From Isolated Islands to *Pangea*: Unifying Semantic Space for Human Action Understanding

Yong-Lu Li*, Xiaoqian Wu*, Xinpeng Liu, Zehao Wang, Yiming Dou, Yikun Ji, Junyi Zhang,
Yixing Li, Xudong Lu, Jingru Tan, Cewu Lu†
Shanghai Jiao Tong University
{yonglu_li, enlighten, davidwang200099, douyiming, junyizhang, lyxing0, luxudong2001,
lucewu}@sjtu.edu.cn, {xinpengliu0907, jiyikun2002, tanjingru120}@gmail.com

Abstract

*Action understanding has attracted long-term attention. It can be formed as the mapping from the physical space to the semantic space. Typically, researchers built datasets according to idiosyncratic choices to define classes and push the envelope of benchmarks respectively. Datasets are incompatible with each other like "**Isolated Islands**" due to semantic gaps and various class granularities,* e.g.*,* `do housework` *in dataset A and* `wash plate` *in dataset B. We argue that we need a more principled semantic space to concentrate the community efforts and use all datasets together to pursue generalizable action learning. To this end, we design a structured action semantic space in view of verb taxonomy hierarchy and covering massive actions. By aligning the classes of previous datasets to our semantic space, we gather (image/video/skeleton/MoCap) datasets into a unified database in a unified label system,* i.e.*, bridging "isolated islands" into a "**Pangea**". Accordingly, we propose a novel model mapping from the physical space to semantic space to fully use Pangea. In extensive experiments, our new system shows significant superiority, especially in transfer learning. Our code and data will be made public at https://mvig-rhos.com/pangea.*

1. Introduction

Visual action understanding is an important direction in computer vision and matters to various domains [14, 65]. Generally speaking, it can be formulated as the mapping from the physical space to the semantic space. Here, *physical* space indicates the visual patterns (information carrier) and *semantic* space represents the action semantics (class).

In terms of the physical space, many works were proposed to extract representations from different modalities to capture action cues, such as image [42], video [5], skeleton [29], MoCap [25], RGBD [62], *etc*. However, few efforts have been made to semantic space design. Previous benchmarks [6, 10, 32] are typically designed according to designers' choice and incompatible with each other due to semantic gaps. They have three main weaknesses: (1) **Ambiguity**. Similar actions may have different class names, *e.g.*, `clean`, `wipe`, `scrub`. Though this may strengthen the diversity in visual-language learning [59], it hinders machines from learning the subtle similarities and differences of actions. Besides, the same class may represent different actions, *e.g.*, `address` means either addressing oneself to something or addressing a conference. This phenomenon brings both generalization possibility and challenge. (2) Overlooking **granularity/hierarchy**. The datasets are constructed independently, thus typically overlooking granularity, *e.g.*, `do housework` in dataset A and `clean floor` in dataset B, sometimes even in one dataset. (3) **Integration/transfer difficulty**. Large models need more data. However, due to the "isolated islands", it is hard to integrate datasets and conclude the "few-shotness" and "zero-shotness" of classes. We do not know which

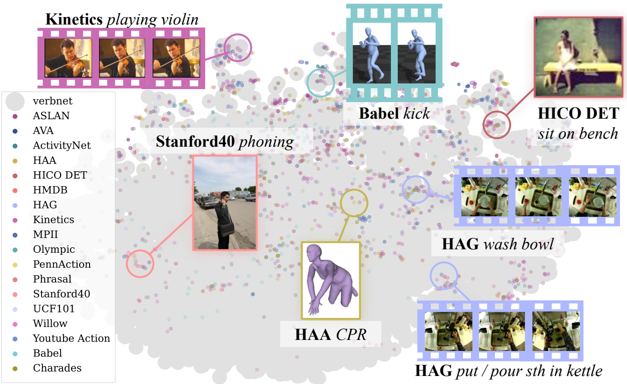


Figure 1. "Isolated islands". The semantic gap brings a great challenge to general action understanding.

*The first two authors contribute equally.
†Corresponding author.

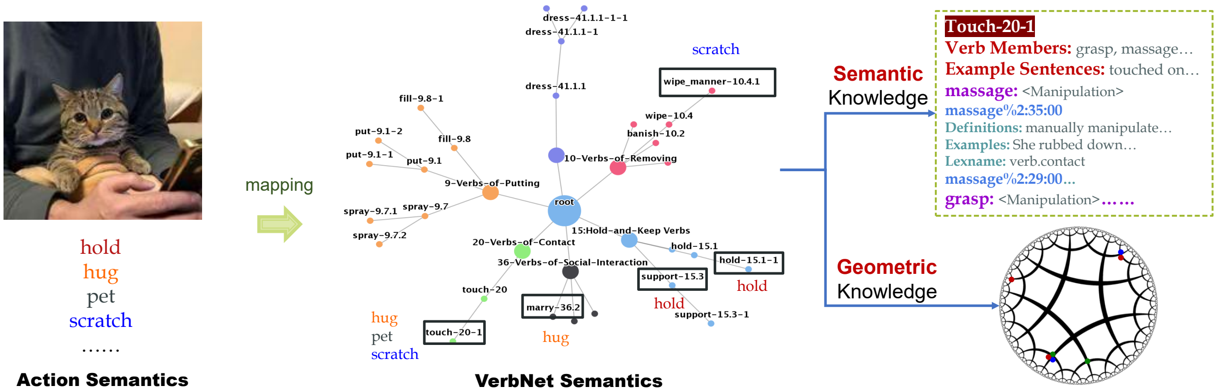


Figure 2. Verb tree. The conventional action semantics (*e.g.*, `hold, hug`) can be mapped into node semantics (*e.g.*, `touch-20-1, support-15.3`). The proposed semantic space has abundant semantic and geometric knowledge.

classes should be enriched or used for transfer learning.

In Fig. 1, we visualize the class word embeddings [21] of 18 datasets via t-SNE. Huge semantic gaps exist. Even for the very large Kinetics-700 [5], there are still many classes beyond its coverage. Here, we first clearly reveal the overlooked "Isolated Islands ($\mathbf{I}^2$)" problem. It brings semantic gaps and impedes cross-dataset learning. Though CLIP [59]-like works alleviate this problem to some extent with the open-vocabulary property, their latent space may be difficult to capture the subtle polysemy, taxonomy, and hierarchy of action semantics. In experiments (Sec. 6), CLIP trained with simply-mixed datasets performs not well.

