predicted. Lastly, our **caption averaging** method involves obtaining the maximal prediction over all the clips of a video is taken for each caption and then averaging over the set of captions in a query. This gives a single prediction for the entire video.

Method	Prediction	R@1	R@5	R@10
Random	-	0.23	1.15	2.32
MCN (ours)	MV-Clip	38.8	67.4	76.8
MCN (ours)	MV-Video	38.8	67.7	78.4
MCN (ours)	Caption Avg.	**53.4**	**75.0**	81.4
Miech *et al.* [33]	Caption Avg.	43.1	68.6	79.1
MIL-NCE* [32]	Caption Avg.	46.6	74.3	**83.7**

Table 4: Comparison of Text-to-Full Video retrieval systems on the YouCook2 dataset. The prediction column denotes the method used to obtain video-level predictions: majority vote over clips (MV-Clip), majority vote over videos (MV-Video), and caption averaging (Caption Avg.).

Loss	YR10	MR10	CTR	MYT-IOU
NCE	39.2	33.5	33.9	21.5
MIL-NCE	40.0	33.0	33.7	21.1
MMS	43.7	32.9	34.3	22.1
MMS + Cluster	44.3	33.7	34.5	22.6
MMS + Cluster + Reconstruct	45.2	33.8	35.1	23.1

Table 5: Ablation study on different loss including the selection of contrastive learning loss, the additional clustering, and reconstruction loss.

We examine the results of the text-to-full video retrieval task on the YouCook2 dataset (Table 4). Of the three methods to obtain full video predictions, the caption averaging achieves better results than both majority voting schemes. Furthermore, we find that our method outperforms prior works on this task with a 6.8% improvement on R@1. Since we obtain full video predictions, we also perform full-video classification on the CrossTask dataset using the set of subtask labels as the set of query captions, where we achieve a top-1 accuracy of 68.7%.

4.6. Ablation Studies

To better understand the contributions of various algorithmic design choices used to build the proposed MCN model, we perform a set of ablation studies on the following downstream tasks: YouCook2 R@10 (YR10), MSR-VTT R@10 (MR10), CrossTask average recall (CTR) and MiningYoutube IOU (MY-IOU). For each setting, we use the same feature extractor for three modalities as described in Sec 4.1 for a fair comparison. More ablations are in the supplement.

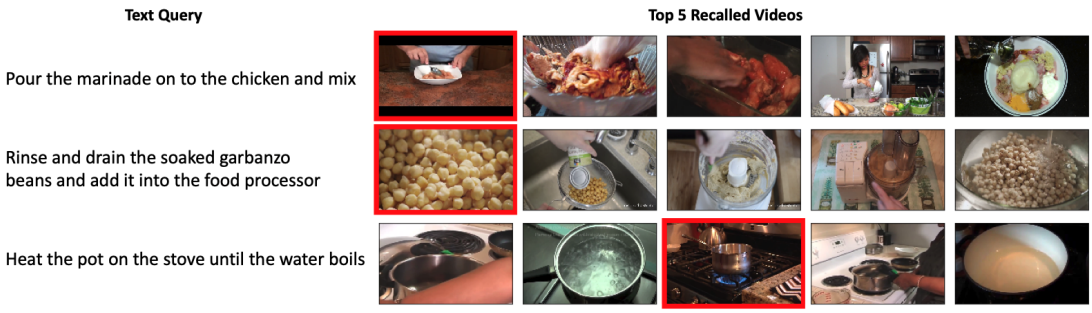
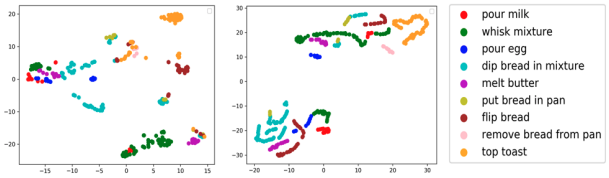


Figure 4: Qualitative results for the text-to-video retrieval task on YouCook2. Top-ranked clips show a high similarity to the described task as well as among each other without being too visually similar.



(a) Visual Feature (b) Multimodal Feature (c) Action Steps

Figure 5: t-SNE visualizations on the CrossTask dataset for the task of "Make French Toast". Best viewed in color.

Method	Target	Labels	YR10	MR10	CTR	MYT-IOU
Sinkhorn	Swap	hard	39.0	33.4	33.6	21.1
Sinkhorn	Swap	soft	41.8	33.9	34.5	22.1
Sinkhorn	Joint	hard	44.4	33.4	34.6	21.1
Sinkhorn	Joint	soft	43.6	32.4	34.1	21.6
K-means	Swap	hard	41.3	32.8	33.2	21.0
K-means	Joint	hard	44.3	33.1	34.6	21.4
K-means	Centroid	hard	45.2	33.8	35.1	23.1

Table 6: Ablation study on different clustering pipelines with various methods, loss prediction target, and label types.

Selection on different losses. In our first set of experiments, we find the proposed clustering is crucial not only for clustering-related tasks but also for retrieval (MSR-VTT) tasks as shown in Table 5. This validates our hypothesis that semantically close instances should be clustered closely in the joint embedding space. Also, the selection of contrastive loss (MMS) shows better results in our model.

