Self-Supervised 3D Hand Pose Estimation from monocular RGB via Contrastive Learning

Adrian Spurr* Aneesh Dahiya* Xi Wang Xucong Zhang Otmar Hilliges
Department of Computer Science, ETH Zurich, Switzerland

Abstract

Encouraged by the success of contrastive learning on image classification tasks, we propose a new self-supervised method for the structured regression task of 3D hand pose estimation. Contrastive learning makes use of unlabeled data for the purpose of representation learning via a loss formulation that encourages the learned feature representations to be invariant under any image transformation. For 3D hand pose estimation, it too is desirable to have invariance to appearance transformation such as color jitter. However, the task requires equivariance under affine transformations, such as rotation and translation. To address this issue, we propose an equivariant contrastive objective and demonstrate its effectiveness in the context of 3D hand pose estimation. We experimentally investigate the impact of invariant and equivariant contrastive objectives and show that learning equivariant features leads to better representations for the task of 3D hand pose estimation. Furthermore, we show that standard ResNets with sufficient depth, trained on additional unlabeled data, attain improvements of up to 14.5% *in PA-EPE on FreiHAND and thus achieves state-of-the-art performance without any task specific, specialized architectures. Code and models are available at https://ait.ethz.ch/projects/2021/PeCLR/*

1. Introduction

Estimating the 3D pose of human hands from monocular images alone has many important applications in robotics, Human-Computer Interaction and AR/VR. As such the problem has received significant attention in computer vision literature [12, 14, 15, 26, 31–33, 41]. However, estimating the location of 3D hand joints within an RGB image is a challenging structured regression problem with difficulties that arise from a large diversity in backgrounds, lighting conditions, hand appearances, as well as self-occlusion caused by the high degrees of freedom of the human hand.

Annotated datasets that cover a larger diversity of environments and settings are one possibility to alleviate this issue. However, acquiring 3D labeled data is laborious, cost intensive and typically requires multi-view imagery or some form of user instrumentation. Data collected under such circumstances is often difficult to transfer well to in-the-wild imagery [20, 42]. Therefore, much interest is given to approaches that can leverage auxiliary data, which has either no or only 2D joint annotations. For example, such data can be used to outperform many supervised approaches via making use of weak-supervision [3, 4], the integration of

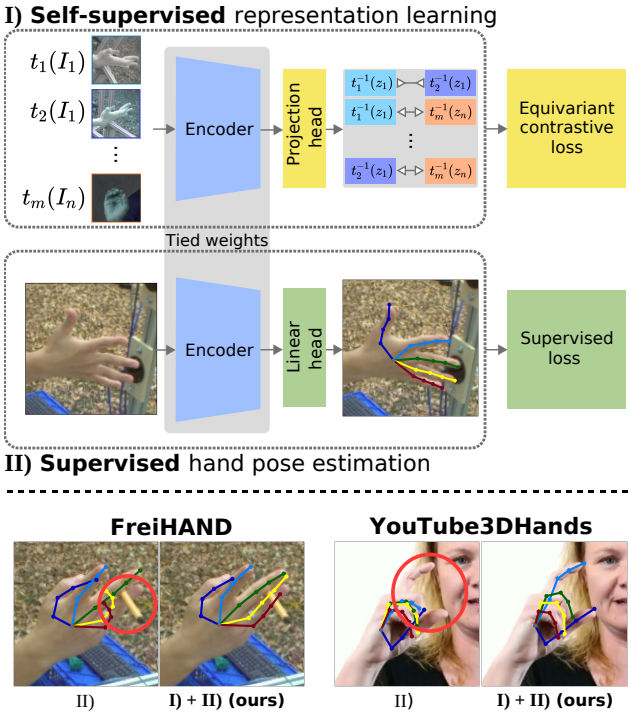


Figure 1: We propose a two-stage framework for 3D hand pose estimation. I) An encoder is trained in a self-supervised manner on a large set of unlabeled data using a novel equivariant contrastive objective. II) The pre-trained encoder is fine-tuned with little labeled data. The resulting network is more accurate across datasets.

*Denotes equal contribution

kinematic priors [31], or by exploiting temporal information [14]. Off-the-shelf joint detectors [5] have been leveraged to automatically generated 2D annotations in large quantities [20]. However, the accuracy of models trained on these labels, or on 3D annotations derived from them, are inherently bounded by the label noise. Therefore, the question of how to efficiently leverage unlabeled data for hand pose estimator training remains unanswered.

Recently, self-supervised approaches such as contrastive learning have shown that they can reach parity with supervised approaches on image classification tasks [6, 8]. These methods leverage unlabeled data to learn powerful feature representations. To do so, positive and negative pairs of images are projected into a latent space via a neural network. The contrastive objective encourages the latent space samples of the positive pairs to lie close to each other and pushes negative pairs apart. The resulting pre-trained network can then be applied to downstream tasks. Positive pairs are created by sampling an image and applying two sets of distinct augmentations on it, whereas negative pairs correspond to separate but similarly augmented images. These augmentations include appearance transformations, such as color drop, and geometric transformations, such as rotation. The contrastive objective induces invariance under all of these transformations. However, 3D regressions tasks, such as hand pose estimation, inherently require *equivariance* under *geometric* transformations. Hence, representations learnt from a standard contrastive objective may not effectively transfer to pose estimation.

