Context-Aware Sequence Alignment using 4D Skeletal Augmentation

Taein Kwon[1] Bugra Tekin[2] Siyu Tang[1] Marc Pollefeys[1,2]
[1]Department of Computer Science, ETH Zürich [2]Microsoft MR & AI Lab, Zürich

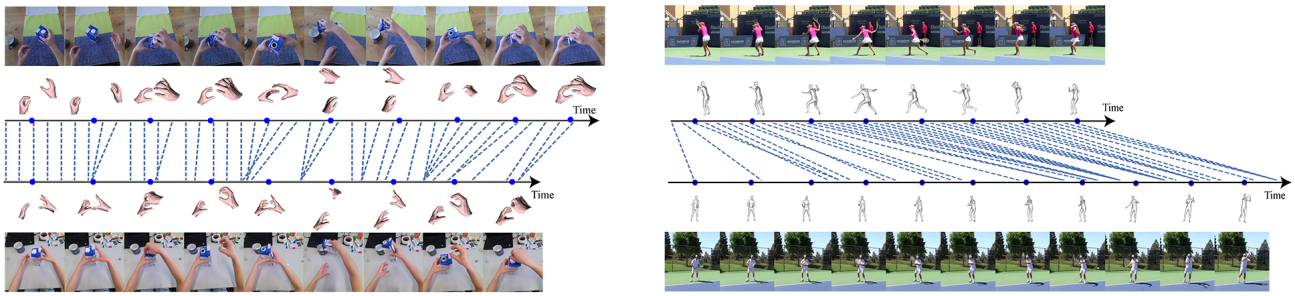


Figure 1. **Sequence Alignment.** We propose a skeletal self-supervised learning approach that uses alignment as a pretext task. Our work can align pose sequences for both hands and bodies, for which examples are shown above on the H2O [29] and PennAction [67] datasets. Our approach to alignment relies on a context-aware attention model that incorporates spatial and temporal context within and across sequences. Pose data provides a valuable cue for alignment and downstream tasks, such as phase classification and phase progression, as it is robust to different camera angles and changes in the background, while being efficient for real-time processing.

Abstract

Temporal alignment of fine-grained human actions in videos is important for numerous applications in computer vision, robotics, and mixed reality. State-of-the-art methods directly learn image-based embedding space by leveraging powerful deep convolutional neural networks. While being straightforward, their results are far from satisfactory, the aligned videos exhibit severe temporal discontinuity without additional post-processing steps. The recent advancements in human body and hand pose estimation in the wild promise new ways of addressing the task of human action alignment in videos. In this work, based on off-the-shelf human pose estimators, we propose a novel context-aware self-supervised learning architecture to align sequences of actions. We name it CASA. Specifically, CASA employs self-attention and cross-attention mechanisms to incorporate the spatial and temporal context of human actions, which can solve the temporal discontinuity problem. Moreover, we introduce a self-supervised learning scheme that is empowered by novel 4D augmentation techniques for 3D skeleton representations. We systematically evaluate the key components of our method. Our experiments on three public datasets demonstrate CASA significantly improves phase progress and Kendall's Tau scores over the previous state-of-the-art methods.

1. Introduction

Temporal alignment of human activities in videos aims to identify sequential per-frame correspondence between two video instances of the same action as shown in Fig. 1. This is challenging due to large variation in speed of actions, severe self-occlusion, and diverse backgrounds across different videos. Furthermore, an accurate temporal alignment of human activities requires semantic understanding of human motion and causal reasoning of the action stages. When it comes to hand-centric fine-grained activities under first-person views, the challenges are amplified by the varying viewpoints and embodied movement of camera wearers. State-of-the-art methods leverage large-scale datasets and powerful deep convolution neural networks to learn image-based representation to perform temporal video alignment [16, 22]. Despite rapid progress in terms of accuracy and advanced learning schemes, the results are still far from applicable to real-world applications.

Recent advancements and growing availability of head-mounted devices (e.g. Microsoft HoloLens [56]) enable new ways of communication and collaboration. For instance, the built-in hand tracking system of HoloLens provides real-time accurate hand pose estimation of the camera wearer. Such systems promise a revolution in how hand motion and actions can be captured, modeled, and analyzed. Consequently, they point towards a new way to align fine-grained hand-centric actions in videos based on 3D skeleton motion extracted from off-the-

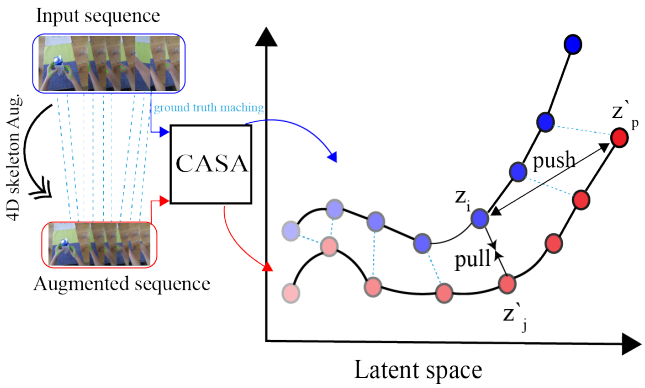


Figure 2. **Self-supervised learning using 4D augmentation.** Given a sequence and its augmentation in 4D, we optimize our latent space such that the distances between the features of matching frames (z_i and z'_j) are minimized, while those of non-matching frames (z_i and z'_p) are encouraged to be further apart.

shelf pose estimators.

One appealing application of this setting is to utilize mixed reality headsets to close the skills gap between experts and learners. Traditionally, transferring skills from experts to learners is not easy. Experts often have to stay close to learners to teach and inspect individuals. Given the videos shared by experts from their point-of-view and on-device hand pose estimation, an accurate temporal alignment method that provides dense correspondences between the fine-grained hand actions performed by the expert and the learners will enable significantly more efficient and precise skill transfer guidance.