Thus, we rethink the action semantic space design and take a step towards a principled semantic space. We propose a new system to pave a promising way to address the I^2 problem. Our core idea is to use a *structured* action semantic space to replace the existing hand-crafted ones. We build this semantic space according to the linguistic structure knowledge of VerbNet [61]. VerbNet is a network linking the syntactic and semantic patterns of verbs. It is a domain-independent tree-structure lexicon and has a clear hierarchy covering most verbs. We visualize the verb tree in Fig. 2. To maximize the potential of our semantic space, we gather many datasets (image/video/skeleton/MoCap) to build a database and align their classes to our semantic space easily, *i.e.*, linking the "isolated islands" into a "*Pangea*". Then, we can use the continuous hyperbolic space together with the semantic-geometric prompt to embed the structured knowledge.

Our space has four-fold superiority: (1) **Unambiguous** verb nodes correlating all related verbs, *e.g.*, `pat, nudge, massage` with similar meaning are shared by the node `touch-20-1`. (2) **Rich knowledge**. Besides the thematic role, syntactic, semantic description, and selectional preferences of verbs, VerbNet has mappings to other knowledge bases (WordNet [53], PropBank [33], FrameNet [1]). We can conveniently adopt Large Language Models [3] to extract meaningful language representations to advance learning. (3) **Hierarchy** to represent actions from abstract to specific granularity, *e.g.*, `sports, ball sports, basketball, dunk`. (4) **Extensive coverage**. It contains about 5,800 verbs. In Fig. 1, our space not only covers all datasets but also spans the semantics a lot.

To fully use *Pangea*, we propose a compact mapping system to conduct action understanding, which effectively maps multi-modal physical patterns to the structured semantic space. In experiments, our method armed with *Pangea* demonstrates representative and transfer ability. On multi-modal benchmarks, it brings decent improvements.

Our contributions are: 1) We propose a structured semantic space to bridge the "isolated islands". 2) We build the *Pangea* database gathering 28 multi-modal datasets. 3) A physical-to-semantic mapping model is proposed given *Pangea* and shows significant transfer ability.

2. Related Work

Action Understanding has achieved progress recently. There are mainly image [6, 26, 77], video [15, 24, 32, 66], skeleton [46], and 3D body [57] datasets. The common tasks are action recognition and temporal/spatial localization/detection. Early benchmarks focus on classifying an image or a short video into classes [6, 66, 77]. Recently, benchmarks that require both accurate recognition and active subject detection are emerging [7, 24, 26]. Moreover, few/zero-shot action learning [8] also attracts attention. Many methods have been proposed to push this direction forward. For image tasks, 2D CNN is the dominant architecture, while knowledge like part state [41, 50], 2D/3D human [39, 40], and language prior [2, 28, 42, 56] is used too. For video tasks, 2D-CNN [13, 43, 78], two-stream network [17, 64], and 3D-CNN [5, 18] are the major architectures adopted. For skeleton tasks, both GCN [38, 47, 75] and 2D-CNN [9, 74] are widely used. Recently, with the success of Transformer [71], besides directly importing it into action detection [4, 68], visual-language contrastive learning [59] has changed this direction a lot.

In terms of action semantic space, most datasets [6, 24, 26, 35, 36] overlook action hierarchy. While some works consider hierarchy [15, 48, 63]. For example, ActivityNet [15] defines 200+ action classes belonging to 7 high-level classes (*e.g.*, personal care, household) based on activity scenarios; FineGym [63] organizes hierarchical actions from gymnasium videos; VerSe [22] augments COCO [44] and TUHOI [37] with verb sense labels to provide finer-grained action semantics on 3.5 K images. However, they are scale/class/domain-limited and built with manually-picked classes. Instead, we choose to cover the hierarchy based on well-defined linguistic works such as VerbNet [61], WordNet [53], FrameNet [1], *etc*.

3. Preliminary

In this section, we first introduce the preliminaries of the physical and semantic space.

Multi-Modal Physical Space. Here, we adopt two modalities for physical space P: 2D and 3D. For 2D, we adopt CNN or Transformer (*e.g.*, ResNet [27], CLIP [59]) to extract representation from image/video. For 3D, we use the widely-used model SMPL [49] to embed 3D humans.

Structured Semantic Space. Intuitively, the ambiguity of objects is relatively smaller, thus objects/nouns are easier to label. Things are different for actions/verbs which are more ambiguous. Previous works typically design semantic space manually and optionally. Instead, we build the structured semantic space S via the hierarchical verb tree from VerbNet [61] (Fig. 2). Here, we define the *nodes* as the *classes* of our semantic space. Compared with conventional design [15], our space has elegant characteristics: **(1)** Due to the lack of a unified naming standard, classes of previous datasets have ambiguity. For example, different datasets may have feast, eating, and dining respectively, where a common semantic is shared. Instead, in our S, actions with **shared** meanings are connected with their *common* nodes. **(2)** Each node is equipped with **abundant knowledge**. In Fig. 2, touch-20-1 node is explained by: a) Verb members, *e.g.*, grasp; b) Example sentences as instantiations of the node semantics; c) Each verb member is explained via connections with other lexical resources (*e.g.*, WordNet [53], FrameNet [1]). In Fig. 2, the verb massage is explained by its frame in FrameNet [1] (manipulation) and the corresponding items in WordNet [53] (massage%2:35:00, massage%2:29:00). **(3) Hierarchy** reveals semantic connections between nodes and provides structured knowledge. The nodes are numbered according to shared semantics and syntax. Nodes sharing a high-level number (9-109) have semantic relations [61], *e.g.*, banish-10.2 and wipe-10.4 share a parent node as they are all about removing. Though some works [15, 63] consider hierarchy too, they are either of limited coverage or defined empirically according to scenes. Instead, our verb semantics are more explicit. **(4)** Our S covers **5,800+** verbs which is broader than previous works.

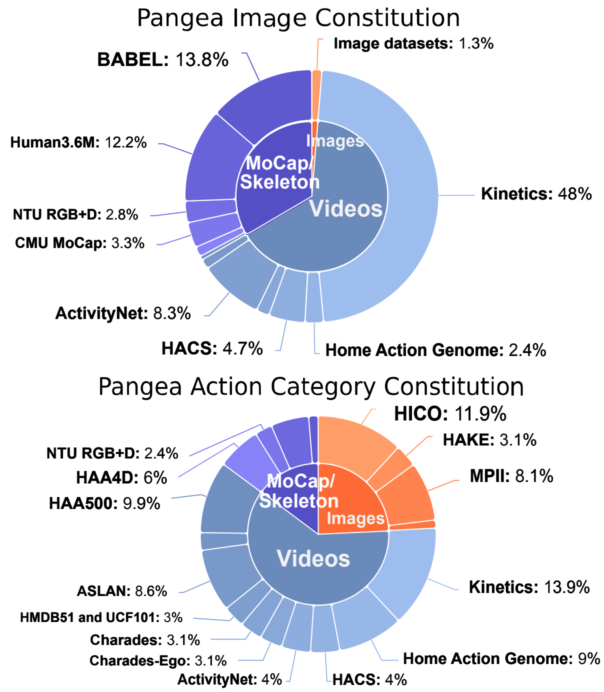


Figure 3. Gathered datasets in *Pangea*.

