

Unified Multi-modal Unsupervised Representation Learning for Skeleton-based Action Understanding

Shengkai Sun*
Zhejiang Gongshang University

Daizong Liu*
Peking University

Jianfeng Dong†
Zhejiang Gongshang University
Zhejiang Key Lab of E-Commerce

Xiaoye Qu
Huazhong University of Science and
Technology

Junyu Gao
Institute of Automation, Chinese
Academy of Sciences

Xun Yang
University of Science and Technology
of China

Xun Wang
Zhejiang Gongshang University
Zhejiang Key Lab of E-Commerce

Meng Wang
Hefei University of Technology

ABSTRACT

Unsupervised pre-training has shown great success in skeleton-based action understanding recently. Existing works typically train separate modality-specific models (*i.e.*, joint, bone, and motion), then integrate the multi-modal information for action understanding by a late-fusion strategy. Although these approaches have achieved significant performance, they suffer from the complex yet redundant multi-stream model designs, each of which is also limited to the fixed input skeleton modality. To alleviate these issues, in this paper, we propose a Unified Multimodal Unsupervised Representation Learning framework, called *UmURL*, which exploits an efficient early-fusion strategy to jointly encode the multi-modal features in a single-stream manner. Specifically, instead of designing separate modality-specific optimization processes for uni-modal unsupervised learning, we feed different modality inputs into the same stream with an early-fusion strategy to learn their multi-modal features for reducing model complexity. To ensure that the fused multi-modal features do not exhibit modality bias, *i.e.*, being dominated by a certain modality input, we further propose both intra- and inter-modal consistency learning to guarantee that the multi-modal features contain the complete semantics of each modal via feature decomposition and distinct alignment. In this manner, our framework is able to learn the unified representations of uni-modal or multi-modal skeleton input, which is flexible to different kinds of modality input for robust action understanding in practical cases. Extensive experiments conducted on three large-scale datasets, *i.e.*, NTU-60, NTU-120, and PKU-MMD II, demonstrate that *UmURL* is highly efficient, possessing the approximate complexity

with the uni-modal methods, while achieving new state-of-the-art performance across various downstream task scenarios in skeleton-based action representation learning. Our source code is available at <https://github.com/HuiGuanLab/UmURL>.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision representations**; **Activity recognition and understanding**.

KEYWORDS

Multi-modal Learning, Unsupervised Representation Learning, Action Understanding

ACM Reference Format:

Shengkai Sun, Daizong Liu, Jianfeng Dong, Xiaoye Qu, Junyu Gao, Xun Yang, Xun Wang, and Meng Wang. 2023. Unified Multi-modal Unsupervised Representation Learning for Skeleton-based Action Understanding. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, October 29–November 3, 2023, Ottawa, ON, Canada. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3581783.3612449>

1 INTRODUCTION

Human action understanding [9, 12, 19, 27, 28, 34, 44, 46, 50] is one of the fundamental and important tasks within the realm of multimedia, which demonstrates extensive applicability across diverse domains, including human-computer interaction [17, 21–24, 31, 53, 54, 56, 62], intelligent surveillance, and sports analysis, etc. Recently, skeleton-based action understanding [6, 36, 51, 59] that represents the human major joints with 3D coordinates has garnered considerable research interest, on account of its lightweight, appearance-robust, and privacy-preserving advantages in comparison to RGB videos [15, 39]. Despite their achieved impressive performance, these approaches rely on a large amount of labeled training data that are time-consuming and arduous to acquire. To address this limitation, unsupervised representation learning [45, 55, 60, 61] from unlabeled data has been introduced into the skeleton-based action understanding task.

Early unsupervised learning attempts for skeleton-based action understanding were primarily focused on devising pretext tasks for generating the supervision signals, such as skeleton reconstruction [16, 61], motion prediction [5], and skeleton colorization [52].

*Co-first authors.

†Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0108-5/23/10...\$15.00
<https://doi.org/10.1145/3581783.3612449>

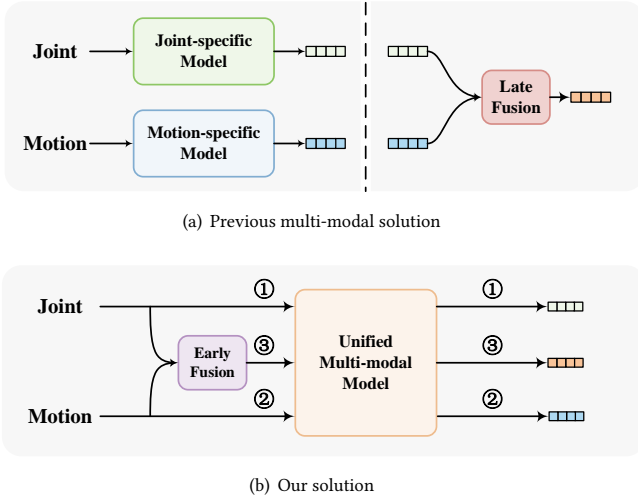


Figure 1: (a) In the context of unsupervised skeleton-based action understanding, previous methods require numerous modality-specific models with a late fusion strategy for multi-modal comprehension. (b) Different from them, our model supports inputs from multiple modalities in a single yet unified multi-modal model, reducing the model complexity.

Due to the complex pipeline of hand-craft pretext tasks and limited performance, these methods have been out of fashion gradually. Recent unsupervised approaches [18, 29, 45, 58] tend to employ advanced contrastive learning techniques [2, 3, 7, 11, 49, 64], and achieve strong generalization capabilities to varying downstream tasks. Although great efforts have been devoted to skeleton-based action understanding, existing methods are typically designed for a specific modality of skeletons. As skeletons can be readily represented as multiple modalities, such as joint, motion, and bone, uni-modal methods based on a specific modality are suboptimal. One simple strategy for extending uni-modal methods to multi-modal ones is late fusion [10, 63], as illustrated in Figure 1(a). Given multiple pre-trained modal-specific models, their prediction results are ensemble via late fusion. Despite these methods achieving strong performance, they still suffer from two indispensable problems: (1) Complicated and redundant design. They require training separate models for encoding each modality, leading to a significant increase in computational overheads on pre-training and downstream tasks. (2) Inflexible inference. Since their modality-aware features are dis-unified (that is, different modalities are separately encoded from fixed models), they require to prepare appropriate models for matching the input modalities in the inference stage.

Considering the above issues, we propose to learn a unified multi-modal representation by jointly learning features of uni-modal and multi-modal inputs, as shown in Figure 1(b). Such a single-stream encoder significantly reduces the model complexity of previous unimodal-ensemble frameworks. Moreover, this unified representation learner is flexible to the input formats of different modalities, and is able to effectively produce representative features via a modality-agnostic encoder. It is worth noting that for the multi-modal input learning, one can directly utilize an early fusion strategy [38, 47] before the feature encoding. Unfortunately, relying

solely on this straightforward modification may not be appropriate for unsupervised learning due to partial feature domination, which could potentially lead to performance degradation. We attribute the cause to the gap between pre-training and downstream objectives. That is, there will be such a suboptimal scenario that a certain modality is easier to learn according to the unsupervised objective during pre-training, but it does not possess informative enough features for downstream tasks, which eventually leads to the model being biased towards one modality and does not adequately exploit other available modal information.