Inspired by these observations, we propose to align 3D skeletons extracted from videos for human action alignment tasks. Instead of using 2D features used in [16, 22, 41], we propose CASA, Context-Aware Sequence Alignment, a novel context-aware self-supervised learning framework for 3D skeletons using 4D augmentation. As shown in Fig. 2, our framework reasons about the context through the attention module and performs self-supervised learning with our novel 4D augmentation strategies. From the ground-truth matching between the augmented and original sequences, we can learn powerful representation to perform downstream tasks.

Furthermore, our 3D skeleton-based alignment method not only works for hand action analysis but also can be applied to full body-related actions where we extract 3D human bodies from videos using off-the-shelf body estimators. Although, in such cases the reconstructed 3D human bodies can be less accurate than the hand tracking result from mixed reality devices, our method still generalizes well due to the novel context-aware network architecture and the self-supervised learning framework enabled by the powerful 4D augmentation schemes.

We perform extensive experiments to validate the effectiveness and applicability of CASA on three public datasets: Penn Action [67], IKEA ASM [2], and H2O [29]. CASA achieves the best performance in most phase classification tasks of three datasets. Furthermore, in terms of phase progress and Kendall's tau, our method significantly outperforms the previous state-of-the-art methods [16, 22]. The results demonstrate the importance of knowing the context of action and the applicability of utilizing 3D poses for fine-grained video alignment tasks.

Contributions. In summary, our contributions are: **(1)** we propose a novel attention-based and context-aware dense alignment framework for fine-grained human action analysis; **(2)** we introduce novel 4D augmentation strategies for 3D skeletons in self-supervised learning that consider both temporal and spatial augmentation; **(3)** to the best of our knowledge, it is the first work to perform 3D skeleton-based fine-grained video alignment using self-supervised learning. We prove the utility of our 3D skeleton-based temporal alignment methods by largely outperforming the state-of-the-art in three public datasets.

2. Related Work

Self-Supervised Learning. Several image-based self-supervised learning methods have been proposed recently that rely on different hand-crafted pretext tasks. For example, recent work used image colorization [30], solving jigsaw puzzles [38, 62], rotation prediction [19] or image inpainting [26] as pretext tasks to train self-supervised models. These hand-crafted tasks rely on particular adhoc heuristics, which limits their generalization power. Alternatively, contrastive learning approaches learn representations by contrasting positive pairs against negative pairs [15, 23, 35, 55, 64, 71]. Notably, Chen et al. [7] demonstrated that composition of multiple data augmentation operations is crucial in defining the contrastive prediction tasks that yield effective representations for single image data. Inspired by the success of self-supervised methods in image domain, recently several self-supervised learning methods were proposed for videos, either using pretext tasks, such as predicting future frames [1, 12, 48, 59], clip order [18, 31, 36, 65], pace [3, 8, 61, 66] or arrow of time [40, 63], or focusing on instance-based contrastive learning techniques [11, 17, 25, 42].

Compared to image and video-based self-supervised learning, skeleton-based self-supervised learning started to emerge as an active field only recently. Proxy tasks such as skeleton inpainting [68] and motion prediction [50] have been proposed by recent work. However, such methods do not explicitly account for the spatio-temporal dependencies of skeletal representations. Skeletal self-supervised learning techniques that rely on neighborhood consistency [47], fusion of multiple pretext tasks [32] and motion continuity [51] have also shown the promise of self-supervised techniques for learning skeletal sequence representations. Unlike previous approaches, we propose a self-supervised learning framework with a compo-

sition of 4D data augmentation strategies. We consider both temporal and spatial transformations of the data, globally for the skeletal motion, and locally for individual joints.

Transformer. After the success of the Transformer architecture [58] in Natural Language Processing (NLP), there has been a surge in interest in its application for computer vision. Several Transformer-based architectures have been proposed for image classification [14], object detection [5], and semantic segmentation [60]. More closely related to our work, Sun et al. and Sarlin et al. [45,53] proposed Transformers for the task of image alignment. While Transformers have been actively used within supervised learning contexts, recent work also has shown the potential of self-supervised pretraining of a standard Vision Transformer model for several downstream tasks [6]. Correspondingly, in this work, we propose a self-supervised Transformer architecture for fine-grained alignment of videos.

Sequence Alignment. Dynamic Time Warping (DTW) has become the *de facto* standard for unsupervised sequence matching due to its simplicity and generality for different types of modalities [4]. Cuturi and Blondel [10] proposed a differentiable approximation of DTW which allows for pairing it with neural networks and training sequence models. Canonical Time Warping [70] and Generalized Time Warping [69] generalized DTW and enabled alignment of signals with different dimensionality. As an alternative to DTW, Su et al. [49], relied on optimal transport to match two sequences frame-by-frame, while regularizing the loss such that temporal information is preserved in the matching process. While focusing on the alignment problem, these approaches do not aim at feature learning for sequence matching unlike our work.

Closely related to the sequence alignment problem, metrics for assessing human motion similarity have been actively explored by previous studies [4, 9, 13, 33, 34, 37, 52, 54, 54]. The assessment of the similarity between two sequences of poses or motion is a non-trivial problem since human motion varies across sequences due to a number of different factors such as speed, anthropometric variations, and subject-specific pose patterns. Conventional approaches for measuring similarity of human motion sequences are based on estimating the L2 displacement error [13, 34] or DTW [4]. However, these metrics disregard contextual information in the time dimension, which limits their application for human motion analysis. To overcome the limitations of standard metrics, deep metric learning methods have been proposed by [9, 37, 52, 54].

In the context of self-supervised learning-based video alignment [16, 20, 22, 46], Time Contrastive Networks (TCN) [46] used synchronized frames with contrastive learning to align frames from different points of view. Temporal Cycle Consistency (TCC) method [16] learned an embedding space that maximizes one-to-one mapping of cycle-consistent points across pairs of video sequences. Learning by Aligning Videos (LAV) [22] adopted soft-DTW [10] as a self-supervised temporal alignment loss. Unlike our work that aims at self-supervised skeletal sequence learning, these works all focused on matching images across videos.

