Contrastive Representation Learning for Hand Shape Estimation

Christian Zimmermann, Max Argus(✉), and Thomas Brox

University of Freiburg, Freiburg im Breisgau, Germany
argus@cs.uni-freiburg.de
https://lmb.informatik.uni-freiburg.de/projects/contra-hand/

Abstract. This work presents improvements in monocular hand shape estimation by building on top of recent advances in unsupervised learning. We extend momentum contrastive learning and contribute a structured collection of hand images, well suited for visual representation learning, which we call *HanCo*. We find that the representation learned by established contrastive learning methods can be improved significantly by exploiting advanced background removal techniques and multi-view information. These allow us to generate more diverse instance pairs than those obtained by augmentations commonly used in exemplar based approaches. Our method leads to a more suitable representation for the hand shape estimation task and shows a 4.7% reduction in mesh error and a 3.6% improvement in F-score compared to an ImageNet pretrained baseline. We make our benchmark dataset publicly available, to encourage further research into this direction.

Keywords: Hand shape estimation · Self-supervised learning · Contrastive learning · Dataset

1 Introduction

Leveraging unlabeled data for training machine learning is a long standing goal in research and its importance has increased dramatically with the advances made by data-driven deep learning methods. Using unlabeled data is an attractive proposition, because more training data usually leads to improved results. On the other hand, label acquisition for supervised training is difficult, time-consuming, and cost intensive.

While the use of unsupervised learning is conceptually desirable the research community struggled long to compete with simple transfer learning approaches using large image classification benchmark datasets. For a long time, there has been a substantial gap between the performance of these methods and the results of supervised pretraining on ImageNet. However, recent work [14] succeeded to surpass ImageNet based pretraining for multiple downstream tasks. The core innovation was to use a consistent, dynamic and large dictionary of embeddings in combination with a contrastive loss, which is a practice we are following in this work as well. Similarly, we make use

C. Zimmermann and M. Argus—These authors contributed equally.

© Springer Nature Switzerland AG 2021
C. Bauckhage et al. (Eds.): DAGM GCPR 2021, LNCS 13024, pp. 250–264, 2021.
https://doi.org/10.1007/978-3-030-92659-5_16

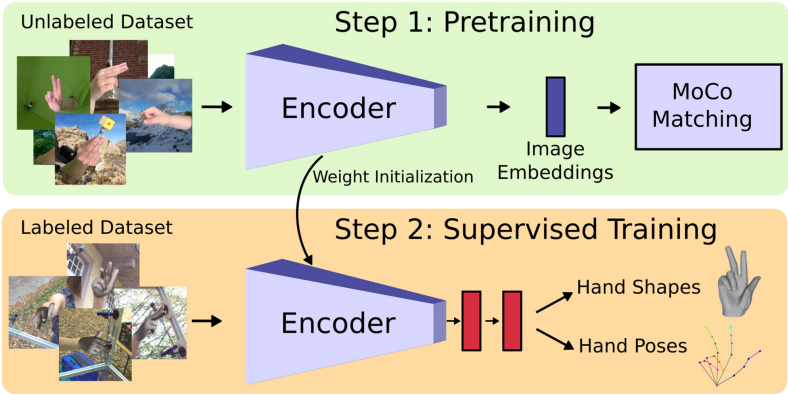

Fig. 1. We train a self-supervised feature representation on our proposed large dataset of unlabeled hand images. The resulting encoder weights are then used to initialize supervised training based on a smaller labeled dataset. This pretraining scheme yields useful image embeddings that can be used to query the dataset, as well as increasing the performance of hand shape estimation.

of strong geometric and color space augmentations, like flipping, cropping, as well as modification of hue and brightness, to generate positive pairs at training time.

Additionally, we find that large improvements lie in the use of strategies that extend the simple exemplar strategy of using a single image and heavy augmentation to generate a positive pair of samples. More precisely, we explore sampling strategies that exploit the structure in *HanCo* that is available without additional labeling cost. The data was captured in an controlled multi-view setting as video sequences, which allows us to easily extract foreground segmentation masks and sample correlated hand poses by selecting simultaneously recorded images of neighboring cameras. This allows us to generate more expressive positive pairs during self-supervised learning.

Hand shape estimation is a task where it is inherently hard to acquire diverse training data at a large scale. This stems from frequent ambiguities and its high dimensionality, which further raises the value of self-supervision techniques. Its need for large amounts of training data also makes hand shape estimation an ideal testbed for developing self-supervision techniques. Most concurrent work in the field of hand shape estimation follows the strategy of weak-supervision, where other modalities are used to supervise hand shape training indirectly. Instead, we explore an orthogonal approach: pretraining the network on data of the source domain which eliminates both the need for hand shape labels as well as additional modalities.