To this end, in this paper, we propose a novel Unified Multi-modal Unsupervised Representation Learning (UmURL) framework, which efficiently encodes unified uni-modal or multi-modal features through a modality-agnostic single-stream for skeleton-based action understanding. Specifically, we build a simple yet effective early-fusion pipeline that exploits single-stream to handle the multi-modal inputs. To guarantee that the extracted multi-modal features contain the complete semantics of each modal, we further decompose the features into each uni-modal domain for both intra- and inter-modal semantic consistency learning. In this way, the learned multi-modal representations are unified with its contained individual modality features, sharing the same intra-modal semantics while complementing inter-modal contexts for robust action recognition. Thanks to this unified multi-modal representation, our framework is also flexible to different kinds of modality inputs.

In summary, this paper makes the following contributions:

- We propose a novel and practical multi-modal unsupervised framework, *i.e.*, UmURL, for skeleton-based action understanding. This framework learns the unified representations of uni-modal or multi-modal skeleton inputs, which is efficient and flexible to different kinds of modality input for robust action understanding.
- To effectively realize the above proposal, we propose to guarantee that the unified multi-modal representations contain the complete semantics of its individual modality features. In particular, we decompose the unified representations into each uni-modal domain for both intra- and inter-modal semantic consistency learning.
- Extensive experiments on three datasets verify the effectiveness and transferability of our proposed framework. With a much more efficient multi-modal network than previous multi-modal solutions, we achieve new state-of-the-art performance in multiple downstream tasks.

2 RELATED WORK

2.1 Uni-modal USARL Methods

Uni-modal USARL methods typically utilize a specific modality (*e.g.*, joint) as input, and focus on designing pretext tasks suitable for skeleton data. In early works, skeleton reconstruction is a prevalent pretext task [16, 30, 40, 61]. For instance, Zhang *et al.* [61] and Kundu *et al.* [16] respectively reconstruct the skeleton from the latent features of original and corrupted skeletons. In [30], Nie *et al.* disentangle the skeleton into pose-dependent and view-dependent features, and then reconstruct the skeleton from the disentangled features. Additionally, numerous innovative pretext tasks have been also introduced. Su *et al.* [42] propose to colorize

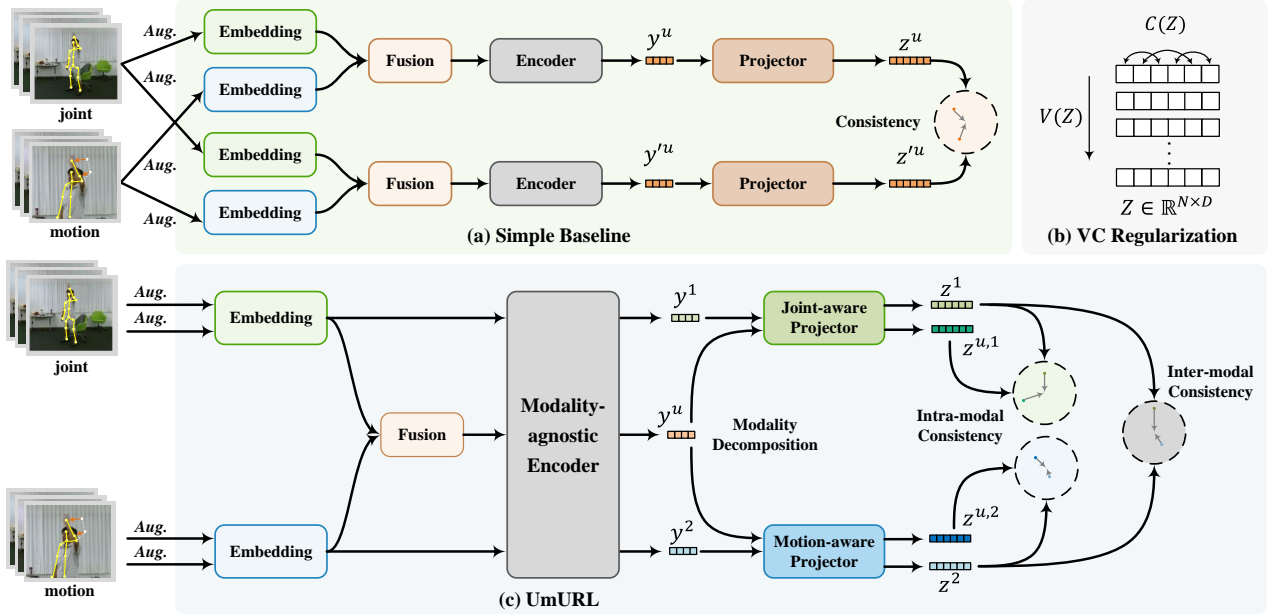


Figure 2: (a) A simple multi-modal baseline, which utilizes early fusion to learn the multi-modal features in a single stream. (b) VC regularization is separately applied to all projected features to prevent model collapse. (c) To learn unified representations, we train a modality-agnostic encoder by aligning both intra-modal and inter-modal consistent semantics. Note that the encoder in the simple baseline and the modality-agnostic encoder in UmURL are of the same structure, and their different names are due to their different roles in the corresponding framework.

each joint of human skeletons based on their temporal and spatial orders, and adopt the skeleton colorization prediction as the pretext task. In [13], a multi-interval pose displacement prediction pretext task is proposed for unsupervised learning.

Recently, we observe an increasing use of contrastive learning [11] as the pretext task due to its simple mechanism and promising performance [8, 32, 45]. The key idea of these contrastive methods is to learn the skeleton representations that are invariant to transformations. Typically, they utilize data augmentation to generate multiple views of the input skeleton sequences and subsequently train an encoder to minimize the distance between positive pairs (*i.e.*, views of the same skeleton sequence) and simultaneously maximize the distance between negative pairs (*i.e.*, views of different skeleton sequence) in the feature space. Rao *et al.* [32] is the first to migrate contrastive learning from image representation learning to unsupervised skeleton-based action representation learning. Since then, a number of uni-modal works have concentrated on enhancing contrastive learning in the context of skeleton-based action understanding [18, 41, 45, 57, 63]. Some works [4, 20, 42] combine contrastive learning with other pretext tasks to learn discriminative skeleton representations. [10, 58, 63] investigate data augmentation strategies for skeleton data. Dong *et al.* [8] encode the skeleton action as multiple representations and then performed hierarchical contrast to generate more supervision. Su *et al.* [41] propose to represent skeleton sequences in a probabilistic embedding space. [18, 29, 57] exploit positive mining and knowledge exchange to alleviate irrational negative samples problem in contrastive learning.

2.2 Multi-modal USARL Methods

As skeletons can be represented as multiple modalities, such as joint, motion, and bone, jointly utilizing multiple modalities for representation are usually beneficial. One *de facto* multi-modal solution is to first train multiple uni-modal models for all modalities, and subsequently fuse them via late fusion. Almost all existing works adopt such a solution to extend uni-modal methods to multi-modal ones [10, 18, 29, 57, 63]. However, this solution is of high computation complexity due to the fact that multiple uni-modal models should be pre-trained and then utilized via late fusion for downstream tasks. Our proposed method belongs to the multi-modal USARL method. Instead of using a cumbersome late fusion strategy, this work proposes an efficient multi-modal representation learning framework. It fuses various modalities by early fusion and obtains a unified representation at a lower cost while retaining uni-modal encoding capability.