To the best of our knowledge, for the first time, we investigate self-supervised representation learning techniques for 3D hand pose estimation in this paper. We derive a method named *Pose Equivariant Contrastive Learning (PeCLR)*. One of our core contributions is a novel formulation of a contrastive learning objective that induces equivariance to geometric transformations and we show that this allows to effectively leverage the large diversity of existing hand images without *any* joint labels. These images are used to pre-train a network, which can then be transferred to the final hand pose estimation task via supervised fine-tuning. This provides a promising direction for hand pose estimation and enables an easy transfer of images collected in-the-wild or calibration to a specific domain by fine-tuning a pre-trained network with fewer labels.

Fig. 1 provides an overview of our method. First, we perform self-supervised representation learning. Given an RGB image of the hand, we apply appearance and geometric transformations to generate positive and negative pairs of derivative images. These are used to train an encoder via our proposed equivariant contrastive loss. By undoing the geometric transformation in latent space, we promote equivariance. However, invertion of these transformations is not straightforward. This is because transformations on images should lead to proportional changes in the latent space. Therefore special care needs to be taken due to different magnitudes between latent space and pixel space under learned projection. We propose a latent sample normalization technique that compensates for this difference and we show that the resulting model yields improved pose estimation accuracy (cf. Fig. 1, bottom) compared to both supervised and standard contrastive learning.

In the second stage, the pre-trained encoder is fine-tuned on the task of 3D hand pose estimation using labeled data. The resulting model is evaluated thoroughly in a variety of settings. We demonstrate increased label efficiency for semi-supervision and show that using more unlabeled data is beneficial for the final performance, yielding improvements of up to 43% in 3D EPE in the lowest labeled setting (cf. Fig. 6). Next, we show that this improvement also transfers to the fully supervised case, where using a standard ResNet with sufficient depth in combination with unlabeled data and our proposed pre-training scheme outperforms specialized state-of-the-art architectures (cf. Tab. 2). Finally, we demonstrate that self-supervised pre-training leads to an improvement of 5.6% 3D PA-EPE in cross-data evaluation, indicating that pre-training is beneficial for cross-domain generalization (cf. Tab. 3).

In summary, our contributions are as follows:

1. To the best of our knowledge, we perform the first investigation of contrastive learning to efficiently leverage unlabeled data for 3D hand pose estimation.
2. We propose a contrastive learning objective that encourages *invariance* to appearance transformations and *equivariance* to geometric transformations.
3. We conduct controlled experiments to empirically derive the best performing augmentations.
4. We show that the proposed method achieves better label efficiency in semi-supervised settings and that adding more unlabeled data is beneficial.
5. We empirically show that our proposed method outperforms current, more specialized state-of-the-art methods using standard ResNet models.

Code and models are available for research purposes: https://ait.ethz.ch/projects/2021/PeCLR/.

2. Related work

Hand pose estimation. Hand pose estimation usually follows one of three paradigms. Some work predicts 3D joint skeletons directly [4, 12, 18, 26, 27, 31–33, 37, 41], make use of MANO [30], where the parameters of a parametric hand model are regressed [1–3, 14, 15, 40], or predicts the full mesh model of the hand directly [13, 21, 25]. A staged approach is introduced in [41], where the 2D keypoints are regressed directly and then lifted to 3D. Spurr *et al*. [32] introduces a cross-modal latent space which facilitates better

learning. Mueller *et al*. [27] makes use of a synthetically created dataset and reduces the synthetic/real discrepancy via a GAN. Cai *et al*. [4] makes use of supplementary depth supervision to augment the training set. Proposing a more efficient hand representation, a 2.5D representation is introduced in [18]. Action recognition as well as hand/object pose estimation is performed in [33]. [37] introduces a disentangled latent space, for the purpose of better image synthesis. A graph-based neural network is used to jointly refine the hand/object pose in [12]. Biomechanical constraints are introduced to refine the pose predictions on 2D supervised data [31]. Moon *et al*. [26] predict the pose of both hands and takes their interaction into account.

Templated-based methods such as MANO induce a prior of hand poses, as well as providing a mesh surface. Some methods [1, 3, 40] estimate the MANO parameters directly from RGB, sometimes making use of weak supervision such as hand masks [1, 40] or in-the-wild 2D annotations [3, 40]. A unified approach is introduced to jointly predict MANO as well as the object mesh [15]. Hasson *et al*. [14] builds upon the mentioned framework, by learning from partially labeled sequences via a photometric loss. An alternative to MANO is proposed in [25] by predicting pose and subject dependant correctives to a base hand model. Some methods regress the mesh of a hand directly. However, mesh annotations are difficult to acquire. Ge *et al*. [13] tackles this by introducing a fully mesh-annotated synthetic dataset and performs noisy supervision for real data. With the help of spiral convolutions, a hand mesh is predicted in [21], supervised using MANO.

Clearly, much work has been dedicated to custom, sometimes highly specialized architectures for hand-pose estimation. In contrast, we explore a purely data-driven approach, utilizing unlabeled data, and an equivariance inducing contrastive formulation to achieve state-of-the-art performance with a standard CNN.

Self-supervised learning. Self-supervised learning aims to learn representation of data without any annotations. Literature defines the pre-text task as the specific strategy to learn the representation in a self-supervised manner. Such tasks include predicting the position of a second patch relative to the first [11], colorizing a grayscale image [39], solving a jigsaw puzzle [28], estimating the motion flow of pixels in a scene [35], predicting positive future samples in audio signals [29], or completing the next sentence based on relations between two sentences [10]. However, it is not clear which pretext task would be optimal given a specific downstream task in terms of performance and generalizability.

