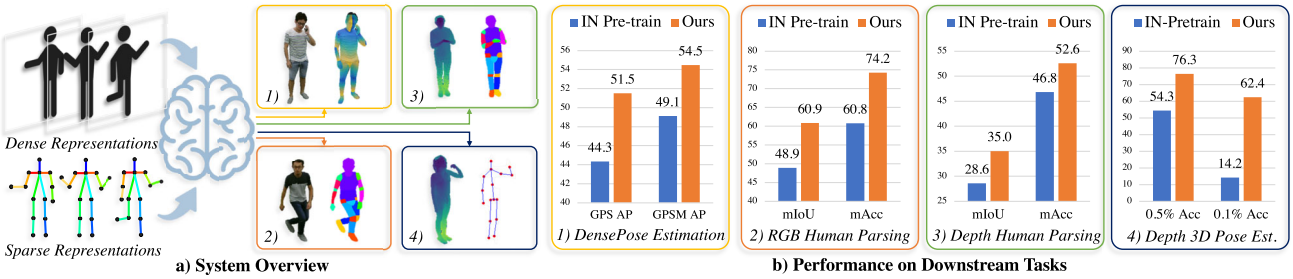
Versatile Multi-Modal Pre-Training for Human-Centric Perception

Fangzhou Hong[1], Liang Pan[1], Zhongang Cai[1,2,3], Ziwei Liu[1]✉
[1]S-Lab, Nanyang Technological University [2]SenseTime Research [3]Shanghai AI Laboratory
{fangzhou001, liang.pan, ziwei.liu}@ntu.edu.sg caizhongang@sensetime.com



a) System Overview **b) Performance on Downstream Tasks**

Figure 1. **An Overview of HCMoCo.** a) We present HCMoCo, a versatile multi-modal pre-training framework that takes multi-modal observations of human body as input for human-centric perception. The pre-train models can be transferred to various human-centric downstream tasks with different modalities. b) Our HCMoCo shows superior performance on all four downstream tasks, especially for data-efficient settings (10% DensePose, 20% RGB/depth human parsing, 0.5/0.1% 3D pose estimation). 'IN' stands for ImageNet.

Abstract

*Human-centric perception plays a vital role in vision and graphics. But their data annotations are prohibitively expensive. Therefore, it is desirable to have a versatile pre-train model that serves as a foundation for data-efficient downstream tasks transfer. To this end, we propose the **H**uman-**C**entric Multi-**Mo**dal **Co**ntrastive Learning framework **HCMoCo** that leverages the multi-modal nature of human data (*e.g. *RGB, depth, 2D keypoints) for effective representation learning. The objective comes with two main challenges: dense pre-train for multi-modality data, efficient usage of sparse human priors. To tackle the challenges, we design the novel Dense Intra-sample Contrastive Learning and Sparse Structure-aware Contrastive Learning targets by hierarchically learning a modal-invariant latent space featured with continuous and ordinal feature distribution and structure-aware semantic consistency. HCMoCo provides pre-train for different modalities by combining heterogeneous datasets, which allows efficient usage of existing task-specific human data. Extensive experiments on four downstream tasks of different modalities demonstrate the effectiveness of HCMoCo, especially under data-efficient settings (7.16% and 12% improvement on DensePose Estimation and Human Parsing). Moreover, we demonstrate the versatility of HCMoCo by exploring cross-modality supervision and missing-modality inference, validating its strong ability in cross-modal association and reasoning. Codes are available at https://github.com/hongfz16/HCMoCo.*

✉ Corresponding author

1. Introduction

As a long-standing problem, human-centric perception has been studied for decades, ranging from sparse prediction tasks, such as human action recognition [8, 27, 42, 50], 2D keypoints detection [2, 26, 43, 48] and 3D pose estimation [22, 31, 40], to dense prediction tasks, such as human parsing [7, 11, 12, 25] and DensePose prediction [14]. Unfortunately, to train a model with reasonable generalizability and robustness, an enormous amount of labeled real data is necessary, which is extremely expensive to collect and annotate. Therefore, it is desirable to have a versatile pre-train model that can serve as a foundation for all the aforementioned human-centric perception tasks.

With the development of sensors, the human body can be more conveniently perceived and represented in **multiple modalities**, such as RGB, depth, and infrared. In this work, we argue that *the multi-modality nature of human-centric data can induce effective representations* that transfer well to various downstream tasks, due to three major **advantages**: **1)** Learning a modal-invariant latent space through pre-training helps efficient task-relevant mutual information extraction. **2)** A single versatile pre-train model on multi-modal data facilitates multiple downstream tasks using various modalities. **3)** Our multi-modal pre-train setting bridges heterogeneous human-centric datasets through their common modality, which benefits the generalizability of pre-train models.

We mainly explore two groups of modalities as shown in Fig. 1 a): dense representations (*e.g*. RGB, depth, infrared) and sparse representations (*e.g*. 2D keypoints, 3D pose). Dense representations can provide rich texture and/or 3D geometry information. But they are mostly low-level and noisy. On the contrary, sparse representations obtained by off-the-shelf tools [4, 9] are semantic and structured. But the sparsity results in insufficient details. We highlight that it is non-trivial to integrate these heterogeneous modalities into a unified pre-training framework for the following two main **challenges**: **1)** learning representations suitable for dense prediction tasks in the multi-modality setting; **2)** using weak priors from sparse representations effectively for pre-training.

Challenge 1: Dense Targets. Existing methods [21, 30] perform contrastive learning densely on pixel-level features to achieve view-invariance for dense prediction tasks. However, those methods require multiple views of a static 3D scene [10], which is inapplicable for human-centric applications with only single view. Furthermore, it is preferable to learn representations that are continuously and orderly distributed over the human body. In light of this, we generalize the widely used InfoNCE [33] and propose a dense intra-sample contrastive learning objective that applies a soft pixel-level contrastive target, which can facilitate learning ordinal and continuous dense feature distributions.

Challenge 2: Sparse Priors. To employ priors in contrastive learning, previous works [3, 23, 46] mainly use the supervision to generate semantically positive pairs. However, these methods only focus on the sample-level contrastive learning, which means each sample is encoded to a global embedding. It is not optimal for human dense prediction tasks. To this end, we propose a sparse structure-aware contrastive learning target, which uses semantic correspondences across samples as positive pairs to complement positive intra-sample pairs. Particularly, leveraging sparse human priors leads to an embedding space where semantically corresponding parts are aligned more closely.

To sum up, we propose **HCMoCo**, a **H**uman-**C**entric multi-**Mo**dal **Co**ntrastive learning framework for versatile multi-modal pre-training. To fully leverage multi-modal observations, HCMoCo effectively utilizes both dense measurements and sparse priors using the following three-levels hierarchical contrastive learning objectives: **1)** sample-level modality-invariant representation learning; **2)** dense intra-sample contrastive learning; **3)** sparse structure-aware contrastive learning. As an effort towards establishing a comprehensive multi-modal human parsing benchmark dataset, we label human segments for RGB-D images from NTU RGB+D dataset [42], and contribute the **NTURGBD-Parsing-4K** dataset. To evaluate HCMoCo, we transfer our pre-train model to four human-centric downstream tasks using different modalities, including DensePose estimation (RGB) [14], human parsing using RGB [22] or depth frames, and 3D pose estimation (depth) [16]. Under full set and data-efficient training settings, HCMoCo constantly achieves better performance than training from scratch or pre-train on ImageNet. To name a few, as shown in Fig. 1 b), we achieve **7.16%** improvement in terms of GPS AP on 10% training data of DensePose estimation; **12%** improvement in terms of mIoU on 20% training data of Human3.6M human parsing. Moreover, we evaluate the modal-invariance of the latent space learned by HCMoCo for dense prediction on NTURGBD-Parsing-4K with two settings: cross-modality supervision and missing-modality inference. Compared against conventional contrastive learning targets, our method improves the segmentation mIoU by **29%** and **24%** for the two settings, respectively. To the best of our knowledge, we are the first to study multi-modal pre-training for human-centric perception.

The main **contributions** are summarized below: **1)** As the first endeavor, we provide an in-depth analysis for human-centric pre-training, which is formulated as a challenging multi-modal contrastive learning problem. **2)** Together with the novel hierarchical contrastive learning objectives, a comprehensive framework HCMoCo is proposed for effective pre-training for human-centric tasks. **3)** Through extensive experiments, HCMoCo achieves superior performance than existing methods, and meanwhile shows promising modal-invariance properties. **4)** To benefit multi-modal human-centric perception, we contribute an RGB-D human parsing dataset, NTURGBD-Parsing-4K.

2. Related Work

Human-Centric Perception. Many efforts have been put into human-centric perception in decades. Lots of work in 2D keypoint detection [2,26,43,48] has achieved robust and accurate performance. 3D pose estimation has long been a challenging problem and is approached from two aspects, lifting from 2D keypoints [22, 31, 40] and predicting from depth map [16, 49]. Human parsing can be defined in two ways. The first one parses garments together with visible body parts [11, 12, 25]. The second one only focuses on parsing human parts [7, 20, 22]. In this work, we focus on the second setting because the depth and 2D keypoints do not contain the texture information needed for garment parsing. There are a few works [19, 32] about human parsing on depth maps. However, the data and annotations are too coarse or unavailable. To further push the accuracy of human-centric perception, DensePose [14, 44] is proposed to densely model each human body surface point. The cost of DensePose annotation is enormous. Therefore, we also explore data-efficient learning of DensePose.

Multi-Modal Contrastive Learning. Multi-modality naturally provides different views of the same sample which fits well into the contrastive learning framework. CMC [45]

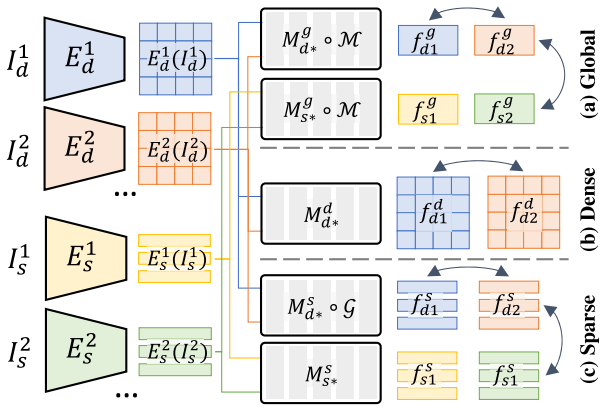


Figure 2. **Illustration of the general paradigm of HCMoCo.** We group modalities of human data into dense I_d^* and sparse representations I_s^*. Three levels of embeddings are extracted (3.1). Combining the nature of human data and tasks (3.2), we present contrastive learning targets for each level of embedding (3.3).

proposes the first multi-view contrastive learning paradigm which takes any number of views. CLIP [39] learns a joint latent space from large-scale paired image-language dataset. Extensive studies [1, 15, 17, 34, 35, 41] focus on video-audio contrastive learning. Recently, 2D-3D contrastive learning [21, 29, 30] has also been studied with the development in 3D computer vision. In this work, aside from commonly used modalities, we also explore the potential of 2D keypoints in human-centric contrastive learning.

3. Our Approach

In this section, we first introduce the general paradigm of HCMoCo (3.1). Following the design principles (3.2), hierarchical contrastive learning targets are formally introduced (3.3). Next, an instantiation of HCMoCo is introduced (3.4). Finally, we propose two applications of HCMoCo to show the versatility (3.5).

3.1. HCMoCo

As shown in Fig. 2, HCMoCo takes multiple modalities of perceived human body as input. The target is to learn human-centric representations, which can be transferred to downstream tasks. The input modalities can be categorized into dense and sparse representations. Dense representations I_d^* are the direct output of imaging sensors, *e.g.* RGB, depth, infrared. They typically contain rich information but are low-level and noisy. Sparse representations are structured abstractions of the human body, *e.g.* 2D keypoints, 3D pose, which can be formulated as graph $I_s^* = G(V, E)$. Different representations of the same view of a human should be spatially aligned, which means intra-sample correspondences can be obtained for dense contrastive learning. HCMoCo aims to pre-train multiple encoders E_d^* and E_s^* that produce embeddings of dense representations and sparse representations for downstream tasks transfer.