3 METHOD

In this section, we first introduce a simple yet powerful multi-modal baseline model by extracting multi-modal features that thoroughly integrate information across all modalities via an early fusion. Different from previous heavy late fusion methods, this baseline is able to reduce the computational load associated with independent uni-modal optimization and subsequent late decision fusion. Then, to ensure that the extracted multi-modal features do not exhibit modality bias, *i.e.*, being dominated by a certain modality, we further extend the baseline model with unified representation learning. Specifically, considering that a well-learned multi-modal feature

should contain the complete semantics of each contained modality, we decompose the extracted multi-modal feature into separate uni-modal domains for distinct intra- and inter-modal semantic consistency learning. By jointly learning uni-modal and multi-modal unified representations, our framework is robust and flexible to different kinds of modality inputs, achieving better performance. In the following, we elaborate the simple multi-modal baseline and our unified multi-modal unsupervised representation learning.

3.1 Simple Multi-modal Baseline

As shown in Figure 2(a), the simple baseline extracts the multi-modal features with an efficient early-fusion strategy. Unlike previous works that separately train different backbones for individual modality feature encoding and then interaction, our baseline solely utilizes single-stream models for efficient multi-modal representation learning.

Multi-modal Input. Generally, an input skeleton sequence is represented as $x \in \mathbb{R}^{T \times C \times V}$, where T , C , and V denote the number of frames, channels, and joints. Other skeleton modalities like bone and motion information can be additionally extracted through the linear transformation over raw 3D coordinates [35, 36], to provide complementary spatio-temporal information to the original joint modality. Based on this, an input multi-modal action can be formally represented as $x^m = \{x^1, x^2, \dots, x^k\}$, which contains information from k different modalities. Similar to prevalent unsupervised methods, our baseline is also designed to learn feature representations that are invariant to data transformations without manually annotated labels. To achieve this, we generate augmented views of the corresponding modality by applying augmentations.

Modality-specific Embedding. Before fusing the multi-modal inputs, we first map each heterogeneous modality data into the embedding space of the same dimension. Concretely, given the augmented input data of modality m , we first flatten it with the temporal dimension kept, and then employ a modality-specific embedding module (MSEM) to embed the input into a space of dimension D_h . After employing MSEM to all modalities, we obtain $h^m \in \mathbb{R}^{T \times D_h}$ by:

$$h^m = \text{MSEM}_m(t(x^m)), \quad m \in \{1, 2, \dots, k\}, \quad (1)$$

where MSEM_m indicates the corresponding modality-specific embedding module which is implemented by a multi-layer perception, and t denotes a random augmentation operation.

Multi-modal Fusion and Encoding. After embedding all inputs of different modalities into the uniform representations, we fuse them at the early stage via a simple averaging operation followed by a linear transformation. Formally, given k modalities, the fused representation is obtained as:

$$h^u = \text{Linear}\left(\frac{1}{k} \sum_{m=1}^k h^m\right), \quad m \in \{1, 2, \dots, k\}, \quad (2)$$

where $\text{Linear}(\cdot)$ is a learnable linear transformation. As multi-modal fusion is not the focus of this work, we employ this fusion for simplicity but it can be replaced by more advanced fusion ways.

To obtain the final multi-modal representation, a multi-modal encoder is further employed over the fused representation. Formally,

the final multi-modal representation $y^u \in \mathbb{R}^{D_h}$ is obtained as:

$$y^u = \text{Encoder}(h^u), \quad (3)$$

where Encoder can be a sequence encoder layer, such as Transformer Layer. This mechanism of extracting multi-modal features after early fusing modality-specific embeddings not only preserves the unique semantics of each modality to a certain extent, but also reduces model complexity compared to adopting a fully independent encoding structure for all modalities.

Unsupervised Learning. To implement the baseline under the unsupervised setting, a straightforward idea is using contrastive learning that is commonly adopted in the existing skeleton representation works [8, 32, 45]. However, contrastive learning tends to be costly, requiring large batch sizes or memory banks [3, 11]. Instead of using contrastive learning, we utilize an information maximization method VICREG proposed in [1] considering its high computation efficiency and promising performance. VICREG mainly consists of semantic-consistent regularization and Variance-Covariance (VC) regularization.

Semantic-consistent Regularization. Suppose the feature of additional view of the multi-modal input is $y^{u'}$, we employ a projection head g_u to map the features of different augmented multi-modal features to the same space, obtaining $z^u = g_u(y^u)$, $z^{u'} = g_u(y^{u'})$. By processing the data in batches of size N , we obtain the projected feature as $Z^u \in \mathbb{R}^{N \times D}$ and $Z^{u'} \in \mathbb{R}^{N \times D}$. The semantic-consistent regularization is employed to encourage the semantic consistency between two views of multi-modal input. To this end, we minimize the mean square error (MSE) loss over the projected features, and the loss is defined as:

$$\mathcal{L}_{\text{consistency}}(Z^u, Z^{u'}) = \text{MSE}(Z^u, Z^{u'}) = \frac{1}{N} \sum_{i=1}^N \|z_i^u - z_i^{u'}\|^2. \quad (4)$$

where $z_i^u, z_i^{u'}$ are the i -th vector in Z^u and $Z^{u'}$.

VC Regularization. To prevent model collapse, a VC regularization consisting of a variance term and a covariance term is further introduced. Given a batch of embeddings $Z \in \mathbb{R}^{N \times D}$, the variance term forces the embedding vectors of samples within a batch to be different. It is implemented by maintaining the variance of each embedding dimension above a threshold, which is defined as:

$$V(Z) = \frac{1}{D} \sum_{j=1}^D \max(0, \gamma - \sqrt{\text{Var}(Z_{:,j})} + \epsilon), \quad (5)$$

where γ is the variance threshold, ϵ is a small scalar preventing numerical instabilities, and $\text{Var}(Z_{:,j})$ indicates the variance of j -th embedding dimension vector $Z_{:,j}$. Additionally, the covariance term is designed to decorrelate the variables of each embedding, ensuring that each feature dimension encodes different information by:

$$C(Z) = \frac{1}{D} \sum_{i \neq j} [\text{Cov}(Z)]_{i,j}^2, \quad (6)$$

where $\text{Cov}(Z)$ is the auto-covariance matrix of Z . By combining the above two kinds of terms, as shown in Figure 2(b), the final VC regularization loss can be formulated as:

$$\mathcal{L}_{VC}(Z) = \mu V(Z) + C(Z), \quad (7)$$

where μ is a hyper-parameter to balance two terms.

The VC regularization is separately applied to both projected features Z^u and $Z^{u'}$ of two views of the input. Finally, the total loss function of the baseline for unsupervised learning is as follows:

$$\mathcal{L} = \lambda \mathcal{L}_{consistent}(Z^u, Z^{u'}) + \mathcal{L}_{VC}(Z^u) + \mathcal{L}_{VC}(Z^{u'}) \quad (8)$$

where λ is a hyper-parameter coefficient.

3.2 Unified Multi-modal Unsupervised Representation Learning

Although the above simple baseline incorporates the semantics of multi-modal input, it still suffers from the underlying modality bias issues, *i.e.*, the learned multi-modal features may be dominated by a certain modality during the pre-training process (validated in Section 4.3), leading to a worse multi-modal representation compared to independent training and then fusion. To alleviate this issue, we propose to learn the multi-modal representation that contains the complete semantics of every modality-specific input. Our hypothesis is that a good multi-modal representation should contain comprehensive information of the input modalities. Concretely, we propose a novel UmURL model as illustrated in Figure 2(c). Note that the pipeline of obtaining the multi-modal representation y^u in UmURL is the same as that in the baseline, and the main difference between the two methods is the way of representation learning. In our UmURL, we first learn to decompose the multi-modal features into different modality domains. Then, by extracting the original uni-modal features as guidance via the same modality-agnostic encoder, we introduce two consistency losses to guarantee the intra-modal semantic as same as possible while aligning the inter-modal semantic for representation learning.