4. Constructing *Pangea*

Data Curation. With the structured S, we collect data with diverse modalities, formats, and granularities, and adapt them into a unified form. Our database *Pangea* contains a large range of data including image, video, and skeleton/MoCap. We process and formulate them as follows:

1) **Semantic Consistency.** The class definitions of datasets are various, but they can be mapped to our semantic space with the fewest semantic damages. The mapping is completed via manual annotation with the help of word embedding [59] distances and OpenAI GPT-3.5. Manual annotation is the most accurate and most expensive, while word embedding comparison is the least. Thus, we adopt a hybrid method: potential class-node mapping is first filtered out roughly by comparing word embedding, then selected via GPT-3.5 prompting, and finally checked by human annotators. As more and more classes are aligned and covered, the process would be faster and faster with synonyms checking. As shown in Fig. 1, our semantic space covers a broad range of semantics, verifying this mapping.

2) **Temporal Consistency.** Some videos [5] only have sparse labels for a whole clip instead of each frame. For these sparse datasets, we sample the clip with 3 FPS and give frames the label of their belonged clip. We provide both frame- and clip-level labels.

3) **Spatial Consistency.** There are both instance (boxes) [7] and image [6] level labels. We merge the instance labels of each image/frame into image/frame-level labels. For demands of instance-level training, we can use

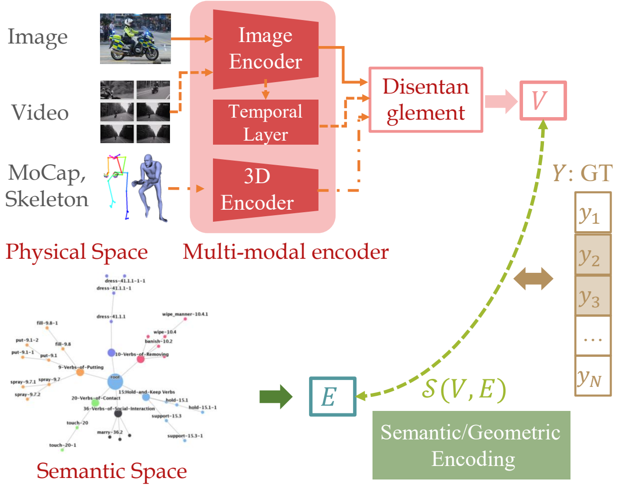


Figure 4. P2S mapping. Given a sample, we obtain its representation V via encoders. V is then aligned with node representations G under the supervision of Y. v_{raw} to V is omitted for clarity.

the original instance labels [7, 24] and detectors [4, 60] to get instance boxes even masks [34] for future annotation.

4) **3D Format Consistency.** 3D datasets typically have various formats, *e.g.*, SMPL [49] has 24 keypoints while CMU MoCap [25] has 31 keypoints. To keep consistency, we transform all of them into SMPL via a fitting procedure.

5) **2D-3D Consistency.** Image/video datasets mostly contain only 2D labels without GT 3D humans. Aside from the GT 3D humans from 3D datasets [57], we recover 3D humans from 2D data as pseudo 3D labels via ROMP [67] and EFT [31]. We use both GT and pseudo 3D humans in 3D action recognition. Though the reconstruction is sometimes noisy, we use the pseudo 3D humans as noisy data *augmentation* to supplement 2D learning. In tests, we find that 2D and 3D learning are complementary.

Analysis. With the large data collection and unified semantic space, we build *Pangea* as shown in Fig. 3. It contains 19.5 M images, 1.1 M videos, and 840 K 3D humans over 28 datasets, with coverage of 4 K+ action classes of original datasets. *Pangea* covers the semantics of 513 verb nodes over all the 898 nodes of VerbNet [61] and includes 290 leave nodes carrying fine-grained semantics.

5. Methodology

5.1. Overview of P2S Mapping

First, we introduce the Physical-to-Semantic Space (P2S) mapping (Fig. 4). We aim to propose a multi-modal, concise, and practical model as the baseline to inspire future work. Given a sample of the physical space P, we obtain its representation V via different encoders according to its modality. For images, we use a CNN/Transformer-based image encoder. For videos, we first input them to the image encoder for frame encoding and then use a temporal layer for temporal encoding. For SMPLs, we covert them into point clouds and use a PointNet++[58] as the encoder.

In the semantic space S, we define N target verb nodes. For each node, two types of information are provided by VerbNet [61]: 1) **semantic** one to describe its meaning, *e.g.*, example sentences, WordNet definitions; 2) **geometric** one to locate it in the tree and reveal its connection with the other nodes. The semantic and geometric information can be encoded via the verb node representation $E = \{e_i\}_{i=1}^N$ (detailed in Sec. 5.3). The ground-truth (GT) label for the sample is $Y = \{y_i | y_i \in \{0, 1\}\}_{i=1}^N$. P2S mapping is a multi-label classification, where a physical sample is mapped to multi-node of the semantic space (**one-to-many** mapping). The similarity $\mathcal{S}(V, E)$ between V and E is bound with the GT label Y, and the loss function is derived in Sec. 5.3. In Sec. 5.2, we discuss how to facilitate such one-to-many mapping with semantic disentangle and augmentation. We summarize the training and inference in Sec. 5.4.

5.2. Semantic Disentanglement and Augmentation

A person typically performs multi-action simultaneously, *e.g.*, standing while eating. Such entanglement of multi-action semantics increases the annotation and learning difficulty. It is a challenge to annotate all the ongoing actions of a person in previous datasets due to the limited coverage and ambiguity of their classes. Besides, as *Pangea* has a broader semantic space, after the *action*→*node* mapping in Sec. 4, we face a **partial-label** learning problem. Moreover, in the mapping, inevitably, some labels are early filtered out and a few of them should have been annotated as True. Also, errors of omission may exist within the labels because of annotators' bias. Thus, each sample theoretically has a partial annotation $Y = \{y_i | y_i = 1, 0, \emptyset\}_{i=1}^N$, where $1, 0$ are certain positive/negative labels, and $\emptyset$ is uncertain. Though it is nearly impossible to supplement the labels of all N verb nodes in *Pangea* for all samples (images/videos/MoCap), we can conduct flexible weakly-supervised learning with partly-labeled data with representation disentanglement.

