Proceedings of Machine Learning Research 148:255-271, 2021 NeurIPS 2020 Preregistration Workshop

Exploring self-supervised learning techniques for hand pose estimation

Aneesh Dahiya ADAHIYA@STUDENT.ETHZ.CH
Adrian Spurr ADRIAN.SPURR@INF.ETHZ.CH
Otmar Hilliges OTMAR.HILLIGES@INF.ETHZ.CH
Department of Computer Science, ETH Zurich, Switzerland

Abstract

3D hand pose estimation from monocular RGB is a challenging problem due to significantly varying environmental conditions such as lighting or variation in subject appearances. One way to improve performance across-the-board is to introduce more data. However, acquiring 3D annotated data for hands is a laborious task, as it involves heavy multi-camera setups leading to lab-like training data which does not generalize well. Alternatively, one could make use of unsupervised pre-training in order to significantly increase the training data size one can train on. More recently, contrastive learning has shown promising results on tasks such as image classification. Yet, no study has been made on how it affects structured regression problems such as hand pose estimation. We hypothesize that the contrastive objective does not extend well to such downstream task due to its inherent invariance and instead propose a relation objective, promoting equivariance. Our goal is to perform extensive experiments to validate our hypothesis.

Keywords: Hand pose estimation, contrastive learning

1. Introduction

Given a monocular RGB image, estimating the location of hand joints is a challenging structured regression problem. Amongst others, conditions that significantly contribute to the difficulty are large diversity in backgrounds, lighting conditions and hand appearances, as well as self-occlusion.

One straightforward way of improving the performance of a learning-based model is to include more training data. However, acquiring 3D labeled data is laborious and expensive as it requires large lab-like settings whose data does not translate well to in-the-wild imagery (Zimmermann et al., 2019; Kulon et al., 2020b). The community has been relying increasingly more on supplementary 2D annotated data to tackle this and demonstrated that inclusion of this additional data leads to better prediction accuracy. For example, Spurr et al. (2020) showed that one can outperform many supervised approaches by using weakly-supervised data more effectively via appropriate priors. Although easier to acquire, 2D annotations do not come for free. To tackle this, works exist (Kulon et al., 2020b) that use automatically generated 2D annotations with the help of OpenPose (Cao et al., 2019). However, there is no guarantee that these poses are indeed correct and the accuracy one can achieve with such an approach is bounded by the performance of the OpenPose model.

Alternatively, one could resort to using unlabeled data directly with the help of self-supervision. Recently, approaches such as (Chen et al., 2020a,b) have shown that they are close to reaching parity or even outperform supervised baseline models with the help of contrastive learning on tasks such as image classification. This raises an interesting question: *Does the contrastive self-supervised learning capability extend to structured regression tasks as well?* We hypothesize that features learned during contrastive-based training may not readily transfer to regression-based tasks, as the former results in features being *invariant* to the respective transformations. However, structured regression-based task require *equivariant* features. For example, given two images of the same hand, one being the rotated form of the other, the keypoints predicted on one hand should be the rotated version of the other. Yet, the objective function of contrastive learning encourages the features of both images to lie as closely as possible from one another, possibly inhibiting performance.

To tackle this, we propose a relative loss where the relative transformation from one image to the other is predicted. Our assumption is that this novel task pushes the model to learn a representation that is equivariant to the transformations applied. Coming back to our previous example of the two rotated hand poses, the relative loss requires the model to be able to predict the relative rotation between both the images. We hypothesize that doing so results in equivariant features, as the representation learned needs to be informative to infer the the applied transformation.

In this paper, we propose to explore self-supervised learning approaches for hand pose estimation by analyzing the currently prevalent method of contrastive learning. Our goal is to validate the hypothesis that the contrastive objective is not an effective way to leverage self-supervision and that by forcing the model to learn equivariant features, we can improve the performance of hand pose estimation approaches across the board. We want to compare our proposed loss with the original contrastive learning objective on the downstream task of hand pose estimation

We envision that the knowledge gained through the thorough evaluation of self-supervised methods in the context of structured regression problems will be valuable for communities such as hand and body pose. In the interest of reproducibility and contributing to the research community, we will be releasing the code and trained network model.

2. Related Work

Self-supervised learning has gained interest in recent years as a form of unsupervised pre-training. Generally these rely on solving a pretext task which is not of interest to the actual task at hand. However, by solving the task, a good representation is learned as a by-product which can be used in downstream tasks.