Decomposing Multi-modal Features. In order to decompose the multi-modal representation into different modality domains for mining the independent semantics of each modality, we utilize k modality-aware projectors that are expected to extract modality-specific patterns. Formally, given the multi-modal representation y^u , the decomposed modality-specific features are obtained as:

$$z^{u,m} = g_m(y^u), \quad m \in \{1, 2, \dots, k\}. \quad (9)$$

where g_m is the modality-aware projector for modality m , which is implemented by a multi-layer perception.

Extracting Original Uni-modal Features. To constrain the decomposed feature learning, we also extract the original modality-specific features as guidance. Different from previous works [18, 29] that typically utilize modality-specific encoder, here we develop a modality-agnostic encoder to extract the original modality-specific features for all modalities. Note that the modality-agnostic encoder is the same encoder for multi-modal representation. Such a design allows our model flexible to different kinds of modalities during inference. Formally, given a skeleton sequence of modality m , its original modality-specific representation is obtained as:

$$y^m = \text{Encoder}(\text{MSEM}_m(t(x^m))), \quad m \in \{1, 2, \dots, k\}. \quad (10)$$

where *Encoder* denotes the modality-agnostic encoder, t is the random augmentation operation. Subsequently, to make the original and composed modality-agnostic feature comparable in the same space, modality-aware projectors g_m are also utilized, obtaining the projected original uni-modal feature as $z^m = g_m(y^m)$. The corresponding batch of these features are denoted as $Z^m \in \mathbb{R}^{N \times D}$.

Learning Unified Representations. To make the decomposed multi-modal representation semantic-consistent with the original features of individual modalities, we aim to learn uni-modal and multi-modal unified representations in an unsupervised manner. To achieve this goal, we propose a **intra-modal consistency learning** to encourage the decomposed modality-specific features and the original uni-modal features consistent. A **inter-modal consistency learning** is further introduced to learn more representative uni-modal features which in turn provides better constraints for intra-modal consistency learning.

Intra-modal Consistency Learning. As for intra-modal consistency learning, we force the decomposed modality-agnostic features to share the same semantics as the corresponding uni-modal features by adding a regularization that penalizes inconsistency between decomposed features and uni-modal features. Here, we use MSE regularization defined in Eq. 4, and employ it on each modality. The final loss is the summation of the MSE over all modalities:

$$\mathcal{L}_{intra} = \sum_{m=1}^k \text{MSE}(Z^{u,m}, Z^m). \quad (11)$$

Inter-modal Consistency Learning. The baseline model in Sec.3.1 severely relies on joint multi-modal augmentation within two identical streams for representation learning. However, this process not only suffers from the coarse alignment between complex multi-modal features, but also fails to explore the complementary contexts between the cross-modal features. To this end, we reformulate such joint multi-modal contrastive process into a detailed cross-modal one, which aligns more fine-grained semantics between different uni-modal features for better capturing their action-specific consistency and enhancing the action-aware representative features. Therefore, given the uni-modal features Z^i, Z^j of different modalities, we also utilize the MSE loss to minimize the pairwise distance between different modalities of the same skeleton. The constraint is employed between any two modalities, and the corresponding loss is defined as:

$$\mathcal{L}_{inter} = \sum_{i \neq j} \text{MSE}(Z^i, Z^j), \quad i, j \in \{1, 2, \dots, k\}. \quad (12)$$

Overall Optimization Losses. In this manner, we are able to generate unified uni-modal or multi-modal features sharing the representative information for downstream tasks. In addition to the above two distinct consistency losses, we also employ the VC regularization like Eq. 7 to prevent the model collapse for uni-modal and decomposed individual feature learning as:

$$\mathcal{L}_{reg} = \sum_{m=1}^k \mathcal{L}_{VC}(Z^m) + \mathcal{L}_{VC}(Z^{u,m}), \quad m \in \{1, 2, \dots, k\}. \quad (13)$$

Overall, the total learning objectives of the model are as follows:

$$\mathcal{L} = \lambda(\mathcal{L}_{intra} + \mathcal{L}_{inter}) + \mathcal{L}_{reg} \quad (14)$$

where λ denotes the hyper-parameter coefficient.

Table 1: Comparisons to the state-of-the-art methods for skeleton-based action recognition downstream task on NTU-60, NTU-120 and PKU-MMD II. Our proposed UmURL achieves the best balance between model performance and computational complexity. J: Joint, M: Motion, B: Bone.

Method	Publication	Modality	FLOPs/G	NTU-60		NTU-120		PKU-MMD II
				x-sub	x-view	x-sub	x-setup	x-sub
ISC [45]	ACM MM'21	J	5.76	76.3	85.2	67.1	67.9	36.0
AimCLR [10]	AAAI'22	J	1.15	74.3	79.7	63.4	63.4	-
PTSL [63]	AAAI'23	J	1.15	77.3	81.8	66.2	67.7	49.3
CrosSCLR [18]	CVPR'21	J	5.76	77.3	85.1	67.1	68.6	41.9
GL-Transformer [13]	ECCV'22	J	118.62	76.3	83.8	66.0	68.7	-
CPM [57]	ECCV'22	J	2.22	78.7	84.9	68.7	69.6	48.3
CMD [29]	ECCV'22	J	5.76	79.8	86.9	70.3	71.5	43.0
<i>UmURL</i>	This work	J	1.74	82.3	89.8	73.5	74.3	52.1
3s-HiCLR [58]	AAAI'23	J+M+B	7.08	78.8	83.1	67.3	69.9	-
3s-AimCLR [10]	AAAI'22	J+M+B	3.45	78.9	83.8	68.2	68.8	39.5
3s-PSTL [63]	AAAI'23	J+M+B	3.45	79.1	83.8	69.2	70.3	52.3
3s-CrosSCLR [18]	CVPR'21	J+M+B	17.28	82.1	89.2	71.6	73.4	51.0
3s-CPM [57]	ECCV'22	J+M+B	6.66	83.2	87.0	73.0	74.0	51.5
3s-CMD [29]	ECCV'22	J+M+B	17.28	84.1	90.9	74.7	76.1	52.6
<i>UmURL</i>	This work	J+M+B	2.54	84.2	90.9	75.2	76.3	54.0
<i>3s-UmURL</i>	This work	J+M+B	5.22	84.4	91.4	75.9	77.2	54.3

4 EXPERIMENTS

4.1 Experimental setup

4.1.1 Datasets. Following the previous works [18, 29, 45], we evaluate our method on three skeleton-based action datasets, *i.e.*, NTU-60 [33], NTU-120 [25], and PKU-MMD II [26].

4.1.2 Performance Metric. Following the previous works [29, 45], we adopt the top-1 accuracy as the performance metric for all downstream tasks.

4.2 Comparison to the State-of-the-art

In this section, we compare our approach with state-of-the-art methods in the context of two downstream tasks: skeleton-based action recognition and skeleton-based action retrieval. It is worth noting that after our model has been trained, it can be selectively employed using a single modality or multiple modalities during inference for downstream tasks. By contrast, previous works should train multiple models using a specific modality if multiple modalities are used for downstream tasks.