Contrastive learning is a powerful paradigm for self-supervised, task-independent learning. At the core of contrastive learning lies a concept emerging from distance metric learning, where a pair of data is encouraged to be close in latent space if they are connected in a meaningful way, while unrelated data are pushed apart. One of the appeals of contrastive learning lie in the numerous amounts of data that is available for training. General representations are learned through this paradigm and have been successfully used in many downstream tasks such as image and video classification [6, 8, 34], object detection [17, 36], and speech classification [29]. However, contrastive learning has not been investigated for the task of hand pose estimation.

Contrastive learning has been explore in works such as Contrastive Predictive Coding (CPC) [17, 29], Contrastive Multiview Coding (CMC) [34], and SimCLR [6, 7]. CPC learns to extract representations by predicting future representations in latent space. Autoregressive models are used to enable predictions of many steps in the future. While CPC learns from the two views of the past and future, CMC extends this idea to multi-view learning. It aims to learn view-invariant representations by maximizing mutual information among different views of the same content. The most relevant framework for contrastive learning is a simple yet effective approach [6]. It largely benefits from data augmentation and its learnt representation achieves performance that is on par with supervised models on the image classification task. However, the learned transformation-invariant features are not suited for structured regression tasks such as hand pose estimation as these require an equivariant representation with respect to geometric transformations. In this work, we extend SimCLR by differentiating between appearance and geometric transformations, and propose a model that can successfully learn representations dedicated for both transformations.

3. Method

We start by reviewing SimCLR [6]. We then introduce the overall framework of pre-training and finetuning. Next, we identify an issue with SimCLRs contrastive formulation when applied to hand pose estimation, motivating our proposed equivariant contrastive objective. Lastly, we present our hand pose estimation model and the method used for 3D keypoint estimation during supervised training.

Notation. In the following, we denote the set of all transformations used as T. It contains *appearance* transformations t^a (e.g color jitter), *geometric* transformations t^g (e.g. scale, rotation and translation) as well as compositions of them. For a given transformation $t_i \in T$, t_i^a, t_i^g correspond to the appearance or geometric component of the transformation t_i. Fig. 4 shows all transformation used in this study.

3.1. SimCLR

The idea of the SimCLR [6] framework is to maximize the agreement in latent space between the representations of samples that are similar, while repelling dissimilar pairs. The positive pairs are artificially generated by applying various augmentations on an image. Given a set of sam-

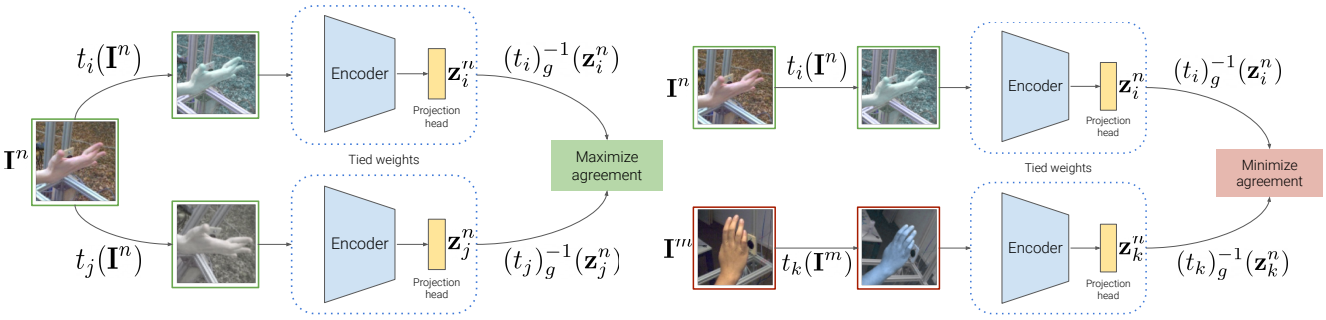


Figure 2: **Method overview**. An augmentation $t = t_g \circ t_a$ is applied to input image $\boldsymbol{I}^n$. Here t_g and t_a denote the geometric and appearance components of the augmentation $t \in T$, respectively. The model then generates the projections $\boldsymbol{z}^n$ for each augmented input. Geometric augmentations are *reversed* in *projection space* before optimizing the contrastive objective. The agreement between projections from the same input image is maximized (left) and agreements amongst projections from different input images are minimized (right).

ples $\{\boldsymbol{I}^n\}_{n=1}^N$, we consider two augmented views $\{\boldsymbol{I}_i^n, \boldsymbol{I}_j^n\}$, where $\boldsymbol{I}_i^n = t_i(\boldsymbol{I}^n)$, $\boldsymbol{I}_j^n = t_j(\boldsymbol{I}^n)$, $t_i, t_j \in T$.

The framework consists of an encoder E and a projection head $g(\cdot)$. The overall model $f = g \circ E$ maps an image $\boldsymbol{I}$ to a latent space sample $\boldsymbol{z} \in \mathbb{R}^k$, i.e. $\boldsymbol{z}_i^n = f(\boldsymbol{I}_i^n)$. It is trained using a contrastive loss function that maximizes the agreement between all positive pairs of projections $\{\boldsymbol{z}_i^n, \boldsymbol{z}_j^n\}_{i \neq j}$, which are extracted from two augmented views of the same image $\boldsymbol{I}^n$. Simultaneously, it also minimizes the agreement amongst negative pairs of projections $\{\boldsymbol{z}_i^n, \boldsymbol{z}_k^m\}$, where $\boldsymbol{z}_k^m$ are extracted from different images.