3. Method

Fig. 3 shows an overview of our proposed pipeline. We propose a self-supervised skeletal representation learning approach that uses skeletal alignment as a pretext task. Our model relies on an attention-based context-aware framework for sequence alignment. Our self-supervised loss, inspired by the success of image-based contrastive learning [7], relies on minimizing the difference between a skeletal sequence and its augmentation in 4D, that is, in 3D space and time. Our framework learns a representative latent space, which is effective in downstream tasks and can be used to align two skeletal sequences via nearest-neighbor search.

Notations. Each 3D skeleton of a sequence is defined as $s_i \in \mathbb{R}^{J\times 3}$ with J skeleton joints in x, y, z locations. Each k-th ($1 \leq k \leq L$) sequence of skeletons is shown with $S_k = \{s_1, s_2, .., s_M\}$ and its augmentation is shown with $S'_k = \{s'_1, s'_2, .., s'_N\}$. The embedding of a skeletal sequence is computed as $(U_k, U'_k) = \Phi(S_k, S'_k; \Omega)$, where Φ is our framework's encoder network with the parameters, Ω. The embedding of the original sequence, U_k, is denoted with $\{u_1, u_2, ...u_M\}$ and that of the augmented sequence, U'_k, is denoted with $\{u'_1, u'_2, ...u'_N\}$. Our latent space in which we optimize an alignment loss is denoted with ($Z_k = \{z_1, z_2, ...z_M\} = P(U_k)$ and $Z'_k = \{z'_1, z'_2, ...z'_N\} = P(U'_k)$), where $P(\cdot)$ is a projection head [7]. Note that we use upper-case notations for sequence-level processing and lower-case notations for per-frame processing.

3.1. Preliminaries

3D human body representation. We use SMPL [39] on the Penn Action dataset and Keypoint RCNN [24] body joints representation, $s_{rcnn}, \in \mathbb{R}^{17\times 3}$ on the IKEA dataset. Pose parameters, $\theta_{smpl} \in \mathbb{R}^{72}$, store angles for 22 skeleton joints along with a global rotation and translation vector. We remap the 22 SMPL skeleton joints to the skeleton representation of FrankMocap [44], $s_{smpl} \in \mathbb{R}^{25\times 3}$, to be able to use the FrankMocap estimator. We recover 3D body skeleton, s_{smpl}, based on $SMPL(\beta_{smpl}, \theta_{smpl})$, where $SMPL(\cdot)$ is the function that calculates 3D skeleton, given shape, β_{smpl}, and pose, θ_{smpl}, parameters.

3D hand representation. We use MANO [43] 3D hand skeleton representation $s_{mano} \in \mathbb{R}^{42\times 3}$ in the H2O dataset. MANO contains human hand shape parameters, $\beta_{mano} \in \mathbb{R}^{20}$, and pose parameters, $\theta_{mano} \in \mathbb{R}^{34\times 3}$, storing angles for 30 skeleton joints, 2 global rotation and 2 translation vectors for both hands. We recover 3d hand skeleton s_{mano} from $MANO(\beta_{mano}, \theta_{mano})$, where $MANO(\cdot)$ is a function that calculates 3D hand skeletons given shape (β_{mano}) and pose parameters (θ_{mano}).

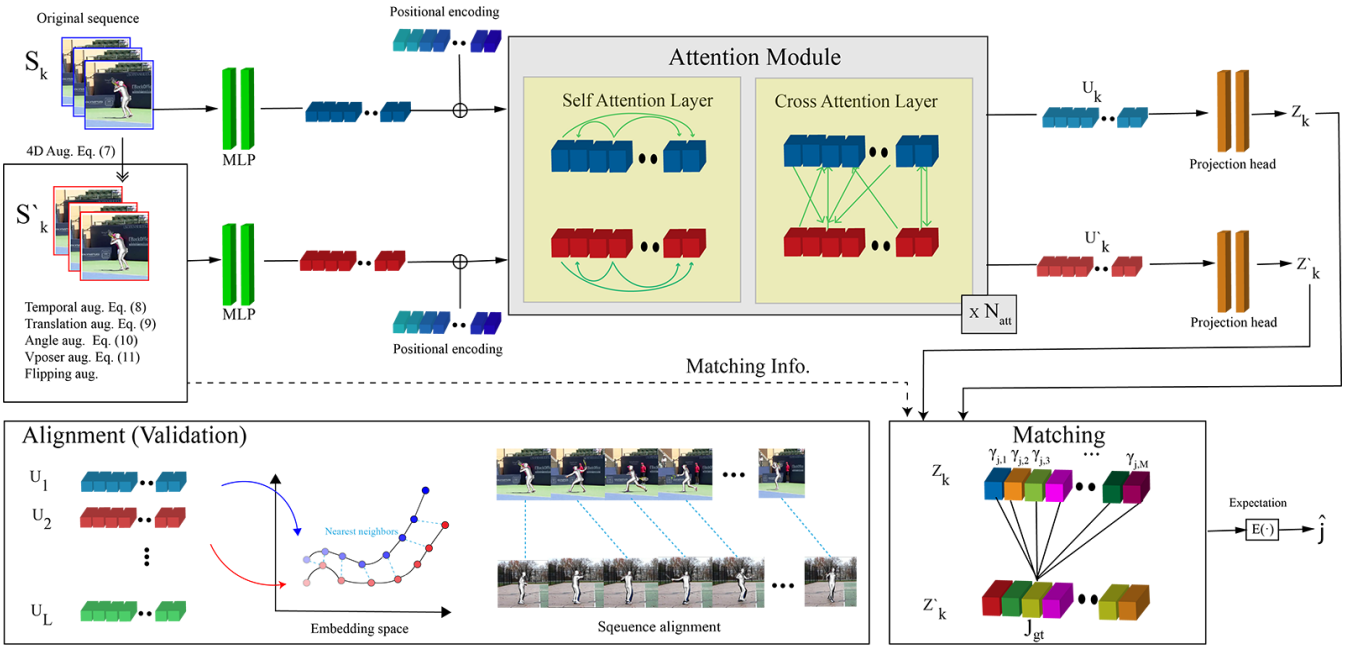


Figure 3. **Overview of our pipeline.** The proposed framework takes as input a skeleton sequence S_k along with its spatio-temporally augmented version S'_k. Both sequences are encoded by temporal positional encodings. Self- and cross-attentional layers learn contextual information within and across sequences with the help of temporal positional encoding. We employ a projection head to improve our representation quality [7] . We use a *contrastive regression loss* that matches a pose sequence with its 4D augmented version. For the downstream tasks and alignment, we use the embeddings before the projection head stage.