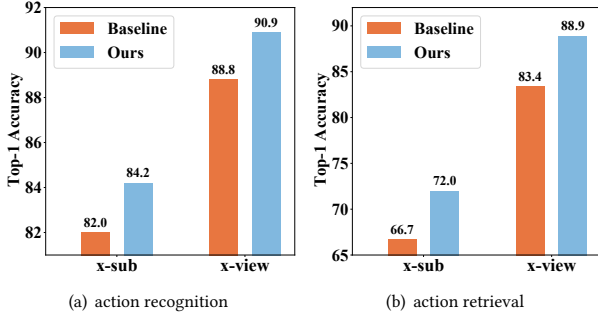
Skeleton-based Action Recognition. Following the standard practice from previous works [18, 29, 45], we train a linear classifier on top of the frozen encoder pre-trained with our proposed method. Table 1 summarizes the results on NTU-60, NTU-120, and PKU-MMD II datasets, where the results are split into two groups according to the modality used during the inference. Besides the model performance, for each model we also compute the computational complexity in terms of FLOPs it takes to encode given a skeleton sequence.

In the first group of using the joint modality during inference, our proposed method outperforms all competitors with significant margins. We attribute it to the fact that our model utilizes multiple modalities during training, which helps one modality absorb information from other modalities. Among the competitors, CrosSCLR [18] and CMD [29] also utilize multiple modalities during training, but our model performs better and shows much lower FLOPs. In the second group, all models utilize the joint, motion, and bone modalities for training and inference. In this scenario, the competitors first train three models using a specific modality individually, and then fuse the results from the three models. Comparing the results in the first group, all the compared methods achieve clear performance gains but at the cost of higher computational complexity (The computational complexities of multi-modal variants are three times higher than the uni-modal counterparts). By contrast, our proposed method with a unified multi-modal representation learning framework has the best balance between model performance and computational complexity. Additionally, we also report the results of our model using three-stream networks by late fusion, our model achieves further performance gain.

Skeleton-based Action Retrieval. In this experiment, the action representations obtained by pre-training unsupervised models are directly employed for retrieval without fine-tuning. Given an action query, the nearest neighbor in the representation space is retrieved using cosine similarity. Table 2 shows a comparison of various methods on the NTU-60 and NTU-120 datasets. With the joint modality as input for inference, our proposed UmURL performs better than the previous works. Moreover, our method achieves a

Table 2: Comparisons to the state-of-the-art methods for skeleton-based action retrieval on NTU-60 and NTU-120.

Method	Modality	NTU-60		NTU-120	
		x-sub	x-view	x-sub	x-setup
LongT GAN [61]	J	39.1	48.1	31.5	35.5
P&C [40]	J	50.7	76.3	39.5	41.8
AimCLR [10]	J	62.0	71.5	-	-
ISC [45]	J	62.5	82.6	50.6	52.3
HiCLR [58]	J	67.3	75.3	-	-
HiCo [8]	J	68.3	84.8	56.6	59.1
CMD [29]	J	70.6	85.4	58.3	60.9
<i>UmURL</i> (This work)	J	71.3	88.3	58.5	60.9
<i>UmURL</i> (This work)	J+M+B	72.0	88.9	59.5	62.2

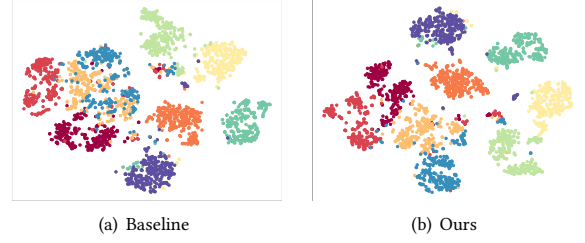
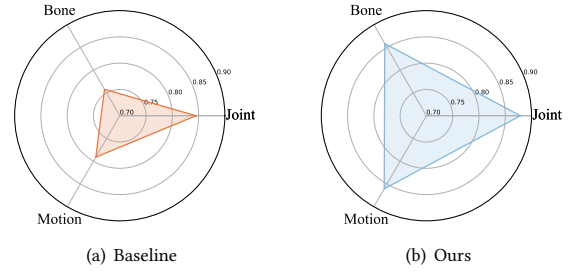
**Figure 3: Comparison to the simple baseline of multi-modal unsupervised representation learning in the context of skeleton-based (a) action recognition and (b) action retrieval downstream tasks.**

clear performance improvement when all three modalities are utilized. The results further demonstrate that the action representation obtained by our method is more discriminative.

4.3 Comparison to the Simple Baseline

To further verify the effectiveness of our framework, we further compare it to the simple baseline described in Section 3.1. The comparisons are conducted on NTU-60 in the context of the skeleton-based action recognition and action retrieval downstream tasks.

Results. Recall that the simple baseline is the direct implementation of multi-modal unsupervised representation learning, which can be roughly regarded as a special case of our proposed framework without our modality decomposition, intra-modal consistency learning and inter-modal consistency learning modules. As illustrated in Figure 3, our proposed framework consistently outperforms the simple baseline with clear margins on both downstream tasks. It further verifies the effectiveness of our proposed modules. Additionally, we also visualize their learned action representations via t-SNE [48]. Compared to the dots in Figure 4(a), dots of the same colors (e.g., blue and yellow dots) in Figure 4(b) are more clustered, and dots of different colors are more separated. The results demonstrate that our proposed framework allows it to learn more discriminative multi-modal representation.

**Figure 4: t-SNE visualization of the learned multi-modal action representations obtained by (a) simple baseline and (b) our proposed UmURL on NTU-60. 10 classes from the testing set are randomly selected for visualization. Dots with the same color indicate actions belonging to the same class.****Figure 5: Modality contribution to the final multi-modal representation. For the baseline, the *joint* modality is more dominant in the final multi-modal representation, while each modality contributes more balanced in our UmURL.**

Analysis. To further investigate our unified multi-modal representation learning framework, we try to analyze how much each modality contributes to the final multi-modal representation. We measure the modality contribution via the dependency between the obtained representation and the corresponding modality input, which can be computed by the distance correlation proposed by [43]. Note that, the higher correlation indicates more contribution to the final multi-modal representation. As shown in Figure 5, we provide the contribution results of both the simple multi-modal baseline and our UmURL framework. For the baseline model, the *joint* modality is more dominant in the final multi-modal representation since it is easier than other modalities to learn during the unsupervised training. However, this will miss complementary information from other modalities, thus does not possess informative enough features for downstream tasks. Instead, our UmURL framework introduces two consistency constraints to learn the unified representations of uni-modal and multi-modal input, achieving balanced contribution among different modalities during the feature learning.

4.4 Ablation Study

In this section, we study the effectiveness of intra-modal and inter-modal consistency learning. As the intra-modal consistency learning module is employed on multiple modalities, we also explore its influence on individual modalities. The experiments are conducted on NTU-60 in the context of action recognition using unified multi-modal representation, and the results are shown in Table 3. The

Table 3: The ablation study on intra-modal and inter-modal consistency learning.

Intra-modal			Inter-modal	x-sub	x-view
Joint	Bone	Motion			
-	-	✓	-	78.9	84.3
-	✓	-	-	82.4	89.4
✓	-	-	-	82.8	89.8
✓	✓	✓	-	83.9	90.6
✓	✓	✓	✓	84.2	90.9

Table 4: Comparisons to the state-of-the-art methods with transfer learning.