To support dense downstream tasks, other than the usual sample-level global embeddings used in [5,6,13,18,28,45], we propose to consider different levels of embeddings *i.e.* global embeddings f^g, sparse embeddings f^s and dense embeddings f^d [1], which are defined as follows: **1)** For dense representations I_d, the global embedding is obtained by applying a mapper network M_d^g to the mean pooling $\mathcal{M}$ of the corresponding feature map, which is formulated as $f_d^g = M_d^g \circ \mathcal{M} \circ E_d(I_d)$. Similarly, for sparse representations I_s, the global embedding is defined as $f_s^g = M_s^g \circ \mathcal{M} \circ E_s(I_s)$. **2)** Sparse embeddings have the same size as that of sparse representations. Formally, for sparse representations $I_s = G(V, E)$, where $V \in \mathbb{R}^{J \times K}$, the corresponding sparse embedding is defined as $f_s^s = M_s^s \circ E_s(I_s)$, where $f_s^s \in \mathbb{R}^{J \times K'}$, M_s^s is a mapper network. For dense representations, the corresponding sparse features are pooled from the dense feature map using the correspondences $\mathcal{G}$. Then the sparse features are mapped to sparse embeddings as $f_d^s = M_d^s \circ \mathcal{G} \circ E_d(I_d)$. **3)** Dense embeddings are only defined on dense representations, which is formulated as $f_d^d = M_d^d \circ E_d(I_d)$. With three levels of embeddings defined, we formulate the overall learning objective as

$$\mathcal{L} = \lambda_g \mathcal{L}_g(f^g) + \lambda_d \mathcal{L}_d(f^d) + \lambda_s \mathcal{L}_s(f^s), \quad (1)$$

which is analyzed and explained as follows.

3.2. Principles of Learning Targets Design

In this subsection, we analyze the intuitions when designing learning targets, which makes the following three principles. **1) Mutual Information Maximization**: Inspired by [36,47], we propose to maximize the lower bound on mutual information, which has been proved by many previous works [5, 6, 18, 45] to be able to produce strong pre-train models. **2) Continuous and Ordinal Feature Distribution**: Inspired by the property of human-centric perception, it is desirable for the feature maps of the human body to be continuous and ordinal. The human body is a structural and continuous surface. The dense predictions, *e.g.* human parsing [11, 12, 25], DensePose [14], are also continuous. Therefore, such property should also be reflected in the learned representations. Besides, for an anchor point on human surfaces, closer points have higher probabilities of sharing similar semantics with the anchor point than that of far away points. Therefore, the learned dense representations should also align with such ordinal relationship. **3) Structure-Aware Semantic Consistency**: Sparse representations are abstractions of the human body, which contains valuable structural semantics about the human body. Instead of identity information, the human pose and struc-

[1] For easier understanding of the notations, the superscripts of f and M stand for the kind of embeddings. The subscripts stand for the kind of representations ('g' for 'global'; 'd' for 'dense'; 's' for 'sparse').

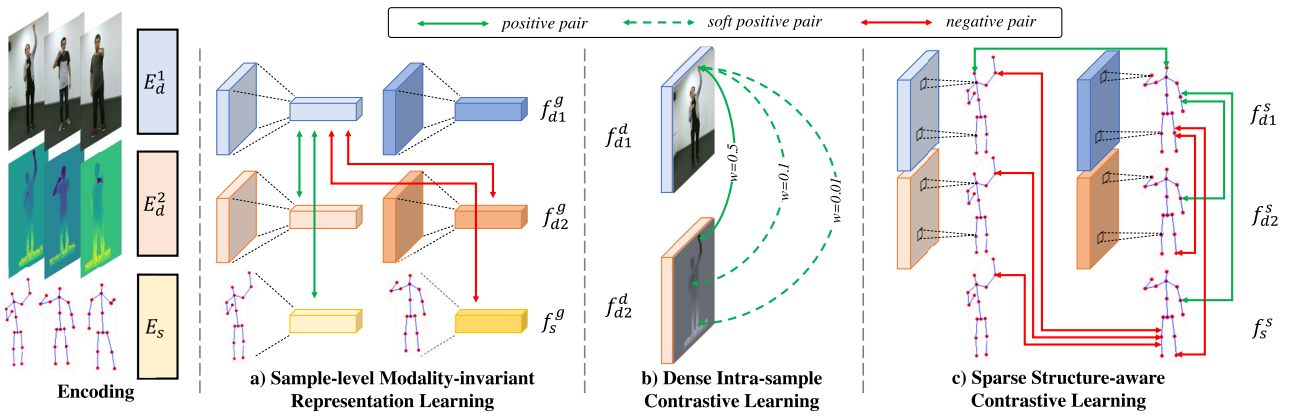


Figure 3. **Our Proposed Instantiation of HCMoCo.** For dense representations, we choose to use RGB and depth. For sparse representations, 2D keypoints are used for its convenience to obtain. a) At sample-level, the global embeddings are used for modality-invariant representation learning. b) Between paired dense embeddings, soft contrastive learning target is proposed for continuous and ordinal feature learning. c) Using human prior provided by sparse representations, intra- and inter-sample contrastive learning targets are proposed.

ture understanding are the keys to our target downstream tasks. Therefore, it is reasonable to eliminate the identity information and enhance the structure information by enforcing structure-aware semantic consistency where semantically close embeddings (*e.g.* embeddings of left hands from different samples) are pulled close and vice versa.

3.3. Hierarchical Contrastive Learning Targets

Based on the above three principles, we formally define hierarchical contrastive learning targets in this subsection.

Sample-level modality-invariant representation learning aims at learning a joint latent space at the sample level using global embeddings, which fulfills the first principle. Inspired by [45], the learning target can be formulated as

$$\mathcal{L}_g = -\mathop{\mathbb{E}}_{\substack{F_1^g, F_2^g \in S_g \\ f_1^g \in F_1^g}} \left[\log \frac{\exp(f_1^g \cdot \bar{f}_2^g / \tau)}{\sum_{f_2^g \in F_2^g} \exp(f_1^g \cdot f_2^g / \tau)} \right], \quad (2)$$

where F_*^g is a set of global embeddings of one modality, S_g is the set of F_*^g of all modalities, $\bar{f}_2^g$ is the embedding of the paired view of that of f_1^g, τ is the temperature. It should be noticed that f_1^g can be sampled from the global embeddings of either dense or sparse representations.

Dense intra-sample contrastive learning is operated on the paired dense representations. For any two paired dense embeddings $f_{d1}^d, f_{d2}^d \in \mathbb{R}^{H \times W \times K'}$, to simultaneously satisfy the first and the second principle, the dense intra-sample contrastive learning target between them is defined in a 'soft' way as

$$\mathcal{L}_d^{12} = -\mathop{\mathbb{E}}_{\substack{x,y \\ m,n}} \left[\mathcal{W}_{xy}^{mn} \log \frac{\exp(f_{d1}^d(x,y) \cdot f_{d2}^d(m,n))/\tau)}{\sum\limits_{x',y'} \exp(f_{d1}^d(x,y) \cdot f_{d2}^d(x',y')/\tau)} \right], \quad (3)$$

where $\mathcal{W}_{xy}^{mn}$ is the weight, τ is the temperature, $(x,y), (m,n), (x',y')$ are coordinates on the dense representation, $1 \le x, x', m \le H$, $1 \le y, y', n \le W$. The above equation is a generalized version of InfoNCE [33]. InfoNCE is a special case when $\mathcal{W}_{xy}^{mn}$ is set to 1 if $x = m$ and $y = n$ else 0. We use the normalized distances as the weights, which is formulated as

$$\mathcal{W}_{xy}^{mn} = \frac{\exp(\sqrt{(x-m)^2 + (y-n)^2})}{\sum_{x',y'} \exp(\sqrt{(x-x')^2 + (y-y')^2})}. \quad (4)$$

For each pair of dense representations, the above learning target is calculated between each pair of dense embeddings. Therefore, the whole learning target is defined as

$$\mathcal{L}_d = \mathop{\mathbb{E}}_{\substack{F_1^d, F_2^d \in S_d \\ f_{d1}^d, f_{d2}^d \in F_1^d, F_2^d}} \mathcal{L}_d^{12}(f_{d1}^d, f_{d2}^d), \quad (5)$$

where F_*^d is a set of dense embeddings of one modality, S_d is the set of all F_*^d, f_{d1}^d and f_{d2}^d are two paired embeddings. It should be noticed that the 'soft' learning target cannot guarantee an ordinal feature distribution. Instead, it serves as a computationally efficient relaxation of the requirement of ordinal distribution.

Sparse structure-aware contrastive learning takes two sparse representations f_1^s and f_2^s as inputs. The paired features f_{1j}^s and f_{2j}^s (*i.e.* features of the j-th joint) should be pulled close while unpaired features are pushed away. The two sparse representations can be sampled from the same or different modalities, intra- or inter-sample. The intra-sample alignment satisfies the first principle. The inter-sample alignment follows the third principle. The sparse structure-aware contrastive learning target is formulated as

$$\mathcal{L}_s = -\mathop{\mathbb{E}}_{\substack{F_1^s, F_2^s \in S_s \\ j; f_1^s, f_2^s \in \{F_1^s, F_2^s\}}} \left[\log \frac{\exp(f_{1j}^s \cdot f_{2j}^s / \tau)}{\sum\limits_{j'; f_i^s \in \{F_1^s, F_2^s\}} \exp(f_{1j}^s \cdot f_{ij'}^s / \tau)} \right], \quad (6)$$

where F_*^s is a set of sparse embeddings of one modality, S_s is the set of F_*^s, τ is the temperature, f_1^s, f_2^s are sam-

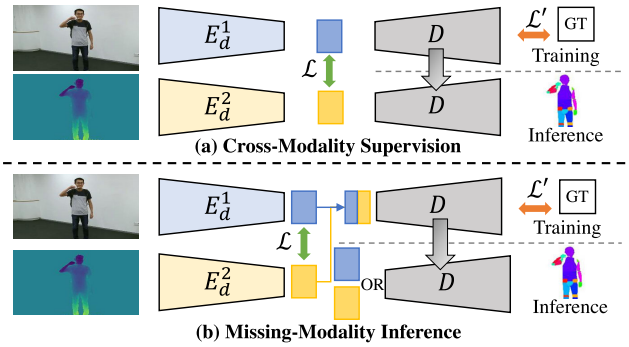


Figure 4. **Pipelines of Two Applications of HCMoCo.**

pled from the union of F_1^s and F_2^s. To conclude, the overall learning target is formulated as Eq. 1, where λ_* are the weights to balance the targets.

3.4. Instantiation of HCMoCo

In this section, we introduce an instantiation of HCMoCo. As shown in Fig. 3, for dense representations, we use RGB and depth. Large-scale paired human RGB and depth data is easy to obtain with affordable sensors *e.g.* Kinect. These two modalities are the most commonly encountered in human-centric tasks [7, 11, 12, 22, 25]. Moreover, a proper pre-train model for depth is highly desired. Therefore, RGB and depth are reasonable choices of human dense representations, both of which are easy to acquire and important to downstream tasks. For sparse representations, 2D keypoints are used, which provide positions of human body joints in the image coordinate. Off-the-shelf tools [4, 9] are available to quickly and robustly extract human 2D keypoints given RGB images. Using 2D keypoints as the sparse representation is a good balance between the amount of human prior and acquisition difficulty.

For RGB inputs I_d^1, an image encoder E_d^1 [43] is applied to obtain feature maps $E_d^1(I_d^1)$. Similarly, for depth inputs I_d^2, an image encoder [43] or 3D encoder [37, 38] E_d^2 can be applied to extract feature maps $E_d^2(I_d^2)$. 2D keypoints I_s are encoded by a GCN-based encoder [51] E_s to produce sparse features $E_s(I_s)$. Mapper networks comprise a single linear layer and a normalization operation.

As for the implementation of contrastive learning targets, we choose to use a memory pool to store all the global embeddings which are updated in a momentum way. Sparse and dense embeddings cannot all fit in memory. Therefore, for the last two types of contrastive learning targets, the negative samples are sampled within a mini-batch.

3.5. Versatility of HCMoCo

On top of the pre-train framework HCMoCo, we propose to further extend it on two direct applications: cross-modality supervision and missing-modality inference. The extensions are based on the key design of HCMoCo: dense intra-sample contrastive learning target. With the feature maps of different modalities aligned, it is straightforward to implement the two extensions, which are shown in Fig. 4.

Cross-Modality Supervision is a novel task where we train the network on the source modality, while test on the target modality. This is a practical scenario where people transfer the knowledge of some single modality dataset to other modalities. At training time, an additional downstream task head (*e.g.* segmentation head) D is attached to the backbone of the source modality. The hierarchical contrastive learning targets $\mathcal{L}$ together with downstream task loss $\mathcal{L}'$ are used for end-to-end training. At inference time, D is attached to the backbone of the target modality. The extracted feature maps of the target modality are passed to D for prediction.

Missing-Modality Inference is another novel task where we train the network using multi-modal data and inference on single modality. Multi-modal data collection in practice would inevitably result in data with incomplete modalities, which brings the requirement of missing-modality inference. At training time, the feature maps of multiple modalities are fused using max-pooling and fed to a downstream task head D. Similarly, hierarchical contrastive learning targets $\mathcal{L}$ and downstream task loss $\mathcal{L}'$ are used for co-training. At inference time, the feature map of a single modality is passed to D for missing-modality inference.

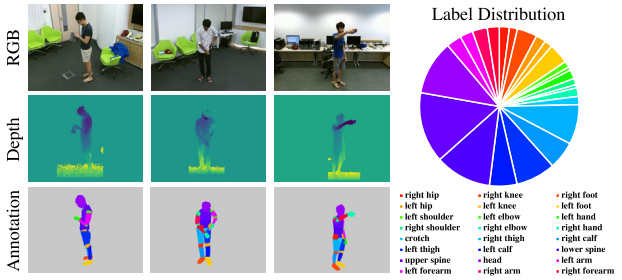


Figure 5. Illustration of the RGB-D human parsing dataset **NTURGBD-Parsing-4K**.