To facilitate the one-to-many P2S mapping and address the partial-label learning problem, we propose disentangling a physical representation into **node-specific** representations. We use v_{raw} as the entangled physical feature. Thanks to our unambiguous verb node definition, we disentangle the input v_{raw} into N representations supervised by N verb nodes respectively. Thus, the gradients of verb nodes (True/False labeled clearly) were disentangled during training from the uncertain ones. As is illustrated in Fig. 5, a model is trained to transform the entangled physical representation $v_{raw} \in \mathbf{R}^d$ (d: dimension) into node-specific representation $V = \{v_i\}_{i=1}^N \in \mathbf{R}^{N \times d}$ (i: verb node index) as conditions. To get $V = \{v_i\}_{i=1}^N \in \mathbf{R}^{N \times d}$, we first define the verb node-specific disentangling mapping function

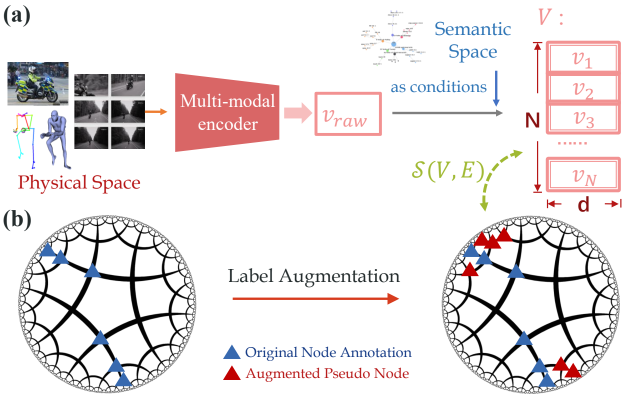


Figure 5. Semantic disentanglement and augmentation. (a) Entangled physical representation v_{raw} is mapped into node-specific $V = \{v_i\}_{i=1}^{N} \in \mathbf{R}^{N\times d}$ (i: node index as *conditions*). (b) Getting pseudo node labels via language priors and structure knowledge.

f_i for the i-th node, which is in practice a learnable MLP. Then, given v_{raw}, $f_i(\cdot)$ transforms it into $v_i = f_i(v_{raw})$. To apply P2S mapping with this disentangled physical representation, the similarity between V and G is measured as

$$\mathcal{S}(V,E) = \{\mathcal{S}(v_i, e_i)\}_{i=1}^{N}, \tag{1}$$

If not disentangled, it goes like

$$\mathcal{S}(v_{raw},E) = \{\mathcal{S}(v_{raw}, e_i)\}_{i=1}^{N}, \tag{2}$$

i.e., physical representation is shared by all verb nodes.

Aside from disentanglement, the partial-label learning problem can be alleviated by augmenting the GT label Y. As the verb nodes in our structured semantic space have clear semantic and geometric relations in a tree, we propose a label augmentation method to generate pseudo node labels for missing ones via language priors and structure knowledge. For more details, please refer to Suppl. Sec. 3.

5.3. Verb Node Encoding and Alignment

5.3.1 Semantic Encoding

Next, we discuss how to use text representation to encode semantic information of nodes into $E = \{e_i\}_{i=1}^{N} \in \mathbf{R}^{N\times d}$. As mentioned in Sec. 3, a verb node is composed of several actions with shared meanings. The node semantic information includes: 1) verb members; 2) example sentences; 3) WordNet [53] definition and FrameNet [1] mapping for each verb member. Following CLIP [59] text encoder, we get E via inputting these texts into a Transformer encoder.

Different from CLIP [59] where the text is short (up to 77 tokenized symbols, or equally 30 words approximately), our node description can be longer when the node contains many verb members. It is inefficient, unstable, and memory-costly to input such long text into the encoder directly. Thus, we sample key texts clarifying the node semantics better. We use TextRank [54] to extract keywords and then take the *summarized* text as the text encoder input.

5.3.2 Geometric Encoding

Next, we encode the geometric information into $E = \{e_i\}_{i=1}^{N} \in \mathbf{R}^{N\times d}$. To encode the hierarchy, parent-child relation, verb tree depth, *etc*., we use hyperbolic representations [11] of the physical representation V and verb node representation E. Besides, to utilize the representative ability of language models [12, 59], we also propose a geometric prompt strategy to strengthen the training. Fig. 6 is the overview of the encoding and $V - E$ alignment processes.

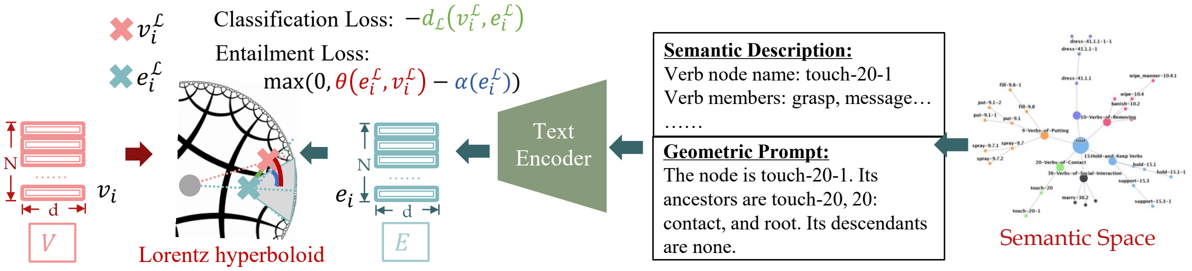
Geometric Prompt. A direct way is to use language descriptions as prompts, *e.g*., the node `touch-20-1` is described as: "The node is `touch-20-1`. Its ancestors are `touch-20, 20`: contact, and `root`. Its descendants are none." We use a text encoder to encode these prompts. In practice, we use text concatenation to integrate the geometric descriptions and those semantic descriptions introduced in Sec. 5.3.1. We concatenate these sentences and input them together into one Transformer encoder to get E.