Method	Modality	Transfer to PKU-MMD II	
		NTU-60	NTU-120
LongT GAN [61]	J	44.8	-
M ² L [20]	J	45.8	-
ISC [45]	J	45.9	-
CrosSCLR [18]	J	54.0	52.8
HiCo [8]	J	56.3	55.4
CMD [29]	J	56.0	57.0
UmURL (This work)	J	58.2	57.6
UmURL (This work)	J+M+B	59.7	58.5

model with the intra-modal consistency learning module on three modalities outperforms the counterparts on a specific modality, which demonstrates the benefit of using the intra-modal consistency learning module on each modality. Besides, integrating the inter-modal consistency learning module achieves a further performance gain. The results not only verify the effectiveness of the inter-modal consistency learning module but also demonstrate the complementary between the intra-modal and inter-modal modules.

4.5 The Potential for Other Downstream Tasks

We further evaluate the learned representation for other downstream tasks, including skeleton-based action recognition in the scenario of semi-supervised learning and transfer learning.

Semi-supervised Skeleton-based Action Recognition. Following the previous works [5, 45], we report the results of using 1% and 5% randomly sampled training data with labels for fine-tuning. Note that the skeleton encoder is firstly pre-trained by our proposed unified multi-modal representation learning using unlabeled data, and fine-tuned with an extra classifier using labeled data. Table 5 summarizes the semi-supervised results on NTU-60. With the single modality of joint or multiple modalities for inference, our proposed method consistently outperforms the previous works by a clear margin. The results demonstrate the potential of our method for semi-supervised action recognition.

Skeleton-based Action Recognition with Transfer Learning. We evaluate the generalizability of the learned representation by transferring knowledge from a source dataset to a target dataset.

Table 5: Comparisons to the state-of-the-art methods with semi-supervised learning on NTU-60 dataset.

Method	Modality	x-sub		x-view	
		1%	5%	1%	5%
ASSL [37]	J	-	57.3	-	63.6
ISC [45]	J	35.7	59.6	38.1	65.7
MCC [42]	J	-	47.4	-	53.3
Hi-TRS [4]	J	39.1	63.3	42.9	68.3
GL-Transformer [13]	J	-	64.5	-	68.5
Colorization [52]	J	48.3	65.7	52.5	70.3
CrosSCLR [18]	J	48.6	67.7	49.8	70.6
HiCo [8]	J	54.4	-	54.8	-
CPM [57]	J	56.7	-	57.5	-
CMD [29]	J	50.6	71.0	53.0	75.3
UmURL (This work)	J	58.1	72.5	58.3	76.8
3s-AimCLR [10]	J+M+B	54.8	-	54.3	-
3s-CMD [29]	J+M+B	55.6	74.3	55.5	77.2
UmURL (This work)	J+M+B	59.6	74.6	60.3	78.6

Concretely, a model is initially pre-trained on a source dataset using unsupervised learning, and subsequently fine-tuned on a target dataset. We use the same setting as previous methods [8, 45], where NTU-60 and NTU-120 are chosen as source datasets, and PKU-MMD II is selected as the target dataset. The evaluation was conducted under the x-sub protocol, and the corresponding results are shown in Table 4. Our proposed method outperforms competitors by a significant margin, demonstrating the good transferability of our learned representation. The results suggest that our proposed framework can effectively learn skeleton representations that generalize to new datasets, which is crucial for real-world applications.

5 CONCLUSION

In this paper, we present a novel unified multi-modal representation learning framework, *i.e.*, UmURL, for skeleton-based action understanding. Compared to the conventional multi-modal approaches via the late-fusion strategy, our proposed UmURL requires less computational overheads on pre-training and downstream tasks, and is more flexible to the input modalities during inference. With a much more efficient multi-modal network than previous multi-modal solutions, we achieve new state-of-the-art performance in multiple downstream tasks, including skeleton-based action recognition and retrieval. We believe that our proposed framework offers an effective alternative to conventional multi-modal approaches in unsupervised skeleton-based representation learning.

Acknowledgements. This work was supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No.2023C01212), Public Welfare Technology Research Project of Zhejiang Province (No. LGF21F020010), National Natural Science Foundation of China (No. 61976188, 62272435, and U22A2094), Young Elite Scientists Sponsorship Program by CAST (No. 2022QNR0001), the open research fund of The State Key Laboratory of Multimodal Artificial Intelligence Systems, and the Fundamental Research Funds for the Provincial Universities of Zhejiang.