In our work, we find that using momentum contrastive learning yields a valuable starting point for hand shape estimation. It can find a meaningful visual representation from pure self-supervision that allows us to surpass the ImageNet pretrained baseline significantly. We provide a comprehensive analysis on the learned representation and show how the procedure can be used for identifying clusters of hand poses within the data or perform retrieval of similar poses.

For the purpose of further research into self-supervised pretraining we release a dataset that is well structured for this purpose it includes a) a larger number of images, b) temporal structure, and c) multi-view camera calibration.

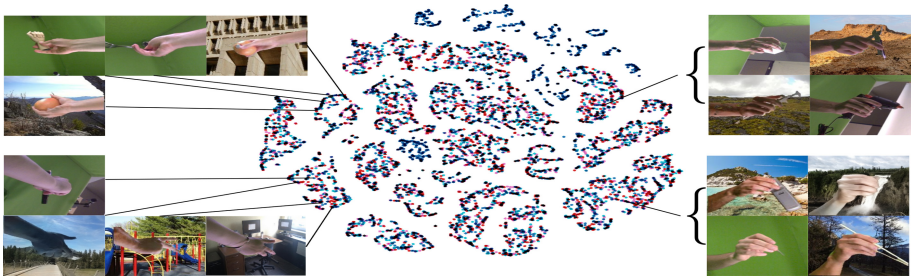

Fig. 2. Shown is a two dimensional *t-SNE* embedding of the image representation found by unsupervised visual representation learning. On the left hand side, we show that similar hand poses are located in proximity to each other. On the right hand side examples of nearly identical hand poses are shown which are approximately mapped to the same point. The color of the dots indicates which pre-processing method has been applied to the sample, a good mixing shows that the embedding focuses on the hand pose instead. Blue represents the unaltered camera-recorded images, red are images that follow the simple cut and paste strategy, while pink and bright blue correspond to images processed with the methods by Tsai *et al*. [33] and Zhang *et al*. [40] respectively.

2 Related Work

Visual Representation Learning. Approaches that aim to learn representations from collection of images without any labels can be roughly categorized into generative and discriminative approaches. While earlier work was targetting generative approaches the focus shifted towards discriminative methods that either leverage contrastive learning or formulate auxiliary tasks using pseudo labels.

Generative approaches aim to recover the input, subject to a certain type of pertubation or regularization. Examples are DCGAN by Radford *et al*. [26], image colorization [39], denoising autoencoders [34] or image in-painting [24].

Popular auxiliary tasks include solving Jigsaw Puzzles [22], predicting image rotations [11], or clustering features during training [5].

In contrast to auxiliary tasks the scheme of contrastive loss functions [12], doesn't define pseudo-labels, but uses a dynamic dictionary of keys that are sampled from the data and resemble the current representation of samples via the encoding network. For training one matching pair of keys is generated and the objective drives the matching keys to be similar, while being dissimilar to the other entries in the dictionary. Most popular downstream task is image classification, where contrastive learning approaches have yielded impressive results [4,6,7]. In this setting pretrained features are evaluated by a common protocol in which these features are frozen and only a supervised linear classifier is trained on the global average pooling of these features. In this work we follow the paradigm of contrastive learning, hence the recent successes, and build on the work by Chen *et al*. [7]. However, we find retraining of the complete convolutional backbone is necessary for hand shape estimation, which is following the transfer learning by fine-tuning [8,37] idea. Furthermore, we extend sampling of matching keys beyond exemplar and augmentation-based strategies.

Fig. 3. Examples showing the cosine-similarity scores for the embeddings produced by pretraining. The first row shows embeddings for the same hand pose with different backgrounds, the embedding learns to focus on the hand and ignore the background. In the second row we show images for similar hand poses, with different background, these also score highly. In the third row different hand poses, with the same background produce low scores.

Hand Shape Estimation. Most recent works in the realm of hand shape estimation rely on a deep neural network estimating parameters of a hand shape model. By far, the most popular choice is MANO presented by Romero *et al.* [28]. There are little approaches using another surface topology; examples are Moon *et al.* [21] and Ge *et al.* [10], but MANO was used by the majority of works [2,13,18,42] and is also used in our work.

Commonly, these approaches are trained with a combination of losses, frequently incorporating additional supervision coming from shape derived modalities like depth [3,21,35], silhouette [1,2,20], or keypoints [2,13,18,21,42], which is referred to as weak-supervision. This allows to incorporate datasets into training that don't contain shape annotation, significantly reducing the effort for label acquisition. Sometimes, these approaches incorporate adversarial losses [17], which can help to avoid degenerate solutions on the weakly labeled datasets. Other works focusing on hand pose estimation propose weak supervision by incorporating biomechanical constraints [30] or sharing a latent embedding space between multiple modalities [31,32].