4. NTURGBD-Parsing-4K Dataset

Although RGB human parsing has been well studied [7, 11, 12, 25], human parsing on depth [19, 32] or RGB-D data has not been fully addressed due to the lack of labeled data. Therefore, we contribute the first RGB-D human parsing dataset: NTURGBD-Parsing-4K. The RGB and depth are uniformly sampled from NTU RGB+D (60/120) [27, 42]. As shown in Fig. 5, we annotate 24 human parts for paired RGB-D data. The partition protocols follow that of [22]. The train and test set both have 1963 samples. The whole dataset contains 3926 samples. Hopefully, by contributing this dataset, we could promote the development of both human perception and multi-modality learning.

Table 1. DensePose Estimation Results on COCO. * randomly initializes the model before pre-training. † initializes the model by ImageNet pre-train before pre-training. All results in [%].

Method	Pre-train Datasets	Full Data				10% Data			
		BBox AP	GPS AP	GPSM AP	IOU AP	BBox AP	GPS AP	GPSM AP	IoU AP
From Scratch	-	57.27	62.03	63.61	65.88	39.38	35.75	41.62	49.92
CMC* [45]	NTURGBD+MPII	60.33	64.97	65.66	66.96	44.92	43.84	47.94	54.00
MMV* [1]	NTURGBD+MPII	59.89	64.23	65.47	67.03	43.24	41.40	45.99	52.52
Ours*	NTURGBD+MPII	**61.33**	**65.89**	**66.92**	**67.66**	**47.76**	**48.47**	**51.65**	**56.15**
IN Pre-train	-	62.66	66.48	67.42	68.63	48.28	44.34	49.11	56.11
CMC† [45]	NTURGBD+MPII	62.76	66.16	67.30	68.06	49.21	48.82	52.57	57.94
MMV† [1]	NTURGBD+MPII	62.97	66.67	67.51	68.29	50.16	50.28	53.54	58.32
Ours†	NTURGBD+MPII	**63.11**	**67.33**	**68.12**	**68.72**	**50.29**	**51.50**	**54.47**	**58.66**
CMC† [45]	NTURGBD+COCO	**63.58**	67.22	67.77	68.46	51.77	53.53	56.18	59.37
Ours†	NTURGBD+COCO	62.95	**67.77**	**68.29**	**68.63**	**52.18**	**54.01**	**56.64**	**59.93**

Table 2. Human Parsing Results on Human3.6M. * randomly initializes the model before pre-training. † initializes the model by ImageNet pre-train before pre-training. All results in [%].

Method	Full Data			20% Data			10% Data			1% Data		
	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc
From Scratch	44.13	58.88	98.82	42.41	56.25	98.81	32.61	43.76	98.52	7.27	10.97	97.45
CMC* [45]	54.33	68.01	99.09	52.10	65.65	99.03	48.37	61.18	98.95	14.61	20.07	98.07
MMV* [1]	52.69	65.82	99.06	50.66	63.55	99.01	46.23	58.52	98.90	12.86	17.10	97.94
Ours*	**61.36**	**75.09**	**99.25**	**59.17**	**73.44**	**99.19**	**57.08**	**71.75**	**99.13**	**16.55**	**22.27**	**98.18**
IN Pre-train	56.90	69.94	99.14	48.86	60.75	98.97	44.55	56.86	98.87	14.65	20.22	98.09
CMC† [45]	58.93	71.70	99.20	57.41	70.13	99.17	54.35	67.47	99.09	17.77	23.77	98.20
MMV† [1]	59.08	71.57	99.20	57.28	69.69	99.17	53.86	66.46	99.08	17.66	23.54	98.20
Ours†	**62.50**	**75.84**	**99.27**	**60.85**	**74.23**	**99.23**	**58.28**	**71.99**	**99.17**	**20.78**	**27.52**	**98.34**

5. Experiments

5.1. Experimental Setup

Implementation Details. The default RGB and depth encoders are HRNet-W18 [43]. The default datasets for pre-train are NTU RGB+D [27] and MPII [2]. The former provides paired indoor human RGB, depth, and 2D keypoints, The latter provides in-the-wild human RGB and 2D keypoints. Mixing human data from different domains helps our pre-train models adapt to a wilder domain.

Downstream Tasks. We test our pre-train models on four different human-centric downstream tasks, two on RGB images and two on depth. **1)** DensePose estimation on COCO [14]: DensePose aims at mapping pixels of the observed human body to the surface of a 3D human body, which is a highly challenging task. **2)** RGB human parsing on Human3.6M [22]. Human3.6M provides pure human part segmentation, which aligns with our objectives. We uniformly sample 2fps of the video for training and evaluation. **3)** Depth human parsing on NTURGBD-Parsing-4K. **4)** 3D pose estimation from depth maps on ITOP [16] (only side view). For all the above downstream tasks, we use the pre-train backbones for end-to-end fine-tune.

Comparison Methods. Since there are few previous human-centric multi-modal pre-train methods, we propose to use general multi-modal contrastive learning methods CMC [45] and MMV [1] as the baselines. Although there are other multi-modal contrastive learning works, they either require the multi-view calibration [21] or focus on multi-modal downstream tasks [17, 29] and therefore are not suitable for comparison. In addition, for RGB tasks, we also experiment under two settings, one initializes encoders with supervised ImageNet [24] (IN) pre-train while the other does not.

5.2. Performance on Downstream Tasks

DensePose Estimation. As shown in Tab. 1, we test DensePose estimation [14] under two settings: full and 10% of the training data. The trained models are tested on the full validation set of DensePose. Firstly, if not using IN pre-train, our pre-train model significantly outperforms both 'From Scratch' and two baseline methods. Especially under 10% of training data, **12.7%** improvement in terms of GPS AP is observed. And our pre-train model even outperforms that using IN pre-train by **4.13%** in terms of GPS AP. When we use IN pre-train as initialization, which is a common practice for 2D tasks, our method still outperforms all the baselines. Our method surpasses IN pre-train by **7.2%** and **5.4%** in terms of GPS/GPSM AP under 10% setting. To further test the performance of in-domain transfer, we also

Table 3. Ablation Study on Densepose/ Human3.6M/ ITOP/ NTURGBD-Parsing-4K. All results in [%].

Method	DensePose 10%				ITOP 0.1%/ 0.2%		Human3.6M 10%		NTURGBD 20%	
	BBox	GPS	GPSM	IoU	Acc	Acc	mIoU	mAcc	mIoU	mAcc
Sample-level Mod-invariant	49.21	48.82	52.57	57.94	57.73	50.08	54.35	67.47	30.40	51.54
+ Hard Dense Intra-sample	49.40	49.14	52.49	57.30	56.43	54.05	55.36	68.43	31.26	51.54
+ Soft Dense Intra-sample	50.21	50.25	53.42	57.70	62.33	51.50	56.35	69.26	32.20	51.06
+ Sparse Structure-aware	**50.29**	**51.50**	**54.47**	**58.66**	**65.83**	**62.36**	**58.28**	**71.99**	**35.01**	**52.55**

pre-train models using training sets of NTU RGB+D and COCO. The performance gain under 10% setting further improves to **9.7%** and **7.5%** in terms of GPS/GPSM AP.

RGB Human Parsing. As shown in Tab. 2, we test four settings on Human3.6M [22]: full, 20%, 10% and 1% training data. In all settings, our method outperforms all baselines in all metrics. On full training data, we outperform IN pre-train by **5.6%** in terms of mIoU. The performance gain increases with the amount of training data decreases. It is worth noticing that with only 10% of training data, our method outperforms IN pre-train with full training data.

Table 4. Human Parsing Results on NTURGBD-Parsing-4K [%].

Method	Full Data			20% Data		
	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc
IN Pre-train	37.49	57.52	98.36	28.56	46.81	98.10
CMC [45]	38.20	58.73	98.39	30.40	51.54	98.02
MMV [1]	38.09	58.49	98.37	30.41	50.62	98.07
Ours	**39.32**	**58.79**	**98.47**	**35.01**	**52.55**	**98.53**

Depth Human Parsing. As shown in Tab. 4, we test the pre-train depth backbone on our proposed Dataset NTURGBD-Parsing-4K with all training data and 20% training data. We outperform all baselines on two settings. Especially, only using 20% of training data, we surpass IN pre-train by **6.4%** and MMV [1] by **4.6%** in terms of mIoU.

Table 5. 3D Pose Estimation Results on ITOP. All results in [%].

Method	100%	10%	1%	0.5%	0.2%	0.1%
IN Pre-train	85.19	83.44	77.20	54.31	13.27	14.21
CMC [45]	87.08	85.36	79.49	75.07	57.73	50.08
MMV [1]	86.13	83.49	**79.70**	71.70	60.83	54.44
Ours	**87.19**	**85.49**	78.71	**76.34**	**65.83**	**62.36**

3D Pose Estimation. As shown in Tab. 5, we test the pre-train depth backbone on ITOP [16] with six different ratios of training data. Our pre-train model outperforms all baselines on most settings. With only 10% training data, the accuracy of our method outperforms that of IN pre-train with all training data. It is also worth noticing that 0.1% of training data are 17 samples, which makes this a few-shot learning setting. With such limited training data, IN pre-train barely produce meaningful results, while our method improves the accuracy by **48.2%**.

5.3. Ablation Study

In this subsection, we perform a thorough ablation study on HCMoCo to justify the design choices. As shown in Tab. 3, we firstly report the results of only applying sample-level modality-invariant representation learning. Then we add dense intra-sample contrastive learning and sparse structure-aware contrastive learning in order. To further demonstrate the effect of the 'soft' design in dense intra-sample contrastive learning, we also report results of the 'hard' learning target, which takes the form of a classic InfoNCE [33]. We report the results of the ablation study on all four downstream tasks under data-efficient settings.

For DensePose estimation, it is important to learn feature maps that are continuously and ordinally distributed, which is the expected result of soft dense intra-sample contrastive learning. The performance gain of the soft learning target over the hard counterpart justifies the observation and the learning target design. The dense intra-sample contrastive learning also shows superiority on three other downstream tasks, which shows the importance of fine-grained contrastive learning targets for dense prediction tasks.

Explicitly injecting human prior into the network through sparse structure-aware contrastive learning also proves its effectiveness by further improving the performance on DensePose. Thanks to the strong hints provided by 2D keypoints, the performance of 3D pose estimation is improved. Moreover, the sparse structure-aware contrastive learning boosts the performance of human parsing both on RGB and depth maps by **1.9%** and **2.8%** respectively in terms of mIoU. Although 2D keypoints are sparse priors, they still provide the rough location of each part of the human body, which facilitate the feature alignment of same body parts. To summarize, the sparse and dense learning targets both contribute to the performance of our methods, which is in line with our analysis.

5.4. Performance on HCMoCo Versatility

Cross-Modality Supervision. We test the cross-modality supervision pipeline on the task of human parsing on NTURGBD-Parsing-4K because it has two modalities and respective dense annotations. Two baseline methods are adopted: 1) using CMC [45] contrastive learning target; 2) no contrastive learning target. For a fair comparison, the backbones of all methods are initialized by CMC [45] pre-train. At training time, the target modality of training data

Table 6. Cross-Modality Supervised Human Parsing Results on NTURGBD-Parsing-4K. All results in [%].

Method	RGB → Depth			Depth → RGB		
	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc
No Contrastive	3.94	4.36	92.24	3.71	4.03	91.63
CMC [45]	3.86	5.59	86.81	3.85	4.27	91.75
Ours	**33.19**	**54.38**	**94.70**	**26.80**	**48.80**	**92.84**

is not available. We experiment on two settings where we supervise on RGB, test on depth (RGB → Depth), and vice versa (Depth → RGB). As shown in Tab. 6, our method outperforms both baselines under two settings. Specifically, our method improves the mIoU of both settings by **29.2%** and **23.0%**, respectively. Even compared to methods with direct supervision, we can achieve comparable results.

Table 7. Missing-Modality Human Parsing Results on NTURGBD-Depth. All results in [%].

Method	Only RGB			Only Depth		
	mIoU	mAcc	aAcc	mIoU	mAcc	aAcc
No Contrastive	13.45	14.77	93.35	24.41	30.49	95.27
CMC [45]	19.62	28.19	92.94	16.58	19.83	93.94
Ours	**43.88**	**64.27**	**96.15**	**43.98**	**63.66**	**96.34**

Missing-Modality Inference. For missing-modality inference, we report the experiments on the same dataset and same baselines as above. As shown in Tab. 7, with no pixel-level alignment, the two baseline methods struggle in two missing-modality settings *i.e.* 'Only RGB' and 'Only Depth'. While our method improves the segmentation mIoU by **24.3%** and **19.6%** on two settings.