REFERENCES

- [1] Adrien Bardes, Jean Ponce, and Yann Lecun. 2022. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. In *International Conference on Learning Representations*. 1–1.
- [2] Haibo Chen, Zhizhong Wang, Huiming Zhang, Zhiwen Zuo, Ailin Li, Wei Xing, Dongming Lu, et al. 2021. Artistic style transfer with internal-external learning and contrastive learning. *Advances in Neural Information Processing Systems* 34 (2021), 26561–26573.
- [3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*. 1597–1607.
- [4] Yuxiao Chen, Long Zhao, Jianbo Yuan, Yu Tian, Zhaoyang Xia, Shijie Geng, Ligong Han, and Dimitris N Metaxas. 2022. Hierarchically Self-Supervised Transformer for Human Skeleton Representation Learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 185–202.
- [5] Yi-Bin Cheng, Xipeng Chen, Junhong Chen, Pengxu Wei, Dongyu Zhang, and Liang Lin. 2021. Hierarchical transformer: Unsupervised representation learning for skeleton-based human action recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*. 1–6.
- [6] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. 2022. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20186–20196.
- [7] Jianfeng Dong, Xiaoman Peng, Zhe Ma, Daizong Liu, Xiaoye Qu, Xun Yang, Jixiang Zhu, and Baolong Liu. 2023. From Region to Patch: Attribute-Aware Foreground-Background Contrastive Learning for Fine-Grained Fashion Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1273–1282.
- [8] Jianfeng Dong, Shengkai Sun, Zhonglin Liu, Shujie Chen, Baolong Liu, and Xun Wang. 2023. Hierarchical Contrast for Unsupervised Skeleton-based Action Representation Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 1–1.
- [9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slow-fast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6202–6211.
- [10] Tianyu Guo, Hong Liu, Zhan Chen, Mengyuan Liu, Tao Wang, and Runwei Ding. 2022. Contrastive Learning from Extremely Augmented Skeleton Sequences for Self-supervised Action Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 762–770.
- [11] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9729–9738.
- [12] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2012. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (2012), 221–231.
- [13] Boeun Kim, Hyung Jin Chang, Jungcho Kim, and Jin Young Choi. 2022. Global-local Motion Transformer for Unsupervised Skeleton-based Action Learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 209–225.
- [14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [15] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: a large video database for human motion recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2556–2563.
- [16] Jogendra Nath Kundu, Maharshi Gor, Phani Krishna Uppala, and Venkatesh Babu Radhakrishnan. 2019. Unsupervised feature learning of human actions as trajectories in pose embedding manifold. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1459–1467.
- [17] Kun Li, Jiaxiu Li, Dan Guo, Xun Yang, and Meng Wang. 2023. Transformer-based Visual Grounding with Cross-modality Interaction. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 6 (2023), 1–19.
- [18] Linguo Li, Minsi Wang, Bingbing Ni, Hang Wang, Jiancheng Yang, and Wenjun Zhang. 2021. 3d human action representation learning via cross-view consistency pursuit. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4741–4750.
- [19] Yicong Li, Xun Yang, Xindi Shang, and Tat-Seng Chua. 2021. Interventional video relation detection. In *Proceedings of the 29th ACM International Conference on Multimedia*. 4091–4099.
- [20] Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. 2020. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2490–2498.
- [21] Daizong Liu, Xiang Fang, Wei Hu, and Pan Zhou. 2023. Exploring optical-flow-guided motion and detection-based appearance for temporal sentence grounding. *IEEE Transactions on Multimedia* (2023), 1–14.
- [22] Daizong Liu and Wei Hu. 2022. Skimming, locating, then perusing: A human-like framework for natural language video localization. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4536–4545.
- [23] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. 2021. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11235–11244.
- [24] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4070–4078.
- [25] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. 2019. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 10 (2019), 2684–2701.
- [26] Jiaying Liu, Sijie Song, Chunhui Liu, Yanghao Li, and Yueyu Hu. 2020. A benchmark dataset and comparison study for multi-modal human action analytics. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 2 (2020), 1–24.
- [27] Zhengguang Liu, Haoming Chen, Runyang Feng, Shuang Wu, Shouling Ji, Bailin Yang, and Xun Wang. 2021. Deep dual consecutive network for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 525–534.
- [28] Zhengguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, and Li Cheng. 2019. Towards natural and accurate future motion prediction of humans and animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10004–10012.
- [29] Yunyao Mao, Wengang Zhou, Zhenbo Lu, Jiajun Deng, and Houqiang Li. 2022. CMD: Self-supervised 3D Action Representation Learning with Cross-modal Mutual Distillation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 734–752.
- [30] Qiang Nie, Ziwei Liu, and Yunhui Liu. 2020. Unsupervised 3d human pose representation with viewpoint and pose disentanglement. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 102–118.
- [31] Xiaoye Qu, Pengwei Tang, Zhikang Zou, Yu Cheng, Jianfeng Dong, Pan Zhou, and Zichuan Xu. 2020. Fine-grained iterative attention network for temporal language localization in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4280–4288.
- [32] Haocong Rao, Shihao Xu, Xiping Hu, Jun Cheng, and Bin Hu. 2021. Augmented skeleton based contrastive action learning with momentum lstm for unsupervised action recognition. *Information Sciences* 569 (2021), 90–109.
- [33] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1010–1019.
- [34] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. 2019. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*. 279–287.
- [35] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7912–7921.
- [36] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. 2019. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12026–12035.
- [37] Chenyang Si, Xuecheng Nie, Wei Wang, Liang Wang, Tieniu Tan, and Jiashi Feng. 2020. Adversarial self-supervised learning for semi-supervised 3d action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 35–51.
- [38] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. 2020. Stronger, faster and more explainable: A graph convolutional baseline for skeleton-based action recognition. In *proceedings of the 28th ACM international conference on multimedia*. 1625–1633.
- [39] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).
- [40] Kun Su, Xiulong Liu, and Eli Shlizerman. 2020. Predict & cluster: Unsupervised skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9631–9640.
- [41] Yukun Su, Guosheng Lin, Ruizhou Sun, Yun Hao, and Qingyao Wu. 2021. Modeling the Uncertainty for Self-supervised 3D Skeleton Action Representation Learning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 769–778.
- [42] Yukun Su, Guosheng Lin, and Qingyao Wu. 2021. Self-Supervised 3D Skeleton Action Representation Learning With Motion Consistency and Continuity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13328–13338.
- [43] Gabor J Székely, Maria L Rizzo, and Nail K Bakirov. 2007. Measuring and testing dependence by correlation of distances. *Annals of statistics* 35, 6 (2007), 2769–2794.
- [44] Yi Tan, Yanbin Hao, Xiangnan He, Yinwei Wei, and Xun Yang. 2021. Selective dependency aggregation for action classification. In *Proceedings of the 29th ACM International Conference on Multimedia*. 592–601.

- [45] Fida Mohammad Thoker, Hazel Doughty, and Cees GM Snoek. 2021. Skeleton-Contrastive 3D Action Representation Learning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1655–1663.
- [46] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6450–6459.
- [47] Neel Trivedi and Ravi Kiran Sarvadevabhatla. 2023. PSUMNet: Unified Modality Part Streams are All You Need for Efficient Pose-based Action Recognition. In *European Conference on Computer Vision Workshops*. 211–227.
- [48] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605.
- [49] Rukai Wei, Yu Liu, Jingkuan Song, Yanzhao Xie, and Ke Zhou. 2023. Deep debiased contrastive hashing. *Pattern Recognition* 139 (2023), 109483.
- [50] Junbin Xiao, Xindi Shang, Xun Yang, Sheng Tang, and Tat-Seng Chua. 2020. Visual relation grounding in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI* 16. Springer, 447–464.
- [51] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI Conference on Artificial Intelligence*. 7444–7452.
- [52] Siyuan Yang, Jun Liu, Shijian Lu, Meng Hwa Er, and Alex C Kot. 2021. Skeleton cloud colorization for unsupervised 3d action representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13423–13433.
- [53] Xun Yang, Jianfeng Dong, Yixin Cao, Xun Wang, Meng Wang, and Tat-Seng Chua. 2020. Tree-augmented cross-modal encoding for complex-query video retrieval. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 1339–1348.
- [54] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1–10.
- [55] Xun Yang, Xueliang Liu, Meng Jian, Xinjian Gao, and Meng Wang. 2020. Weakly-supervised video object grounding by exploring spatio-temporal contexts. In *Proceedings of the 28th ACM international conference on multimedia*. 1939–1947.
- [56] Xun Yang, Shanshan Wang, Jian Dong, Jianfeng Dong, Meng Wang, and Tat-Seng Chua. 2022. Video moment retrieval with cross-modal neural architecture search. *IEEE Transactions on Image Processing* 31 (2022), 1204–1216.
- [57] Haoyuan Zhang, Yonghong Hou, Wenjing Zhang, and Wanqing Li. 2022. Contrastive Positive Mining for Unsupervised 3D Action Representation Learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 36–51.
- [58] Jiahang Zhang, Lilang Lin, and Jiaying Liu. 2023. Hierarchical Consistent Contrastive Learning for Skeleton-Based Action Recognition with Growing Augmentations. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 1–1.
- [59] Yuhang Zhang, Bo Wu, Wen Li, Lixin Duan, and Chuang Gan. 2021. STST: Spatial-Temporal Specialized Transformer for Skeleton-based Action Recognition. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3229–3237.
- [60] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. 2020. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5741–5750.
- [61] Nenggan Zheng, Jun Wen, Risheng Liu, Liangu Long, Jianhua Dai, and Zhefeng Gong. 2018. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32. 2644–2651.
- [62] Qi Zheng, Jianfeng Dong, Xiaoye Qu, Xun Yang, Yabing Wang, Pan Zhou, Baolong Liu, and Xun Wang. 2023. Progressive localization networks for language-based moment localization. *ACM Transactions on Multimedia Computing, Communications and Applications* 19, 2 (2023), 1–21.
- [63] Yujie Zhou, Haodong Duan, Anyi Rao, Bing Su, and Jiaqi Wang. 2023. Self-supervised Action Representation Learning from Partial Spatio-Temporal Skeleton Sequences. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 1–1.
- [64] Zhiwen Zuo, Lei Zhao, Ailin Li, Zhizhong Wang, Zhanjie Zhang, Jiafu Chen, Wei Xing, and Dongming Lu. 2023. Generative Image Inpainting with Segmentation Confusion Adversarial Training and Contrastive Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 3888–3896.