One specific way of supervision between modalities is leveraging geometric constraints of multiple camera views, which was explored in human pose estimation before:

Rhodin *et al*. [27] proposed to run separate 3D pose detectors per view and constrain the estimated poses with respect to each other. Simon *et al*. [29] presents an iterative bootstrapping procedure of self-labeling with using 3D consistency as a quality criterion. Yao *et al*. [36] and He *et al*. [16] supervise 2D keypoint detectors by explicitly incorporating epipolar constraints between two views.

In our work the focus is not on incorporating constraints by adding weak-supervision to the models' training, but we focus on finding a good initialization for supervised hand shape training using methods from unsupervised learning.

3 Approach

Our approach to improve monocular hand shape estimation consists of two steps and is summarized in Fig. 1: First, we are pretraining the CNN encoder backbone on large amounts of unlabeled data using unsupervised learning on a pretext task. Second, the CNN is trained for hand shape estimation in a supervised manner, using the network weights from the pretext task as initialization.

Momentum Contrastive Learning. *MoCo* [14] is a recent self-supervised learning method that performs contrastive learning as a dictionary look-up. *MoCo* uses two encoder networks to encode different augmentations of the same image instance as query and key pairs. Given two images $\boldsymbol{I}_i \in \mathbb{R}^{H \times W}$ and $\boldsymbol{I}_j \in \mathbb{R}^{H \times W}$ the embeddings are calculated as

$$q = f(\theta,\ \boldsymbol{I}_i) \quad \text{and} \tag{1}$$

$$k = f(\tilde{\theta},\ \boldsymbol{I}_j). \tag{2}$$

which yields a query q and key k. The same function f is used in both cases, but parameterized differently. The query function uses θ which is directly updated by the optimization, while the key function $\tilde{\theta}$ is updated indirectly. At a given optimizations step n it is calculated as

$$\tilde{\theta}_n = m \cdot \tilde{\theta}_{n-1} + (1 - m) \cdot \theta_n \tag{3}$$

using the momentum factor m, which is chosen close to 1.0 to ensure a slowly adapting encoding of the key values k.

During training a large queue of dictionary keys k is accumulated over iterations which allows for efficient training as a large set of negative samples has been found to be critical for contrastive training [14]. Following this methodology *MoCo* produces feature representations that transfer well to a variety of downstream tasks. As training objective the InfoNCE loss [23] is used, which relates the inner product of the matching key-query-pair to the inner products of all negative pairs in a softmax cross-entropy kind of fashion.

At test time the similarity of a key-value-pair can be computed using cosine similarity

$$\text{cossim}(q,\ k) = \frac{q \cdot k}{\|q\|_2 \cdot \|k\|_2} \tag{4}$$

which can return values ranging from 1.0 in the similar case to -1.0 in the dissimilar case.

MoCo relies entirely on standard image space augmentations to generate positive pairs of image during representation learning. A function $g(.)$, subject to a random vector ζ, is applied to the same image instance $\boldsymbol{I}$ two times to generate different augmentations

$$\boldsymbol{I}_i = g(\boldsymbol{I},\ \zeta_1) \tag{5}$$

$$\boldsymbol{I}_j = g(\boldsymbol{I},\ \zeta_2) \tag{6}$$

that are considered as the matching pair. The function $g(.)$ performs randomized: crops of the image, color jitter, grayscale, and conversion and Gaussian blur. We omit randomized image flipping as this augmentation changes the semantic information of the hand pose.

The structured nature of *HanCo* allows us going beyond these augmentation-based strategies. Here we are looking for strategies that preserve the hand pose, but change other aspects of the image. We investigate three different configurations a) background randomization, b) temporal sampling and c) multi-view sampling.

HanCo consists of short video clips that are recorded by multiple cameras simultaneously at 5 Hz. The clips are up to 40 s long and have an average length of 14 s. For simplicity and without loss of generality we describe the sampling methods for a single sequence only. Extending the approach towards multiple sequences is straight forward, by first sampling a sequence and then applying the described procedure. Formally, we can sample from a pool of images $\boldsymbol{I}_t^c$ recorded at a timestep t from camera c.

During background randomization we use a single image $\boldsymbol{I}_t^c$ with its foreground segmentation as source to cut the hand from the source image and paste it into a randomly sampled background image. Example outputs of background randomization are shown in the first row of Fig. 3.

For temporal sampling we exploit the fact, that our unlabeled dataset stems from a video stream which naturally constraints subsequent hand poses to be highly correlated. A positive pair of samples is generated by sampling two neighboring frames $\boldsymbol{I}_t^c$ and $\boldsymbol{I}_{t+1}^c$ for a given camera c. Due to hand movement the hand posture is likely to change slightly from t to $t+1$, which naturally captures the fact that similar poses should be encoded with similar embeddings.