In each iteration, SimCLR samples both positive and negative pairs. For a given batch of N images, two augmentations are applied on each sample, resulting in $2N$ augmented images. Hence, for every augmented image $\boldsymbol{I}_i^n$, there is one positive sample $\boldsymbol{I}_j^n$, and $2(N-1)$ negative samples $\{\boldsymbol{I}_k^m\}_{m \neq n}$. The model is trained to project positive samples close to each other, whereas keeping negative samples far apart. This is achieved via the following loss function, termed as *NT-Xent* in [6]:

$$\mathcal{L}_{i,j} = -\log \frac{\exp\left(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau\right)} \tag{1}$$

Here τ is a temperature parameter, $\text{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^T\boldsymbol{v}/||\boldsymbol{u}||||\boldsymbol{v}||$ is the cosine similarity between $\boldsymbol{z}_i^n$, $\boldsymbol{z}_j^n$ and $\mathbb{1}_{[k \neq i]}$ is the indicator function.

3.2. Equivariant contrastive representations

Inspecting Eq. 1, we observe that the objective function promotes invariance under all transformations. Given a sample $\boldsymbol{I}_j^n = t_j(\boldsymbol{I}^n)$ and its positive sample $\boldsymbol{I}_i^n = t_i(\boldsymbol{I}^n) = t_i(t_j^{-1}(\boldsymbol{I}_j^n)) = \tilde{t}_i(\boldsymbol{I}_j^n)$, the numerator in Eq. 1 is minimized if $f(\boldsymbol{I}_j^n) = \boldsymbol{z}_j^n = \boldsymbol{z}_i^n = f(\tilde{t}_i(\boldsymbol{I}_j^n))$. Hence, a model that satisfies Eq. 1 needs to be invariant to all transformations in T. However, hand pose estimation requires equivariance with respect to geometric transformations as these change the displayed pose. Hence, we require:

$$t_i^g f(\boldsymbol{I}_j^n) = f(t_i^g(\boldsymbol{I}_j^n)). \tag{2}$$

Inverting transformations in latent space. To fulfill Eq. 2, we first note that it is equivalent to $f(\boldsymbol{I}_j^n) = (t_i^g)^{-1} f(t_i^g(\boldsymbol{I}_j^n)) \leftrightarrow \boldsymbol{z}_j^n = (t_i^g)^{-1}\boldsymbol{z}_i^n$. This leads us to the following equivariant modification of *NT-Xent*:

$$\mathcal{L}_{i,j} = -\log \frac{\exp\left(\text{sim}((\tilde{\boldsymbol{z}}_i, \tilde{\boldsymbol{z}}_j)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\text{sim}(\tilde{\boldsymbol{z}}_i, \tilde{\boldsymbol{z}}_k)/\tau\right)}, \tag{3}$$

where $\tilde{\boldsymbol{z}}_i = (t_i^g)^{-1}\boldsymbol{z}_i$ and $\boldsymbol{z}_i \in \mathbb{R}^{m \times 2}$. In order to minimize the numerator in Eq. 3 it must hold that $\tilde{\boldsymbol{z}}_i = \tilde{\boldsymbol{z}}_j$, which leads to the desired property of Eq. 2. Further details can be found in the supplementary. As t_i^g is an affine transformation, its inverse can be easily computed. However, whereas scaling and rotation are transformations that are performed relative to the image size, translation is performed in terms of an absolute quantity. In other words, if we translate an image $\boldsymbol{I}^n$ by x pixels, we need to translate its latent space projection $\boldsymbol{z}^n$ by a proportional quantity. Therefore, we translate $\boldsymbol{z}^n$ by a quantity proportional to its magnitude. To achieve this, we obtain the translation proportional to the image size and scale it up by a factor proportional to the range spanned by the projections in latent space. To this end, we normalize the translation vector $\hat{\boldsymbol{v}}$ before applying its inverse to a latent space sample $\boldsymbol{z}_i$ to undo the transformation. The normalized vector $\hat{\boldsymbol{v}}$ is computed as follows:

$$\hat{\boldsymbol{v}} = \frac{\boldsymbol{v}}{L} L_z \tag{4}$$

Where $L_z = \max(\boldsymbol{z}_i) - \min(\boldsymbol{z}_i)$ and L is the image length. The intuition behind L_z is that it corresponds to the mag-

nitude of latent space values. Hence, the resulting translation vector is proportional in magnitude. Lastly, we note here that due to the cosine similarity used in Eq. 3, the effect of scaling is effectively removed (*i.e.* $\text{sim}(a\boldsymbol{z}_i, b\boldsymbol{z}_j) = \text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)$, for $a, b \in \mathbb{R}$). The complete equivariant contrastive learning framework is visualized in Fig. 2.

From pre-training to fine-tuning. After having performed pre-training using our proposed loss function, we fine-tune the encoder supervised on the task of hand pose estimation. To this end, following [6] we remove the projection layer g from the model and replace it with a linear layer. The entire model is then trained end-to-end using the losses as described next, in Sec. 3.3.

