

UniHPR: Unified Human Pose Representation via Singular Value Contrastive Learning

Zhongyu Jiang¹ Wenhao Chai¹ Lei Li² Zhuoran Zhou¹ Cheng-Yen Yang¹
Jenq-Neng Hwang¹

¹University of Washington, ²University of Copenhagen

{zyjiang, wchai, zhouz47, ccyang, hwang}@uw.edu, lilei@di.ku.dk

Abstract—In recent years, there has been a growing interest in developing effective alignment pipelines to generate unified representations from different modalities for multi-modal fusion and generation. As an important component of Human-Centric applications, Human Pose representations are critical in many downstream tasks, such as Human Pose Estimation, Action Recognition, Human-Computer Interaction, Object tracking, *etc.* Human Pose representations or embeddings can be extracted from images, 2D keypoints, 3D skeletons, mesh models, and lots of other modalities. Yet, there are limited instances where the correlation among all of those representations has been clearly researched using a contrastive paradigm. In this paper, we propose UniHPR, a unified Human Pose Representation learning pipeline, which aligns Human Pose embeddings from images, 2D and 3D human poses. To align more than two data representations at the same time, we propose a novel singular value-based contrastive learning loss, which better aligns different modalities and further boosts performance. To evaluate the effectiveness of the aligned representation, we choose 2D and 3D Human Pose Estimation (HPE) as our evaluation tasks. In our evaluation, with a simple 3D human pose decoder, UniHPR achieves remarkable performance metrics: MPJPE 49.9mm on the Human3.6M dataset and PA-MPJPE 51.6mm on the 3DPW dataset with cross-domain evaluation. Meanwhile, we are able to achieve 2D and 3D pose retrieval with our unified human pose representations in Human3.6M dataset, where the retrieval error is 9.24mm in MPJPE.

Index Terms—Human Pose Estimation, Representation Learning

I. INTRODUCTION

As an important component of human-centric applications, human pose representations (HPRs) are critical in many downstream tasks, such as human pose estimation, action recognition, human-computer interaction, object tracking, *etc.* Recently, aligning text and human pose sequences (human motion) [1], [2] has been widely discovered. However, there are many more data representations that can be used to denote human poses, including images, 2D keypoints, 3D skeletons, mesh models and *etc.* From the perspective of representation learning, many previous methods have been dedicated to mapping the representation of human pose sequences into the corresponding text space [1], [2]. On the other hand, in this paper, we propose **UniHPR**, a **Unified Human Pose Representation** learning framework, which aims to align RGB images, 2D and 3D human poses in the shared feature space. In order to evaluate the quality of the proposed learned

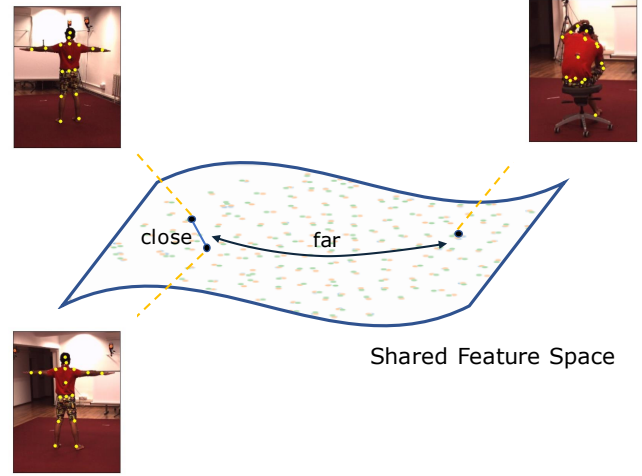


Fig. 1. RGB image, 2D and 3D human pose embeddings extracted by corresponding encoders in the shared feature space. After conducting contrastive learning during the pre-training stage, the embeddings extracted from these three different data representations of the same training sample are close to each other and away from other negative samples.

representation, we choose human pose estimation (HPE) as our evaluation task. By conducting task-specific fine-tuning, UniHPR can achieve the SOTA performance on both 2D and 3D HPE tasks.

Learning joint embeddings across more than two data representations (or modalities) is quite challenging. Inspired by Contrastive Language-Image Pre-Training (CLIP) [3], which proposes to learn aligned visual features with natural language supervisions trained on web-scale image-text paired data. We claim that alignment among RGB image, 2D and 3D human pose representations can also benefit from contrastive learning on large-scale and diverse datasets (e.g., Human3.6M [4], MPI-INF-3DHP [5], *etc.*).