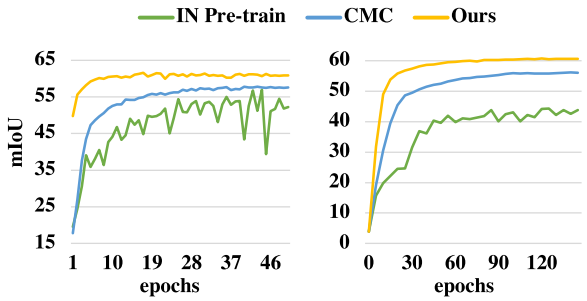


Figure 6. Validation mIoU Changes with Training Epochs Increase. Left: Human3.6M human parsing full training set. Right: Human3.6M human parsing 20% training set.

5.5. Further Analysis

Faster Convergence. One of the advantages of pre-training is the fast convergence speed when transferred to downstream tasks. Our HCMoCo also shows superiority in this feature. We log the validation mIoU of Human3.6M human parsing at different training epochs. As shown in Fig. 6, compared with IN pre-train and CMC [45], our pre-train model is able to converge within a few training epochs in both the full training data and data-efficient settings.

Table 8. Experiments on changing the backbone. * stands for 'IN Pre-train' for DensePose and 'From Scratch' for NTURGBD-Parsing-4K. All results in [%].

Method	DensePose 10%				NTURGBD 20%	
	BBox	GPS	GPSM	IoU	mIoU	mAcc
*	**55.10**	54.60	57.60	61.73	45.36	59.51
CMC [45]	53.88	54.62	57.46	61.14	48.74	62.94
Ours	54.55	**55.80**	**58.36**	**61.75**	**49.43**	**63.52**

Changing Backbone. So far our experiments are all performed on HRNet-W18. To further demonstrate HCMoCo's performance on other backbones, for the 2D backbone, we also experiment with HRNet-W32 [43]. For the depth backbone, we choose to test with PointNet++ [38]. For the RGB pre-train model, we experiment on the 10% DensePose estimation. For the depth pre-train model, we experiment on the 20% NTURGBD-Parsing-4K. As shown in Tab. 8, our method outperforms its pre-train counterparts by a reasonable margin, which is in line with our previous experimental results.

6. Discussion and Conclusion

In this work, we propose the first versatile multi-modal pre-training framework HCMoCo specifically designed for human-centric perception tasks. Hierarchical contrastive learning targets are designed based on the nature of human datasets and the requirements of human-centric downstream tasks. Extensive experiments on four different human downstream tasks of different modalities demonstrated the effectiveness of our pre-training framework. We contribute a new RGB-D human parsing dataset NTURGBD-Parsing-4K to support research of human perception on RGB-D data. Besides downstream task transfer, we also propose two novel applications of HCMoCo to show its versatility and ability in cross-modal reasoning.

Potential Negative Impacts & Limitations. Usage of large amounts of data and long training time might negatively impact the environment. Moreover, even though we did not collect any new human data in this work, human data collection could happen if our framework is used in other applications, which potentially raises privacy concerns. As for the limitations, due to limited resources, we could only experiment with one possible instantiation of HCMoCo. And for the same reason, even though the theoretical possibility exists, we do not have the chance to further scale up the amount of human dataset and network size.

Acknowledgments This work is supported by NTU NAP, MOE AcRF Tier 2 (T2EP20221-0033), and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s).

References

[1] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 2(6):7, 2020. 3, 6, 7, 15, 16, 17

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 1, 2, 6

[3] Mahmoud Assran, Nicolas Ballas, Lluis Castrejon, and Michael Rabbat. Supervision accelerates pre-training in contrastive semi-supervised learning of visual representations. *arXiv preprint arXiv:2006.10803*, 2020. 2

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 2, 5

[5] 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. 3

[6] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 3

[7] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1978, 2014. 1, 2, 5

[8] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13359–13368, 2021. 1

[9] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020. 2, 5

[10] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2

[11] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 770–785, 2018. 1, 2, 3, 5

[12] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 932–940, 2017. 1, 2, 3, 5

[13] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020. 3

[14] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018. 1, 2, 3, 6, 12, 13

[15] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *arXiv preprint arXiv:2010.09709*, 2020. 3

[16] Albert Haque, Boya Peng, Zelun Luo, Alexandre Alahi, Serena Yeung, and Li Fei-Fei. Towards viewpoint invariant 3d human pose estimation. In *European Conference on Computer Vision*, pages 160–177. Springer, 2016. 2, 6, 7, 12

[17] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131, 2020. 3, 6

[18] 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. 3

[19] Antonio Hernández-Vela, Nadezhda Zlateva, Alexander Marinov, Miguel Reyes, Petia Radeva, Dimo Dimov, and Sergio Escalera. Graph cuts optimization for multi-limb human segmentation in depth maps. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 726–732. IEEE, 2012. 2, 5

[20] Fangzhou Hong, Liang Pan, Zhongang Cai, and Ziwei Liu. Garment4d: Garment reconstruction from point cloud sequences. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 2

[21] Ji Hou, Saining Xie, Benjamin Graham, Angela Dai, and Matthias Nießner. Pri3d: Can 3d priors help 2d representation learning? *arXiv preprint arXiv:2104.11225*, 2021. 2, 3, 6

[22] Catalin Ionescu, Dragos 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*, 36(7):1325–1339, jul 2014. 1, 2, 5, 6, 7, 12, 13, 14

[23] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020. 2

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 6

[25] Jianshu Li, Jian Zhao, Yunchao Wei, Congyan Lang, Yidong Li, Terence Sim, Shuicheng Yan, and Jiashi Feng. Multiple-human parsing in the wild. *arXiv preprint arXiv:1705.07206*, 2017. 1, 2, 3, 5

[26] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2

[27] 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. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019. 1, 5, 6

[28] Songtao Liu, Zeming Li, and Jian Sun. Self-emd: Self-supervised object detection without imagenet. *arXiv preprint arXiv:2011.13677*, 2020. 3

[29] Yunze Liu, Qingnan Fan, Shanghang Zhang, Hao Dong, Thomas Funkhouser, and Li Yi. Contrastive multimodal fusion with tupleinfonce. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 754–763, 2021. 3, 6

[30] Yunze Liu, Li Yi, Shanghang Zhang, Qingnan Fan, Thomas Funkhouser, and Hao Dong. P4contrast: Contrastive learning with pairs of point-pixel pairs for rgb-d scene understanding. *arXiv preprint arXiv:2012.13089*, 2020. 2, 3

[31] 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*, pages 2640–2649, 2017. 1, 2

[32] Kaichiro Nishi and Jun Miura. Generation of human depth images with body part labels for complex human pose recognition. *Pattern Recognition*, 71:402–413, 2017. 2, 5

[33] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2, 4, 7

[34] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, Joao F Henriques, Geoffrey Zweig, and Andrea Vedaldi. Multi-modal self-supervision from generalized data transformations. *arXiv preprint arXiv:2003.04298*, 2020. 3

[35] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, João F Henriques, Geoffrey Zweig, and Andrea Vedaldi. On compositions of transformations in contrastive self-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9577–9587, 2021. 3

[36] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR, 2019. 3

[37] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 5

[38] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017. 5, 8, 13

[39] 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. *arXiv preprint arXiv:2103.00020*, 2021. 3

[40] N Dinesh Reddy, Laurent Guigues, Leonid Pishchulin, Jayan Eledath, and Srinivasa G Narasimhan. Tessetrack: End-to-end learnable multi-person articulated 3d pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15190–15200, 2021. 1, 2

[41] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, et al. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*, 2020. 3

[42] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 1, 2, 5, 12

[43] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019. 1, 2, 5, 6, 8, 12

[44] Feitong Tan, Danhang Tang, Mingsong Dou, Kaiwen Guo, Rohit Pandey, Cem Keskin, Ruofei Du, Deqing Sun, Sofien Bouaziz, Sean Fanello, et al. Humangps: Geodesic preserving feature for dense human correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1820–1830, 2021. 2

[45] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. 2, 3, 4, 6, 7, 8, 13, 15, 16, 17

[46] Longhui Wei, Lingxi Xie, Jianzhong He, Jianlong Chang, Xiaopeng Zhang, Wengang Zhou, Houqiang Li, and Qi Tian. Can semantic labels assist self-supervised visual representation learning? *arXiv preprint arXiv:2011.08621*, 2020. 2

[47] 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. 3

[48] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 1, 2

[49] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou, and Junsong Yuan. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 793–802, 2019. 2, 12, 13

[50] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action

recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 1

[51] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3425–3435, 2019. 5

Supplementary Material

In this supplementary material, we provide more implementation details of HCMoCo, downstream tasks and two applications (Sec. 1). More detailed quantitative and qualitative results of downstream tasks are also illustrated (Sec. 2 and Sec. 3).

1. Implementation Details

1.1. HCMoCo

Network Details. For the proposed instantiation of HCMoCo, we implement the sample-level modality-invariant representation learning target by maintaining a memory pool, which is adapted from an open-sourced implementation [2]. The memory pool is updated in a momentum style with the momentum of 0.5. For global embeddings, we sample 16384 negative samples from the memory pool. For other hyper-parameters, we use a batch size of 224, a learning rate of 0.03, a temperature of 0.07 for all three contrastive learning targets. For the pre-train, 4 NVIDIA V100 GPUs are used. The training process is divided in two steps. The first step only pre-train the model using sample-level modality-invariant representation learning target for 100 epochs. The second stage adds the other two learning targets and trains for another 100 epochs. The whole training process takes approximately 48 hours.

Mixing Heterogeneous Datasets. Since we mix several heterogeneous human datasets for pre-train, we need to mask out the missing modalities. For example, when we use NTU RGB+D and MPII for pre-train. The former dataset has all the required modalities, while the latter one misses depth maps. Therefore, for the hierarchical contrastive learning targets, we mask out the missing depth embeddings of MPII for all the positive pairs sampling. By using the masking technique, it is possible to combine multiple heterogeneous datasets into this pre-train paradigm as long as there are at least two common modalities.

Datasets for Pre-train. For NTU RGB+D, we only use the version with 60 actions [42]. With the provided RGB-D videos, we uniformly sample one frame from every 30 frames, which makes 143648 samples. The RGB and depth frames are calibrated by the correspondences provided by the 2D keypoints positions on RGB and depths. For MPII and COCO, we use the full training sets for pre-train.

1.2. DensePose Estimation

For the DensePose [14] estimation, we use the official open-sourced implementation [3]. For the full training set, we train the network for 130000 iterations with a batch size of 16, a learning rate of 0.01 on 4 NVIDIA V100 GPUs, which takes around 80 hours to train. For the 10% training set, we train the network for 13000 iterations with a learning rate of 0.005 and other settings the same, which takes 9 hours to train. The 10% training set is uniformly sampled from the default ordered training set.

1.3. RGB Human Parsing

For the RGB human parsing on Human3.6M [22], we use the official HRNet [43] semantic segmentation implementation [4]. Different ratios of training settings are uniformly sampled from the default ordered full training set. For the full training set, we train the network for 50 epochs with a learning rate of 0.007, a batch size of 40 on 2 NVIDIA V100 GPUs. For other data-efficient settings, we train the network for 150 epochs with other settings the same. We use the the standard dataset split protocol, where the subjects $1, 5, 6, 7, 8$ are for training and the subjects 9 and 11 are for evaluation.

1.4. Depth Human Parsing

For the depth human parsing on NTURGBD-Parsing-4K, we use the same implementation as that of RGB human parsing. To use the HRNet to encode depth maps, we repeat the depth dimension for three times to fit the RGB input, which is also how HCMoCo deals with depth inputs. For all training settings, we train the network for 150 epochs with a learning rate of 0.007, a batch size of 80 on 2 NVIDIA V100 GPUs. Even though the encoder is used to deal with depth inputs, we still initialize it using ImageNet pre-train for that it might help with the performance proved by some previous works [49].

1.5. 3D Pose Estimation on Depth

For the 3D pose estimation from depth maps on ITOP [16], we choose to adapt the official implementation [5] of A2J [49]. The original implementation uses ResNet as the backbone. And we switch to HRNet. Since the original implementation only provides validation scripts, we re-implement the whole training pipeline. We change the original normalization method where a global mean and variance is counted for a global normalization. Instead, we perform an online instance normalization where we only centralize each depth pixel to zero mean but do not normalize its variance, since its a better way to prevent the over-fitting to the relatively small dataset. We train the network for 50 epochs with a learning rate of 0.00035 and a batch size of 12 on one NVIDIA V100 GPU. As for the dataset, we use the side-view of ITOP since the depth maps in pre-train are side views. Following the official dataset split, there are 17991 samples for training and 4863 for testing. Following

[2] https://github.com/HobbitLong/PyContrast

[3] https://github.com/facebookresearch/detectron2

[4] https://github.com/HRNet/HRNet-Semantic-Segmentation

[5] https://github.com/zhangboshen/A2J

the practice of A2J [49], we initialize the encoders using ImageNet pre-train.