We will call, s_{smpl}, s_{rcnn}, and s_{mano} as s and omit β from the following equations for simplicity. Also, we will use the transform function $T(\cdot)$ instead of $MANO(\cdot)$ and $SMPL(\cdot)$. Accordingly, we will transform each pose parameter to 3D skeletons using $S_k = T(\Theta_k)$.

3.2. Model architecture

Our model consists of multi-layer perceptrons (MLP), positional encodings, an attention module, and projection heads. In the following, we will explain each part of the model.

MLP. We use two nonlinear layers of a fully connected network with the same input dimension to extract features from our 3D joint representation before feeding them input to the attention module.

Attention module. Transformer [58] has received a lot of attention due to its impressive performance in the NLP field, as summarized in Section 2. To leverage the power of Transformers in temporal understanding, we employ self- and cross- attention layers, that efficiently capture temporal context as compared to methods that compute features from single-images [16,22]. We model self-attention to learn dependencies within the skeletons in the same sequence and cross-attention to learn the inter-dependencies between the original sequences and their 4D augmentations. To reduce the computational complexity of attention layers, we adopt Linear Transformer [28] architecture. By contrast to the earlier Transformer-based works [45, 53], our attention module is integrated into a self-supervised learning framework that uses 4D augmentations for sequence matching.

Temporal positional encoding. We inject temporal information into our framework using positional encodings [58]. Using positional encodings, our model reasons about temporal locations of each skeleton frame. Such information is crucial in understanding temporal dependencies between skeletons. Different from other vision-based tasks [14,53], we only need 1D positional encodings as the order of joints in the skeleton is fixed. We choose sinusoidal positional encoding as it is proved to be effective in machine translation, which can be conceptually similar to aligning two skeletal sequences from the same activity.

$$PE_i = \begin{cases} sin(w_l \cdot i), i = 2l \\ cos(w_l \cdot i), i = 2l+1 \end{cases}, \quad (1)$$

where $w_l = \frac{1}{5000^{(2l/d)}}$ and d is the dimension of the skeleton joints, i is an index for the temporal frame location in the sequence. We choose 5000 for the denominator of w_l since the maximum length of the sequences in our case is bounded by 5000.

Projection head. To improve the quality of our representation, we employ a projection head as in [7]. As shown by [7], without the projection head, the learned model is more likely to overfit to the optimization task. While we optimize

for the alignment, we aim to have representative features for downstream tasks that address fine-grained action recognition. Therefore we use a projection head, $p(\cdot)$, in the form of an MLP with one hidden layer.

$$z_i = p(u_i) = W^2\sigma(W^1 u_i), \tag{2}$$

where σ is a ReLU layer and W^1 and W^2 are fully connected layers. We show that the projection head improves our accuracy in downstream tasks in Section 4.4.

Matching and loss. Given an original sequence and its augmentation in the time dimension, the temporal correspondences between two sequences are already known and preserved. Note also that 3D geometric augmentation will not affect the correspondences between two sequences as the 3D perturbations we use for data augmentation are time-independent. Our self-supervised learning framework, inspired by recent advances in contrastive learning [7,21], learns representations by maximizing the agreement between positive pairs, which we take, in our case, as a skeletal sequence and its 4D augmentation. We formulate the contrastive loss for positive pairs, (i, j), using the following equation:

$$\mathcal{L}_{i,j} = -log\frac{exp(-\left\|z_i - z_j\right\|/\lambda_{temp})}{\sum_{m=1}^{N} exp(-\left\|z_i - z_m\right\|/\lambda_{temp})}, \tag{3}$$

where λ_{temp} is a temperature parameter. However, the classification-based loss can not reason about how far the prediction for a matched frame is, from the ground-truth alignment. Therefore, instead of using Equation 3, we adopt a regression loss [16] to penalize nearby frames less by accounting for the temporal relationships of neighboring frames. The difference from [16] is that we compute this loss for every frame to gather contextual information from the whole sequence, instead of using frames only from a local neighborhood. The probability of a frame, i, in the original sequence, being a match to a frame j, in the augmented sequence, is denoted with $\gamma_{j,i}$ and computed by

$$\gamma_{j,i} = \frac{e^{-\left\|z'_j - z_i\right\|/\lambda_{temp}}}{\sum_{m=1}^{M} e^{-\left\|z'_j - z_m\right\|/\lambda_{temp}}}, \tag{4}$$

where $\gamma_{j,i}$ is i-th value of probability γ_j. We then predict the target frame index, $\hat{j}$, by weighing the frame indices with their corresponding probabilities, as follows:

$$\hat{j} = \sum_{i}^{M}(\gamma_{j,i} \cdot i), \tag{5}$$

The final loss $\mathcal{L}$ will be the mean squared error between the predicted frame index $\hat{j}$ and ground truth frame index j_{gt}, which is already known and preserved after data augmentation.

$$\mathcal{L} = \frac{1}{N}\sum_{j}^{N}\left\|j_{gt} - \hat{j}\right\|^2, \tag{6}$$

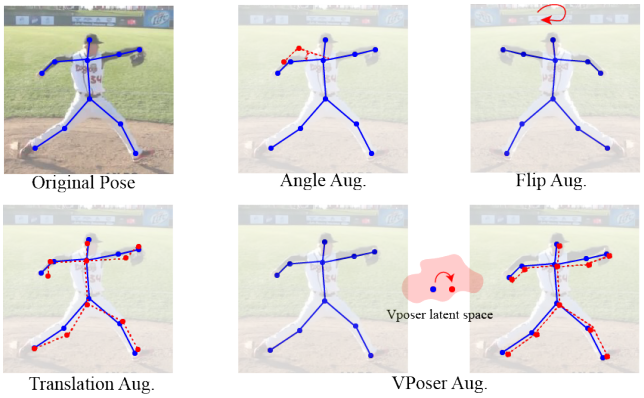


Figure 4. **Different types of 3D geometric augmentation.** Translation augmentation tackles the noisy estimates of off-the-shelf pose estimators. We observe that different augmentation strategies produce feasible poses that provide positive matches for the self-supervised learning framework.