Such pretext tasks can take any form. The most recent include Contrastive Multiview Coding (CMC) (Tian et al., 2019), Contrastive Predictive Coding (CPC) (Oord et al., 2018), simCLR (Chen et al., 2020a), MoCo (Chen et al., 2020b), whereas earlier works include (Vincent et al., 2008; Zhang et al., 2016; Pathak et al., 2016; Zhang et al., 2017; Doersch et al., 2015; Noroozi and Favaro, 2016; Wang and Gupta, 2015; Pathak et al., 2017). Alternatively, adversarial losses (Goodfellow et al., 2014) can also be utilized for unsupervised representation learning (Donahue et al., 2016; Donahue and Simonyan, 2019).

Novotny et al. (2018) tackle self-supervision by learning geometrically stable pixel level descriptors across a range of objects with probabilistic objective. Whereas Zhang et al. (2019) and Rocco et al. (2017), estimate geometric features by predicting parameters for relative geometric transformation applied to the image and Gidaris et al. (2018) estimates it by predicting one out of four angels used to rotate the input image. Differently, we propose to use geometric as well as appearance transformations. Lastly, none of the mentioned related work compares the contrastive with the pairwise relative loss formulation and does not report results on tasks such as hand pose.

In this paper, we focus on contrastive learning for self-supervision. However to the best of our knowledge, contrastive learning has not yet been applied to downstream tasks such as structured regression problems like that of hand pose estimation. One reasons for this could be that the resulting features may be invariant to the respective transformations, instead of equivariant. Our goal is to validate this hypothesis.

3. Methodology

Here we briefly recap the original contrastive formulation as was proposed by Chen et al. (2020a) and our proposed relative objective.

3.1. Recap on contrastive learning

We show an overview of the contrastive framework in Fig. 1 (left). Contrastive learning enables a neural network f to learn features in an unsupervised manner by encouraging similar looking images to lie close in feature space. As such, it creates similar looking pairs of images by applying transformations $t_i : \mathcal{R}^n \rightarrow \mathcal{R}^n$ on a source image $\mathbf{x} \in \mathcal{R}^n$ and optimizing a neural networks weights to output similar features $f(t_i(\mathbf{x})) = \mathbf{h}_i$. These features are projected into a latent space $g(\mathbf{h}_i) = \mathbf{z}_i$ via a projection head g, on which the contrastive loss is applied to. To use the trained network f on a downstream task, the projection head is discarded and a linear classifier is trained on the features $\mathbf{h}$.

$$\sum_{i,j,i\neq j} -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{i\neq k} \exp \text{sim}((\mathbf{z}_i, \mathbf{z}_k)/\tau)} \tag{1}$$

Where $\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T\mathbf{v}/||\mathbf{u}||_2||\mathbf{v}||_2$ computes a similarity and τ is a temperature parameter. Although impressive performance was achieved via this method, it is unclear how well these translate to structured regression problems. Although Chen et al. (2020a) report that they were capable of predicting the rotation angle with 67.6% accuracy, there are still issues: 1) It is unclear how the contrastive representation affects the structured regression-based downstream tasks 2) The classification was done by predicting an angle out of four. We hypothesise that there is more potential performance to be gained by reformulating the contrastive task to a relative one.

3.2. Proposal

Instead of contrasting an image pair, we propose to predict their relative transformations. Concretely, given a family of parameterized transformations $\mathcal{T}$ (e.g rotations), two randomly

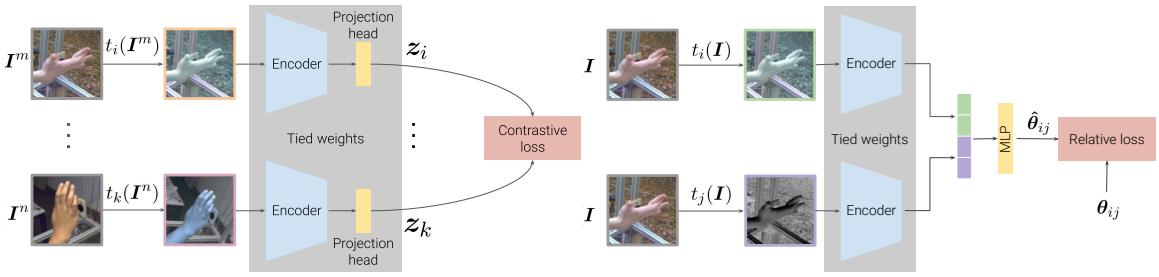


Figure 1: Models

sampled parameters θ_i, θ_j (e.g rotation angles), we first compute the transformed sample pair via $\mathbf{x}_i = t(\mathbf{x}; \theta_i)$, where $t \in \mathcal{T}$, before passing it into the network f to obtain their respective features $\mathbf{h}_i$. These are fed into a transformation-specific projection head g_t to predict their relative transformation $\theta_{ij} = \theta_j - \theta_i$. As such, the objective changes from contrastive to relative. Hence, we reformulate Eq. 1 to following for one augmentation:

$$L_t = \sum_{i,j,i \neq j} ||\theta_{ij} - g_t(\mathbf{h}_i, \mathbf{h}_j)|| \tag{2}$$