1.6. Cross-Modality Supervision

To experiment with the cross-modality supervision, we choose the downstream task of human parsing on NTURGBD-Parsing-4K. The modalities to experiment with are RGB and depth. To make the experiment fair and the networks to converge faster, the backbones are initialized by CMC [45] pre-train. The following descriptions are for the setting of 'RGB→Depth', where the source modality is RGB and the target modality is depth. To implement 'Depth→RGB', one can simply switch the source and target modalities. At training time, a randomly initialized segmentation header, which is the same one used for human parsing experiments, is attached to the dense mapper network of RGB. Then the network is trained with both the hierarchical contrastive learning targets $\mathcal{L}$ and a cross-entropy loss $\mathcal{L}'$ for the supervision of the segmentation. For the 'No Contrastive' baseline, we only train with $\mathcal{L}'$. As for the 'CMC' baseline, the network is supervised by both the learning target proposed by CMC [45] $\mathcal{L}$ and the segmentation loss $\mathcal{L}'$. Note that, during the whole training time, including the CMC pre-train, the target modality of NTURGBD-Parsing-4K is not exposed to better simulate the application scenario. In order to build the connection between RGB and depth during training time, we mix the NTURGBD-Parsing-4K with NTU RGB+D which is the same one used for our pre-train. At inference time, we attach the trained segmentation head to the mapper network of depth. Since the dense embeddings of RGB and depth are aligned thanks to our hierarchical contrastive learning targets, it is reasonable for the segmentation head to be able to handle the dense embeddings of depth.

1.7. Missing-Modality Inference

We also use human parsing on NTURGBD-Parsing-4K to experiment with our extension of missing-modality inference. The basic setup is the same as that of the cross-modality supervision experiments. At training time, we take the dense embeddings of both RGB and depth together for a max pooling operation for a simple feature-level fusion. Then the fused dense embedding is passed to a segmentation header, which is the same one used by the human parsing experiment, to produce the segmentation prediction. The network is supervised with both the hierarchical contrastive learning targets $\mathcal{L}$ and a cross-entropy loss $\mathcal{L}'$ for segmentation supervision. Similarly, the 'No contrastive' baseline does not use any contrastive learning targets. The 'CMC' baseline uses the contrastive learning target proposed in CMC [45] as $\mathcal{L}$. At inference time, if RGB is missing, then the dense embedding of depth is passed to the trained segmentation header for prediction. Since the dense embeddings of RGB and depth are aligned and the segmentation header is trained with the fusion of both embeddings, missing one of them will still produce reasonable predictions.

2. More Quantitative Results

DensePose Estimation. Due to the page limitation, we could not report all metrics for the DensePose [14] estimation. Therefore, we report them in this supplementary material. As shown in Tab. 9, detailed results of all settings mentioned in the main paper are listed. Specifically, for the initialization of the network, we test with the network randomly initialized ('From Scratch') and the network initialized by ImageNet pre-train ('IN Pre-train'). As for the ratio of training data, we test with the full training set and 10% of the training set. As for the pre-train datasets, we test with two combinations: NTU RGB+D + MPII and NTU RGB+D + COCO. As for the backbone, we test with HRNet-W18 and HRNet-W32. Compared with the baseline and two other state-of-the-art pre-train counterparts, our method outperforms them in most of the metrics. Especially, our method has advantages in GPS and GPSM, which are two critical metrics for DensePose quality. Additionally, we also report full results of the ablation study. The detailed results further validates the analysis in the main paper.

RGB Human Parsing. We further report detailed RGB human parsing results on Human3.6M [22] that could not fit into the main paper. As shown in Tab. 10, we report the per-class IoU for all the settings reported in the main paper. Similarly, for the initialization of the network, we test with the network randomly initialized ('From Scratch') and the network initialized by ImageNet pre-train ('IN Pre-train'). As for the ratio of training data, we test with the full training set, 20%, 10% and 1% of the training set. The pre-train datasets are NTU RGB+D + MPII. In most classes, our method outperforms comparison methods. Moreover, we also report per-class IoU for the four settings in ablation study, which are in line with our analysis in the main paper.

Depth Human Parsing. We report detailed depth human parsing results on NTURGBD-Parsing-4K. As shown in Tab. 11, we report the per-class IoU for all the settings reported in the main paper. We initialize the networks using ImageNet pre-train. Two ratios of the training set, *i.e.* full and 20%, are tested. We also change the backbone to PointNet++ [38] ('PN++'). Since it is a point-based backbone, the 'background' class is ignored and not included in the calculation of mIoU. The per-class IoU results also agree with the conclusion in the main paper that our method is superior than other comparison methods.

Cross-Modality Supervision. As shown in Tab. 12, we report detailed per-class IoU for the experiments of cross-modality supervision. In both 'RGB→Depth' and 'Depth→RGB' settings, our method outperforms other

baseline methods in all classes. Especially, other baseline methods barely make correct predictions while ours makes a huge improvement.

Missing-Modality Inference. As shown in Tab. 12, we list detailed per-class IoU for the experiments of missing-modality inference. In both 'Only RGB' and 'Only Depth' settings, our method outperforms baseline methods in most classes. Therefore, the detailed results further validates the conclusions made in the main paper.

3. More Qualitative Results

More qualitative results of RGB human parsing on Human3.6M [22] and depth human parsing on NTURGBD-Parsing-4K are shown in Fig. 7, Fig 8 and Fig. 9. We choose to visualize both the full training set and 10% training set for RGB human parsing. The segmentation results produced by our pre-train model are superior than those of other comparison methods, especially in data-efficient settings. For challenging classes like hands and elbows, our method is capable of producing correct predictions constantly while other methods struggle. The depth map is a challenging modality for the dense prediction task like semantic segmentation. Our method manages to produce reasonable predictions that are better than those of other comparison methods.

Table 9. Detailed DensePose Estimation Results on COCO. 'Ratio' stands for the ratio of training data for downstream tasks transfer. * randomly initializes the model before pre-training. † initializes the model by ImageNet pre-train before pre-training. 'Ablation1' is 'Sample-level Mod-invariant'; 'Abation2' is '+ Hard Dense Intra-sample'; 'Ablation3' is '+ Soft Dense Intra-sample'; 'Ablation4' is '+ Sparse Structure-aware'. ‡ uses HRNet-W32 while others all use HRNet-W18. All results in [%].

Methods	Pre-train Datasets	Ratio	BBox						GPS										GPSM										IoU									
			AP	AP50	AP75	APs	APm	APl	AP	AP50	AP75	APm	APl	AR	AR50	AR75	ARm	ARl	AP	AP50	AP75	APm	APl	AR	AR50	AR75	ARm	ARl	AP	AP50	AP75	APm	APl	AR	AR50	AR75	ARm	ARl
From Scratch	-	100%	57.27	85.44	61.37	28.31	53.64	72.46	62.03	89.93	69.57	58.29	62.89	69.00	93.05	76.06	60.78	69.55	63.61	91.89	74.56	57.18	64.71	69.24	95.05	79.36	59.72	69.88	65.88	93.86	77.95	58.43	67.04	71.11	96.12	82.97	60.35	71.83
CMC* [45]	NTURGBD+MPII	100%	57.27	85.44	61.37	28.31	53.64	72.46	62.03	89.93	69.57	58.29	62.89	69.00	93.05	76.06	60.78	69.55	63.61	91.89	74.56	57.18	64.71	69.24	95.05	79.36	59.72	69.88	65.88	93.86	77.95	58.43	67.04	71.11	96.12	82.97	60.35	71.83
MMV* [1]	NTURGBD+MPII	100%	60.33	86.82	66.23	30.37	57.29	75.59	64.97	91.10	72.99	59.59	65.90	71.57	93.80	79.00	61.42	72.25	65.66	92.22	77.31	57.81	66.81	70.80	94.96	81.77	59.72	71.54	66.96	93.80	79.14	59.73	67.97	71.83	95.72	83.68	61.42	72.52
Ours*	NTURGBD+MPII	100%	**61.33**	**87.81**	**66.48**	**31.80**	**58.27**	**76.30**	**65.89**	**92.20**	**75.36**	**61.12**	**66.87**	**72.52**	**94.69**	**81.01**	**62.91**	**73.17**	**66.92**	**93.27**	**78.72**	**59.75**	**67.97**	**71.87**	**95.50**	**82.92**	**61.63**	**72.56**	**67.66**	**94.41**	**80.14**	**61.08**	**68.71**	**72.62**	**96.21**	**84.80**	**62.62**	**73.29**
IN Pre-train	-	100%	62.66	**89.28**	68.47	**35.78**	59.59	76.47	66.49	92.11	75.47	**64.45**	67.43	73.41	94.87	81.28	**66.24**	73.89	67.42	93.20	79.72	**62.11**	68.52	72.63	95.90	83.82	**63.97**	73.20	68.63	94.64	81.81	**62.56**	69.72	73.45	96.30	85.64	**64.11**	74.08
CMC† [45]	NTURGBD+MPII	100%	62.76	88.32	68.32	33.67	60.22	76.95	66.17	92.45	74.79	62.79	66.96	72.75	94.69	80.34	64.68	73.29	67.31	93.51	79.71	61.55	68.27	72.23	95.68	83.73	63.55	72.80	68.07	94.68	81.41	61.89	69.01	72.90	96.30	85.51	63.62	73.53
MMV† [1]	NTURGBD+MPII	100%	62.97	88.75	68.91	34.62	**60.83**	76.87	66.67	92.42	76.24	63.09	67.64	73.03	94.78	81.28	64.04	73.63	67.51	93.44	79.72	60.91	68.50	72.29	95.81	83.10	61.99	72.97	68.29	94.52	**81.84**	61.04	69.20	72.95	**96.34**	85.42	62.20	73.67
Ours†	NTURGBD+MPII	100%	**63.11**	88.66	**69.64**	34.53	60.80	**77.01**	**67.33**	**93.20**	**76.27**	62.63	**68.20**	**73.77**	**95.23**	**81.59**	63.83	**74.44**	**68.12**	**94.09**	**79.90**	61.35	**68.99**	**72.94**	**96.08**	**83.95**	62.48	**73.64**	**68.72**	**94.74**	81.67	61.50	**69.77**	**73.47**	96.26	**85.73**	62.70	**74.19**
CMC† [45]	NTURGBD+COCO	100%	**63.58**	**88.94**	**69.69**	**35.24**	**61.37**	**77.46**	67.22	92.68	76.27	**64.61**	68.20	73.75	95.10	81.59	**65.60**	74.29	67.77	**93.65**	79.56	62.62	68.81	72.78	**95.99**	83.55	63.83	73.38	68.46	93.93	**81.80**	**62.60**	69.50	**73.37**	95.94	**85.78**	**63.76**	**74.02**
Ours†	NTURGBD+COCO	100%	62.95	88.78	68.83	34.77	60.59	77.02	**67.77**	**93.18**	**77.13**	64.02	**68.63**	**74.15**	**95.23**	**82.39**	65.18	**74.75**	**68.29**	93.60	**80.53**	**62.77**	**69.23**	**73.21**	**95.99**	**84.44**	**64.11**	**73.81**	**68.63**	**94.53**	81.53	62.36	**69.51**	73.30	**96.17**	85.69	63.33	73.97
From Scratch	-	10%	39.38	72.29	37.98	12.03	35.59	55.16	35.75	73.78	30.07	27.19	37.28	45.32	81.41	43.69	31.49	46.25	41.62	80.25	38.81	30.71	43.33	49.67	87.34	50.74	35.60	50.61	49.92	85.96	54.21	38.56	51.61	57.90	91.62	65.14	43.19	58.88
CMC* [45]	NTURGBD+MPII	10%	44.92	76.41	46.34	16.09	41.58	60.88	43.84	79.94	42.97	40.31	45.00	52.25	85.69	54.21	42.55	52.90	47.94	84.22	50.23	41.42	49.33	55.00	89.43	59.61	43.90	55.74	54.00	88.65	61.20	46.16	55.45	60.96	92.73	69.82	48.94	61.77
MMV* [1]	NTURGBD+MPII	10%	43.24	74.53	44.27	13.49	40.43	59.22	41.40	77.54	39.26	34.54	42.93	50.45	84.22	51.67	37.02	51.35	45.99	81.83	48.11	37.22	47.64	53.58	88.19	58.18	39.72	54.51	52.52	87.73	59.10	43.45	54.01	59.80	92.15	68.26	46.17	60.71
Ours*	NTURGBD+MPII	10%	**47.76**	**78.66**	**50.56**	**18.59**	**45.19**	**63.12**	**48.47**	**82.74**	**51.27**	**45.21**	**49.51**	**56.41**	**87.56**	**61.21**	**47.73**	**56.99**	**51.65**	**85.50**	**56.06**	**45.86**	**52.76**	**58.27**	**90.33**	**64.20**	**48.44**	**58.91**	**56.15**	**89.63**	**64.34**	**49.27**	**57.43**	**62.92**	**93.36**	**72.54**	**51.13**	**63.71**
IN Pre-train	-	10%	48.29	80.07	50.90	19.55	44.54	63.64	44.34	78.77	44.83	36.85	46.49	54.54	86.49	57.38	39.29	55.56	49.11	84.02	51.91	40.57	51.38	57.78	90.82	63.17	43.40	58.73	56.11	88.10	63.14	46.91	58.16	63.73	92.56	72.94	49.50	64.69
CMC† [45]	NTURGBD+MPII	10%	49.21	80.43	52.01	21.66	47.22	63.71	48.82	83.85	51.39	43.97	50.05	57.21	88.72	61.35	46.67	57.92	52.57	86.67	57.21	45.69	53.89	59.52	91.26	66.21	48.65	60.25	57.94	**90.82**	66.11	50.11	59.30	64.80	**94.07**	74.50	52.34	65.64
MMV† [1]	NTURGBD+MPII	10%	50.16	**81.33**	**53.68**	22.17	47.73	64.76	50.28	84.02	53.95	44.06	51.68	58.43	89.03	63.40	45.74	59.28	53.54	86.66	59.25	44.91	55.06	60.35	91.22	67.59	46.74	61.25	58.32	90.15	67.10	49.21	59.89	**64.94**	93.62	74.90	50.92	**65.88**
Ours†	NTURGBD+MPII	10%	**50.29**	80.94	53.62	**22.74**	**47.91**	**65.25**	**51.50**	**84.89**	**54.64**	**45.63**	**52.68**	**59.03**	**89.17**	**63.58**	**47.80**	**59.78**	**54.47**	**87.13**	**60.26**	**46.82**	**55.71**	**60.66**	**91.44**	**67.99**	**49.29**	**61.42**	**58.66**	90.32	**68.82**	**50.63**	**59.93**	**64.94**	93.89	**75.57**	**52.48**	65.77
CMC† [45]	NTURGBD+COCO	10%	51.77	81.86	55.93	23.64	49.54	66.50	53.53	86.23	58.30	46.79	54.78	60.81	90.15	66.47	50.21	61.52	56.18	88.58	64.41	48.38	57.39	62.35	92.33	71.73	**52.13**	63.04	59.37	90.74	68.92	**52.20**	60.62	65.83	93.85	76.42	**54.68**	66.58
Ours†	NTURGBD+COCO	10%	**52.18**	**82.71**	**56.57**	**24.92**	**49.96**	**66.86**	**54.01**	**86.41**	**58.31**	**48.15**	**55.14**	**61.53**	**90.55**	**67.10**	**50.99**	**62.24**	**56.64**	**88.67**	**64.79**	**48.78**	**57.88**	**63.05**	**92.69**	**72.18**	51.70	**63.81**	**59.93**	**91.31**	**69.99**	52.01	**61.08**	**66.50**	**94.34**	**77.40**	54.54	**67.29**
Ablation1	NTURGBD+MPII	10%	49.21	80.43	52.01	21.66	47.22	63.71	48.82	83.85	51.39	43.97	50.05	57.21	88.72	61.35	46.67	57.92	52.57	86.67	57.21	45.69	53.89	59.52	91.26	66.21	48.65	60.25	57.94	**90.82**	66.11	50.11	59.30	64.80	**94.07**	74.50	52.34	65.64
Ablation2	NTURGBD+MPII	10%	49.40	80.93	51.61	21.94	47.56	63.72	49.14	82.17	51.83	44.53	50.43	57.65	87.83	62.42	46.88	58.37	52.49	85.69	58.42	46.56	53.84	59.61	90.68	66.87	49.15	60.31	57.30	89.93	65.38	50.58	58.74	64.29	93.80	73.96	**53.12**	65.03
Ablation3	NTURGBD+MPII	10%	50.21	**81.02**	53.39	**22.86**	47.78	64.75	50.25	84.20	53.58	44.74	51.41	57.95	88.50	62.91	47.66	58.64	53.42	86.41	58.95	46.27	54.68	59.85	90.68	66.87	49.01	60.56	57.70	89.75	67.40	**50.95**	58.89	64.40	93.31	74.72	52.91	65.17
Ablation4	NTURGBD+MPII	10%	**50.29**	80.94	**53.62**	22.74	**47.91**	**65.25**	**51.50**	**84.89**	**54.64**	**45.63**	**52.68**	**59.03**	**89.17**	**63.58**	**47.80**	**59.78**	**54.47**	**87.13**	**60.26**	**46.82**	**55.71**	**60.66**	**91.44**	**67.99**	**49.29**	**61.42**	**58.66**	90.32	**68.82**	50.63	**59.93**	**64.94**	93.89	**75.57**	52.48	**65.77**
IN Pre-train‡	-	10%	**55.10**	**84.95**	**59.81**	**27.33**	52.01	**69.65**	54.60	86.42	59.80	48.63	55.93	62.48	90.68	68.12	50.43	63.29	57.60	88.62	66.39	49.61	59.15	64.09	92.55	72.49	51.63	**64.92**	61.73	**91.74**	72.52	53.24	**63.10**	**67.99**	**94.69**	78.73	55.11	**68.86**
CMC‡ [45]	NTURGBD+MPII	10%	53.88	83.61	58.47	26.16	51.24	68.64	54.62	86.83	58.93	47.45	55.86	62.05	90.73	67.19	49.01	62.93	57.46	88.57	65.55	49.59	58.71	63.57	92.33	72.05	51.21	64.39	61.14	91.65	72.58	53.46	62.39	67.21	94.25	78.73	54.96	68.03
Ours‡	NTURGBD+MPII	10%	54.55	83.77	58.83	26.94	**52.56**	68.78	**55.80**	**87.37**	**61.22**	**51.83**	**56.75**	**62.70**	**90.91**	**68.48**	**53.12**	**63.34**	**58.36**	**89.29**	**67.31**	**52.03**	**59.39**	**64.11**	**92.87**	**73.52**	**53.55**	64.82	**61.75**	91.59	**72.79**	**54.77**	62.87	67.72	94.38	**79.00**	**56.38**	68.48