During the evaluation, UniHPR serves as an encoder with additional downstream task decoders for 2D or 3D HPE. Therefore, the whole pipeline consists of image, 2D and 3D human pose encoders, and 2D and 3D human pose decoders. The embedding features of these three data representations are aligned and shared. To be specific, we first encode the images by HRNet [6], 2D and 3D human poses by shallow Transformers [7] respectively to get the corresponding embeddings,

respectively. We then conduct contrastive learning to align the embeddings from these three different data representations of the same training sample in the shared feature space for the unified representation learning. However, aligning embeddings from more than two data representations is challenging, and therefore, we propose a singular value based supervised contrastive learning loss to align three data representations at the same time. After that, during the training stage, we jointly train encoders and decoders with contrastive learning and multi-task learning simultaneously. During inference, since the embeddings are aligned in the same feature space, UniHPR can simultaneously support 2D human pose estimation and 3D human pose estimation, both lifting-based and image-based, in the same pipeline.

Our contributions can be summarised as follows:

- We propose the singular value based InfoNCE loss for supervised contrastive learning to effectively align embedding of more than two data representations at the same time.
- UniHPR aligns the embedding of Human Pose Representations from three distinctive data representations, i.e., images, 2D and 3D human poses.
- With a simple additional diffusion-based decoder, UniHPR achieves SOTA performance on frame-based 3D HPE tasks, e.g., MPJPE 49.9mm on the Human3.6M dataset with image-3D branch and PA-MPJPE 51.6mm with 2D-3D branch on the 3DPW dataset for the 3D human pose estimation task.

II. RELATED WORKS

A. Lifting Method for 3D HPE

2D-3D lifting [8]–[11] methods aim to infer 3D human pose under the assistance of the 2D joint detector. Thus, the relations between 2D and 3D human poses have captivated the attention of numerous researchers in computer vision and human motion analysis. Though the internal correspondence is tight, it is rather challenging to align their representations in the embedding space as they contain varying spatial information, and ambiguities in depth may also cause severe one-to-many 2D-3D mappings.

B. Image-based Method for 3D HPE

The other approach for estimating 3D human poses is building an end-to-end network designed to predict the 3D joint coordinates of the poses or SMPL [12] parameters directly from RGB images. Those methods can be categorized into two main classes: heatmap-based [13], [14] and regression-based [15]–[21] methods. Following the architecture of 2D human pose estimation, heatmap-based methods generate a 3D likelihood heatmap for each individual joint, and the joint’s position is ascertained by identifying the peak within the heatmap. On the other hand, the regression-based methods detect the root location and regress the relative locations of other joints in two branches. In contrast, the SMPL regression methods focus on regressing SMPL parameters from image or video input. Kolotouros et al. [16] propose SPIN, which takes advantage

of an optimization-based 3D pose estimation method, i.e., SMPLify [22], to achieve semi-supervised learning on 2D pose only datasets. VIBE [17] utilizes temporal information and a discriminator pretrained on a large 3D pose dataset.

III. METHODOLOGY

We build a unified human pose representation learning pipeline. During training, for any triplet of the cropped human image, $I \in \mathbb{N}^{H \times W \times 3}$, 2D and 3D human poses, $P_{2D/3D} \in \mathbb{R}^{J \times 2/3}$, UniHPR aligns the embeddings from all three representations and utilizes 2D and 3D pose decoders for downstream tasks.

A. Framework Architecture

Image encoder. The extraction of embedding from an RGB image is based on the HRNet [6], which is a convolution-based backbone for various visual recognition tasks. We concatenate and flatten the average pooled features from the last stage and pass it through a linear projection layer to obtain a 1-D embedding as our image representation.

2D/3D pose encoders. We adopt two Transformer-based [7] encoders to extract the embeddings from 2D and 3D human poses, respectively. We conduct bounding box normalized keypoint-wise patch embedding and retain the spatial information of each keypoint via adding learnable spatial position embedding. Then, the pose tokens prepended with a $[CLS]$ token and a bounding box token are fed into standard transformer encoder layers, including multi-head self-attention, feed-forward layers, and normalization layers. After that, we use the $[CLS]$ tokens as 2D and 3D pose embedding, respectively, which effectively aggregates the information of the other tokens and can be regarded as general prior.

2D and 3D pose decoders. We try several different architectures for our task specific decoder, including an MLP, a transformer and a diffusion based model. The diffusion decoder provides the best results in decoding the embedding to generate 2D and 3D human poses. We treat the decoders following the Score Matching paradigm [23]. To be specific, the encoded embedding is added with time embedding as well as a data representation token, which indicates the source of the embedding (e.g., from an image, 2D or 3D pose) in the diffusion network as a condition embedding and is used to generate the final 2D and 3D poses. The detailed architectures of all decoders are in the supplemental material.