Appendix

This appendix contains the following contents which are not included in the paper due to space limits:

- More results including the actual running time comparison (Section A.1) and visualization of learned representation (Section A.2).
- Additional ablation studies including modality selection, fusion ways, and architecture designs (Section B).
- Implementation details including the descriptions of used datasets, model structure and training details (Section C).

A ADDITIONAL RESULTS

A.1 Actual running time comparison

Besides the theoretical analysis in terms of FLOPs, we also compare our proposed UmURL to the recent state-of-the-art method 3s-CMD [29] in terms of the actual running time consumption during the pre-training and downstream inference. For a fair comparison, the two models have been pre-trained with 450 epochs. The models are trained and evaluated under x-sub protocol on NTU-120. All results are obtained in the same environment using one RTX 3090 GPU. As demonstrated in Table 6, our proposed UmURL runs significantly faster than 3s-CMD [29] when using the same multiple modalities. Besides, our proposed model achieves better accuracy. The results demonstrate both efficiency and effectiveness of our method.

Table 6: Actual running time comparison with 3s-CMD [29] that also uses multiple modalities. Our proposed model is more efficient during both the pre-training and inference stages, and also performs better.

Methods	Pre-training	Inference	Accuracy
3s-CMD	71h 57m	66s	74.7
UmURL	12h 23m	14s	75.2

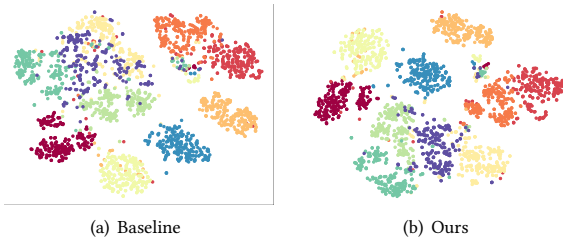


Figure 6: t-SNE visualization of the multi-modal action representations obtained under x-sub protocol on NTU-60 by (a) simple baseline and (b) our proposed UmURL.

A.2 Additional Visualization Results

In addition to the visualization presented under the x-view protocol in Section 4.3, we extend our visualization of the learned action representation using t-SNE [48] under the x-sub protocol on NTU-60. Similarly, we randomly select 10 classes from the testing set for

visualization. Dots of identical color represent actions belonging to the same class. As shown in Figure 6, dots corresponding to our proposed UmURL (e.g., purple dots) appear more clustered. These results further prove that our proposed method is capable of learning a more discriminative multi-modal representation than the baseline.

B ADDITIONAL ABLATION STUDIES

In this section, all experiments are conducted on NTU-60 in the context of action recognition using our proposed unified multi-modal representation.

B.1 Effects of modality selection.

We evaluate the performance of the unified multi-modal representation obtained via pre-training with different selections of skeleton modalities. It is worth noting that the joint modality is consistently preserved as a fundamental modality, given its better performance relative to other modalities. For the uni-modal baseline, we utilize the same optimization method as in the simple multi-modal baseline but with inputs from a single modality only. Table 7 summarizes the results, we can find that using additional modalities enhances the performance of our proposed UmURL.

Table 7: Performance of UmURL with different modality selection on NUT-60. Jointly using three modalities performs the best.

Modality	x-sub	x-view
Joint	81.7	88.9
Joint+Motion	83.7	90.3
Joint+Bone	83.3	90.4
Joint+Motion+Bone	84.2	90.9

B.2 Effects of different fusion ways.

We investigate various fusion ways for different modality embeddings before encoding including weighted sum with learned scalar weights, averaging, averaging followed by a linear transformation, and concatenation followed by a linear transformation. Three modalities are jointly used in this experiment. As shown in Table 8, the fusion operation of averaging followed by linear transformation slightly performs better than the others. The results demonstrate that our proposed UmURL is not very sensitive to fusion ways.

Table 8: Performance of UmURL using different fusion ways on NUT-60. Our UmURL is not very sensitive to fusion ways.

Fusion	x-sub	x-view
Weighted sum	83.9	90.3
Averaging	84.0	90.6
Averaging+linear	84.2	90.9
Concatenation+linear	84.3	90.7

B.3 Effects of different architecture designs.

To preserve the unique semantics and extract modality-specific patterns, we utilize independent modules of embedding and projector for each modality. We also experiment with replacing the independent module with a shared one. As shown in Table 9, both shared embeddings and projector designs lead to performance degradation. The results verify the effectiveness of our modality-specific architecture design.

Table 9: Performance of UmURL with different architecture designs on NUT-60. Compared to shared ones, modality-specific (MS) embeddings and projectors are beneficial.

MS embedding	MS projector	x-sub	x-view
-	✓	83.2	90.1
✓	-	83.4	90.2
✓	✓	84.2	90.9

C IMPLEMENTATION DETAILS

In this section, we describe the implementation details. The proposed model is implemented using PyTorch.

C.1 Datasets

NTU-60 [33] is a large-scale action recognition dataset, which contains 56,880 action samples collected from 40 subjects, with a total of 60 categories. There are two recommended standard evaluation protocols: 1) x-sub: the data are split according to the subjects, where samples from half of the subjects are used as training data, and the rest subjects are used for testing. 2) x-view: the data are split according to camera views, where samples captured by cameras 2 and 3 are used for training, and samples captured by camera 1 are used for testing.

NTU-120 [25] is an extended version of *NTU-60*, containing 120 action categories and 114,480 samples. Action samples of 106

subjects are captured using 32 different setups according to the camera distances and background. There are also two recommended standard evaluation protocols: 1) x-sub: similar to *NTU-60*, samples from 53 subjects are used as training data and the rest 53 subjects are used as testing data. 2) x-setup: samples having even setup IDs are used for training, and samples having odd setup IDs are used for testing.

PKU-MMD II [26] is a popular benchmark for skeleton-based human action understanding. It contains 41 action categories, and each category is performed by one or two subjects, with 5,339 skeleton samples for training and 1,613 for testing. *PKU-MMD II* is challenging due to its larger view variation. Following prior works, we evaluate our method with the *PKU-MMD II* under recommended x-sub protocol.

C.2 Model Structure.

We use the transformer encoder to process multimodal information. Following [8, 59], we simultaneously model skeleton sequences in both spatial and temporal dimensions, utilizing a single-layer encoder with 1024 hidden units for each dimension. The spatial input is obtained by directly reshaping the original skeleton sequence. The final representation is produced by concatenating the features from both dimensions. The projector is composed of two fully-connected layers with batch normalization and ReLU, and a third linear layer with the output size of 4096. The model's inputs are temporally downsampled to 64 frames.

C.3 Training Details.

For the optimizer, we employ the Adam algorithm [14] with a weight decay of $1e-5$. The mini-batch size is set to 512. Following the pre-training scheme in [29], the model is trained for 450 and 1000 epochs for *NTU-60/120* and *PKU-MMD II* datasets, respectively. The initial learning rate is set to $5e-4$, and it is reduced to $5e-5$ at epoch 350 and 800 for *NTU-60/120* and *PKU-MMD II* respectively. We adopt the same data augmentation strategies employed in [29, 45] for a fair comparison. The γ is set to 1 following [1]. For other hyper-parameters, the λ and μ are set to 5 and 5, respectively.