3.3. 3D Hand Pose Estimator

Our hand pose estimation model makes use of the 2.5D representation [18]. Given an image, the network predicts the 2D keypoints $\boldsymbol{J}^{2D} \in \mathbb{R}^{21\times 2}$ and the root-relative depth $\boldsymbol{d}^r \in \mathbb{R}^{21}$ of the hand. As such, our hand pose model is trained with the following supervised loss functions:

$$\begin{aligned} \mathcal{L}_{\boldsymbol{J}^{2D}} &= |\hat{\boldsymbol{J}}^{2D} - \boldsymbol{J}^{2D}| \\ \mathcal{L}_{\boldsymbol{d}^r} &= |\hat{\mathbf{d}}^r - \boldsymbol{d}^r| \end{aligned} \tag{5}$$

Given the predicted values of $\mathbf{J}^{2D}$ and $\boldsymbol{d}^r$, the depth value of the root keypoint d^{root} can be acquired as detailed in [18]. As a final step, we refine the acquired root depth to increase accuracy and stability as described [31], which yields $d^{\text{root}}_{\text{ref}}$. The resulting 3D pose is acquired as follows:

$$\mathbf{J}^{3D} = \mathbf{K}^{-1}\mathbf{J}^{2D}(\mathbf{d}^r + d^{\text{root}}_{\text{ref}}), \tag{6}$$

where $\mathbf{K}$ is the camera intrinsic matrix.

4. Experiments

Sec. 4.4 investigates the impact of different data augmentation operations and evaluate their effectiveness in the hand pose estimation task. Next, with the self-supervised learnt representation, we demonstrate in Sec. 4.5 how our model efficiently makes use of labeled data in semi-supervised settings. In Sec. 4.6 we compare our method with related works in hand pose estimation and demonstrate that PeCLR can reach state-of-the-art performance on FH. Finally, in Sec. 4.7 we perform a cross-dataset evaluation to show the advantages of the proposed representation learning across domain distributions.

4.1. Implementation

For pre-training, we use ResNet (RN) [16] as encoder, which takes monocular RGB images of size 128×128 as input. We use LARS [38] with ADAM [19] with batches of size 2048 and learning rate of $4.5e$-3 in the representation learning stage. During fine-tuning, we use RGB images of size 128×128 (Sec. 4.4, 4.5) or 224×224 (Sec. 4.6, 4.7). As optimizer we use ADAM with a learning rate of $5e$-4 in the supervised fine-tuning stage. Further training details can be found in the supplementary.

Method	3D EPE ↓ (cm)	AUC ↑	2D EPE ↓ (px)
SimCLR	16.62	0.72	12.05
PeCLR (**ours**)	**16.05**	**0.74**	**10.51**

Table 1: Comparison of SimCLR and PeCLR on FH. The encoders are pre-trained with either SimCLR or PeCLR, and are *frozen* during fine-tuning. Both methods use their optimal set of augmentations, as explained in Sec. 4.4.

4.2. Evaluation Metrics

We report the End-point-error (EPE) and the Area-Under-Curve (AUC). EPE denotes the average euclidean distance between the ground-truth and predicted keypoints. AUC denotes the area under the Percentage-of-correct-Keypoints (PCK) curve for threshold values between 0 and 5 cm in 100 equally spaced increments. Lastly, the prefix PA denotes procrustes-alignment, which globally aligns the ground-truth and prediction using procrustes analysis before computing the metric in question

4.3. Datasets

We use the following datasets in our experiments.

FreiHAND (FH) [42] consists of 32'560 frames captured with green screen background in the training set, as well as real backgrounds in the test set. Its final evaluation is performed online, hence we do not have access to the ground-truth for the test set. We use the FH dataset for all supervised and self-supervised training and report the absolute as well as the procrustes-aligned EPE and AUC.

YouTube3DHands (YT3D) [20] consists of in-the-wild images, with automatically acquired 3D annotations via key point detection from OpenPose [5] and MANO [30] fitting. It contains 47'125 in-the-wild frames. We use the YT3D dataset exclusively for self-supervised representation learning. YT3D contains only 3D vertices and no camera intrinsic information, hence we report the procrustes-aligned EPE and 2D pixel error via weak perspective projection.

4.4. Evaluation of augmentation strategies

To study which set of data augmentations performs best, we first consider various augmentation operations for the representation learning phase. Fig. 4 visualizes the studied transformations in our experiment. We first evaluate individual transformations and then find their best composition.

We conduct the experiment on FH using our own training and validation split (90% as training and 10% as validation set) and use a RN50 as the encoder. We train two