As the data is recorded using a calibrated multi-camera setup and frame capturing is temporally synchronized, views from different cameras at a particular point in time show the same hand pose. This can be used as a powerful method of "augmentation" as different views change many aspects of the image but not the hand pose. Consequently, we generate a positive sample pair $\boldsymbol{I}_t^{c_1}$ and $\boldsymbol{I}_t^{c_2}$ in the multi-view case by sampling neighboring cameras c_1 and c_2 at a certain timestep t. The dataset contains an 8 camera setup, with cameras mounted on each of the corners of a cube, in order to simplify the task of instance recognition, we chose to sample neighboring cameras, meaning those connected by no more than one edge of the cubical fixture.

Hand Shape Estimation. Compared to unsupervised pretraining the network architecture used for shape estimation is modified by changing the number of neurons of the last fully-connected layer from 128 to 61 neurons in order to estimate the *MANO* parameter

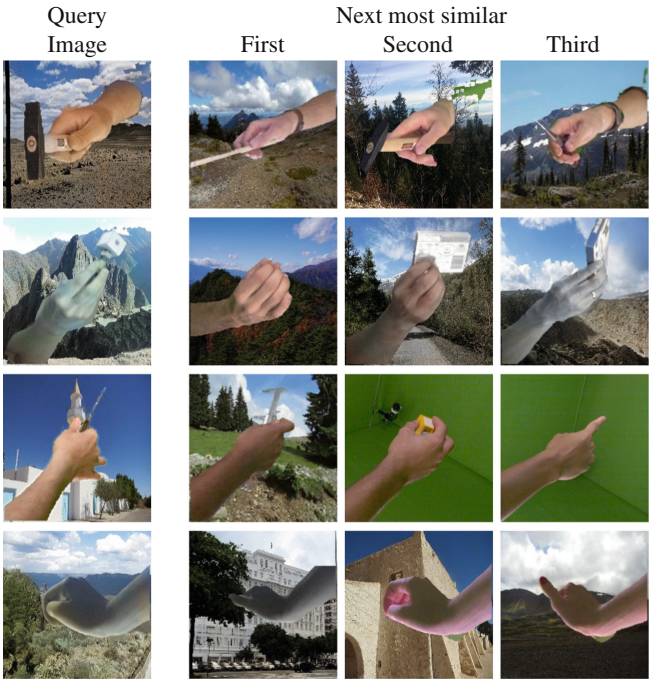

Fig. 4. Given a query image our learned embedding allows to identify images showing similar poses. This enables identifying clusters in the data, without the need of pose annotations. The nearest neighbors are queried from a random subset of 25,000 samples of *HanCo*.

vector. Consequently, the approach is identical to the one presented by Zimmermann *et al*. [42] with the difference being only that the weights of the convolutional backbone are initialized through unsupervised contrastive learning and not ImageNet pretraining. The network is being trained to estimate the *MANO* parameter vector $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{61}$ using the following loss:

$$\begin{aligned} \mathcal{L} = w_{\text{3D}} \left\| \boldsymbol{P} - \tilde{\boldsymbol{P}} \right\| + \\ w_{\text{2D}} \left\| \Pi(\boldsymbol{P}) - \Pi(\tilde{\boldsymbol{P}}) \right\| + \\ w_{\text{p}} \left\| \boldsymbol{\theta} - \tilde{\boldsymbol{\theta}} \right\| . \end{aligned} \tag{7}$$

We deploy L_2 losses for all components and weight with $w_{\text{3D}} = 1000$, $w_{\text{2D}} = 10$, and $w_{\text{p}} = 1$ respectively. To derive the predicted keypoints $\tilde{\boldsymbol{P}}$ from the estimated shape $\tilde{\boldsymbol{\theta}}$ in a differentiable way the MANO *et al*. [28] model implementation in *PyTorch* by Hasson *et al*. [13] is used. $\boldsymbol{P} \in \mathbb{R}^{21}$ is the ground truth 3D location of the hand joints and $\boldsymbol{\theta}$ the ground truth set of MANO parameter, both of which are provided by the training dataset. Denoted by $\Pi(.)$ is the projection operator mapping from 3D space to image pixel coordinates.

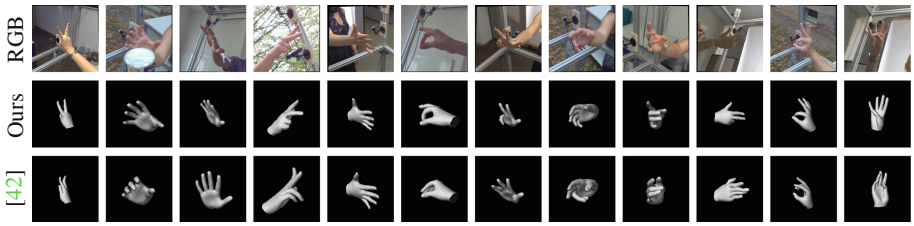

Fig. 5. Qualitative comparison of MANO predictions between *Ours-Multi view* and the approach by Zimmermann *et al*. [42] showing improvements in hand mesh predictions yielded by our self-supervised pretraining on the evaluation split of *FreiHAND* [42]. Generally, our predictions look seem to capture the global pose and grasp of the hand more accurately, which results into a lower mesh error and higher F@5 score.