In presence of $|\mathcal{T}|$ number of augmentations we minimize the loss described in Eq. 3, where the loss from each augmentation $t \in \mathcal{T}$ is scaled by a trainable parameter σ_t (Kendall et al., 2018).

$$L = \sum_{k \in \mathcal{T}} (L_k / \sigma_k + \log \sigma_k) \tag{3}$$

As each task family $t \in \mathcal{T}$ is different, each g_t will be an independent network, but share the features $\mathbf{h}$ produced by the network f. Our hypothesis is that by predicting relative transformation parameters, the features learned will be equivariant to these transformations. This can be helpful for structured regression task where transformation of input also transforms the keypoints.

This intuition stems from the following. Given $\mathbf{x}$, we produce two samples $\mathbf{x}_1$ and $\mathbf{x}_2$ via $\mathbf{x}_i = t(\mathbf{x}; \theta_i)$. As such, our proposed objective is $||\theta_{21} - g_t(\mathbf{h}_2, \mathbf{h}_1)||_2$, where $\theta_{21} = \theta_2 - \theta_1$ is the target label. Transforming a new sample $\mathbf{x}_3$ via $\theta_3 = \theta_2 + \Delta\theta$ results in the new target label:

$$\theta_{31} = \theta_3 - \theta_1 = \theta_2 + \Delta\theta - \theta_1 = \theta_{21} + \Delta\theta. \tag{4}$$

Hence the target changes in accordance to the change in parameters. We postulate that this induces equivariance in the features h.

4. Experimental protocol

In order to provide fair comparison of our proposal with the contrastive loss, we will closely follow the experimental protocol outlined in Chen et al. (2020a). The goal of the experiment section is to first verify which transformation benefits from which self-supervised loss. Next, we want to identify which composition results in the most beneficial feature representation. Lastly, we explore cross-dataset generalization and compare with fully supervised methods.

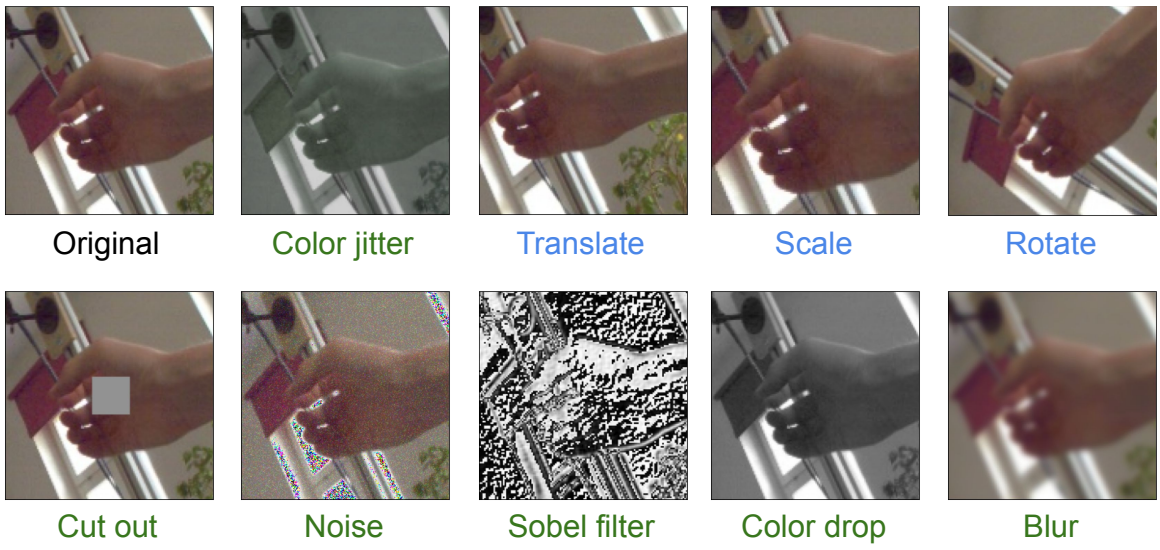


Figure 2: Example of transformations used in this paper. Samples taken from FreiHAND (Zimmermann et al., 2019).