B. Unified Representation Learning via Contrastive Learning

During the representation learning stage, we aim to align the embeddings from images, 2D and 3D human poses via the supervised contrastive learning. Given a batch of data, we have the RGB images, $I \in \mathbb{N}^{B \times H \times W \times 3}$, 2D poses, $P_{2D} \in \mathbb{R}^{B \times J \times 2}$, and 3D poses, $P_{3D} \in \mathbb{R}^{B \times J \times 3}$, where B, H, W, J are batch size, image height and width, and number of human body keypoints, respectively. The image, 2D, and 3D pose encoders E_{img}, E_{2D}, E_{3D} are trained by maximizing the similarity between image embedding $x_{img} \in \mathbb{R}^{B \times D}$, 2D

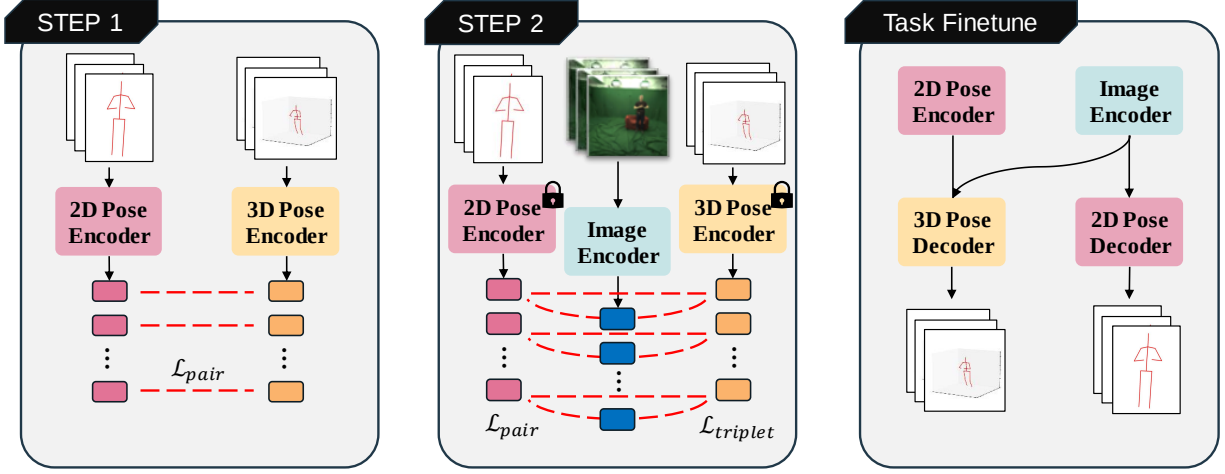


Fig. 2. The training scheme of **UniHPR**. Steps 1 and 2 are representation learning stages, and Task-Specific Finetune is the finetuning stage for any specific task. During Step 1, the 2D and 3D pose embedding alignment is trained first with $\mathcal{L}_{pair}$, and in Step 2, the image encoder is aligned with frozen 2D and 3D pose encoders via $\mathcal{L}_{pair}$ and $\mathcal{L}_{triplet}$. In the Task-Specific Finetuning stage, encoders and decoders are trained jointly via both contrastive learning, $\mathcal{L}_{pair}$ and $\mathcal{L}_{triplet}$, and task loss.

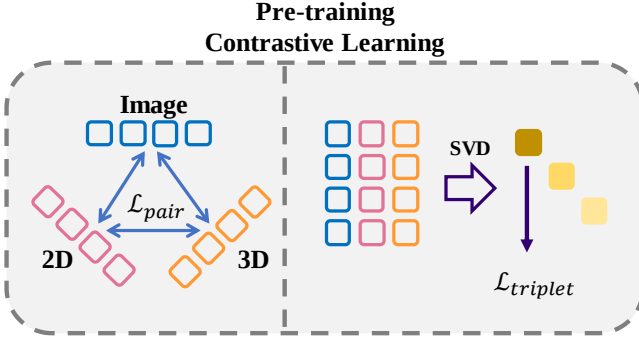


Fig. 3. $\mathcal{L}_{pair}$ is applied three times for contrastive learning and the singular value based $\mathcal{L}_{triplet}$ focuses on aligning three representations at the same time.

pose embedding $x_{2D} \in \mathbb{R}^{B \times D}$, and 3D pose embedding $x_{3D} \in \mathbb{R}^{B \times D}$, where D is the dimension of the embedding, which is the same over all three data representations. The most intuitive approach to aligning three embeddings is to apply three pair-wise contrastive losses. For embeddings, x_S, x_T , from any pair of data representations, the contrastive learning loss is

$$\mathcal{L}_{pair} = -\log \frac{\exp(x_S \cdot x_T^+ / \tau)}{\sum_{i=1}^B \exp(x_S \cdot x_{T,i} / \tau)}, \quad (1)$$

where τ is the learnable temperature initialized by τ_0 .