4 Experiments

Dataset. Our experiments are conducted on data recorded by Zimmermann *et al*. [42], which the authors kindly made available to us. The images show 32 subjects which are recorded by 8 cameras simultaneously that are mounted on a cubical aluminum fixture. The cameras face towards the center of the cube and are calibrated. One part of the data was recorded against a green background, which allows extracting of the foreground segmentation automatically and to perform background randomization without any additional effort. Another part of the data was intended for evaluation and was recorded without green backgrounds. For a subset of both parts there are hand shape labels, which were created by the annotation method [42]. The set of annotated frames was published before as *FreiHAND* [42], which we use for evaluation of the supervised hand shape estimation approaches.

For compositing the hand foreground with random backgrounds we have collected 2193 background images from Flickr, that are showing various landscapes, city scenes and indoor shots. We manually inspect these images, to ensure they don't contain shots targeted at humans. There are three different methods to augment the cut and paste version: The colorization approach by Zhang *et al*. [40] is used in both its automatic and sampling-based mode. Also we use the deep harmonization approach proposed by Tsai *et al*. [33] that can remove color bleeding at the foreground boundaries. The background post-processing methods are also reflected by the point colors in the *t-SNE* embedding plot (Fig. 2).

The complete dataset is used for visual representation learning and we refer to it as *HanCo*, which provides $107,538$ recorded time instances or poses, each recorded by 8 cameras. All available frames are used for unsupervised training, which results into $860,304$ frames, while a subset of $63,864$ frames contains hand shape annotations and are recorded against green screen, which we use for supervised training of the monocular hand shape estimation network.

Training Details. For training of neural networks the *PyTorch* framework is used and we rely on ResNet-50 as convolutional backbone [15,25].

During unsupervised training we follow the procedure by Chen *et al*. [7] and train with the following hyper-parameters: a base learning rate of 0.015 is used, which is annealed following a cosine schedule over 100 epochs. We train with a batch size of 128 and an image size of 224×224 pixels. We follow the augmentations of Chen *et al*. [7], but skip image flipping. For supervised training we follow the training schedule by Zimmermann *et al*. [42]. The network is trained for $500,000$ steps with a batch size of 16. A base learning rate of 0.0001 is used, and decayed by a factor of 0.5 after $220,000$ and $340,000$ steps.

Evaluation of Embeddings. First, we perform a qualitative evaluation of the learned embedding produced by pretraining. For this purpose, we sample a random subset of $25,000$ images from the unlabeled dataset. The query network is used to compute a 128 dimensional feature vector for each image. Using the feature vectors we find a *t-SNE* [19] representation that is shown in Fig. 2. It is apparent, that similar poses cluster closely together, while different poses are clearly separated.

Pairs of images, together with their cosine similarity scores (4) are shown in Fig. 5, the cosine similarity scores reveal many desirable properties of the embedding: The representation is invariant to the picked background (first row), small changes in hand pose only result in negligible drop in similarity (second row). Viewing the same hand pose from different directions results in high similarity scores, though this can be subject to occlusion (third row). Large changes in hand pose induce a significant drop of the score (last row). This opens up the possibility to use the learned embedding for retrieval tasks, which is shown in Fig. 4. Given a query image it is possible to identify images showing similar and different hand poses, without an explicit hand pose annotation.

Hand Shape Estimation. For comparison we follow [42] and rely on the established metrics *mesh error* in cm and F-score evaluated at thresholds of 5 mm and 15 mm. All of them being reported for the Procrustes aligned estimates as calculated by the online Codalab evaluation service [41].

The results of our approach are compared to literature in Table 1. The results reported share a similar architecture that is based on a ResNet-50 and the differences can be attributed to the training objective and data used. The approach presented by Boukhayma *et al*. [2] performs pretraining on a large synthetic dataset and subsequent training on a combination of datasets containing real images with 2D, 3D and hand segmentation annotation. The datasets used are *MPII+NZSL* [29], *Panoptic* [29] and *Stereo* [38]. Another rendered dataset is proposed and used for training by Hasson *et al*. [13], which is combined with real images from Garcia *et al*. [9]. Zimmermann *et al*. [42] use only the real images from the *FreiHAND* dataset for training, which is the setting we are also using for the results reported.