4.1. Protocol

We briefly describe the datasets and transformations used, metrics reported and the setting assumed for all experiments.

Datasets. We will benchmark our performance on two hand pose dataset. The first is FreiHAND (FH) Zimmermann et al. (2019) which contains 32560 samples of single hand pose with green screen backgrounds. Using synthetic background imagery, these are extended to $130k$ samples. The second is the InterHands2.6M (IH) dataset, of which we focus on the single-hand split which contains $688k$ samples.

Transformations. Following Chen et al. (2020a), we investigate the following transformations: crop, cutout, color jitter, sobel filter, noise, blur, rotate, scale and translate. A sample of these can be seen in figure 2. Due to chirality in hands, crop and flip is not used as an augmentation, instead we include random 2D translation of the hand as seen in figure 2 *Translate*. All the images are pre-processed by cropping the hand and resizing it to 128×128 RGB images. Images in Fig. 2 *Scale* are cropped from a larger image to isolate the hand, similar to Chen et al. (2020a).

In our proposed relative objective we investigate translation, color jitter and rotation as they have meaningful parameters that can be regressed. The translation parameters are (x, y) coordinates of crop box center. Rotation is characterized by an angle θ, around which the image is rotated. Color jitter is characterized by $h, s, (\alpha, \beta)$ parameters which change hue, saturation and value of the image pixels respectively. Augmentations like cutout and sobel filter are not included in the prediction since their parameters are trivial to regress. Relative augmentation parameters of gaussian blur, noise and cutout are ambiguous to predict, they are therefore not included for our proposed relative objective. We emphasize

here that the images are still augmented with these augmentations, but we do not estimate their relative parameters.

Metrics. We report the mean per joint error (MPJE), as well as median per joint error (median PJE) on the downstream task of hand pose estimation. More specifically, given the self-supervised pre-trained network f, we follow the linear evaluation protocol (Zhang et al., 2016; Oord et al., 2018; Bachman et al., 2019; Kolesnikov et al., 2019), where we train a linear layer on top of the frozen pre-trained network to regress the 2.5D hand pose representation. This allows us to quantify the feature representation learned in our proposed pretext task.

Setting. Following Su et al. (2019), we use a ResNet-18 (He et al., 2016) backbone network to facilitate training with bigger batch sizes as it was reported to improve performance (Chen et al., 2020a,b). We use a 2-layer MLP projection head and a 128-dimensional latent space. All models are trained using the ADAM optimizer. Inspired by Chen et al. (2020a), the learning rate is scheduled using LARS (You et al., 2017) with an initial warmup phase to stabilize training for large batches. The learning rate is scaled using the square root of the batch size n_{bs}, i.e. $lr = 0.0001 \times \sqrt{n_{bs}}$. For the downstream task of hand pose estimation, we discard the projection head and replace it with a linear layer. The optimal parameters are chosen using random grid search. For Sec. 4.5, we change the backbone network to that of Iqbal et al. (2018), but keep the training scheme the same.

4.2. Data augmentation specific objective function

Before we attempt to investigate the ideal series of transformations, we need to answer a question: *Given our downstream task of regression, will all transformation yield a boost with our proposed relative objective? Could certain tasks relate to the contrastive loss as opposed to the regressive loss?* For example, given the color augmentation where the color channels are augmented, it would perhaps be more beneficial to require the feature representation to be *invariant* as opposed to *equivariant*. In order to answer this question, we first perform an initial ablative study to inspect which transformation benefits from a relative objective as opposed to a contrastive one. To this end, we report the downstream task performance for each individual transformation, using either the contrastive or relative loss.

4.3. Data augmentation compositions

As was highlighted in Chen et al. (2020a), the composition of transformation operations is crucial to the final performance achieved in the downstream task. Since the downstream task here is regression, it is not clear if the same combination of transformations reported to be superior in Chen et al. (2020a) for classification will still remain as such in our downstream task. Therefore it is vital to determine which combination of transformations performs the best. To this end, we perform an exhaustive search. For each augmentation, we first pick the best performing pretext objective function, as determined in Sec. 4.2. Then, we inspect all possible combinations of augmentations and determine which composition performs best, based on the downstream task.

4.4. Cross-dataset generalization

Generalization is an important concept in any deep learning network. One simple way to cross the domain gap is to train on data of the target domain. However, fully labeled data is often only available in constrained lab environments. In this section, we want to explore if our effectively self-supervised learning can be used to cross the domain gap. To this end, we perform self-supervised pre-training on IH and FH, but fine-tune the last linear layer only on FH. The final evaluation is done on IH to quantify if a reasonable improvement can be gained. In order to have a comparison, we train a fully supervised model solely on FH and compare the two results.