Different choices of clustering methods. We evaluate the performance of (1) Selection of different clustering methods such as Sinkhorn clustering [6] and K-means [4]. (2) Different prediction targets such as using swap prediction, which uses the pseudo label of other modalities for prediction target as [11, 2]. Or using the mean feature pseudo label as a joint prediction for three modalities. Also, using the centroid of the cluster as the target. (3) Different prediction labels, including hard labels (one-hot) or soft labels (continuous). Detailed descriptions are included in the supplement. As shown in Table 6, our method encourages each modality feature to move closer to the semantic centroid, which improves performance by explicitly encouraging semantically close features from different domains to cluster together.

4.7. Qualitative Analysis

We perform a qualitative analysis with the model's ability to do zero-shot text-to-video retrieval shown in Figure 4. Given an open-vocabulary caption, our model can retrieve the correct corresponding video segment. We also visualize the efficacy of using multimodal embeddings (concatenated video and audio representations) over using only visual embeddings. Representations from the CrossTask dataset are visualized using t-SNE plots. We observe that with multimodal features as Figure 5 (b), semantically related instances (based on ground truth classes) tend to be more tightly related than uni-modal visual features trained from contrastive loss (a) that appear more spread out. Also, multimodal features are clearly more separable for different actions.

5. Conclusions

We have developed a novel self-supervised multimodal clustering network that learns a common embedding space by processing local (via a contrastive loss) and global (via a clustering loss) semantic relationships present in multimodal data. The multimodal clustering network is trained on a large corpus of narrated videos without any manual annotations. Our extensive experiments on multiple datasets show that creating a joint video-audio-language embedding space with a clustering loss is essential for self-supervised learning of good video representations. Our approach can be extended to more modalities such as optical flow or sentiment features and applied to other multimodal datasets for learning joint representation spaces without human annotation.

Acknowledgments: We thank IBM for the donation to MIT of the Satori GPU cluster. This work is supported by IARPA via DOI/IBC contract number D17PC00341. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

[1] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. In *NeurIPS*, 2020. 1, 2, 6

[2] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *NeurIPS*, 2020. 1, 2, 3, 8

[3] Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. In *AAAI*, 2021. 2, 6

[4] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, 2006. 8

[5] Yuki Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. In *NeurIPS*, 2020. 2, 3, 7

[6] Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. Self-labelling via simultaneous clustering and representation learning. In *ICLR*, 2020. 2, 3, 8

[7] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. In *arXiv preprint arXiv:1706.00932*, 2017. 1

[8] Angie Boggust, Kartik Audhkhasi, Dhiraj Joshi, David Harwath, Samuel Thomas, Rogerio Feris, Dan Gutfreund, Yang Zhang, Antonio Torralba, Michael Picheny, et al. Grounding spoken words in unlabeled video. In *CVPRW*, 2019. 2

[9] Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Weakly supervised action labeling in videos under ordering constraints. In *ECCV*, 2014. 6

[10] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *ECCV*, 2018. 3, 5

[11] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020. 2, 3, 8

[12] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2, 5

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3

[14] Ting Chen and Lala Li. Intriguing properties of contrastive losses. In *arXiv preprint arXiv:2011.02803*, 2020. 5

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[16] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE, 2021. 2

[17] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020. 2

[18] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, 2010. 2

[19] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *CVPR*, 2006. 2

[20] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, 2018. 5

[21] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *ECCV*, 2018. 1, 2, 5

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[24] Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. Large-scale representation learning from visually grounded untranscribed speech. In *CoNLL)*, 2019. 4

[25] Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all. In *arXiv preprint arXiv:1706.05137*, 2017. 1

[26] Hilde Kuehne, Ahsan Iqbal, Alexander Richard, and Juergen Gall. Mining youtube-a dataset for learning fine-grained action concepts from webly supervised video data. In *CVPR*, 2019. 2, 6, 7

[27] Lei Le et al. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. In *NeurIPS*, 2018. 5

[28] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, 2021. 2

[29] Junnan Li, Pan Zhou, Caiming Xiong, Richard Socher, and Steven CH Hoi. Prototypical contrastive learning of unsupervised representations. In *ICLR*, 2021. 2

[30] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, 2019. 2

[31] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. Univilm: A unified video and language pre-training model for multimodal understanding and generation. In *arXiv preprint arXiv:2002.06353*, 2020. 2

[32] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020. 2, 3, 6, 7

[33] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 2, 5, 6, 7

[34] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *arXiv preprint arXiv:1301.3781*, 2013. 5

[35] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, João Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *ICLR*, 2021. 2, 6

[36] AJ Piergiovanni, Anelia Angelova, and Michael S Ryoo. Evolving losses for unsupervised video representation learning. In *CVPR*, 2020. 1

[37] Andrew Rouditchenko, Angie Boggust, David Harwath, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, et al. Avlnet: Learning audio-visual language representations from instructional videos. In *Interspeech*, 2021. 2

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. In *IJCV*, 2015. 2

[39] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS, 2018. 2

[40] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *ICCV*, 2019. 2

[41] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. Scan: Learning to classify images without labels. In *ECCV*, 2020. 3

[42] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, 2016. 2, 5

[43] Xueting Yan, Ishan Misra, Abhinav Gupta, Deepti Ghadiyaram, and Dhruv Mahajan. Clusterfit: Improving generalization of visual representations. In *CVPR*, 2020. 3

[44] Luowei Zhou, Xu Chenliang, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 2, 5

[45] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, pages 8746–8755, 2020. 2, 6, 7

[46] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *CVPR*, 2019. 2, 6, 7