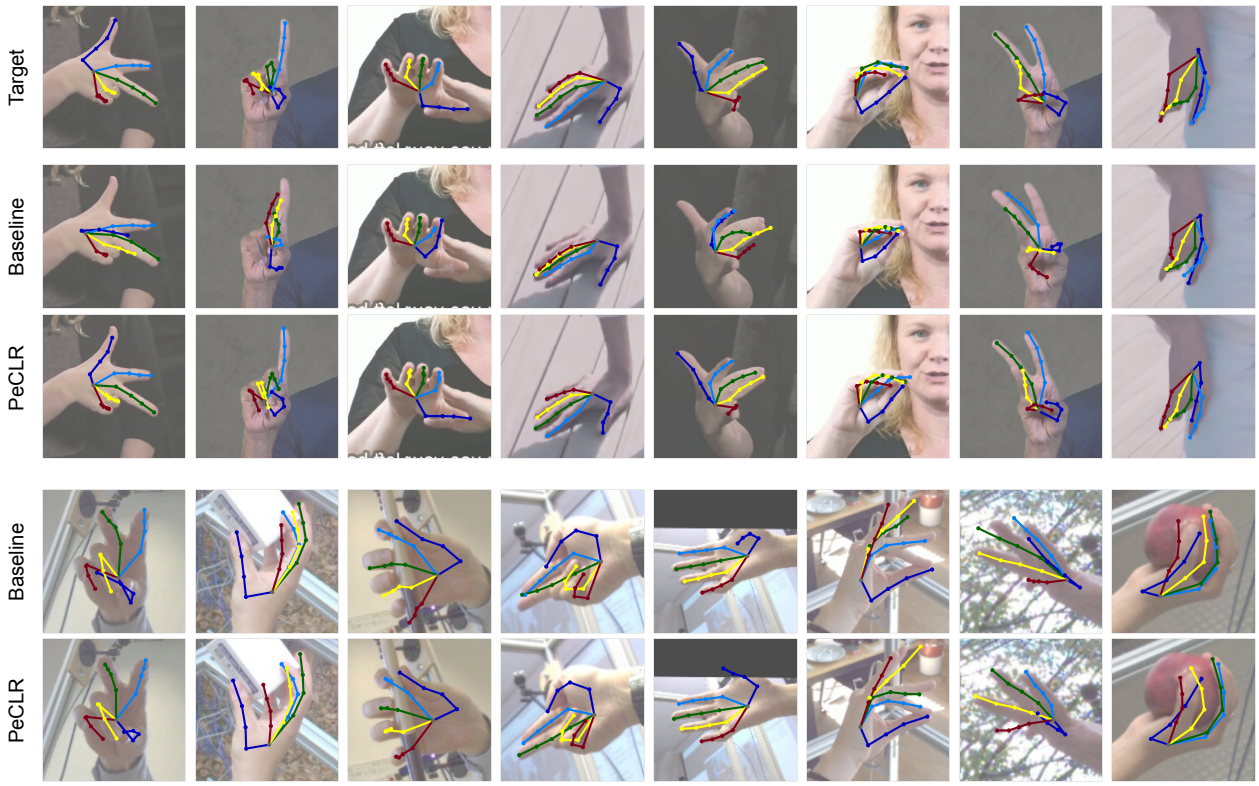


Figure 3: Predictions are shown on the test sets of YT3D (top) and FH (bottom) using either RN152 (Baseline) or RN152 + PeCLR. Note that the ground truth of the test set is not publicly available for FH, thus we only visualize the predictions.

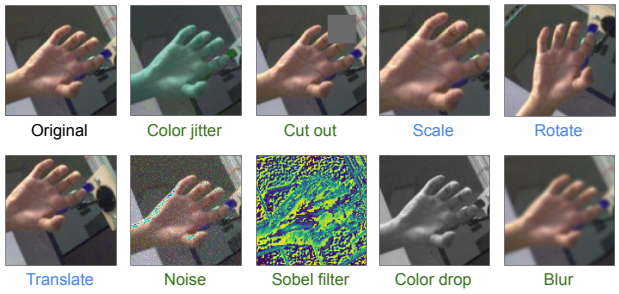
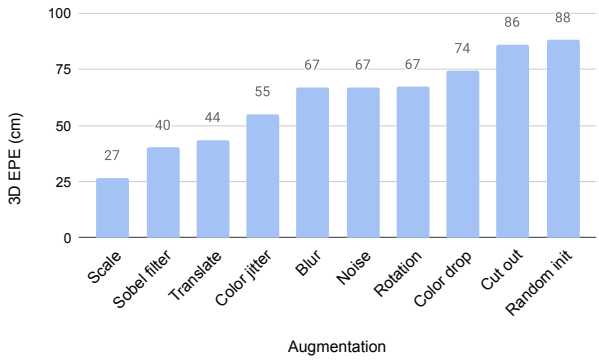


Figure 4: Visualization of transformations evaluated for contrastive learning. Geometric transformations are written in blue whereas appearance transformations are in green. The original sample is taken from FH.

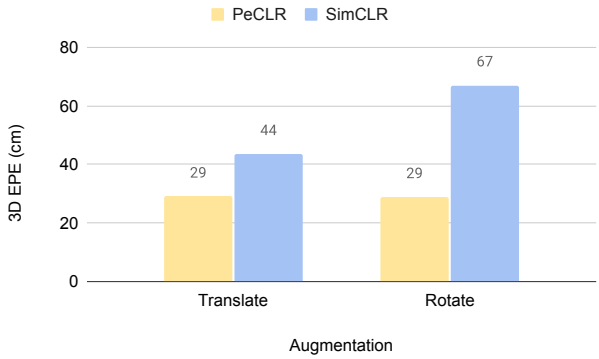
encoders with different objective functions, one using NT-Xent (Eq.1) as proposed in SimCLR, and another one making use of our proposed contrastive formulation (Eq.3). To evaluate the learned feature representation, we freeze the encoder and train a two-layer MLP in a fully-supervised manner on 3D hand labels as described in Sec. 3.3.

Individual augmentation. Fig. 5 shows the performance errors when individual augmentation is applied. Here the SimCLR framework is used. We observe that encoders trained with transformations perform better than random initialization. However, we see that rotation transformation leads to particularly bad performance. As motivated in Sec. 3.2, SimCLR promotes invariance under all transformations, including geometric transformation. We hypothesize that the poor performance stems from this invariance property. To verify this, we compare the performance using the equivariant contrastive loss proposed in PeCLR and SimCLR's contrastive formulation under two geometric transformations, namely translation and rotation. We emphasize here again that due to the cosine similarity, the effect of scale is eliminated. Fig. 5b shows that for both translation and rotation, PeCLR yields significant improvements of 34% and 57% relative to SimCLR, respectively. This results in scale, translation and rotation having the best feature representation as evaluated by the final MLP's accuracy with PeCLR. Note that we only promote equivariance for geometric transformation. Therefore, all other appearance-related transformations yield the same performance for PeCLR and SimCLR.