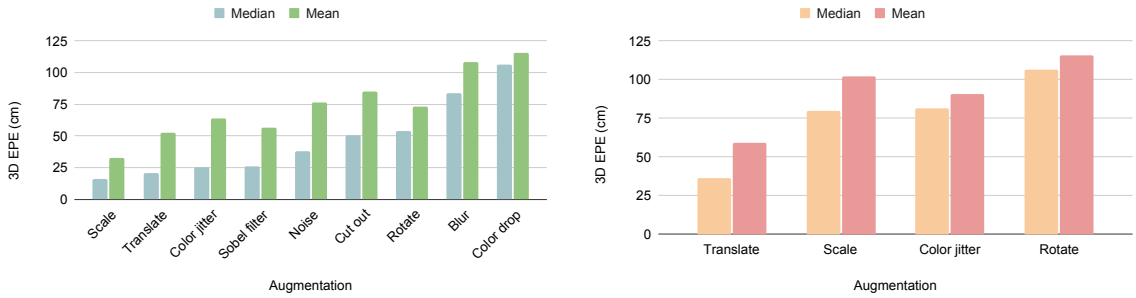
4.5. Comparison with supervised model

Given the best performing self-supervised objective and augmentation composition, we compare the performance against the current state-of-the-art hand pose estimator Iqbal et al. (2018). For this we replace the ResNet-18 encoder used in previous experiments with the hourglass model used in Iqbal et al. (2018). During the self-supervision phase, the 2D backbone network output is vectorized and passed through a non-linear projection layer, like in the prior experiments. During the downstream task training, we train the frozen network like that of Iqbal et al. (2018). The goal of this section is to investigate how state-of-the-art models perform in the context of self-supervision.

5. Results

Augmentation ablation: We conduct this experiment on FreiHAND using our own training and validation split (90% for training and 10% for validation). Figure 3 shows the mean and median 3D error on the validation set. We observed that the choice of augmentation used in pretraining has a direct influence on the hand pose estimation performance. The pre-training with *scale* and *translate* augmentations performed relatively well, for both relative and contrastive objective, when compared to the other augmentations. The relative objective performed poorly for all the augmentations when compared with its contrastive counterpart. This is true for the augmentations like *translate* and *rotate* where the contrastive objective is at risk of learning invariance to these augmentations. This means that the contrastive objective is able to learn rich feature representation for pose estimation, despite the possibility of learning geometric invariance during contrastive training. One reason for the poor performance of the model pre-trained with relative objective could be that the features learned by the model to fulfil the pre-text task of parameter prediction may not be informative to the task of hand pose estimation. For example, in order to regress rotation from two images, the model merely has to fixate on two specific points in the image.

Augmentation composition: We conduct an ablative study with several combinations of augmentations used during pre-training. We use the super set of *scale*, *translate*, *color jitter* and *rotate* to perform an exhaustive search for the best augmentation combination. This selection of augmentations is chosen for two reasons. First, it ensures that the comparison with both the objectives is possible for each augmentation combination explored, as some augmentations are not regressible for relative objective (*gaussian noise* etc). Second, it lim-



(*a*) Contrastive objective (*b*) Relative objective

Figure 3: **Augmentation ablative**: Performance of individual augmentation using contrastive objective (left) and relative objective(right). The encoder is frozen after pre-training and a linear layer is finetuned on FreiHAND (FH). The model is then evaluated on FH.

Model	MPJE (cm)↓	median PJE (cm)↓
ResNet-18 + Contrastive	**26.16**	**12.05**
ResNet-18 + Relative	64.11	23.31

Table 1: **Augmentation composition ablative** : The pre-trained encoder is frozen and a linear layer is fine-tuned on FreiHAND. *Color jitter, translate* and *scale* perform best for the contrastive objective, whereas *translate* and *scale* perform best for relative objective. The results for exhaustive augmentations are in supplementary.

its the numbers of experiments to a manageable level because the number of experiments scales exponentially with respect to the set of augmentations chosen. If n augmentations are chosen, then an exhaustive search will require $2^n - 1$ experiments . Table 1 shows the performance of the best augmentation composition for both the objectives. We observe that the best set of augmentations for contrastive objective is *scale color jitter* and *translate*. The relative objective performs best when *translate* and *scale* are used together. Results for the remaining experiments are included in supplementary. We select these augmentations based on their median PJE on the hand pose estimation This is because compared to MPJE, median PJE is more robust to outliers. Based on these results, we discard the proposed relative objective in favour of the contrastive objective. This is because the relative objective did not perform better than the contrastive objective under any augmentation composition. For all further experiments, we use the optimal augmentation composition for the contrastive objective derived from this experiment.