Hyperbolic Representation The proposed semantic space is hierarchical, revealing connections between nodes and providing structured knowledge (Sec. 3). The text description of nodes implicitly reveals the hierarchy. For example, a node with a text description "The node is `put-9.1.1`. Its ancestors are `put-9.1, 9`: putting, and root. Its descendants are none. Its verbs are: apply, insert, install ..." would be closer to `put-9.1` (more generic concepts) and `put-9.1.2` (neighbor). Besides, P2S is a verb node multi-label classification. Thus, one physical representation can be aligned with both generic concepts which are closer to the *root* node of the hierarchy (*e.g*., `10:removing`), and specific concepts which are closer to the *leaf* node (*e.g*., `banish-10.2`, `wipe-10.4`). Thus, Euclidean space is not suitable for our task, which applies the same distance metric to all embedded points.

Here, we leverage the hyperbolic representation [55] which can capture hierarchy to embed V and E. Specifically, we adopt a Lorentz model of hyperbolic geometry [11]. Thus, similar to [11], the semantic hierarchy emerges in the representation space. We can thus align each disentangled physical representation to its corresponding multiple node representations. For a detailed formulation of the Lorentz model, please refer to Suppl. Sec. 3.

There are two objectives in the alignment: classification loss and entailment loss. Fig. 6 illustrates the calculation.

Classification Loss. We have the disentangled physical representation $V = \{v_i\}_{i=1}^{N}$, node representation $E = \{e_i\}_{i=1}^{N}$, and GT label $Y = \{y_i | y_i \in \{0,1\}\}_{i=1}^{N}$. For each i, v_i and e_i are first mapped into $v_i^{\mathcal{L}}$ and $e_i^{\mathcal{L}}$ in the Lorentz hyperboloid via *exponential map*. The similarity $\mathcal{S}(v_i, e_i)$



Semantic Description:
Verb node name: touch-20-1
Verb members: grasp, message…
……

Geometric Prompt:
The node is touch-20-1. Its ancestors are touch-20, 20: contact, and root. Its descendants are none.

Figure 6. Verb node encoding and $V - E$ alignment. 1) The *right* part: encoding semantic and geometric information via a text encoder. 2) The *left* part: $V - E$ alignment in a Lorentz hyperboloid with two training objectives: classification loss and entailment loss.

Method	Full	Rare	Non-Rare
CLIP	28.25	16.90	37.87
P2S	34.01	21.37	44.72
P2S-aug	**34.25**	**21.56**	**45.00**

Table 1. Verb node classification results on *Pangea* test set.

is measured via the negative of Lorentzian distance $d_{\mathcal{L}}(\cdot,\cdot)$ between $v_i^{\mathcal{L}}$ and $e_i^{\mathcal{L}}$. Thus, the classification loss is:

$$\mathcal{L}_{cls} = \mathcal{L}_{BCE}(\{Sigmoid(\gamma \cdot -d_{\mathcal{L}}(v_i^{\mathcal{L}}, e_i^{\mathcal{L}}))\}_{i=1}^{N}, Y), \quad (3)$$

where γ is a scaling factor. For multi-label classification, the output is processed by a Sigmoid function and bound with Binary Cross Entropy (BCE) loss.

Entailment Loss. In addition to the classification loss, an entailment loss is applied to enforce partial order relationships between the node representation $e_i^{\mathcal{L}}$ and physical representation $v_i^{\mathcal{L}}$. If $y_i = 1$, the physical representation $v_i^{\mathcal{L}}$ should lie inside the entailment cone [19] of the node representation $e_i^{\mathcal{L}}$. As is illustrated in Fig. 6, it is measured by comparing the exterior angle $\theta(e_i^{\mathcal{L}}, v_i^{\mathcal{L}})$ and the half-aperture $\alpha(e_i^{\mathcal{L}})$. Thus, the entailment loss is :

$$\mathcal{L}_{ent} = \frac{1}{sum(Y)} \sum_{i:y_i=1} max(0, \theta(e_i^{\mathcal{L}}, v_i^{\mathcal{L}}) - \alpha(e_i^{\mathcal{L}})). \quad (4)$$

The loss functions as a further constraint besides the classification loss.

5.4. Training and Inference

In training, the total loss $\mathcal{L}_{total} = \mathcal{L}_{cls} + \omega \mathcal{L}_{ent}$, where ω balances the loss weight (here $\omega = 0.01$). In inference, P2S outputs probabilities of verb nodes $\mathcal{S}_{node} = \{Sigmoid(\gamma \cdot -d_{\mathcal{L}}(v_i^{\mathcal{L}}, e_i^{\mathcal{L}}))\}_{i=1}^{N}$ from Eq. 3. We evaluate node classification with $\mathcal{S}_{node}$ on *Pangea* test set (Sec. 6.2). For transfer learning, we pretrain P2S on *Pangea* and finetune it on downstream datasets. To get the action class score $\mathcal{S}_{act}$ of the downstream dataset, we fix the node prediction and use a small learnable MLP to transform $\mathcal{S}_{node}$ to $\mathcal{S}_{act}$.

6. Experiment

6.1. Dataset and Implementation

Dataset. *Pangea* is adopted to evaluate verb node classification. We also conduct transfer learning on several multi-modal benchmarks: HICO [6], HAA [10], HMDB51 [36], Kinetics-400 [32], BABEL [57], and HAA4D [70].

Implementation. **(1)** P2S training: we use 19.5 M 2D images/frames and 840 K 3D humans. **(2)** P2S transfer learning: P2S pretrained on *Pangea* with node classification is a knowledgeable backbone. To make the transfer learning strict, in pretraining, we **exclude** the val & test set data of the downstream dataset from *Pangea* train set. Then the pretrained backbone is finetuned and tested on downstream datasets. For different modalities, we use their corresponding data path. To make our pipeline efficient, we do not adopt complex temporal encoding and video augmentation. Instead, we use *simple* strategies to implement the temporal encoding similar to [73], *e.g.*, mean pooling, a temporal transformer, average prediction of frames, *etc*. P2S is a multi-modal and lite method that is different from the ad-hoc models for sole-modal tasks. Thus we can use it as a **plug-and-play** method, *i.e.*, fusing it with SOTA models in downstream tasks. As P2S is trained in much broader semantic coverage on large-scale data, its learned bias is different from ad-hoc models. So P2S is complementary to these SOTA models and can improve their performances in the cooperation. Moreover, we test different ways to fuse 2D and 3D to mine the potential of multi-modal learning. The simplest late fusion (fusing logits) performs best in our tests (Suppl. Sec. 9). Thus, we use late fusion as the default. For data with one human per image/frame, we fuse the 2D and 3D results. For data with more than one human per image/frame, we first conduct max pooling on the 3D results of multi-human then perform late fusion with 2D. All experiments are conducted on 4 RTX 3090 GPUs.