3.3. 4D Augmentation

To be able to create positive pairs of skeletal sequences with known correspondences, we propose to augment the skeletal sequences in 3D space and time. We illustrate our proposed 4D augmentation strategies in Fig. 4. We propose 5 different augmentation schemes: temporal augmentation, joint angle augmentation, translation augmentation, skeleton flipping, and augmenting the latent space of skeletons based on VPoser [39]. We perform the augmentation by adding noise to each skeletal joint translation or angle. To be able to generate realistic augmentations of skeletons that would reflect different variations of motion, we propose to add temporally smoothed noise across the sequence, using a multivariate normal distribution, which has a covariance matrix that contains high correlations along the diagonal such that temporally closer points are highly correlated. We provide further details about the distribution we employ for temporally smoothed noise in our supplemental material. We apply temporally smoothed noise for augmenting joint angles and the latent space obtained by VPoser. This strategy overall enables the motion to be continuous and smooth over time. The augmentation function $G(\cdot)$ is defined as:

$$S'_k = G_{temp,trans,flip}(T(G_{VPoser,angle}(\Theta_k))), \tag{7}$$

In what follows, we describe our different augmentation strategies in more detail.

Temporal augmentation. We randomly select N frames in the original M frames. Through this step, our self-supervised learning framework learns the different and variable speeds of action within a sequence.

$$\{s'_1, s'_2, ..., s'_N\} = G_{temp}(\{s_1, s_2, .., s_M\}), \tag{8}$$

Translation augmentation. We employ translation augmentation to deal with noise coming from inaccuracies in 3D pose

estimation.

$$S'_k = G_{trans}(S_k) = S_k + \mathcal{N}(\sigma), \tag{9}$$

where $\mathcal{N}(\sigma)$ produces a uniform distribution noise with a standard deviation of σ.

Flipping. As our body is mirror-symmetric, we propose a flipping strategy. The flipping function $G_{flip}(\cdot)$ flips left body joints to right in the spatial coordinates and vice versa.

Angle augmentation. To perform data augmentation, on joint angles we compute

$$\Theta'_k = G_{angle}(\Theta_k) = \Theta_k + \mathcal{MN}(C), \tag{10}$$

$\mathcal{MN}(C)$ denotes a multivariate normal distribution with covariance matrix C which contains high correlations along the diagonal, as explained above.

VPoser Augmentation. VPoser [39] presents a method to learn an embedding space of plausible human poses. We leverage this latent space to further generate matching pairs of skeletal sequences by data augmentation. To this end, we map our pose with VPoser into the latent space and sample nearby location in the latent space. The augmented latent space is then decoded back to the human pose.

$$\Theta'_k = G_{vposer}(\Theta_k) = V_{dec}(V_{enc}(\Theta_k) + \mathcal{MN}(C)), \tag{11}$$

Here, we use the same distribution, $\mathcal{MN}$, for angle augmentation. $V_{enc}(\cdot)$ and $V_{dec}(\cdot)$ correspond to the encoder and decoder of VPoser, respectively.

3.4. Implementation Details

To be robust to different skeleton sizes, we scale the bone length between the chest joint and pelvis joint to the unit length and resize all the other limb lengths accordingly. We set the chest as the origin of our coordinate system to normalize for translation. We align the bone between the chest and pelvis to the z-axis and the bone between the chest and right shoulder to the y-axis, to account for variations in rotation. We perform a similar normalization for hand skeletons.

We rely on the TCC [16] code to reproduce their results for the experiments on the H2O dataset and pose-based alignment, following the same hyperparameters, described in [16]. We provide further details for the parameters of our framework in the supplemental material.

4. Evaluation

In this section, we first describe the datasets and the corresponding evaluation protocols. We then provide a detailed analysis of our approach, CASA, and compare our approach against the state-of-the-art methods.

4.1. Datasets

We verify our model on Penn Action [67], IKEA ASM [2], and H2O [29] datasets. Penn Action is a sports activity dataset. Following previous work [16, 22], we use the subset of 13 activities for evaluation. We precisely follow earlier work [16, 22] for training and test splits. IKEA ASM [2] dataset consists of 371 videos that demonstrate the assembly of four different furniture types. Similarly with LAV [22], we conduct our experiments using *Kallax_Drawer_Shelf* assembly videos (61 for training and 29 for validation). H2O [29] is a recent egocentric action recognition and hand-object interaction dataset that provides ground-truth 3D poses for left & right hands and 6D object poses, along with interaction labels. On this dataset, we select video sequences from the activity, *pouring milk*, which contains monotonic sub-actions. Among 10 subjects performing the action, we select 7 for the training set (27 videos) and 3 for the validation set (11 videos). The sequences have up to 865 frames, and we annotate 10 different phases based on the original action labels, which are only used for the evaluation purposes. We will make these new labels for sequence alignment publicly available. While we use the full-body pose as our input modality in Penn Action and IKEA ASM datasets, we use hand pose as input for the H2O dataset. Particularly for the H2O dataset, our method demonstrates an application of skeletal alignment for hands from an egocentric view, which is highly relevant for augmented reality scenarios. Since the Penn Action dataset does not provide 3D human poses, we estimate the 3D joints of the body using a state-of-the-art body pose estimator [27, 44].

4.2. Evaluation Metrics

Following literature [16, 22], we use three different metrics for our evaluation. We first train our network on the training set without using any labels and then evaluate the performance of our approach using the trained embeddings.

Phase Classification Accuracy is the per-frame classification accuracy for fine-grained action recognition. To evaluate this metric, we train an SVM classifier on a limited subset of the training data to predict phase labels.