Cross data generalization: Table 2 shows the performance of the ResNet-18 encoder pre-trained on FH and IH2.6M with the contrastive objective. The pre-trained encoder is frozen and a linear layer is fine-tuned on FH. The performance is evaluated on IH2.6M. We

Finetuning on FH and evaluation on IH2.6M			
Model	MPJE (cm) ↓	median PJE (cm) ↓	AUC ↑
ResNet-18	**15.5**	**11.5**	**0.718**
+ Contrastive (frozen)	51.8	47.9	0.128
+ Contrastive (unfrozen)	28.6	26.4	0.459

Table 2: **Cross-dataset generalization**: The ResNet-18 encoder is pretrained with contrastive objective on FreiHAND and InterHand2.6M combined. The encoder(frozen/unfrozen) with a linear layer is fine-tuned on FreiHAND and evaluated on InterHand2.6M. This model is then compared with its supervised counterpart trained on FreiHAND.

observe that the fully supervised ResNet-18 model outperforms the pre-trained frozen encoder. In a followup experiment, we also fine-tune the entire pre-trained ResNet-18 encoder along with the linear layer on FH. Compared with the frozen encoders performance there is a significant improvement. This improvement however does not surpass the fully supervised performance. This indicates that pre-training negatively influenced the performance of hand pose estimation.

In addition to the analysis with the ResNet-18 encoder, we perform the cross dataset experiment with a HRNet (Wang et al., 2020) encoder and fine-tune the entire model after pre-training to get further meaningful insights. The HRNet encoder is pre-trained on FH and IH2.6M with contrastive objective. This is followed by fine-tuning the entire network on FH, then evaluating on IH2.6M. We also perform the same experiment with switched datasets, by fine-tuning on IH2.6M and then evaluating on FH. Table 3 shows the performance of the model under these settings. We report Procrustes Aligned (PA) metrics for the analysis using HRNet, as the hand pose community reports these numbers.

We observe that in contrast to the experiment with ResNet-18, HRNet was able to capture features across modality during the contrastive pre-training, as the pre-trained model performed better when compared to the supervised for both IH2.6M and FH. In case of FH we see an improvement of 6.8% and for IH2.6M we see an improvement of 12.5% in terms of MPJE. We attribute this to the increased model capacity, an observation also made by Chen et al. (2020a).

Comparison with supervised model We investigate the benefits of contrastive learning in the fully supervised setting. To this end, we use HRNet Wang et al. (2020) as the encoder. We pre-train the model on both FH and IH2.6M combined, followed by finetuning and evaluation on a target dataset. Here we inspect both FH and IH2.6M as target datasets respectively. Table 4 shows the result. We observe that for FH, we gain an improvement of 12% for MPJE. On the other hand, no improvements were observed for IH2.6M. We hypothesize that this is due to the amount of training labels for IH2.6M ($\approx 372K$) being significantly more than those for FH ($\approx 130K$). We perform additional comparison to related work in the supplementary.

Finetuning: FreiHAND. Evaluation: InterHand2.6M			
Model	PA MPJE (cm)↓	PA median PJE (cm)↓	PA AUC ↑
HRNet	3.2	2.5	0.938
+ Contrastive	**2.8**	**2.1**	**0.944**
Finetuning: InterHand2.6M. Evaluation: FreiHAND			
Model	PA MPJE (cm)↓		PA AUC ↑
HRNet	2.51	-	0.515
+ Contrastive	**2.34**	-	**0.550**

Table 3: **Cross-dataset generalization**: The HRNet encoder is pretrained with contrastive objective on FreiHAND and InterHand2.6M combined. The encoder fine-tuned on FreiHAND is then evaluated on InterHand2.6M (top) and the encoder fine-tuned on InterHand2.6M is evaluated on FreiHAND (bottom)

FreiHAND			
Model	PA MPJE (cm) ↓		PA AUC ↑
HRNet	1.08	-	0.787
+ Contrastive	**0.95**	-	**0.813**
InterHand2.6M			
Model	PA MPJE (cm) ↓	PA median PJE (cm) ↓	PA AUC ↑
HRNet	1.30	0.60	0.976
+ Contrastive	1.30	0.60	0.976

Table 4: **Supervised comparison** on FreiHAND and InterHand2.6M. An HRNet encoder is pre-trained with contrastive objective on both FreiHAND and InterHand2.6M. The pre-trained model is then finetuned and evaluated on the target dataset (FH or IH2.6M). We compare against a model trained from scratch.