Table 10. Detailed Human Parsing Results on Human3.6M. 'Ratio' stands for the ratio of training data for downstream tasks transfer. * randomly initializes the model before pre-training. † initializes the model by ImageNet pre-train before pre-training. 'Ablation1' is 'Sample-level Mod-invariant'; 'Abation2' is '+ Hard Dense Intra-sample'; 'Ablation3' is '+ Soft Dense Intra-sample'; 'Ablation4' is '+ Sparse Structure-aware'. All results in [%].

Methods	Ratio	bg	right hip	right knee	right foot	left hip	left knee	left foot	left shoulder	left elbow	left hand	right shoulder	right elbow	right hand	crotch	right thigh	right calf	left thigh	left calf	lower spine	upper spine	head	left arm	left forearm	right arm	right forearm	mIoU
From Scratch	100%	99.53	29.66	32.28	52.94	25.86	30.71	49.59	35.14	20.78	29.46	34.33	19.01	30.61	37.24	54.53	65.77	54.35	64.12	50.56	57.39	77.57	40.71	35.83	38.14	36.99	44.13
CMC* [45]	100%	99.55	44.95	48.96	54.05	37.57	41.55	53.41	47.98	33.35	37.05	46.01	32.28	38.18	50.81	64.84	69.95	62.22	68.16	67.76	71.78	81.23	56.32	46.75	55.92	47.53	54.33
MMV* [1]	100%	99.54	39.92	45.88	53.18	34.04	34.77	52.97	45.35	31.22	34.10	45.08	31.92	35.26	53.11	63.96	69.41	60.81	66.78	67.53	72.35	79.96	55.63	43.48	56.07	44.99	52.69
Ours*	100%	**99.61**	**51.09**	**58.38**	**61.22**	**40.81**	**48.02**	**59.65**	**54.21**	**44.75**	**48.57**	**52.78**	**43.66**	**48.68**	**56.70**	**71.43**	**75.89**	**67.98**	**73.73**	**71.63**	**76.79**	**83.89**	**64.64**	**57.51**	**64.47**	**57.97**	**61.36**
From Scratch	20%	99.52	29.07	32.15	50.17	23.11	24.90	48.12	34.98	17.26	23.58	33.20	17.75	25.39	34.76	54.76	65.22	51.44	62.03	55.16	60.63	78.27	39.47	30.30	37.31	31.84	42.41
CMC* [45]	20%	99.53	41.21	42.49	49.99	39.44	39.37	49.24	46.58	30.02	33.19	45.35	30.08	34.48	53.06	62.33	66.06	61.24	64.88	67.02	71.59	80.35	54.80	42.67	54.03	43.61	52.10
MMV* [1]	20%	99.51	38.19	43.71	49.45	34.44	32.47	49.59	42.88	28.15	30.74	43.23	28.81	32.13	52.26	62.83	66.82	59.55	63.94	66.44	71.66	79.42	54.10	40.48	54.03	41.73	50.66
Ours*	20%	**99.59**	**46.76**	**51.17**	**56.35**	**43.52**	**48.66**	**55.20**	**53.79**	**42.58**	**43.76**	**52.41**	**42.29**	**44.58**	**60.19**	**68.57**	**71.95**	**67.80**	**70.67**	**70.56**	**75.74**	**82.68**	**62.90**	**52.71**	**62.44**	**52.48**	**59.17**
From Scratch	10%	99.39	18.84	19.95	35.96	19.55	15.23	32.38	23.23	12.43	15.57	20.73	12.46	18.64	28.46	40.41	45.51	37.85	42.10	51.25	58.41	74.93	25.77	20.47	23.29	22.50	32.61
CMC* [45]	10%	99.49	35.71	37.05	44.08	36.65	33.37	42.36	44.43	23.56	29.01	43.30	24.51	30.05	52.29	60.05	60.32	58.81	58.03	65.01	70.68	79.19	52.11	37.89	51.56	39.64	48.37
MMV* [1]	10%	99.46	33.84	34.77	42.58	33.47	27.94	41.83	37.89	21.06	26.17	38.65	23.36	27.40	50.05	58.14	59.71	56.35	57.01	64.47	69.99	77.38	50.43	35.00	51.34	37.35	46.23
Ours*	10%	**99.56**	**42.65**	**48.60**	**52.82**	**44.08**	**45.44**	**51.58**	**52.21**	**40.08**	**41.70**	**51.70**	**40.35**	**42.73**	**59.06**	**66.85**	**69.09**	**66.78**	**67.47**	**68.73**	**74.82**	**81.39**	**61.15**	**49.08**	**60.39**	**48.64**	**57.08**
From Scratch	1%	98.39	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	18.12	0.00	15.94	0.00	18.71	30.70	0.00	0.00	0.00	0.00	0.00	7.27
CMC* [45]	1%	98.95	0.00	0.00	4.08	0.00	0.00	2.69	0.00	0.00	0.00	0.00	0.00	0.00	7.38	22.13	20.40	22.19	17.72	48.99	52.65	68.09	0.00	0.00	0.00	0.00	14.61
MMV* [1]	1%	98.62	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	22.03	12.54	18.33	9.34	46.48	54.31	59.75	0.00	0.00	0.00	0.00	12.86
Ours*	1%	**99.02**	0.00	0.00	**9.95**	0.00	0.00	**8.70**	0.00	0.00	0.00	0.00	0.00	0.00	**9.97**	**26.15**	**25.71**	**24.97**	**24.00**	**57.31**	**58.61**	**69.24**	0.00	0.00	0.00	0.00	**16.55**
IN Pre-train	100%	99.60	37.98	52.94	56.49	32.24	46.97	56.29	46.05	47.55	42.50	42.81	47.05	42.88	46.57	66.69	**75.49**	64.51	**74.23**	61.85	67.29	82.69	58.41	57.98	56.56	**58.90**	56.90
CMC† [45]	100%	99.60	44.36	51.60	57.77	41.26	**51.46**	54.51	53.20	42.49	42.28	51.82	41.57	41.95	55.85	68.72	72.62	67.34	72.18	69.21	74.36	82.78	63.95	54.88	63.38	54.21	58.93
MMV† [1]	100%	99.60	44.58	54.97	59.49	40.01	48.30	56.47	51.56	43.01	43.60	50.11	42.12	42.97	56.61	69.42	74.69	67.35	72.90	68.54	73.46	82.52	63.02	55.04	62.54	54.13	59.08
Ours†	100%	**99.62**	**50.34**	**57.06**	**60.78**	**42.96**	50.52	**57.75**	**56.66**	**48.56**	**48.11**	**56.50**	**47.64**	**47.86**	**60.06**	**72.46**	75.23	**70.20**	73.58	**73.16**	**78.24**	**84.36**	**66.48**	**58.91**	**66.96**	58.60	**62.50**
IN Pre-train	20%	99.53	36.33	37.57	49.56	32.73	33.11	47.01	42.29	29.79	34.31	39.52	29.84	33.40	46.71	55.73	60.20	54.27	58.36	64.67	68.74	81.83	49.21	43.69	48.51	44.49	48.86
CMC† [45]	20%	99.58	44.54	49.05	54.20	42.95	**47.77**	52.52	50.89	40.65	40.08	49.67	39.46	39.85	56.26	68.09	70.69	66.45	69.76	69.23	73.95	81.87	62.20	52.63	61.35	51.45	57.41
MMV† [1]	20%	99.58	44.94	51.80	56.68	38.96	45.12	53.29	50.16	40.35	40.29	48.96	39.99	39.31	56.14	68.21	72.31	66.00	70.07	68.50	73.38	81.62	61.54	52.27	61.23	51.28	57.28
Ours†	20%	**99.60**	**48.34**	**54.49**	**58.57**	**43.35**	47.57	**55.65**	**55.46**	**45.50**	**46.55**	**55.34**	**44.65**	**46.47**	**61.25**	**71.12**	**73.61**	**69.45**	**71.37**	**71.44**	**76.74**	**83.52**	**64.86**	**56.40**	**64.23**	**55.73**	**60.85**
IN Pre-train	10%	99.52	31.06	30.86	42.62	31.49	28.61	39.59	40.24	24.58	28.15	38.36	25.25	26.65	47.38	50.07	53.83	49.90	52.64	63.70	68.21	80.72	44.98	35.24	44.19	35.80	44.55
CMC† [45]	10%	99.55	38.50	45.93	48.87	42.27	42.37	47.59	48.30	37.07	36.52	47.46	36.72	36.07	55.22	65.35	66.68	65.23	64.92	67.31	72.76	81.25	59.06	48.16	58.33	47.14	54.35
MMV† [1]	10%	99.54	38.12	45.86	50.23	37.23	40.96	48.24	47.57	36.17	36.33	46.99	36.13	35.34	55.36	64.36	66.91	63.40	64.91	67.40	72.44	80.88	58.72	47.61	58.71	46.98	53.86
Ours†	10%	**99.57**	**43.01**	**49.27**	**53.44**	**46.22**	**46.13**	**52.20**	**53.68**	**40.73**	**43.25**	**54.63**	**41.18**	**43.80**	**59.88**	**67.81**	**69.59**	**69.08**	**67.91**	**69.02**	**75.36**	**82.30**	**62.99**	**51.10**	**63.35**	**51.42**	**58.28**
IN Pre-train	1%	99.07	0.00	0.00	8.97	0.00	0.00	6.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	19.38	24.79	19.34	23.11	43.39	48.80	72.37	0.00	0.00	0.00	0.00	14.65
CMC† [45]	1%	99.08	0.00	0.00	17.72	0.00	0.00	14.43	0.00	0.00	0.00	0.00	0.00	0.00	33.14	26.55	24.95	25.94	22.73	48.98	52.21	72.88	2.65	0.00	2.99	0.00	17.77
MMV† [1]	1%	99.10	0.00	0.00	12.83	0.00	0.00	10.97	0.00	0.00	0.00	0.00	0.00	0.00	23.62	26.53	24.72	26.40	24.54	51.18	53.27	72.29	8.15	0.00	7.96	0.00	17.66
Ours†	1%	**99.15**	0.00	0.00	**18.29**	0.00	0.00	**16.99**	**0.07**	0.00	0.00	**0.07**	0.00	0.00	**46.56**	**28.73**	**27.59**	**29.13**	**26.38**	**58.59**	**61.46**	**74.41**	**16.32**	0.00	**15.65**	0.00	**20.78**
Ablation1	10%	99.55	38.50	45.93	48.87	42.27	42.37	47.59	48.30	37.07	36.52	47.46	36.72	36.07	55.22	65.35	66.68	65.23	64.92	67.31	72.76	81.25	59.06	48.16	58.33	47.14	54.35
Ablation2	10%	99.56	40.06	45.94	50.88	40.18	43.91	49.06	50.41	36.73	37.65	49.50	37.33	36.94	57.19	66.32	68.69	67.28	66.98	68.08	73.37	81.95	60.25	48.45	59.80	47.54	55.36
Ablation3	10%	**99.57**	**44.49**	46.38	52.15	44.00	43.85	50.25	51.73	37.19	38.12	51.30	37.24	37.61	57.38	67.16	69.08	66.91	67.32	**69.42**	74.45	82.12	61.36	49.91	60.69	48.97	56.35
Ablation4	10%	**99.57**	43.01	**49.27**	**53.44**	**46.22**	**46.13**	**52.20**	**53.68**	**40.73**	**43.25**	**54.63**	**41.18**	**43.80**	**59.88**	**67.81**	**69.59**	**69.08**	**67.91**	69.02	**75.36**	**82.30**	**62.99**	**51.10**	**63.35**	**51.42**	**58.28**