Phase Progression measures how well the *progress* of a process or action is captured by the embeddings. We follow previous work [16] to use a linear regressor on the embeddings to predict the phase progression values. It is computed as the average $R-$squared measure, given by

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}\,, \tag{12}$$

where y_i is the ground truth phase progress value, $\bar{y}$ is the mean of all y_i and $\hat{y}_i$ is the prediction made by the linear regression model. The maximum value of this measure is 1.

Kendall's Tau [16] is a statistical measure that can determine how well-aligned two sequences are in time. It is in the range

Dataset	Method	Pose	ImageNet pre-trained	% of Labels →		
				0.1	0.5	1.0
Penn Action [67]	SaL [36]	·	✓	74.87	78.26	79.96
	TCN [46]	·	✓	81.99	83.67	84.04
	TCC [16]	·	✓	79.72	81.11	81.35
	LAV [22]	·	✓	83.56	83.95	84.25
	TCC [16]	✓	·	79.53	83.75	84.51
	LAV [22]	✓	·	79.83	80.20	80.20
	CASA (ours)	✓	·	**88.55**	**91.87**	**92.20**
IKEA ASM [2]	TCC [16]	·	✓	27.74	25.70	26.80
	LAV [22]	·	✓	**29.78**	29.85	30.43
	TCC [16]	✓	·	11.95	13.53	18.60
	LAV [22]	✓	·	14.52	16.31	18.63
	CASA (ours)	✓	·	21.32	**31.52**	**31.06**
H2O [29]	TCC [16]	·	✓	43.30	52.48	52.78
	LAV [22]	·	✓	23.48	36.41	36.38
	TCC [16]	✓	·	30.40	40.20	42.70
	LAV [22]	✓	·	37.05	39.50	40.45
	CASA (ours)	✓	·	**43.50**	**62.51**	**68.78**

Table 1. **Phase classification results.** We compare our phase classification accuracy to those of both RGB and pose based methods on three different datasets. Our method produces the state-of-the-art results in most cases.

of $[-1, 1]$ where a value of 1 implies that the videos are perfectly aligned, while a value of -1 implies that the videos are aligned in reverse order. Since this metric assumes a strictly monotonic order of the actions, it is evaluated only on the Penn Action dataset.

4.3. Comparison to the State-of-the-Art

We compare our self-supervised skeletal sequence learning approach against several different approaches [16, 22, 36, 46, 54], including the recent self-supervised video representation learning techniques, TCC [16], and LAV [22], that use alignment as a pretext task. Previous approaches do not report results using pose data as input. Therefore we reproduce the results of these baselines to be able to benchmark our results against them by following the implementation details of [16, 22]. Using precisely the same feature extractor for processing poses, we compare our approach against them. For feature extraction, we use two non-linear fully connected layers which have the same dimension with our input to keep the same amount of information. We did our best to make fair comparison by following the same hyperparameters from LAV [22] and TCC [16] except for the learning rate, which we set as 0.00005 for the image, and 0.0005 for pose, as we observed better convergence with these learning rates for different input modalities.

We compare our phase classification accuracy to the state-of-the-art [16, 22, 36, 46] in Table 1. We significantly outperform the existing approaches for all datasets and for all the fractions of labels that are used to train the classifier, except for the case of training with 10% of the labels in the IKEA ASM dataset. Limited performance for 10% of the labels in the IKEA ASM dataset is due to the noisy pose estimates on this dataset resulting from object occlusions, differences in viewpoints (*e.g.* sitting vs standing during furniture assembly),

Method	Pose	ImageNet pre-trained	Progress	τ
TCN [46]	·	✓	0.6762	0.7328
SaL [36]	·	✓	0.5943	0.6336
Pr-VIPE [54]	✓*	·	·	0.7476
TCC [16]	·	·	0.4304	0.4529
LAV [22]	·	·	0.3853	0.4929
TCC [16]	·	✓	0.6638	0.7012
LAV [22]	·	✓	0.6613	0.8047
Hadji [20]	·	✓	·	0.7829
TCC [16]	✓	·	0.6268	0.6267
LAV [22]	✓	·	0.6404	0.6983
CASA (ours)	✓	·	**0.9449**	**0.9728**

Table 2. **Video progress and Kendall's tau results.** We compare our method to other RGB and pose based methods. Note that * uses 2D poses. Our method achieves the best results on the Penn Action dataset.

and missing hand poses that provide informative cues for the assembly task. For TCC [16] and LAV [22], the pose input results in lower accuracy than the image input on the IKEA ASM dataset due to missing contextual information related to object interactions. Yet, our approach achieves better overall accuracy than the existing approaches that either use image or pose as input on this dataset. Our method reasons about fine-grained actions, both, by accounting for contextual information through our transformer-based self- and cross-attention mechanism, and, by exploiting 3D poses, which provide a detailed understanding of subtle human motions.

In Table 2, we further report our phase progression and Kendall's tau results as compared to the state-of-the-art. Remember that these metrics respectively measure how well the progress of an action is and how well aligned two sequences are in time. Our approach outperforms earlier approaches on these metrics by a large margin (0.27 improvement on phase progression and 0.17 improvement on Kendall's tau). We attribute this to the fact our method exploits positional encodings to encode temporal frame location which is a valuable cue for understanding the progress and alignment of actions. Using an attention-based architecture, our method gathers contextual information from the whole sequence during alignment which results in superior accuracy than previous approaches that rely on *only* local context. We provide further quantitative results of our approach when using different fractions of the frames from the full sequence as well as results for online sequence alignment in our supplemental material.

4.4. Ablation Studies

In Table 3, we provide an ablation study on the Penn Action dataset to analyze the influence of different network components. All our design choices consistently improve our overall accuracy. The improvement is particularly pronounced for positional encodings and attention layers. While positional encodings provide local information about frame location, atten-

Method	Classification(%)	Progress	τ
w/o positional encoding	69.01	0.3361	0.3415
w/o projection head	89.87	0.8852	0.9713
w/o self attention layers	91.24	0.9193	0.9310
w/o cross attention layers	92.04	0.9316	0.9616
All	**92.20**	**0.9449**	**0.9728**

Table 3. **Influence of different components of our model.** We ablate on the Penn Action dataset to analyze our different design choices.