6.2. Action Recognition

6.2.1 Verb Node Classification

To evaluate the verb node classification, we build a *Pangea* test set with 178 K images. To evaluate few/zero-shot learning, we split the 290 leave nodes into two sets and evaluate them separately: *rare* (133 leave nodes) and *non-rare* (157 leave nodes). We report the results in Tab. 1. For baseline CLIP, we load the vanilla CLIP pretrained model [59] as the backbone and train it on *Pangea* train set for node classification. We use visual-language contrastive learning

Method	mAP
R*CNN [23]	28.50
Mallya *et al.* [52]	36.10
Pairwise [16]	39.90
RelViT [51]	40.12
CLIP [59]	46.35
CLIP-*Pangea*	45.09
P2S	**47.74**

Table 2. Results on the image benchmark HICO [6].

Method	Top-1 Accuracy (%)
I3D [30]	33.53
TPN [76]	50.53
TSN [72]	55.33
EVL [45]	76.40
CLIP [59]	66.33
CLIP [59]-*Pangea*	68.27
P2S	71.40
P2S + EVL [45]	**80.87**

Table 3. Results on the video benchmark HAA [10].

Method	Top-1 Accuracy (%)(all)
TSN [72]	69.40
RGB-I3D [30]	74.30
Two-stream I3D [30]	80.90
EVL [45]	83.68
CLIP [59]	67.47
CLIP [59]-*Pangea*	67.69
P2S	68.37
P2S + EVL [45]	**85.09**

Table 4. Results on the video benchmark HMDB51 [36].

Method	Top-1 Acc (%)	Top-5 Acc (%)
TSN [72]	73.90	91.10
VideoMAE [69]	87.40	97.60
EVL [45]	87.64	97.71
CLIP [59]	72.82	91.68
CLIP [59]-*Pangea*	70.45	89.14
P2S	73.80	92.01
P2S + EVL [45]	**90.22**	**98.26**

Table 5. Results on the video benchmark Kinetics-400 [32].

Methods	Top-1%	Top-1-norm%
2s-AGCN [57]	40.00	16.00
PointNet++ [58]	42.26	24.73
CLIP [59]	32.42	9.84
PointNet++ [58]-*Pangea*	45.79	30.52
CLIP [59]-*Pangea*	48.53	32.74
P2S	**49.69**	**33.87**

Table 6. Results on the 3D benchmark BABEL-120 [57].

Methods	Top-1%
SGN [70]	53.3
PointNet++ [58]	38.6
CLIP [59]	38.0
PointNet++ [58]-*Pangea*	45.6
CLIP [59]-*Pangea*	49.3
P2S	**54.1**

Table 7. Results on the 3D benchmark HAA4D [70].

P2S	HICO[6] mAP	HAA [10] Acc (%)	HMDB51 [36] Acc (%)
✗	41.32	68.87	70.80
✓	**46.91**	**70.87**	**71.09**

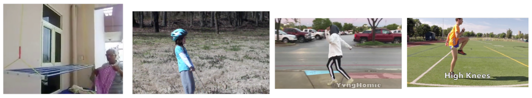
Table 8. Results of P2S + MLLM [20] on several datasets.

in training and use the same texts as P2S in inference. It achieves 28.25 mAP on 290 leave nodes (16.90 mAP for 133 rare nodes, 37.87 mAP for 157 non-rare nodes). Relatively, P2S performs much better with the help of disentanglement and semantic/geometric information. It achieves 34.25 mAP (21.56 for rare nodes, and 45.00 for non-rare nodes). Moreover, with label augmentation, P2S-aug further outperforms P2S on all three tracks.

6.2.2 Transfer Learning

We refer to the downstream benchmark as the *target*. For a fair comparison, we design several baselines: (1) CLIP: finetuning the vanilla CLIP pretrained model on the target train set and testing it on the target test set. The output is activity predictions $\mathcal{S}_{act}$, and the loss is contrastive loss $\mathcal{L}_{act}$; (2) CLIP-*Pangea*: finetuning the vanilla CLIP pretrained model on *Pangea* train set with $\mathcal{L}_{act}$, then finetuning it on the target train set, where $\mathcal{S}_{act}$ is used for evaluation on the target test set. (3) P2S: detailed in Sec. 5.4, where the output $\mathcal{S}_{act}$ is fused with the better one from CLIP/CLIP-*Pangea*.

Image Benchmark. In Tab. 2, CLIP performs well and even outperforms the ad-hoc SOTA models on HICO [6]. Pretrained on the image-text pairs from *Pangea*, CLIP-*Pangea* is inferior to CLIP because of the large domain gap between activity videos in Pangea and human-object interaction images in HICO [6]. Thus, CLIP-*Pangea* cannot utilize the extensive semantic-geometric knowledge. Relatively, P2S boosts the performance and outperforms RelViT and CLIP with **7.62** and **1.39** mAP respectively.



w/o P2S	washing_clothes	roller-skating_forward	shuffle_dance	sprint_run
w/ P2S	hang_curtain	grass_skiing	moon_walk	high_knees
GT	hanging_clothes	grass_skiing	moon_walk	high_knees

Figure 7. Result analysis of MLLM [20] w/ or w/o P2S.

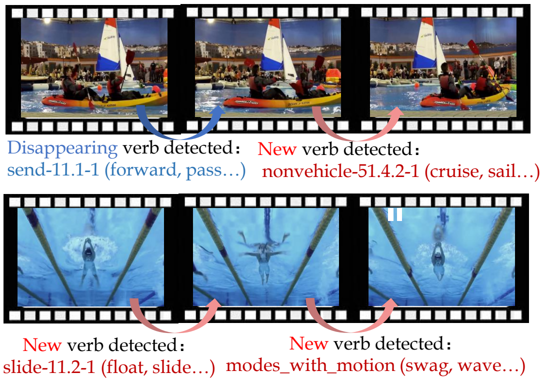

Disappearing verb detected:
send-11.1-1 (forward, pass…)
New verb detected:
nonvehicle-51.4.2-1 (cruise, sail…)
New verb detected:
slide-11.2-1 (float, slide…)
New verb detected:
modes_with_motion (swag, wave…)

Figure 8. Visualization of changed node predictions from 2 videos.