Table 1 summarizes the results of the quantitative evaluation on *FreiHAND*. It shows that training the network from random initialization leads to unsatisfactory results reported by *Ours-Scratch*, which indicates that a proper network initialization is important. *Ours-Fixed* is training only the fully-connected layers starting from the weights found by *MoCo* while keeping the convolutional part fixed. This achieves results that fall far behind in comparison. Additionally, training the parameters associated with batch normalization gives a significant boost in accuracy as reported by *Ours-Fixed-BN*. The entry named *Ours-Augmentation* does not make use of the advanced back-

Table 1. Pretraining the convolutional backbone using momentum contrastive learning improves over previous results by -4.7% in terms of mesh error, 3.6% in terms of F@5 mm and 0.9% in terms of F@15 mm (comparing Zimmermann *et al*. [42] and *Ours-Multi View*). This table shows that fixing the *MoCo* learned convolutional backbone and only training the fully-connected part during hand shape estimation (see *Ours-Fixed*) can not compete with state-of-the-art approaches. *Ours-Fixed-BN* shows that additionally training the batch normalization parameters leads to substantial improvements. Consequently, leaving all parameters open for optimization (*Ours-Augmentation*) leads to further improvements. In *Ours-Scratch* all parameters are trained from random initialization, which performs much better than fixing the convolutional layers, but is still behind the reported results in literature, which illustrates the importance of a good network initialization. Applying, our proposed sampling strategies *Ours-Background*, *Ours-Temporal* or *Ours-Multi View* does improve results over using an augmentation based sampling strategy like Chen *et al*. [7], denoted by *Ours-Augmentation*.

Method	Mesh error in cm ↓	F@5 mm ↑	F@15 mm ↑
Boukhayma *et al*. [2]	1.30	0.435	0.898
Hasson *et al*. [13]	1.32	0.436	0.908
Zimmermann *et al*. [42]	1.07	0.529	0.935
Ours-fixed	2.13	0.299	0.803
Ours-fixed-BN	1.25	0.463	0.914
Ours-scratch	1.24	0.475	0.915
Ours-augmentation	1.09	0.521	0.934
Ours-background	1.04	0.538	0.940
Ours-temporal	1.04	0.538	0.939
Ours-multi view	**1.02**	**0.548**	**0.943**

ground randomization methods, and is the direct application of the proposed *MoCo* approach [7] to our data. In this case all network weights are being trained. It performs significantly better than the fixed approaches and training from scratch, but lacks behind ImageNet based initialization used by Zimmermann *et al*. [42]. Finally, the rows *Ours-Background*, *Ours-Temporal* and *Ours-Multi view* report results for the proposed sampling strategies for positive pairs while training all network parameters. All methods are able to outperform the ImageNet trained baseline by Zimmermann *et al*. [42]. We find the differences between *Ours-Background* and *Ours-Temporal* to be negligible, while *Ours-Multi view* shows an significant improvement over Zimmermann *et al*. [42]. This shows the influence and importance of the proposed sampling strategies.

To quantify the effect of our proposed multi-view pretraining strategy we do an ablation study in which the quantity of training samples for supervised training is reduced and we evaluate how well our proposed multi-view sampling strategy compares to an ImageNet pretrained baseline. This is shown in Fig. 6. In this experiment, the amount of labeled data used during supervised training is varied between 20% and 100% of the full *FreiHAND* training dataset, which corresponds to between 12,772 and 63,864 training samples. Except for varying the training data we follow the approach by Zimmermann *et al*. [42]. The figure shows curves for the minimal, maximal, and mean F@5 scores achieved over three runs per setting for Procrustes aligned predictions.

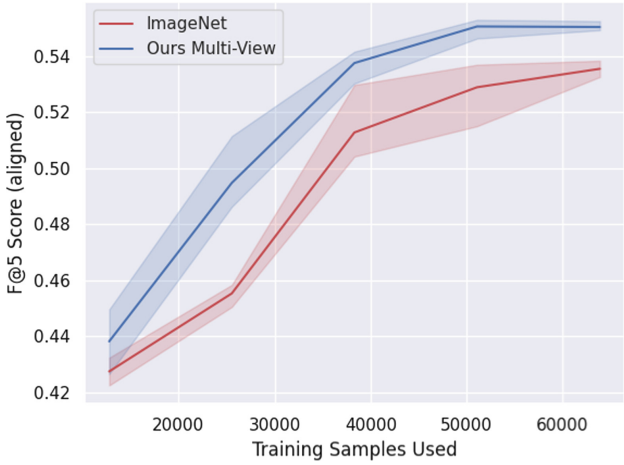

Fig. 6. Comparison between our proposed pretraining method (blue) and an ImageNet baseline (red) for varying fractions of the training dataset used during supervised learning of hand shape estimation. The lines represent the mean result over 3 runs per setting, while the shaded area indicates the range of obtained results. Our proposed *Multi-View* sampling approach consistently outperforms the baseline, with the largest differences occurring when using approximately 40% of the full training dataset, indicating a sweet-spot where pretraining is most beneficial. Learning from very large or small datasets reduces the gains from pretraining, for small datasets we hypothesize that there is not sufficient variation in the training dataset to properly learn the task of hand shape estimation anymore. (Color figure online)

We observe that *Our Multi-View* consistently outperforms the ImageNet pretrained baseline and that more data improves the performance of both methods. The differences between both methods are largest when using 40% of the full training dataset, for a very large or small supervised datasets the differences between the two methods become smaller. We hypothesize, that there is a sweet spot between enough training data, to diminish the value of pretraining, and a minimal amount of training data needed to learn the task of hand shape estimation reasonably well from the supervised data.