Composite augmentations. Finally, we compare different compositions of transformations. To narrow down the



(a)



(b)

Figure 5: a) The feature representation power of individual augmentation as evaluated by an MLP. b) Comparison of PeCLR and SimCLR for translation and rotation, showing a notable improvement of 34% and 56% respectively.

search space, we pick the top-4 performing augmentations from Fig. 5 as candidates. We then conduct an exhaustive search over all combinations of the selected candidates and empirically find that scale, rotation, translation and color jitter deliver the best performance for PeCLR, whereas SimCLR performs best with scale and color jitter.

We compare PeCLR with SimCLR using their respective optimal composition and report the results in Tab. 1. Notice that PeCLR yields better feature than SimCLR, gaining the improvements of 3.4% in terms of 3D EPE and 12.8% in terms of 2D EPE. This demonstrates that PeCLR leads to a more effective representation learning approach for hand pose estimation.

4.5. Semi-supervised learning

In this experiment, we evaluate the efficiency of PeCLR in making use of labeled data. To this end, we perform semi-supervised learning on FH with the pre-trained encoder. We use the optimal data augmentation compositions

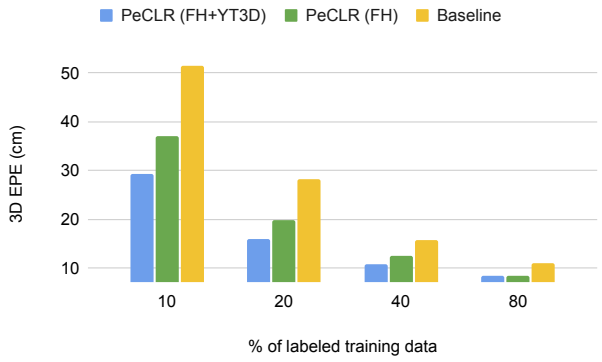


Figure 6: Semi-supervised performance on FH. By pre-training with PeCLR we achieve greater accuracy in contrast to only training supervised. Adding additional unlabeled data increases this effect.

developed in Sec. 4.4. As indicated in [7], deeper neural networks can make better use of large training data. Therefore, we increase our network capacity and use a RN152 as the encoder in the following. Results and discussion of RN50 can be found in supplementary.

Specifically, we pre-train our encoder on FH with the PeCLR. The encoder is then fine-tuned on varying amounts of labeled data on FH. For clarity, we term the resulting model $\mathbf{M}_{FH}$. To quantify the effectiveness of our proposed pre-training strategy, we compare against a baseline method $\mathbf{M}_b$ that is solely trained on the labeled data of FH, excluding the pre-training step. Finally, to demonstrate the advantage of self-supervised representation learning with large training data, we train a third model, pre-trained on both FH and YT3D, named $\mathbf{M}_{FH+YT3D}$.

From the results shown in Fig. 6, we see that $\mathbf{M}_{FH}$, $\mathbf{M}_{FH+YT3D}$ outperform the baseline $\mathbf{M}_b$ regardless of the amount of used labels. This result is in accordance with [7], confirming that the pre-trained models can increase label efficiency for hand pose estimation. Comparing $\mathbf{M}_{FH+YT3D}$ with $\mathbf{M}_{FH}$, we see that increasing the amount of data during the pre-training phase is beneficial and further decreases the errors. These results from $\mathbf{M}_{FH+YT3D}$ and $\mathbf{M}_b$ shed light on label-efficiency of the pre-trained strategy. For example, we see that for 20% of labeled data, $\mathbf{M}_{FH+YT3D}$ performs almost on par with $\mathbf{M}_b$ using 40% of labeled data

4.6. Comparison with state-of-the-art.

With the optimal composition of transformations and representation learning strategy in place, we compare PeCLR with current state-of-the-art approaches on the FH dataset. For our method, we use an increased image resolution of 224×224 pixels and a RN152 as the encoder. The encoder is pre-trained on FH and YT3D with PeCLR

Method	3D PA-EPE (cm) ↓	PA-AUC ↑
Spurr et al [31]	0.90	0.82
Kulon et al [22]	0.84	0.83
Li et al [23]	0.80	0.84
Pose2Mesh [9]	0.77	-
I2L-MeshNet [24]	0.74	-
RN50	0.83	0.84
+ PeCLR (**ours**)	0.71	0.86
RN152	0.74	0.85
+ PeCLR (**ours**)	**0.66**	**0.87**

Table 2: **Comparison with SotA**. Standard ResNet models are unable to outperform state-of-the-art methods. By pre-training using PeCLR, we yield a performance increase of 14.5% / 10.8% for RN50 and RN152 respectively, resulting in state-of-the-art performance for both networks.

FH		
Method	3D EPE ↓ (cm)	AUC ↑
RN152	5.05	0.34
+ PeCLR (**ours**)	**4.56**	**0.36**
Improvement	9.7 %	5.6 %
YT3D		
Method	3D PA-EPE ↓ (cm)	2D EPE ↓ (px)
RN152	3.05	22.1
+ PeCLR (**ours**)	**2.88**	**16.9**
Improvement	5.6 %	23.5 %

Table 3: **Cross-dataset evaluation.** PeCLR model with the RN152 architecture is pre-trained on YT3D and FH and then fine-tuned on FH. The model is then evaluated on both FH (top) and YT3D (bottom) test sets. We observe that similar improvements are gained across both datasets.

and fine-tuned supervised on the FH dataset. In addition, we also have a baseline model that is solely trained on FH in a supervised manner. For completeness, we repeat these experiments with a RN50.