Video Benchmark. The CLIP, CLIP-*Pangea* are with the same setting as above. The conclusion is similar in Tab. 3-5. On HAA and HMDB51, CLIP-*Pangea* weaponized with *Pangea* outperforms CLIP. P2S outperforms CLIP with **5.07**%, **0.90**% and **0.98**% respectively on 3 benchmarks respectively. Moreover, P2S without bells and whistles performs comparably well (*e.g.*, TSN on HMDB51, TSN on Kinetics-400) or even better (*e.g.*, TSN on HAA) compared with ad-hoc SOTA. Lastly, fusing P2S and SOTA models further improves the performance: **4.47**% (HAA), **1.41**% (HMDB51), **2.58**% (Kinetics-400).

3D Benchmark. We set baselines PointNet++ and CLIP and strengthen them with *Pangea* as PointNet++-*Pangea* and CLIP-*Pangea* (Suppl. Sec. 7). Similarly, PointNet++-*Pangea* and CLIP-*Pangea* performs better in Tab. 6, 7. P2S

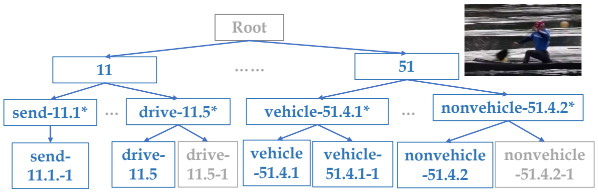


Figure 9. Hierarchical predictions of an image with action `canoeing_sprint`. P2S outputs 898 node predictions for the image, and some nodes among the top 15 highest predictions are shown in the blue blocks. P2S can learn from generic concepts (*e.g.*, `11:sending and carrying`) to finer-grained (*e.g.*, `drive-11.5`) concepts.

outperforms all the baselines, *e.g.* **7.43%** upon PointNet++ on BABEL. Moreover, P2S performs better than the ad-hoc SOTA thanks to the abundant data of *Pangea*. We do not fuse P2S with SOTA due to the modality gap: most SOTA use 3D skeleton while we use point cloud.

Integration with MLLM. As a plug-and-play method, P2S can facilitate recent powerful Multi-Modal Large Language Models (MLLM). We integrate the prediction of P2S with a SOTA MLLM backbone: LLaMA-Adapter V2 [20] on HAA [10], HICO [6] and HMDB [36]. When trained without P2S, the backbone is finetuned on the train set to output captions indicating the activity. Then the top-1 accuracy/mAP is calculated by comparing the semantic distance between output captions and GT actions based on a CLIP text encoder. When trained with P2S, the P2S prediction is converted into a prompt as known information for the MLLM.The results are shown in Tab. 8. The performance improvement shows complementary effectiveness of P2S to enhance MLLM. We also show some cases predicted by MLLM with and without P2S in Fig. 7. In the first column, though MLLM with P2S does not predict the correct action, it does predict the **correct verb** thanks to the knowledge from *Pangea*, making the prediction semantically similar to the ground truth. In other columns, with the help of P2S, MLLM succeeds in giving the correct prediction.

6.3. Further Analysis

Visualization. We analyze changed node predictions in videos in Fig. 8 and show hierarchical predictions in Fig. 9. P2S effectively captures the subtle semantic changes hierarchically. Besides, we can also conduct motion generation given *Pangea*, *i.e.*, **Semantic-to-Physical Space** (**S2P**), to fully represent its efficacy. In Fig. 10, we show the results of inputting verb nodes and use a simple cVAE to generate 3D motions, verifying that S2P is capable of generating reasonable poses for single/multi-node.

Performance Variance Analysis. a) 3D vs. 2D: P2S presents more evident performance improvement on 3D benchmarks because of the smaller-scale train/test set and smaller data domain gaps than 2D image/video benchmarks. b) Image vs. Video Benchmark: P2S performs better in image benchmark as the baseline is a concise, image-based model, rather than a sophisticated video-based one. c) Variations within Video Benchmarks: P2S shows different benefits across video benchmarks mainly because of various sizes of pre-training data and node sample distribution.

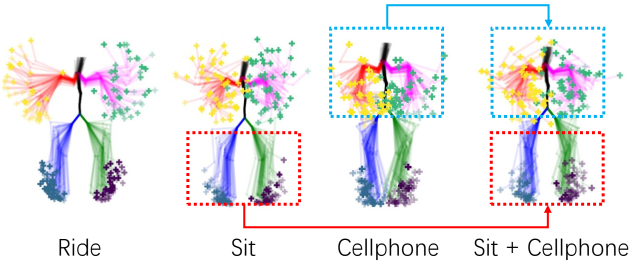


Figure 10. S2P results. `ride` has the elbows away from the spine, while `sit` has the opposite. Adding `cellphone` upon `sit` drives the wrist to distribute around the pelvis more.

Method	Full	Rare	Non-Rare
P2S-aug	**34.25**	**21.56**	**45.00**
w/o Disentanglement	30.09	18.65	39.79
w/o Semantic Augmentation	34.01	21.37	44.72
w/o Text Encoder	31.81	20.05	41.78
w/o Hyperbolic Mapping	32.56	20.49	42.78
w/o Semantic Prompt	33.20	21.00	43.54
w/o Geometric Prompt	33.81	21.20	44.49

Table 9. Ablation studies on the proposed benchmark *Pangea*.

Ablation Study & Discussion. We conduct ablations on *Pangea* to evaluate the P2S components in Tab. 9. Without four key components, P2S shows obvious degradation, which follows the gap between P2S and CLIP-*Pangea*. Moreover, semantic disentanglement matters most to facilitate the one-to-many P2S mapping and weakly-supervised learning. Here, we adopt concise models to verify the efficacy of *Pangea* and quickly trial-an-error with limited GPUs. We believe that larger and more sophisticated models trained with *Pangea* with more computing power would gain more superiority in future work. For additional results and discussions, please also refer to the supplementary.

7. Conclusion

In this work, to bridge the action data "isolated islands", we propose a structured semantic space and accordingly merge multi-modal datasets into a unified *Pangea*. Moreover, to fully use *Pangea*, we propose a concise mapping system to afford multi-modal action recognition showing superiority. We believe our framework paves a new path for future study.

8. Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grants 62306175, National Key Research and Development Project of China (No. 2022ZD0160102, No. 2021ZD0110704), Shanghai AI Laboratory, XPLORER PRIZE grants.