However, we found that simply applying three pairwise InfoNCE loss cannot obtain expected embedding similarity across three representations, as shown in the ablation studies in Section IV-E. Therefore, we propose a singular value-based InfoNCE loss (Triplet-InfoNCE) to address this issue.

We stack the embeddings from three representations to build a normalized embedding matrix, formulated by

$$\mathcal{M}_x = [x_{img} \quad x_{2D} \quad x_{3D}]^T \in \mathbb{R}^{3 \times D}. \quad (2)$$

If we apply singular value decomposition (SVD) to this matrix, $M_x = U\Sigma V^*$, the largest singular value, $\sigma_1 = \Sigma_{11}$, is related to the linear correlation of row vectors. Meanwhile, since the embeddings are normalized, the largest singular value should be in $[-\sqrt{3}, \sqrt{3}]$. Therefore, we can use InfoNCE loss to align any triplet of embeddings by maximizing the σ_1 . However, computing the singular value of a matrix with $3 \times D$, where $3 \ll D$, is time-consuming. Therefore, to accelerate the training procedure, instead of σ_1 , the largest eigenvalue λ_1 of the matrix $\mathcal{M}_x \mathcal{M}_x^T \in \mathbb{R}^{3 \times 3}$ is the optimization target, since $\lambda_1 = \sigma_1^2$. Therefore, by maximizing the λ_1 for positive triplets, which contain three embeddings from the same frame, and minimizing the λ_1 for negative triplets, which contain at least one embedding from a different frame, we are able to align embeddings from three representations jointly.

However, in one minibatch, the number of negative triplets for any positive triplet is $3B^2 - 3B + 1$, and if we use all the negative samples as our denominator in InfoNCE loss, the time consumption is unacceptable. We apply a random sample algorithm to select only $B-1$ negative triplets for each positive triplet. In this case, the singular value based InfoNCE loss can be formulated as,

$$\mathcal{L}_{triplet} = -\log \frac{\exp(\lambda_1^+ / \tau)}{\sum_{i=1}^B \exp(\lambda_{1i} / \tau)}. \quad (3)$$

Overall, our contrastive learning loss is

$$\mathcal{L}_{cl} = \mathcal{L}_{pair} + \alpha \mathcal{L}_{triplet}. \quad (4)$$

where α is the weighted factor.

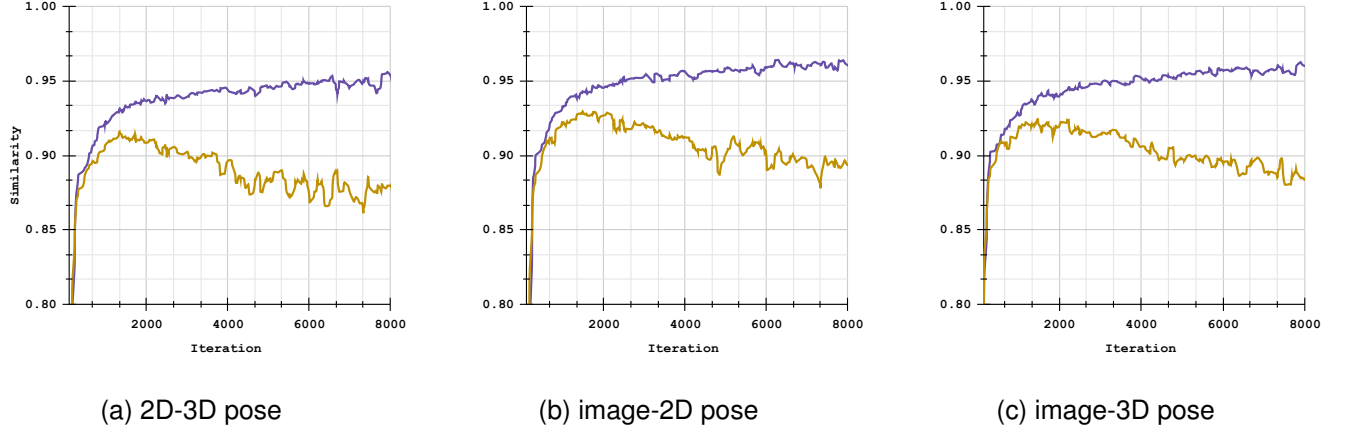


Fig. 4. **Cosine similarities** between different data representations. The yellow line is the one trained only with three pair-wise losses, $\mathcal{L}_{pair}$, and the purple line is the training curve with additional singular value-based InfoNCE loss, $\mathcal{L}_{triplet}$. Our proposed singular value-based InfoNCE loss helps align the feature space.

