# Multiple Transfer Net Based Region Ensemble Network for Deep Hand Pose Estimation

Haoqian Wang* and Da Li† and Xingzheng Wang‡
* Key Laboratory of Broadband Network & Multimedia, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
Shenzhen Institute of Future Media Technology, Shenzhen 518071, China
E-mail: wangyizhai@sz.tisnghua.edu.cn Tel/Fax: +86-18038153065
† Key Laboratory of Broadband Network & Multimedia, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
E-mail: d-li16@tsinghua.edu.cn Tel/Fax: +86-18801051354
‡ Key Laboratory of Broadband Network & Multimedia, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
E-mail: xingzheng.wang@sz.tsinghua.edu.cn Tel/Fax: +86-13613091065

***Abstract*—Deep hand pose estimation from single depth image plays an significant role in human-computer interaction. This paper proposes a novel method based on multiple transfer net to estimate hand pose with single-channel depth photos only. A channel extending process for original single channel depth image is implemented to extend hand and hand palm regions, match the input format of pre-trained network and fully utilize the parameters. A multiple transfer network refinement for the previous convolutional neural network is maken to obtain various different feature maps. And a region ensemble is used to merge all output feature maps and integrate the results. The experimental results demonstrate that proposed method outperforms state-of-art results by a considerable accuracy on NYU [1] and ICVL [2] datasets.**

## I. Introduction

Hand pose estimation is ubiquitously required in many critical applications and it is one of the most important techniques in human-computer interaction like virtual/augmented reality applications. It aims to predict the 3D locations of hand joints [3] from single depth images, which is critical for gesture recognition [4]. Though it has attracted broad research interests in recent years [5], [6], due to the severe occlusions caused by articulate hand pose and noisy input from affordable depth sensors, high accuracy hand pose estimation is very challenging.

Recent years, for deep convolutional networks have great modeling capacity and end-to-end feature learning, they are widely used in several computer vision tasks such as object detection [7], image segmentation [8], and object classification [9]. And they have also been applied on hand pose estimation. Carreira et al. [10] and Haque et al. [11] used CNN-based methods predicting probability heatmaps of each joint, and inferring hand pose from heatmaps. He et al. [12] proposed a sophisticated design with a feedback loop and Chen et al. [8] presented a spatial attention mechanism. And many ensemble methods for deep hand pose estimation have already been proposed recently. A tree-structured Region Ensemble Network (REN) for directly 3D coordinate regression was proposed by Guo et al. [13]. Chen et al. [14] presented a Pose guided structured Region Ensemble Network (Pose-REN) to boost the performance of hand pose estimation.

Recently, ImageNet pre-trained CNNs have been used for chest pathology identification and detection in X-ray and CT modalities [15]. Maxime te al. [16] showed how image representations learned with CNNs on large-scale annotated datasets can be efficiently transferred to other visual recognition tasks with limited amount of training data. Andrej et al. [17] studied multiple approaches for extending the connectivity of a CNN in time domain to take advantage of local spatio-temporal information and suggested a multiresolution for large-scale video classification. AlexandreEmail [18] investigated the possibility of transferring knowledge between CNNs when processing RGB-D data to recognize 3D object. However, the fine-tuning of an ImageNet pre-trained CNN model on datasets of hand pose estimation has not yet been exploited.

In this paper, we achieve further improvements by proposing a multiple transfer CNN based method estimating hand pose only with depth information. We first use a new channel extending method to extend hand and hand palm regions and match pre-trained networks' input format to fully utilize the pre-trained net. And then we refine the previous convolutional neural network with a multiple transfer network. The multiple transfer network can obtain various different feature maps by using different pre-trained convolutional networks. Finally, we optimize our network by using a region ensemble method [13]. All of these works improve the accuracy of the joints' prediction and achieve remarkable performance on the benchmark.

## II. Proposed Method

### A. Channel Extending

As illustrated in Fig. 1, we firstly transform the original single depth images to three channels images. First, we make center crop to the depth image in two different scales and then up-sample the cropped images to the size of original image. Then we will merge them with the original depth image

together and get a three-channel image. For hand palm and the whole hand areas are usually in the central region of the preprocessed images and center crop for single hand is in process of the preprocess and data augmentation we used [19], this channel extending step actually extends hand and hand palm regions in the original images. Furthermore, the