Tab. 2 compares our results to the current state-of-the-art. We see that training a RN model supervised only on FH does not outperform the state-of-the-art, even using large model capacity versions such as RN152. We hypothesize that this is due to the comparably small dataset size of FH and thus lack of sufficient labeled data for training. However, using PeCLR to leverage YT3D in an unsupervised manner improves performance by 14.5% and 10.8% PA-EPE for RN50 and RN152 respectively, outperforming state-of-the-art. Note that all methods in Tab. 2 use highly specialized architectures. In contrast with our formulation, state-of-the-art performance is established in a purely data-driven way. In Fig. 3 (bottom) we visualize qualitative results on both our baseline and PeCLR.

4.7. Cross-dataset analysis

With a large amount of unlabeled training data, we hypothesize that our approach can produce better features that are beneficial for generalization. To verify this, we examine our models of Sec. 4.6 in a cross-dataset setting. More specifically, we investigate the performance of both models on the YT3D dataset. This sheds light on how the models perform under a domain shift. We emphasize here that neither models are trained supervised on YT3D.

The results in Tab. 3 show that PeCLR outperforms the fully-supervised baseline with improvements of 5.6% in 3D EPE and 23.5% in 2D EPE. These improvements can be observed qualitatively in Fig. 3 (top). The results indicate that PeCLR provides indeed a promising way forward in using unlabeled data for representation learning and training a model that can be more easily adapted to other data distributions. We note that cross-dataset generalization is seldom reported in the hand pose literature and it is generally assumed to be very challenging for most existing methods while important for real-world applications.

5. Conclusion

In this paper we investigate self-supervised contrastive learning for hand pose estimation, making use of large unlabeled data for representation learning. We identify a key issue in the standard contrastive loss formulation, where promoting invariance leads to detrimental results for pose estimation. To address this issue, we propose PeCLR, a novel method that encourages equivariance for geometric transformations during representation learning. We thoroughly investigate PeCLR by comparing the resulting feature representation and demonstrate improved performances of PeCLR over SimCLR. We show that our PeCLR has high label efficiency by means of semi-supervision. Finally, our PeCLR achieves state-of-the-art results on the FreiHAND dataset. Lastly, we conduct a cross-dataset analysis on YT3D and show the potential of PeCLR for cross-domain applications. We believe that PeCLR as well as our extensive evaluations can be of benefits to the community, providing a feasible solution to improve generalizability across datasets. We foresee the usage of PeCLR on other tasks such as human body pose estimation.

Acknowledgments. We are grateful to Thomas Langerak for the aid in figure creation and Marcel Bühler for helpful discussions and comments.

References

[1] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019. 2, 3

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Weakly-supervised domain adaptation via GAN and mesh model for estimating 3d hand poses interacting objects. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020.

[3] Adnane Boukhayma, Rodrigo de Bem, and Philip H. S. Torr. 3d hand shape and pose from images in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019. 1, 2, 3

[4] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 3

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019. 2, 5

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, 2020. 2, 3, 4, 5

[7] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. Big self-supervised models are strong semi-supervised learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3, 7

[8] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020. 2, 3

[9] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose, 2020. 8

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL-HLT*, Minneapolis, Minnesota, 2019. 3

[11] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015. 3

[12] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J. Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020. 1, 2, 3

[13] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single RGB image. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019. 2, 3

[14] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020. 1, 2, 3

[15] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019. 1, 2, 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016. 5

[17] Olivier J. Hénaff. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, 2020. 3

[18] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2, 3, 5

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. of ICLR*, 2015. 5

[20] Dominik Kulon, Riza Alp Güler, Iasonas Kokkinos, Michael M. Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020. 1, 2, 5

[21] Dominik Kulon, Riza Alp Güler, Iasonas Kokkinos, Michael M. Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020. 2, 3

[22] Dominik Kulon, Riza Alp Güler, Iasonas Kokkinos, Michael M. Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020. 8

[23] Moran Li, Yuan Gao, and Nong Sang. Exploiting learnable joint groups for hand pose estimation. *arXiv preprint arXiv:2012.09496*, 2020. 8

[24] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image, 2020. 8

[25] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. *arXiv preprint arXiv:2008.08213*, 2020. 2, 3

[26] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. *arXiv preprint arXiv:2008.09309*, 2020. 1, 2, 3

[27] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular RGB. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018. 2, 3

[28] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*. Springer, 2016. 3

[29] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3

[30] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 2017. 2, 5

[31] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints, 2020. 1, 2, 3, 5, 8

[32] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 2018. 2

[33] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+O: unified egocentric recognition of 3d hand-object poses and interactions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019. 1, 2, 3

[34] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 3

[35] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *European Conference on Computer Vision*. Springer, 2016. 3

[36] Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination, 2018. 3

[37] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, 2019. 2, 3

[38] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks, 2017. 5

[39] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*. Springer, 2016. 3

[40] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular RGB image. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 2019. 2, 3

[41] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single RGB images. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, 2017. 1, 2

[42] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan C. Russell, Max J. Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single RGB images. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 2019. 1, 5