TABLE I
IMAGE-BASED 3D HPE PERFORMANCE ON THE 3DPW AND HUMAN3.6M DATASETS UNDER MPJPE AND PA-MPJPE. † INDICATES CROSS-DOMAIN EVALUATION ON 3DPW DATASET.

Method		Representation	3DPW	Human3.6M	
			PA-MPJPE (↓)	MPJPE (↓)	PA-MPJPE (↓)
Temporal	Kanazawa et al. [24]	SMPL	72.6	-	56.9
	Doersch et al. [25]	SMPL	74.7	-	-
	Arnab et al. [26]	SMPL	72.2	77.8	54.3
	DSD [27]	SMPL	69.5	59.1	42.4
	VIBE [17]	SMPL	56.5	65.9	41.5
Frame-based	Pavlakos et al. [15]	SMPL	-	-	75.9
	HMR [28]	SMPL	76.7	88.0	56.8
	NBF [29]	SMPL	-	-	59.9
	DenseRaC [30]	SMPL	-	76.8	48.0
	SPIN [16]	SMPL	59.2	62.5	41.1
	PARE [18]	SMPL	50.9	76.8	50.6
	PyMAF-X [31]	SMPL	47.1	54.2	37.2
	CLIFF [20]	SMPL	43.0	47.1	32.7
	UniHPR-w32† (ours)	Keypoint	65.8	54.5	39.5
	UniHPR-w48† (ours)	Keypoint	64.5	49.9	35.7

C. Task-Specific Finetune

After the representation learning stage, all encoders and decoders are trained jointly. While encoders are trained with $\mathcal{L}_{cl}$, the task losses, $\mathcal{L}_{2D/3D}$, depend on the architectures of decoders. For the diffusion-based decoder, we adopt the loss from the Score Matching Network [37], and for the MLP-based decoder, we utilize $L2$ loss.

Therefore, the overall loss in Task-Specific Finetune is

$$\mathcal{L} = \mathcal{L}_{cl} + \mathcal{L}_{2D} + \mathcal{L}_{3D}. \quad (5)$$

During inference, since the embeddings are well-aligned unified human pose representations in the same feature space, UniHPR can utilize the embedding from any representation and estimate 2D or 3D human poses with shared decoders.

IV. EXPERIMENTS

A. Implementation Details

We implement our proposed framework using PyTorch [38] on a single NVIDIA A100/80G GPU. The representation learning includes two steps: (1) 2D-3D alignment; (2) Image-2D-3D joint alignment (see Fig. 2); followed by a task-specific finetuning stage. In the first step of representation learning, the batch size is 2048, $\tau_0 = 1/14$, and $\tau \in [1/100, 10^4]$, while in the second step, the batch size is 180, $\tau_0 = 1/5$, and $\tau \in [1/10, 10^4]$. During the multi-task training steps, encoders and decoders are trained together with the batch size being 180, $\tau_0 = 1/5$, and $\tau \in [1/5, 10^4]$. For the weight of triplet contrastive loss, $\mathcal{L}_{triplet}$, $\alpha = 1$. The input image size of the image encoder is 192×256 . During both of the two steps, we adopt Adam optimizer with a learning rate of



Fig. 5. **Interpolation** of 3D human pose representations in Human3.6M dataset.

TABLE II

LIFTING-BASED 3D HPE PERFORMANCE ON THE 3DPW AND HUMAN3.6M DATASETS UNDER MPJPE AND PA-MPJPE. THE GROUND TRUTH 2D KEYPOINTS ARE USED ON 3DPW DATASET, WHILE THE DETECTED 2D KEYPOINTS FROM CPN ARE USED ON HUMAN3.6M DATASET. † INDICATES CROSS-DOMAIN EVALUATION ON 3DPW DATASET.