Table 11. Detailed Human Parsing Results on NTURGBD-Parsing-4K. 'Ratio' stands for the ratio of training data for downstream tasks transfer. * randomly initializes the model before pre-training. † initializes the model by ImageNet pre-train before pre-training. 'Ablation1' is 'Sample-level Mod-invariant'; 'Abation2' is '+ Hard Dense Intra-sample'; 'Ablation3' is '+ Soft Dense Intra-sample'; 'Ablation4' is '+ Sparse Structure-aware'. All results in [%].

Methods	Ratio	bg	right hip	right knee	right foot	left hip	left knee	left foot	left shoulder	left elbow	left hand	right shoulder	right elbow	right hand	crotch	right thigh	right calf	left thigh	left calf	lower spine	upper spine	head	left arm	left forearm	right arm	right forearm	mIoU
IN Pre-train	100%	99.24	21.99	19.79	39.62	23.95	20.98	39.61	22.61	14.14	22.11	23.05	12.24	22.52	25.66	47.00	46.85	46.81	48.53	53.26	61.51	61.11	43.45	36.13	46.50	38.56	37.49
CMC† [45]	100%	99.26	22.50	19.49	40.15	24.81	20.65	39.96	**24.64**	14.68	21.79	**25.50**	13.04	23.46	**26.12**	48.93	46.21	49.28	47.68	54.26	62.52	59.24	45.02	37.22	48.32	40.25	38.20
MMV† [1]	100%	99.23	22.64	19.68	39.02	24.66	21.38	38.77	24.40	13.92	22.36	25.15	12.83	23.65	25.78	47.87	46.62	48.12	48.45	53.32	62.46	59.89	45.03	37.80	48.17	40.98	38.09
Ours†	100%	**99.32**	**22.95**	**21.25**	**41.26**	**25.70**	**22.59**	**40.99**	24.47	**15.17**	**23.61**	25.11	**14.26**	**24.87**	25.75	**49.46**	**47.90**	**49.97**	**49.88**	**54.45**	**62.88**	**61.97**	**47.24**	**39.06**	**50.48**	**42.29**	**39.32**
IN Pre-train	20%	99.13	12.39	17.07	31.82	14.08	19.28	32.72	13.68	2.61	12.58	14.59	2.96	10.87	18.16	35.01	33.36	38.41	35.82	46.81	54.45	57.49	31.95	23.39	32.65	22.60	28.56
CMC† [45]	20%	98.98	**13.16**	14.06	29.82	**16.28**	16.13	30.99	17.65	10.09	14.69	17.99	8.23	15.92	18.08	38.79	34.57	41.26	36.41	46.21	54.68	54.60	37.80	26.75	39.11	27.66	30.40
MMV† [1]	20%	99.03	12.92	16.15	30.81	15.62	18.80	32.06	17.75	7.46	15.27	18.73	5.52	15.72	**19.48**	38.29	34.88	40.95	37.82	45.78	54.42	55.21	36.83	25.66	37.85	27.15	30.41
Ours†	20%	**99.40**	12.18	**18.25**	**36.74**	15.35	**21.15**	**38.91**	**18.78**	**11.99**	**20.92**	**20.31**	**10.99**	**20.84**	18.05	**44.73**	**44.45**	**47.86**	**46.87**	**51.41**	**59.57**	**65.76**	**42.43**	**31.07**	**44.67**	**32.65**	**35.01**
Ablation1	20%	99.13	12.39	17.07	31.82	14.08	19.28	32.72	13.68	2.61	12.58	14.59	2.96	10.87	18.16	35.01	33.36	38.41	35.82	46.81	54.45	57.49	31.95	23.39	32.65	22.60	28.56
Ablation2	20%	99.12	12.85	13.79	32.69	**16.99**	16.20	33.29	17.38	8.73	15.12	19.13	7.67	16.71	**19.07**	39.46	36.15	42.08	38.08	48.69	55.53	58.94	38.49	27.62	40.04	27.55	31.26
Ablation3	20%	99.22	**13.40**	14.18	31.37	16.46	17.71	32.88	**19.72**	9.33	16.13	20.19	8.95	17.49	17.48	39.12	37.41	41.55	39.13	49.67	58.23	61.41	39.71	30.95	41.16	32.13	32.20
Ablation4	20%	**99.40**	12.18	**18.25**	**36.74**	15.35	**21.15**	**38.91**	18.78	**11.99**	**20.92**	**20.31**	**10.99**	**20.84**	18.05	**44.73**	**44.45**	**47.86**	**46.87**	**51.41**	**59.57**	**65.76**	**42.43**	**31.07**	**44.67**	**32.65**	**35.01**
From Scratch w/ PN++	20%	-	21.43	30.84	67.53	21.52	29.85	66.76	32.82	23.17	38.37	36.74	23.42	36.23	25.19	55.25	60.25	55.11	60.37	53.34	65.85	88.22	50.77	47.05	52.70	45.88	45.36
CMC* [45] w/ PN++	20%	-	**24.12**	32.89	73.11	23.84	**32.67**	**73.20**	33.43	27.55	44.62	**38.40**	27.59	42.11	26.55	57.82	65.60	57.73	65.02	54.53	66.31	89.16	55.04	51.00	56.66	50.82	48.74
Ours* w/ PN++	20%	-	23.96	**32.90**	**73.30**	**24.16**	32.44	73.10	**34.81**	**29.54**	**45.43**	37.79	**28.16**	**42.89**	**27.83**	**58.25**	**66.16**	**58.64**	**65.51**	**55.60**	**66.92**	**89.51**	**56.33**	**53.05**	**57.87**	**52.29**	**49.43**

Table 12. Detailed Cross-Modality Supervision and Missing-Modality Inference Results on NTURGBD-Parsing-4K. All results in [%].

Methods	Setting	bg	right hip	right knee	right foot	left hip	left knee	left foot	left shoulder	left elbow	left hand	right shoulder	right elbow	right hand	crotch	right thigh	right calf	left thigh	left calf	lower spine	upper spine	head	left arm	left forearm	right arm	right forearm	mIoU
No Contrastive	RGB→Depth	92.87	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22	0.00	0.00	1.08	0.00	0.00	0.25	0.00	0.00	0.00	0.00	2.62	0.00	0.45	0.00	1.11	3.94
CMC [45]	RGB→Depth	89.79	0.00	0.00	0.00	0.00	0.02	0.01	0.00	1.51	0.79	0.00	0.11	0.43	0.00	0.00	0.00	0.01	0.00	1.02	0.20	0.42	1.64	0.37	0.03	0.14	3.86
Ours	RGB→Depth	**96.78**	**12.50**	**23.26**	**37.60**	**16.45**	**21.43**	**40.51**	**22.41**	**19.59**	**22.86**	**21.35**	**17.19**	**25.67**	**17.40**	**36.48**	**35.62**	**37.43**	**34.04**	**49.83**	**59.37**	**61.07**	**33.91**	**24.55**	**36.15**	**26.37**	**33.19**
No Contrastive	Depth→RGB	91.79	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.23	0.00	0.00	0.07	0.00	0.01	0.00	0.00	0.00	0.10	0.00	0.02	0.00	0.50	0.00	0.02	3.71
CMC [45]	Depth→RGB	91.96	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.68	0.32	0.00	0.00	0.21	0.00	0.00	0.00	0.00	0.00	0.46	0.23	0.05	0.49	1.84	0.00	0.01	3.85
Ours	Depth→RGB	**95.15**	**13.70**	**16.04**	**28.70**	**16.59**	**12.15**	**28.12**	**18.11**	**12.78**	**10.51**	**21.06**	**11.12**	**14.46**	**11.02**	**33.71**	**30.90**	**34.03**	**22.93**	**42.62**	**50.99**	**56.12**	**23.51**	**19.89**	**27.29**	**18.50**	**26.80**
No Contrastive	Only RGB	93.55	8.97	6.66	0.42	4.28	0.75	0.02	0.98	1.19	15.22	0.28	2.69	23.56	0.08	0.30	12.35	0.07	0.88	25.48	5.27	57.54	12.63	26.38	7.73	28.90	13.45
CMC [45]	Only RGB	93.80	0.00	11.94	47.69	0.00	12.00	38.76	21.43	0.01	9.58	24.32	13.56	15.38	1.15	1.84	32.30	1.07	18.12	0.59	3.13	36.86	27.04	15.63	42.37	21.90	19.62
Ours	Only RGB	**97.80**	**29.12**	**27.49**	**57.93**	**27.09**	**26.41**	**57.31**	**28.22**	**27.62**	**32.70**	**29.51**	**26.16**	**33.41**	**25.66**	**52.42**	**49.60**	**52.38**	**54.91**	**52.40**	**52.75**	**75.69**	**47.98**	**41.69**	**50.68**	**40.08**	**43.88**
No Contrastive	Only Depth	96.46	**27.81**	7.16	1.46	**33.36**	10.60	2.30	28.64	4.72	1.27	11.49	0.11	7.86	26.16	33.67	39.55	33.21	27.40	46.05	63.16	47.50	25.38	9.99	19.89	5.04	24.41
CMC [45]	Only Depth	94.81	7.25	1.08	0.06	6.82	0.05	0.10	25.47	3.11	2.95	20.80	0.39	2.73	26.79	17.11	17.84	33.96	4.58	10.64	11.32	37.01	41.41	11.08	33.69	3.52	16.58
Ours	Only Depth	**97.89**	23.09	**32.97**	**54.55**	26.44	**32.55**	**57.28**	**34.08**	**19.83**	**25.23**	**32.33**	**21.33**	**28.71**	**30.57**	**54.91**	**59.33**	**53.16**	**59.89**	**56.50**	**65.09**	**61.87**	**49.43**	**36.47**	**50.34**	**35.75**	**43.98**