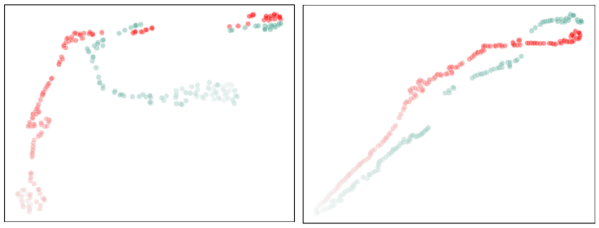
Method	Classification(%)	Progress	τ
No Aug.	89.95	0.8729	0.9653
Temp. Aug.	91.78	0.9446	0.9621
w/o Ang.	92.75	0.9397	0.9719
w/o Trans.	92.64	0.9338	0.9722
w/o Vposer	92.64	0.9379	0.9710
w/o Flip	**92.94**	0.9414	0.9710
All	92.20	**0.9449**	**0.9728**

Table 4. **Ablation study for 4D augmentation.** The best result is depicted in bold. We ablate on the Penn Action dataset to analyze different data augmentation strategies.

tion layers help gather contextual information within the same sequence and across two sequences. The projection head also results in a considerable improvement in accuracy for all the metrics, showing the importance of the nonlinear mapping before applying a self-supervised loss, in line with the recent literature on self-supervised learning [7].

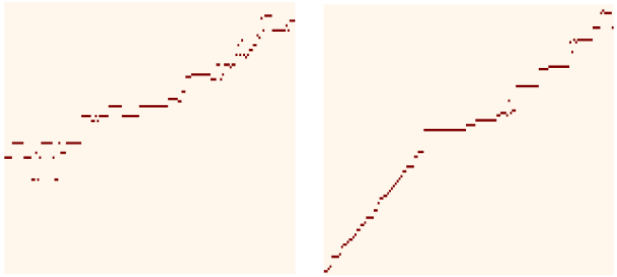
We present further ablation studies on the influence of different types of data augmentation strategies in Table 4. All the augmentation strategies, combined together, results in consistently high accuracies for all the metrics. While the temporal augmentation results in about 2% increase in the phase classification accuracy, 3D spatial augmentation brings in another 1% improvement in phase classification and Kendall's Tau, which demonstrates the individual contributions and complementary nature of different strategies.

We present the t-SNE embeddings [57] of the representation learned by CASA on two different sequences in Fig. 5. Color scale demonstrates the corresponding time frames of a sequence, from start to end. We demonstrate that our approach learns a smooth representation, in which temporally close frames are mapped to nearby positions in the embedding space. Furthermore, the corresponding frames across the two videos are embedded in similar locations. This structure of the embedding space demonstrates the potential and reliability of our method for sequence alignment. We show qualitative examples of alignment between two sequences in Fig. 1. More qualitative results can be found in our supplemental material. We further show frame-wise matches across two sequences, in comparison to TCC, in Fig. 6. We observe that CASA preserves the temporal context and results in smoother alignments.



TCC OURS

Figure 5. **t-SNE visualization of the embedding space learned by CASA.** For this visualization, we select two different sequences from *baseball_pitch*. Our method is able to preserve temporal context and align corresponding frames across videos.



TCC CASA

Figure 6. **Alignment between two sequences.** The x-axis is the time frame of the source sequence and the y-axis is the time frame of the target sequence. We show the closest matching frames across source and target sequences. For visualization, we select two different sequences from *baseball_pitch*. We observe that CASA preserves the temporal context and results in smoother alignments.

5. Conclusion

In this paper, we propose a self-supervised learning framework that uses skeletal sequence alignment as a proxy task. The proposed CASA approach uses the self and cross attention layers in Transformers to transform the local features to be context- and position-dependent, which is crucial for CASA to obtain high-quality sequence alignments. We further propose to augment the skeletal sequences in 3D space and time to generate examples for matching and training a self-supervised loss to minimize alignment score across sequences. Our experiments show that CASA achieves state-of-the-art performances on phase action classification, phase progression, and Kendall's tau scores on multiple datasets.

Our method, CASA, relies on off-the-shelf pose estimators to compute human pose, which is used as an input to our framework for alignment. Wrong predictions of the off-the-shelf pose estimator will result in inaccuracies in sequence alignment, which is a limitation of our approach. End-to-end learning from RGB images for skeletal alignment using a pre-trained pose estimator would be an interesting future direction to overcome this limitation.

Acknowledgements. Taein Kwon was supported by the Microsoft MR & AI Zürich Lab PhD scholarship. The authors thank Jonas Hein, Mihai Dusmanu, Paul-Edouard Sarlin, Luca Cavalli, Yao Feng, and Weizhe Liu for helpful discussions.

References

[1] Unaiza Ahsan, Chen Sun, and Irfan Essa. Discrimnet: Semi-supervised action recognition from videos using generative adversarial networks. *arXiv preprint arXiv:1801.07230*, 2018. 2

[2] Yizhak Ben-Shabat, Xin Yu, Fatemehsadat Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. 2020. 2, 6, 7

[3] Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. Speednet: Learning the speediness in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9922–9931, 2020. 2

[4] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, USA:, 1994. 3

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 3

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 3

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 3, 4, 5, 8

[8] Hyeon Cho, Taehoon Kim, Hyung Jin Chang, and Wonjun Hwang. Self-supervised spatio-temporal representation learning using variable playback speed prediction. *arXiv preprint arXiv:2003.02692*, 3(6):7, 2020. 2

[9] Huseyin Coskun, David Joseph Tan, Sailesh Conjeti, Nassir Navab, and Federico Tombari. Human motion analysis with deep metric learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 667–683, 2018. 3

[10] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *International Conference on Machine Learning*, pages 894–903. PMLR, 2017. 3

[11] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *arXiv preprint arXiv:2101.07974*, 2021. 2