	Method	3DPW	Human3.6M	
		PA-MPJPE (↓)	MPJPE (↓)	PA-MPJPE (↓)
Temporal	VideoPose3D (f=243) [9]	68.0	46.8	36.5
	AdaptPose† [32]	46.5	-	-
	Li et al. [33]	-	43.7	35.2
	MixSTE [34]	-	40.9	32.6
	MPM [35]	-	42.6	34.7
Frame-based	SimpleBaseline† [8]	89.4	62.9	47.7
	SemGCN† [36]	102.0	61.2	47.7
	VideoPose3D† (f=1) [9]	94.6	55.2	42.3
	PoseAug† [10]	58.5	<u>52.9</u>	-
	PoseDA† [11]	<u>55.3</u>	-	-
	UniHPR† (ours)	51.6	52.6	39.9

1×10^{-4} . We train UniHPR on Human3.6M [4] and MPI-INF-3DHP [5] datasets and apply ablation study about the performance difference on different training datasets.

B. Datasets and Performance Metrics

To conduct the quantitative performance evaluation of the proposed UniHPR, we use several widely used 3D human pose datasets to train and evaluate our proposed framework, including Human3.6M [4], MPI-INF-3DHP [5], and 3DPW [39]. We train UniHPR on Human3.6M and MPI-INF-3DHP and evaluate it on Human3.6M and 3DPW.

C. Evaluation of the Unified Human Pose Representation

Quantitative Evaluation of Representation Learning. To better evaluate the quality of learned unified representations, we conduct Pose and Image Retrieval on Human3.6M dataset. The retrieved 3D human pose or image has the most similar 3D pose or image embedding with the image, 2D or 3D pose representation query. For Image Retrieval task, the FPS is set as 1. In Table IV, 2D-3D Pose Retrieval can achieve MPJPE 9.2mm and the MPJPE of Image-3D Pose Retrieval

is 10.4mm, and 2D-Image Image Retrieval can achieve Top-1 Accuracy 95.5%, which illustrate the unified representations are well aligned in images, 2D and 3D human poses. More visualization is included in the supplementary material.

Interpolation of the Unified Representations. Furthermore, UniHPR is capable of interpolating 3D human poses by the corresponding 3D pose representations. As shown in Fig 5, by interpolating 3D representations from two different 3D poses, UniHPR generates smooth and realistic 3D poses in between.

D. Evaluation of Human Pose Estimation

Lifting-based 3D Human Pose Estimation We evaluate the performance of lifting-based 3D HPE tasks on Human3.6M and 3DPW datasets. As shown in Table II, UniHPR archives 51.6 mm in terms of PA-MPJPE on 3DPW dataset and 52.6 mm in terms of MPJPE on Human3.6M dataset, which is the state-of-the-art performance. Since UniHPR is not trained on 3DPW, it is a fair comparison with those cross-domain evaluation methods.

Image-based 3D Human Pose Estimation As for image-based 3D HPE, we also evaluate the performance on Human3.6M and 3DPW datasets. As shown in Table I, UniHPR respectively achieves 49.9 mm and 35.7 mm in terms of MPJPE and PA-MPJPE on Human3.6M dataset, as well as 65.7 mm of PA-MPJPE on 3DPW dataset. Note that we are the only keypoint-based method in Table I, and all the others are SMPL-based. UniHPR achieves comparable performance regarding the number of model parameters and training data with SOTA methods.

E. Ablation Study

In this section, we conduct extensive ablation studies to investigate the importance of each module in the UniHPR, especially how our proposed singular value based loss, $\mathcal{L}_{triple}$, helps the training and improves the performance.

End-to-End training without alignment. We claim that feature alignment, i.e., pre-training via contrastive learning,

TABLE III
QUANTITATIVE EVALUATION OF THE UNIFIED REPRESENTATION. POSE RETRIEVAL ON HUMAN3.6M TEST DATASET.

Retrieval	MPJPE ($\downarrow$)	PA-MPJPE ($\downarrow$)
2D-3D	9.2	7.1
Image-3D	10.4	7.6

TABLE IV
QUANTITATIVE EVALUATION OF THE UNIFIED REPRESENTATION. IMAGE RETRIEVAL ON HUMAN3.6M TEST DATASET WITH 1 FPS.

Retrieval	Top-1 Acc. ($\uparrow$)	Top-3 Acc. ($\uparrow$)
3D-Image	89.2	95.6
2D-Image	95.5	97.6

TABLE V
ABLATION STUDY ON UNIHPR. EVALUATED ON HUMAN3.6M DATASET. $\mathcal{L}_{pair}$ AND $\mathcal{L}_{triplet}$ DENOTES APPLYING THOSE LOSSES ON THE PRE-TRAINING STAGE. $\mathcal{M}$ TOKEN MEANS DECODERS UTILIZE THE REPRESENTATION TOKEN. WE EVALUATE THE PERFORMANCE WITH ADDITIONAL DATA FROM MPI-INF-3DHP DATASET AS WELL.