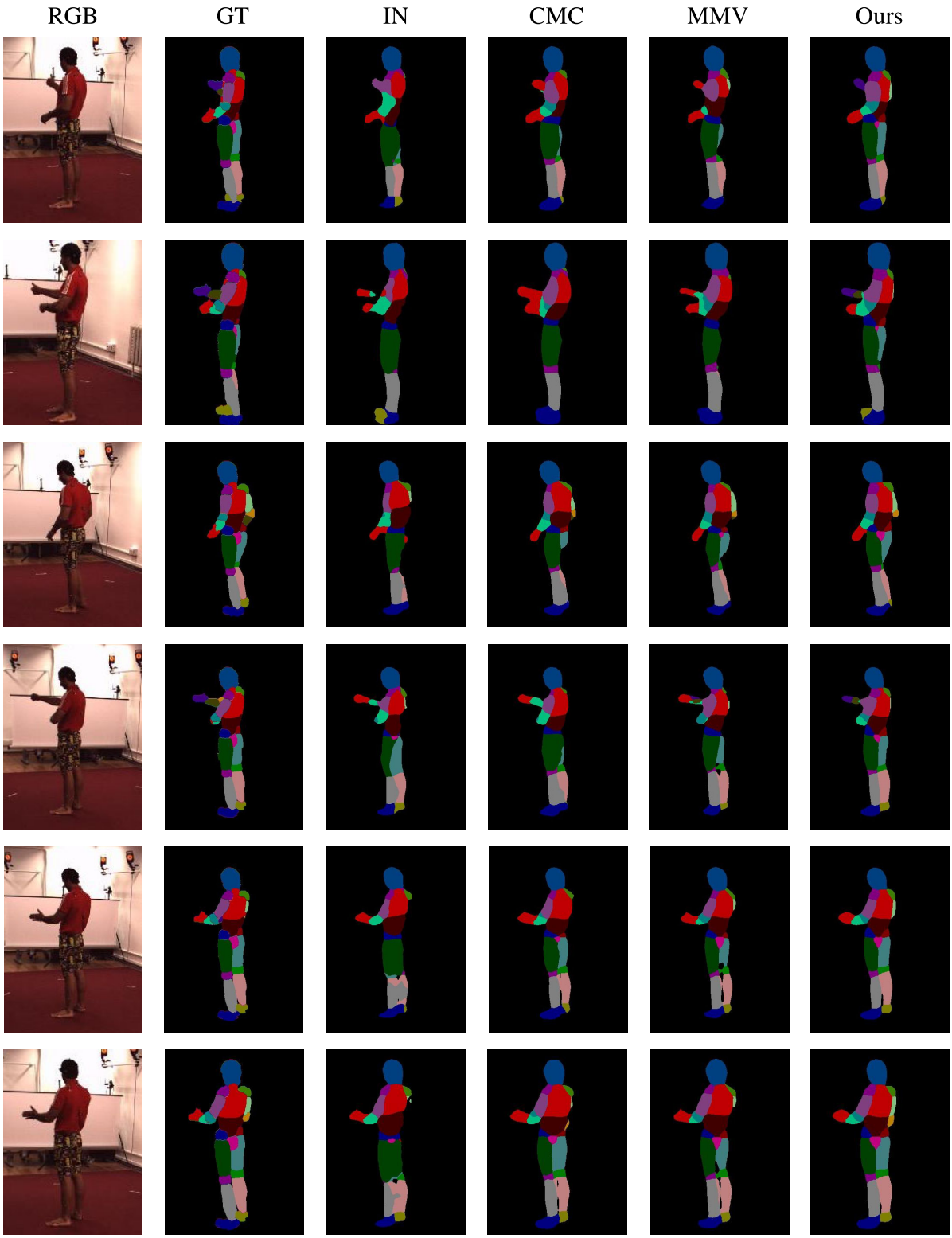


Figure 7. Qualitative Results of RGB Human Parsing on Human3.6M with 10% of the Training Set.

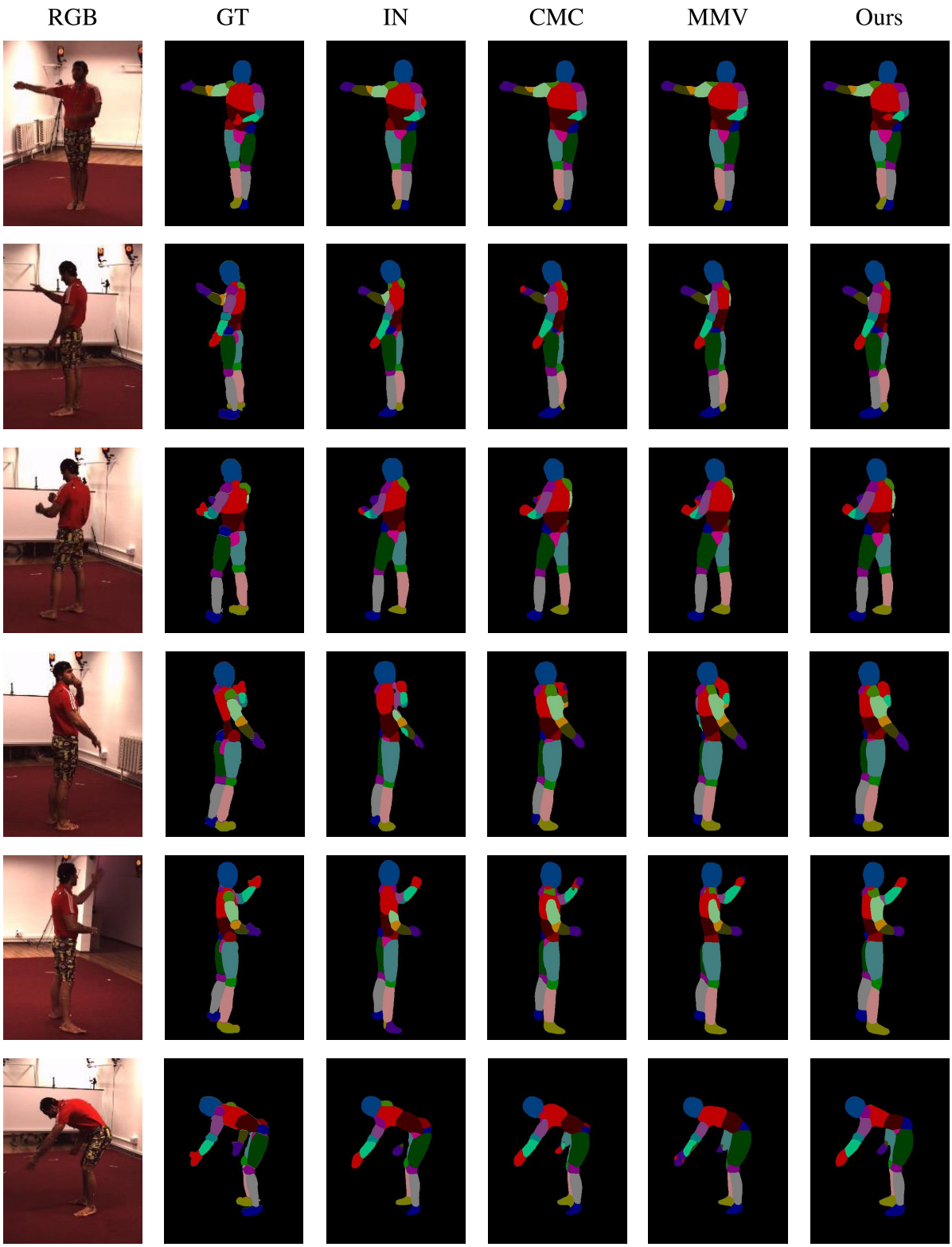


Figure 8. Qualitative Results of RGB Human Parsing on Human3.6M with the Full Training Set.

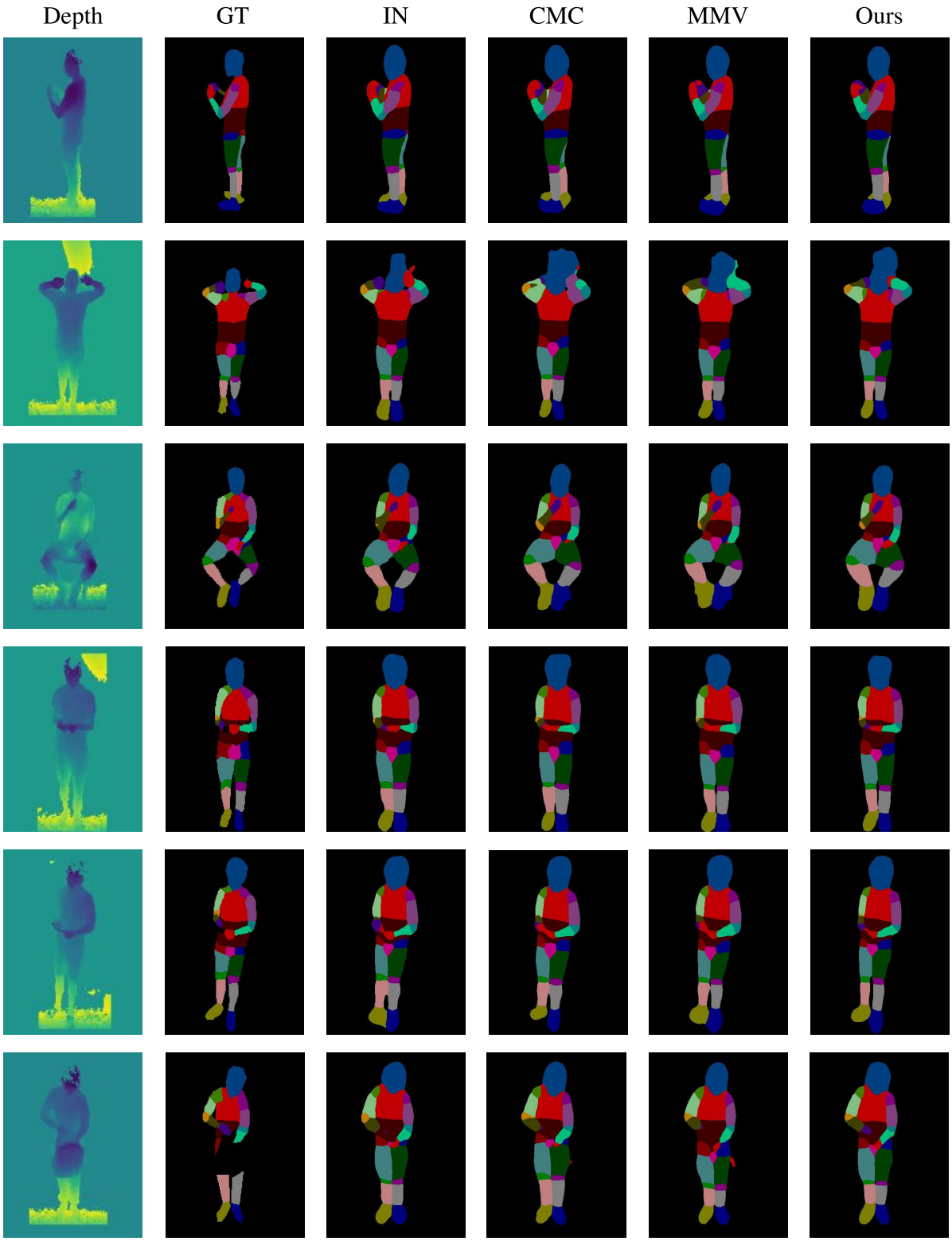


Figure 9. Qualitative Results of Depth Human Parsing on NTURGBD-Parsing-4K with the Full Training Set.