5.1. Modifications

We deviate from the mentioned protocol slightly to gain more insights:

1. We perform additional cross dataset experiments with HRNet as the encoder, without freezing it during fine-tuning. This enables the model to learn new features whilst fine-tuning. This is otherwise impossible if the encoder is frozen. In addition to that we also perform an experiment with supervised fine-tuning on FreiHAND and evaluation on InterHand2.6M. This enables us to show that features learned during pre-training are beneficial in a cross-modality setting. The first version of IH2.6M data released in september 2020 is used for all the experiments because the second and complete version of InterHand2.6M dataset was only released at the end of March 2021. All the experiments were already carried out by then. We use single hand machine and human annotated split for training and evaluation.

2. In section 4.5, instead of using an hourglass shaped heatmap model from Iqbal et al. (2018), we use the heatmap based model from Wang et al. (2020) as the latter is more recent and performs well across a variety of pose estimation tasks.

3. The median PJE is not reported for FreiHAND evaluation. This is due to the fact that the test labels are not public, instead the predictions are evaluated in an online contest where only MPJE, AUC, *Procrustes Aligned* (PA) MPJE and PA AUC are calculated.

4. We report the Procrustes Aligned (PA) metrics for cross data set experiments and supervised experiments as usually PA numbers are reported by the hand pose community. The unaligned numbers are included in the supplementary.

6. Conclusion

In this work, we explored contrastive learning for hand pose estimation tasks. We initially hypothesized that the contrastive objective may be detrimental for hand pose estimation, due to it encouraging a model to learn features that are invariant to all augmentations. We proposed that a relative objective that can regress the relative augmentation parameters may be better suited for these augmentations. However, this was not observed empirically. Contrastive objective learned rich features from images when compared to the relative objective. The hypothesis was not validated, since relative objective performed poorly in comparison to the contrastive objective when pre-trained in presence of selected geometric augmentations. This effect was observed for all augmentation compositions as well. Contrastive objective learned better feature representation than relative objective, irrespective of the augmentation. We continued our analysis with the optimal augmentation composition for contrastive objective in a cross-dataset setup. We observed no improvement for the ResNet-18 pretrained encoder. However upon using a bigger model as encoder, we observed improvement across modality. Furthermore in a supervised experiment setup, models pre-trained with contrastive objective showed improvement over models trained from scratch. This indicates that contrastive learning may be a viable path to travel in search of better performing hand pose estimation models.

References

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15535–15545, 2019.

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning, 2020b.

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.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *Advances in Neural Information Processing Systems*, pages 10542–10552, 2019.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. In *ICLR 2018*, Vancouver, Canada, April 2018. URL https://hal-enpc.archives-ouvertes.fr/hal-01864755.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5d heatmap regression, 2018.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, 2018.

Alexander Kolesnikov, Xiaohua Zhai, and Lucas Beyer. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1920–1929, 2019.

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*, 2020a.

Dominik Kulon, Riza Alp Güler, Iasonas Kokkinos, Michael Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild, 2020b.

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

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.

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, 2020.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

David Novotny, Samuel Albanie, Diane Larlus, and Andrea Vedaldi. Self-supervised learning of geometrically stable features through probabilistic introspection, 2018.

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

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

Deepak Pathak, Ross Girshick, Piotr Dollár, Trevor Darrell, and Bharath Hariharan. Learning features by watching objects move. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2701–2710, 2017.

Ignacio Rocco, Relja Arandjelović, and Josef Sivic. Convolutional neural network architecture for geometric matching, 2017.

Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints, 2020.

Jong-Chyi Su, Subhransu Maji, and Bharath Hariharan. When does self-supervision improve few-shot learning? *arXiv preprint arXiv:1910.03560*, 2019.

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

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition, 2020.

Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2015.

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

Liheng Zhang, Guo-Jun Qi, Liqiang Wang, and Jiebo Luo. Aet vs. aed: Unsupervised representation learning by auto-encoding transformations rather than data, 2019.

Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1058–1067, 2017.

Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images, 2019.

Appendix A. Exhaustive augmentation composition search

In this section, we show the performance of all augmentation compositions considered for the contrastive and relative objective. We use a super set of *scale*, *translate*, *color jitter* and *rotate* to perform the search. Table 5 shows the MPJE and median PJE on the validation set of FreiHAND (FH) (Zimmermann et al., 2019). The encoder is frozen after pre-training and a linear layer is fine-tuned for hand pose estimation. The augmentation with the lowest median PJE is selected as the best augmentation composition for each objective.