$\mathcal{L}_{pair}$	$\mathcal{L}_{triplet}$	$\mathcal{R}$	w. 3DHP	GT 2D		Image	
				MPJPE ($\downarrow$)	PA-MPJPE ($\downarrow$)	MPJPE ($\downarrow$)	PA-MPJPE ($\downarrow$)
<i>baseline</i>				41.3	31.6	91.8	68.7
✓				60.0 (+18.7)	47.5 (+15.9)	65.5 (-26.3)	51.8 (-16.9)
✓	✓			40.9 (-0.4)	31.7 (+0.1)	58.7 (-33.1)	44.4 (-24.3)
✓	✓	✓		39.3 (-2.0)	29.9 (-1.7)	57.5 (-34.3)	42.9 (-25.8)
✓	✓	✓	✓	41.7 (+0.4)	32.6 (+1.0)	54.5 (-37.3)	39.5 (-29.2)

among different representations is the key to success. Therefore, we conduct the ablation studies on skipping the alignment training stages. As shown in Table V, alignment improves the image-based 3D HPE performance significantly on the Human3.6M dataset. As shown in table V, without the 2-step contrastive learning, the performance gap between lifting and image branches shows that the features are not correctly aligned. Furthermore, the combination of $\mathcal{L}_{triplet}$ and $\mathcal{L}_{pair}$ provides the best performance on both lifting and image branches.

Ablation on representation token, $\mathcal{R}$. In UniHPR, we design a representation token when using the 3D pose decoder to estimate 3D human poses. The representation token indicates which representation the features derived from either (e.g. image or 2D pose). As shown in Table V, consistent improvement is observed in using the representation token among lifting-based and image-based 3D HPE tasks on the Human3.6M dataset.

Effectiveness of the $\mathcal{L}_{triplet}$. As shown in Figure 4, compared to simply applying three pairwise InfoNCE loss, $\mathcal{L}_{pair}$, the proposed singular value-based InfoNCE loss, $\mathcal{L}_{triplet}$, significantly better aligns the features from different representations. With the help of $\mathcal{L}_{triplet}$, the embedding cosine similarity between different representation does not decrease after around 1500 iterations and keeps increasing to around 0.95 in 8000 iterations. For quantitative evaluation, in Table V, without the help of $\mathcal{L}_{triplet}$, three $\mathcal{L}_{pair}$ can only achieve MPJPE 65.5mm and 60.0mm for image and keypoint branches, which are 6.8mm and 19.1mm more than the jointly trained model.

Training with additional data. As shown in Table V, it is noted that the distribution of 2D and 3D pose pairs on 3DHP differs from Human3.6M, which increases the robustness



Fig. 6. Failure cases of UniHPR. When there is heavy occlusion, our model may estimate the incorrect pose or the pose of a wrong target.

of the lifting-based branch but decreases the performance slightly on Human3.6M, since the model trained with both Human3.6M and 3DHP achieves the best performance on 3DPW. Furthermore, training with additional data boosts the image-based branch by improving the diversity of image data.

Failure cases. As shown in Figure 6, the image branch of UniHPR fails in the case of large occlusion or low-quality RGB input scenarios. UniHPR is trained on Human3.6M and MPI-INF-3DHP with only one target per frame and a limited amount of occlusion.

V. CONCLUSION

In conclusion, the UniHPR framework represents a significant step forward in unified human pose representation learning by mitigating the gap between image, 2D and 3D human pose representations. Despite its potential limitations in data and computational requirements, UniHPR sets a promising direction for future research, particularly in improving generalization capabilities and multi-modal representation learning. The framework's achievements on benchmark datasets like Human3.6M and 3DPW justify its potential, paving the way for advancements in applications across multiple domains such as text-to-pose and pose-to-image generation.