A qualitative comparison between our method and results by [42] is presented in Fig. 5. It shows that the differences between both methods are visually subtle, but in general hand articulation is captured more accurately by our methods which results into lower *mesh error* and higher F-scores.

5 Conclusion

Our work shows that unsupervised visual representation learning is beneficial for hand shape estimation and that sampling meaningful positive pairs is crucial. We only scratch on the surface of possible sampling strategies and find that sampling needs to be inline with the downstream task at hand.

By making the data available publicly, we encourage the community strongly to explore some of the open directions. These include to extend sampling towards other

sources of consistency that remain untouched by our work, *e.g.*, temporal consistency of hand pose within one recorded sequence leaves opportunities for future exploration.

Another direction points into combining recently proposed weak-supervision methods with the presented pretraining methods and leverage the given calibration information at training time.

References

1. Baek, S., Kim, K.I., Kim, T.: 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, 16–20 June 2019, pp. 1067–1076. Computer Vision Foundation/IEEE (2019). https://doi.org/10.1109/CVPR.2019.00116, http://openaccess.thecvf.com/content_CVPR_2019/html/Baek_Pushing_the_Envelope_for_RGB-Based_Dense_3D_Hand_Pose_Estimation_CVPR_2019_paper.html
2. Boukhayma, A., de Bem, R., Torr, P.H.S.: 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, 16–20 June 2019, pp. 10843–10852. Computer Vision Foundation/IEEE (2019). https://doi.org/10.1109/CVPR.2019.01110, http://openaccess.thecvf.com/content_CVPR_2019/html/Boukhayma_3D_Hand_Shape_and_Pose_From_Images_in_the_Wild_CVPR_2019_paper.html
3. Cai, Y., Ge, L., Cai, J., Yuan, J.: Weakly-supervised 3D hand pose estimation from monocular RGB images. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11210, pp. 678–694. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01231-1_41
4. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. ArXiv abs/2006.09882 (2020)
5. Caron, M., Bojanowski, P., Joulin, A., Douze, M.: Deep clustering for unsupervised learning of visual features. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018. LNCS, vol. 11218, pp. 139–156. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_9
6. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.: Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029 (2020)
7. Chen, X., Fan, H., Girshick, R.B., He, K.: Improved baselines with momentum contrastive learning. CoRR abs/2003.04297 (2020). https://arxiv.org/abs/2003.04297
8. Donahue, J., et al.: Decaf: a deep convolutional activation feature for generic visual recognition. In: ICML (2014)
9. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.: First-person hand action benchmark with RGB-D videos and 3d hand pose annotations. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 409–419. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00050, http://openaccess.thecvf.com/content_cvpr_2018/html/Garcia-Hernando_First-Person_Hand_Action_CVPR_2018_paper.html
10. Ge, L., et al.: 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, 16–20 June 2019, pp. 10833–10842. Computer Vision Foundation/IEEE (2019). https://doi.org/10.1109/CVPR.2019.01109, http://openaccess.thecvf.com/content_CVPR_2019/html/Ge_3D_Hand_Shape_and_Pose_Estimation_From_a_Single_RGB_CVPR_2019_paper.html

11. Gidaris, S., Singh, P., Komodakis, N.: Unsupervised representation learning by predicting image rotations. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018). https://openreview.net/forum?id=S1v4N2l0-
12. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17–22 June 2006, New York, NY, USA, pp. 1735–1742. IEEE Computer Society (2006). https://doi.org/10.1109/CVPR.2006.100
13. Hasson, Y., et al.: Learning joint reconstruction of hands and manipulated objects. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019, pp. 11807–11816. Computer Vision Foundation/IEEE (2019). https://doi.org/10.1109/CVPR.2019.01208, http://openaccess.thecvf.com/content_CVPR_2019/html/Hasson_Learning_Joint_Reconstruction_of_Hands_and_Manipulated_Objects_CVPR_2019_paper.html
14. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020, pp. 9726–9735. IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.00975
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 770–778. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.90
16. He, Y., Yan, R., Fragkiadaki, K., Yu, S.: Epipolar transformers. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, 13–19 June 2020, pp. 7776–7785. IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.00780
17. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 7122–7131. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00744, http://openaccess.thecvf.com/content_cvpr_2018/html/Kanazawa_End-to-End_Recovery_of_CVPR_2018_paper.html
18. Kulon, D., Güler, R.A., Kokkinos, I., Bronstein, M.M., Zafeiriou, S.: 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, 13–19 June 2020, pp. 4989–4999. IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.00504
19. van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)
20. Malik, J., Elhayek, A., Stricker, D.: WHSP-Net: a weakly-supervised approach for 3D hand shape and pose recovery from a single depth image. Sensors **19**(17), 3784 (2019)
21. Moon, G., Shiratori, T., Lee, K.M.: DeepHandMesh: a weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12347, pp. 440–455. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58536-5_26
22. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 69–84. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_5
23. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. CoRR abs/1807.03748 (2018). http://arxiv.org/abs/1807.03748
24. Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: feature learning by inpainting. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, 27–30 June 2016, pp. 2536–2544. IEEE Computer Society (2016). https://doi.org/10.1109/CVPR.2016.278