Augmentation	Relative objective		Contrastive objective	
	MPJE (cm) ↓	**median PJE** (cm) ↓	**MPJE** (cm) ↓	**median PJE** (cm) ↓
CJ	90.54	81.21	63.81	25.18
CJ+S	336.11	69.67	**24.52**	13.05
CJ+S+Ro	95.51	89.30	54.18	26.86
CJ+Ro	108.10	98.82	82.12	46.11
S	101.83	79.55	32.81	15.94
S+Ro	106.51	99.71	72.72	50.59
Ro	115.47	106.21	73.01	54.21
T	58.92	36.60	52.38	21.02
T+CJ	65.80	33.00	31.85	14.86
T+CJ+S	56.50	23.39	26.16	**12.05**
T+CJ+S+Ro	**49.96**	31.37	50.04	24.96
T+CJ+Ro	55.43	41.62	66.59	34.76
T+S	64.11	**23.31**	30.79	14.67
T+S+Ro	53.89	34.26	75.79	54.42
T+Ro	63.78	51.33	70.70	52.85

Table 5: Hand pose estimation performance of encoder trained with contrastive or relative object with various augmentations on FH validation set. A linear layer is fine-tuned with the pre-trained frozen encoder. Here, T, S, CJ and Ro stand for *translate*, *scale*, *color jitter* and *rotate*, respectively. MPJE is sensitive to outliers arising from the error in root depth of 2.5D keypoints, therefore we use median PJE for selecting best augmentation composition.

Appendix B. Cross data experiment

In this section, we report the unaligned numbers for the performance on single hand split of InterHand2.6M (IH2.6M) (Moon et al., 2020) when the model is fine-tuned on FH and vice versa.

Finetuning on FH and evaluation on IH2.6M			
Model	MPJE (cm)↓	median PJE (cm)↓	AUC ↑
HRNet	16.70	12.10	0.695
+ Contrastive	**15.00**	**10.5**	**0.738**
Finetuning on IH2.6M and evaluation on FH			
Model	MPJE (cm)↓		AUC ↑
HRNet	56.62	-	0.01
+ Contrastive	**54.39**	-	**0.02**

Table 6: **Cross-dataset generalization**: The HRNet encoder is pretrained with contrastive objective on FH and IH2.6M combined. The encoder fine-tuned on FH is then evaluated on IH2.6M (top) and the encoder fine-tuned on IH2.6M is evaluated on FH (bottom)

InterHand2.6M			
Method	MPJE (cm) ↓	median PJE (cm) ↓	AUC ↑
HRNet	5.40	3.80	0.890
+ Contrastive	5.40	**3.50**	**0.893**
FreiHAND			
Method	MPJE (cm) ↓		AUC ↑
HRNet	9.66		0.22
+ Contrastive	**8.02**		**0.25**

Table 7: **Supervised comparison** on IH2.6M (top) and FH (bottom). A HRNet encoder is pre-trained with contrastive objective on FH and IH2.6M combined. The pretrained model is then finetuned on IH2.6M. This model is compared with a supervised HRNet model trained only on IH2.6M.

Appendix C. Supervised experiment results

In Table 7, we report the unaligned metrics for the *comparison with supervised model* experiment. The HRNet (Wang et al., 2020) encoder is pretrained on FH and IH2.6M combined and then fine-tuned and evaluated on IH2.6M or FH. Additionally we report the performance on FH and IH2.6M of various recent works in table 8 and 9, respectively. The IH2.6M results are reported on single hand split of human and machine annotated test split of version 0 InterHand2.6M.

Method	PA MPJE (cm) ↓	PA AUC ↑
Spurr et al. (2020)	0.90	0.82
Kulon et al. (2020a)	0.84	0.83
Li et al. (2020)	0.80	0.84
Pose2Mesh (Choi et al., 2020)	0.77	-
I2L-MeshNet (Moon and Lee, 2020)	**0.74**	-
HRNet	1.08	0.79
+ Contrastive	0.95	0.81

Table 8: **Supervised comparison** on FH. A HRNet encoder is pre-trained with contrastive objective on FH and IH2.6M combined. The pretrained model is then finetuned on FH. This model is compared with various other methods.

Method	RA MPJE (cm) ↓	RA AUC ↑
Moon et al. (2020)	**1.26**	-
HRNet	1.50	0.967
+ Contrastive	1.40	0.968

Table 9: **Supervised comparison** on IH2.6M. A HRNet encoder is pre-trained with contrastive objective on FH and IH2.6M combined. The pretrained model is then finetuned on IH2.6M. This model is compared with various other methods. Here RA stands for root aligned.