REFERENCES

- [1] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or, "Motionclip: Exposing human motion generation to clip space," in *European Conference on Computer Vision*. Springer, 2022, pp. 358–374.
- [2] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu, "Motiondiffuse: Text-driven human motion generation with diffusion model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [3] 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 *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.
- [4] Catalin Ionescu, Drago Papava, Vlad Olaru, and Cristian Sminchisescu, "Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 7, pp. 1325–1339, 2013.
- [5] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in *2017 Fifth International Conference on 3D Vision (3DV)*. IEEE, 2017.
- [6] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al., "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.
- [7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- [8] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little, "A simple yet effective baseline for 3d human pose estimation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2640–2649.
- [9] Dario Pavlo, Christoph Feichtenhofer, David Grangier, and Michael Auli, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in *CVPR*, 2019, pp. 7753–7762.
- [10] Kehong Gong, Jianfeng Zhang, and Jiashi Feng, "Poseaug: A differentiable pose augmentation framework for 3d human pose estimation," in *CVPR*, 2021, pp. 8575–8584.
- [11] Wenhao Chai, Zhongyu Jiang, Jenq-Neng Hwang, and Gaoang Wang, "Global adaptation meets local generalization: Unsupervised domain adaptation for 3d human pose estimation," *arXiv preprint arXiv:2303.16456*, 2023.
- [12] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.
- [13] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis, "Coarse-to-fine volumetric prediction for single-image 3d human pose," in *CVPR*, 2017, pp. 7025–7034.
- [14] Diogo C Luvizon, David Picard, and Hedi Tabia, "2d/3d pose estimation and action recognition using multitask deep learning," in *CVPR*, 2018, pp. 5137–5146.
- [15] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis, "Learning to estimate 3d human pose and shape from a single color image," in *CVPR*, 2018, pp. 459–468.
- [16] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis, "Learning to reconstruct 3d human pose and shape via model-fitting in the loop," in *ICCV*, 2019.
- [17] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black, "Vibe: Video inference for human body pose and shape estimation," in *CVPR*, 2020, pp. 5253–5263.
- [18] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black, "Pare: Part attention regressor for 3d human body estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11127–11137.
- [19] Hongwen Zhang, Yating Tian, Xinchu Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun, "Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11446–11456.
- [20] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan, "Cliff: Carrying location information in full frames into human pose and shape estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 590–606.
- [21] Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu, "Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery," *arXiv preprint arXiv:2304.05690*, 2023.
- [22] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black, "Keep it smpl: Automatic estimation of 3d human pose and shape from a single image," in *ECCV*. Springer, 2016, pp. 561–578.
- [23] Yang Song and Stefano Ermon, "Generative modeling by estimating gradients of the data distribution," in *Advances in neural information processing systems*, 2019, vol. 32.
- [24] Angjoo Kanazawa, Jason Y Zhang, Panna Felsen, and Jitendra Malik, "Learning 3d human dynamics from video," in *CVPR*, 2019, pp. 5614–5623.
- [25] Carl Doersch and Andrew Zisserman, "Sim2real transfer learning for 3d human pose estimation: motion to the rescue," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [26] Anurag Arnab, Carl Doersch, and Andrew Zisserman, "Exploiting temporal context for 3d human pose estimation in the wild," in *CVPR*, 2019, pp. 3395–3404.
- [27] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, Yili Fu, and Tao Mei, "Human mesh recovery from monocular images via a skeleton-disentangled representation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5349–5358.
- [28] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik, "End-to-end recovery of human shape and pose," in *CVPR*, 2018, pp. 7122–7131.
- [29] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele, "Neural body fitting: Unifying deep learning and model based human pose and shape estimation," in *2018 international conference on 3D vision (3DV)*. IEEE, 2018, pp. 484–494.
- [30] Yuanlu Xu, Song-Chun Zhu, and Tony Tung, "Denserac: Joint 3d pose and shape estimation by dense render-and-compare," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7760–7770.
- [31] Hongwen Zhang, Yating Tian, Yuxiang Zhang, Mengcheng Li, Liang An, Zhenan Sun, and Yebin Liu, "Pymaf-x: Towards well-aligned full-body model regression from monocular images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
- [32] Mohsen Gholami, Bastian Wandt, Helge Rhodin, Rabab Ward, and Z Jane Wang, "Adaptpose: Cross-dataset adaptation for 3d human pose estimation by learnable motion generation," in *CVPR*, 2022, pp. 13075–13085.
- [33] Wenhao Li, Hong Liu, Runwei Ding, Mengyuan Liu, Pichao Wang, and Wenming Yang, "Exploiting temporal contexts with strided transformer for 3d human pose estimation," *IEEE Transactions on Multimedia*, vol. 25, pp. 1282–1293, 2022.
- [34] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan, "Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video," in *CVPR*, 2022, pp. 13232–13242.
- [35] Zhenyu Zhang, Wenhao Chai, Zhongyu Jiang, Tian Ye, Mingli Song, Jenq-Neng Hwang, and Gaoang Wang, "Mpm: A unified 2d-3d human pose representation via masked pose modeling," *arXiv preprint arXiv:2306.17201*, 2023.
- [36] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas, "Semantic graph convolutional networks for 3d human pose regression," in *CVPR*, 2019, pp. 3425–3435.
- [37] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.
- [38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.
- [39] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll, "Recovering accurate 3d human pose in the wild using imus and a moving camera," in *ECCV*, 2018, pp. 601–617.