[12] Vivek Diba, Ali ad Sharma and Rainer Van Gool, Luc ad Stiefelhagen. Dynamonet: Dynamic action and motion network. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2

[13] Junting Dong, Qing Shuai, Yuanqing Zhang, Xian Liu, Xiaowei Zhou, and Hujun Bao. Motion capture from internet videos. In *European Conference on Computer Vision*, pages 210–227. Springer, 2020. 3

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 4

[15] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27:766–774, 2014. 2

[16] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1801–1810, 2019. 1, 2, 3, 4, 5, 6, 7

[17] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3299–3309, 2021. 2

[18] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3636–3645, 2017. 2

[19] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2

[20] Isma Hadji, Konstantinos G Derpanis, and Allan D Jepson. Representation learning via global temporal alignment and cycle-consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11068–11077, 2021. 3, 7

[21] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006. 5

[22] Sanjay Haresh, Sateesh Kumar, Huseyin Coskun, Shahram N Syed, Andrey Konin, Zeeshan Zia, and Quoc-Huy Tran. Learning by aligning videos in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5548–5558, 2021. 1, 2, 3, 4, 6, 7

[23] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 2

[24] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[25] Kai Hu, Jie Shao, Yuan Liu, Bhiksha Raj, Marios Savvides, and Zhiqiang Shen. Contrast and order representations for video self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7939–7949, 2021. 2

[26] Simon Jenni, Hailin Jin, and Paolo Favaro. Steering self-supervised feature learning beyond local pixel statistics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6408–6417, 2020. 2

[27] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. *3DV*, 2021. 6

[28] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive

transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020. 4

[29] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10138–10148, October 2021. 1, 2, 6, 7

[30] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Learning representations for automatic colorization. In *European conference on computer vision*, pages 577–593. Springer, 2016. 2

[31] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017. 2

[32] Lilang Lin, Sijie Song, Wenhan Yang, and Jiaying Liu. Ms2l: Multi-task self-supervised learning for skeleton based action recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2490–2498, 2020. 2

[33] Jingyuan Liu, Mingyi Shi, Qifeng Chen, Hongbo Fu, and Chiew-Lan Tai. Normalized human pose features for human action video alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11521–11531, 2021. 3

[34] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017. 3

[35] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6707–6717, 2020. 2

[36] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016. 2, 7

[37] Greg Mori, Caroline Pantofaru, Nisarg Kothari, Thomas Leung, George Toderici, Alexander Toshev, and Weilong Yang. Pose embeddings: A deep architecture for learning to match human poses. *arXiv preprint arXiv:1507.00302*, 2015. 3

[38] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 2

[39] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 5, 6

[40] Lyndsey C Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Scholkopf, and William T Freeman. Seeing the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2035–2042, 2014. 2

[41] Senthil Purushwalkam, Tian Ye, Saurabh Gupta, and Abhinav Gupta. Aligning videos in space and time. In *European Conference on Computer Vision*, pages 262–278. Springer, 2020. 2

[42] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021. 2

[43] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), Nov. 2017. 3

[44] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021. 3, 6

[45] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 3, 4

[46] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1134–1141. IEEE, 2018. 3, 7

[47] Chenyang Si, Xuecheng Nie, Wei Wang, Liang Wang, Tieniu Tan, and Jiashi Feng. Adversarial self-supervised learning for semi-supervised 3d action recognition. In *European Conference on Computer Vision*, pages 35–51. Springer, 2020. 2

[48] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852. PMLR, 2015. 2

[49] Bing Su and Gang Hua. Order-preserving wasserstein distance for sequence matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1049–1057, 2017. 3

[50] Kun Su, Xiulong Liu, and Eli Shlizerman. Predict & cluster: Unsupervised skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9631–9640, 2020. 2

[51] Yukun Su, Guosheng Lin, and Qingyao Wu. Self-supervised 3d skeleton action representation learning with motion consistency and continuity. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13328–13338, 2021. 2

[52] Omer Sumer, Tobias Dencker, and Bjorn Ommer. Self-supervised learning of pose embeddings from spatiotemporal relations in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4298–4307, 2017. 3

[53] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8922–8931, 2021. 3, 4

[54] Jennifer J Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In *European Conference on Computer Vision*, pages 53–70. Springer, 2020. 3, 7

[55] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Proceedings of the European Conference on Computer Vision*, pages 776–794. Springer, 2020. 2

[56] Dorin Ungureanu, Federica Bogo, Silvano Galliani, Pooja Sama, Casey Meekhof, Jan Stühmer, Thomas J Cashman, Bugra Tekin, Johannes L Schönberger, Pawel Olszta, et al. Hololens 2 research mode as a tool for computer vision research. *arXiv preprint arXiv:2008.11239*, 2020. 1

[57] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3, 4

[59] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29:613–621, 2016. 2

[60] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2020. 3

[61] Jiangliu Wang, Jianbo Jiao, and Yun-Hui Liu. Self-supervised video representation learning by pace prediction. In *European conference on computer vision*, pages 504–521. Springer, 2020. 2

[62] Chen Wei, Lingxi Xie, Xutong Ren, Yingda Xia, Chi Su, Jiaying Liu, Qi Tian, and Alan L Yuille. Iterative reorganization with weak spatial constraints: Solving arbitrary jigsaw puzzles for unsupervised representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1910–1919, 2019. 2

[63] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018. 2

[64] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 2

[65] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10334–10343, 2019. 2

[66] Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6548–6557, 2020. 2

[67] Weiyu Zhang, Menglong Zhu, and Konstantinos G. Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2013. 1, 2, 6, 7

[68] Nenggan Zheng, Jun Wen, Risheng Liu, Liangqu Long, Jianhua Dai, and Zhefeng Gong. Unsupervised representation learning with long-term dynamics for skeleton based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[69] Feng Zhou and Fernando De la Torre. Generalized time warping for multi-modal alignment of human motion. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1282–1289. IEEE, 2012. 3

[70] Feng Zhou and Fernando Torre. Canonical time warping for alignment of human behavior. *Advances in neural information processing systems*, 22:2286–2294, 2009. 3

[71] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6002–6012, 2019. 2