25. Qiao, S., Wang, H., Liu, C., Shen, W., Yuille, A.L.: Weight standardization. CoRR abs/1903.10520 (2019). http://arxiv.org/abs/1903.10520
26. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: Bengio, Y., LeCun, Y. (eds.) 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, 2–4 May 2016, Conference Track Proceedings (2016). http://arxiv.org/abs/1511.06434
27. Rhodin, H., et al.: Learning monocular 3D human pose estimation from multi-view images. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 8437–8446. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00880, http://openaccess.thecvf.com/content_cvpr_2018/html/Rhodin_Learning_Monocular_3D_CVPR_2018_paper.html
28. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: modeling and capturing hands and bodies together. ACM Trans. Graph. **36**(6), 245:1–245:17 (2017). https://doi.org/10.1145/3130800.3130883
29. Simon, T., Joo, H., Matthews, I.A., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 4645–4653. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.494
30. Spurr, A., Iqbal, U., Molchanov, P., Hilliges, O., Kautz, J.: Weakly supervised 3D hand pose estimation via biomechanical constraints. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12362, pp. 211–228. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58520-4_13
31. Spurr, A., Song, J., Park, S., Hilliges, O.: Cross-modal deep variational hand pose estimation. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, 18–22 June 2018, pp. 89–98. IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00017, http://openaccess.thecvf.com/content_cvpr_2018/html/Spurr_Cross-Modal_Deep_Variational_CVPR_2018_paper.html
32. Theodoridis, T., Chatzis, T., Solachidis, V., Dimitropoulos, K., Daras, P.: Cross-modal variational alignment of latent spaces. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, 14–19 June 2020, pp. 4127–4136. IEEE (2020). https://doi.org/10.1109/CVPRW50498.2020.00488
33. Tsai, Y., Shen, X., Lin, Z., Sunkavalli, K., Lu, X., Yang, M.: Deep image harmonization. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017, pp. 2799–2807. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.299
34. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.: Extracting and composing robust features with denoising autoencoders. In: Cohen, W.W., McCallum, A., Roweis, S.T. (eds.) Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5–9, 2008. ACM International Conference Proceeding Series, pp. 1096–1103. ACM, New York (2008). https://doi.org/10.1145/1390156.1390294
35. Wan, C., Probst, T., Gool, L.V., Yao, A.: Self-supervised 3D hand pose estimation through training by fitting. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019, pp. 10853–10862. Computer Vision Foundation/IEEE (2019). https://doi.org/10.1109/CVPR.2019.01111, http://openaccess.thecvf.com/content_CVPR_2019/html/Wan_Self-Supervised_3D_Hand_Pose_Estimation_Through_Training_by_Fitting_CVPR_2019_paper.html
36. Yao, Y., Jafarian, Y., Park, H.S.: MONET: multiview semi-supervised keypoint detection via epipolar divergence. In: 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pp. 753–762. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00084

37. Yosinski, J., Clune, J., Bengio, Y., Lipson, H.: How transferable are features in deep neural networks? In: NIPS (2014)
38. Zhang, J., Jiao, J., Chen, M., Qu, L., Xu, X., Yang, Q.: 3D hand pose tracking and estimation using stereo matching. CoRR abs/1610.07214 (2016). http://arxiv.org/abs/1610.07214
39. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 649–666. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_40
40. Zhang, R., et al.: Real-time user-guided image colorization with learned deep priors. ACM Trans. Graph. **36**(4), 119:1-119:11 (2017). https://doi.org/10.1145/3072959.3073703. Kindly provide year of the publication for the Ref. [41]
41. Zimmermann, C.: Freihand competition. https://competitions.codalab.org/competitions/21238
42. Zimmermann, C., Ceylan, D., Yang, J., Russell, B.C., Argus, M.J., Brox, T.: 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, pp. 813–822. IEEE (2019). https://doi.org/10.1109/ICCV.2019.00090