References

[1] Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *COLING*, 1998. 2, 3, 5

[2] Ankan Bansal, Sai Saketh Rambhatla, Abhinav Shrivastava, and Rama Chellappa. Detecting human-object interactions via functional generalization. *AAAI*, 2020. 2

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020. 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2, 4

[5] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 1, 2, 3

[6] Yu Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. Hico: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015. 1, 2, 3, 6, 7, 8

[7] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 2, 3, 4

[8] Shizhe Chen and Dong Huang. Elaborative rehearsal for zero-shot action recognition. In *ICCV*, 2021. 2

[9] Vasileios Choutas, Philippe Weinzaepfel, Jérôme Revaud, and Cordelia Schmid. Potion: Pose motion representation for action recognition. In *CVPR*, 2018. 2

[10] Jihoon Chung, Cheng hsin Wuu, Hsuan ru Yang, Yu-Wing Tai, and Chi-Keung Tang. Haa500: Human-centric atomic action dataset with curated videos. In *ICCV*, 2021. 1, 6, 7, 8

[11] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *ICML*, pages 7694–7731. PMLR, 2023. 5

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 5

[13] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 2

[14] Helen L Egger, Geraldine Dawson, Jordan Hashemi, Kimberly LH Carpenter, Steven Espinosa, Kathleen Campbell, Samuel Brotkin, Jana Schaich-Borg, Qiang Qiu, Mariano Tepper, et al. Automatic emotion and attention analysis of young children at home: a researchkit autism feasibility study. *NPJ digital medicine*, 2018. 1

[15] Bernard Ghanem Fabian Caba Heilbron, Victor Escorcia and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 2, 3

[16] Hao Shu Fang, Jinkun Cao, Yu Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *ECCV*, 2018. 7

[17] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *CVPR*, 2016. 2

[18] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, 2019. 2

[19] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *ICML*, 2018. 6

[20] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 7, 8

[21] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021. 2

[22] Spandana Gella, Mirella Lapata, and Frank Keller. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. *arXiv preprint arXiv:1603.09188*, 2016. 3

[23] Georgia Gkioxari, Ross Girshick, and Jitendra Malik. Contextual action recognition with r* cnn. In *ICCV*, 2015. 7

[24] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 2, 4

[25] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *ACMMM*, 2020. 1, 4

[26] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[28] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, 2020. 2

[29] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 2013. 1

[30] Andrew Zisserman Joao Carreira. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 7

[31] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. *arXiv preprint*, 2020. 4

[32] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 2, 6, 7

[33] Paul R Kingsbury and Martha Palmer. From treebank to propbank. In *LREC*, pages 1989–1993, 2002. 2

[34] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 4

[35] O. Kliper-Gross, T. Hassner, and L. Wolf. The action similarity labeling challenge. *TPAMI*, 2012. 2

[36] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 2, 6, 7, 8

[37] Dieu-Thu Le, Jasper Uijlings, and Raffaella Bernardi. Tuhoi: Trento universal human object interaction dataset. In *Proceedings of the Third Workshop on Vision and Language*, 2014. 3

[38] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019. 2

[39] Yong-Lu Li, Siyuan Zhou, Xijie Huang, Liang Xu, Ze Ma, Hao-Shu Fang, Yanfeng Wang, and Cewu Lu. Transferable interactiveness knowledge for human-object interaction detection. In *CVPR*, 2019. 2

[40] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *CVPR*, 2020. 2

[41] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Shiyi Wang, Hao-Shu Fang, Ze Ma, Mingyang Chen, and Cewu Lu. Pastanet: Toward human activity knowledge engine. In *CVPR*, 2020. 2

[42] Yong-Lu Li, Xinpeng Liu, Xiaoqian Wu, Yizhuo Li, Zuoyu Qiu, Liang Xu, Yue Xu, Hao-Shu Fang, and Cewu Lu. Hake: A knowledge engine foundation for human activity understanding. *arXiv preprint arXiv:2202.06851*, 2022. 1, 2

[43] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019. 2

[44] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3

[45] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. *arXiv preprint arXiv:2208.03550*, 2022. 7

[46] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *TPAMI*, 2019. 2

[47] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *CVPR*, 2020. 2

[48] Teng Long, Pascal Mettes, Heng Tao Shen, and Cees GM Snoek. Searching for actions on the hyperbole. In *CVPR*, 2020. 3

[49] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015. 3, 4

[50] Cewu Lu, Hao Su, Yonglu Li, Yongyi Lu, Li Yi, Chi-Keung Tang, and Leonidas J Guibas. Beyond holistic object recognition: Enriching image understanding with part states. In *CVPR*, 2018. 2

[51] Xiaojian Ma, Weili Nie, Zhiding Yu, Huaizu Jiang, Chaowei Xiao, Yuke Zhu, Song-Chun Zhu, and Anima Anandkumar. Relvit: Concept-guided vision transformer for visual relational reasoning. *arXiv preprint arXiv:2204.11167*, 2022. 7

[52] Arun Mallya and Svetlana Lazebnik. Learning models for actions and person-object interactions with transfer to question answering. In *ECCV*, 2016. 7

[53] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 1995. 2, 3, 5

[54] Paco Nathan. Pytextrank, a python implementation of textrank for phrase extraction and summarization of text documents. *[Online] https://github. com/DerwenAI/pytextrank*, 2016. 5

[55] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *NIPS*, 2017. 5

[56] Julia Peyre, Ivan Laptev, Cordelia Schmid, and Josef Sivic. Detecting rare visual relations using analogies. In *ICCV*, 2019. 2

[57] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with english labels. In *CVPR*, 2021. 2, 4, 6, 7

[58] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NIPS*, 2017. 4, 7

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2, 3, 5, 6, 7

[60] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 4

[61] Karin Kipper Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania, 2005. 2, 3, 4

[62] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *CVPR*, 2016. 1

[63] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, 2020. 3

[64] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014. 2

[65] Laura Smith, Nikita Dhawan, Marvin Zhang, Pieter Abbeel, and Sergey Levine. Avid: Learning multi-stage tasks via pixel-level translation of human videos. *arXiv preprint arXiv:1912.04443*, 2019. 1

[66] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[67] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021. 4

[68] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. QPIC: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, 2021. 2

[69] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 7

[70] Mu-Ruei Tseng, Abhishek Gupta, Chi-Keung Tang, and Yu-Wing Tai. Haa4d: Few-shot human atomic action recognition via 3d spatio-temporal skeletal alignment, 2022. 6, 7

[71] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 2

[72] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 7

[73] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 6

[74] An Yan, Yali Wang, Zhifeng Li, and Yu Qiao. Pa3d: Pose-action 3d machine for video recognition. In *CVPR*, 2019. 2

[75] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 2

[76] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 591–600, 2020. 7

[77] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *ICCV*, 2011. 2

[78] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, 2015. 2