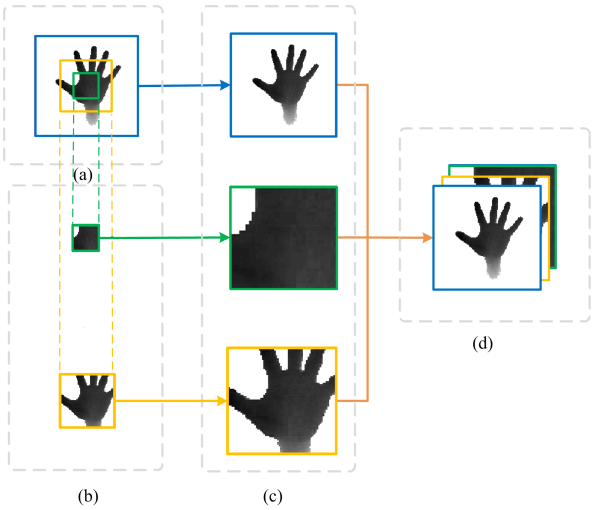


Fig. 1. Extend an original single channel depth image to a three-channel image. (a) is the original single channel depth image. (b) has two images cropped by two different scales. (c) has three channels images resized from (a) and (b). (d) is a three-channel image merged from (c).

input of many common used pre-trained networks like VGG-19 is three-channel RGB image, channel extending can make the single depth data match the input format of the pre-trained network and fully use the pre-trained parameters.

### B. Transfer-Net Ensemble

We use parameters which have already learned from pre-trained network to optimize our network according to the transfer learning method. Compared with training from the beginning, it can decrease the training difficulty and reduce training time. And it can also reduce the overfitting risks. Considering pre-trained networks with more layers have more parameters to learn, which will cost much more time training and have a poor real-time performance while predicting, we firstly choose VGG-11 pre-trained network, which has shallow layers and a few parameters. The network structure is shown in Fig. 2. We use the first eight convolutional layers, whose parameters pre-trained on ImageNet to initialize our convolutional layer parameters. And then fine-tune the fully connected layers defined by ourselves. To get a further improvement, we add a region ensemble [13] layer before fully connected layers. Our experiments show that the convergence speed of this transfer-net is accelerated and the accuracy of prediction achieves state-of-arts.

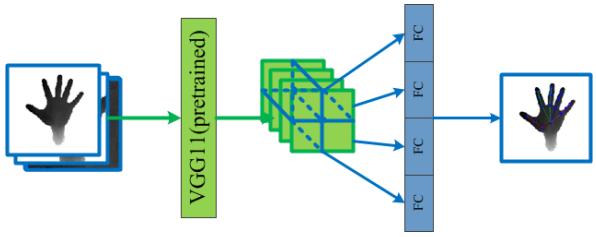


Fig. 2. Network structure of region ensemble single Transfer-net we proposed.

### C. Multiple Transfer Nets Based Region Ensemble

For in the image domain, every activation in the convolutional feature maps is contributed by a receptive field, we can project the multi-view inputs onto the regions of the feature maps. So multi-view voting is equal to utilizing each regions to separately predict the whole hand pose and combining the results. Therefore we use a region ensemble method [13] to optimize the layers which need to be fine-tuning and reach an advance accuracy.

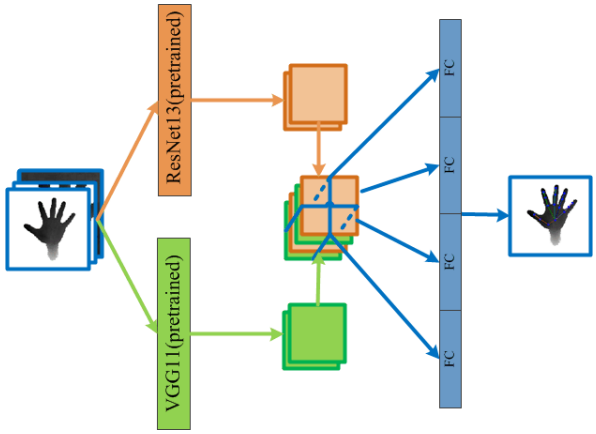


Fig. 3. Network structure of region ensemble multiple transfer net we proposed.

Considering using one pre-trained network only generate a few series of feature maps, we try to use multiple pre-trained models to gain more different feature maps and merge all feature maps together for region ensembling. As is shown in Fig. 3, we use VGG-11 and ResNet-13 pre-trained networks to gain feature maps. Finally upsample all feature maps to 16x16, merge them all together and make a region ensemble to get final outputs.

## III. Experimental Results

We evaluate our multiple transfer net on two public benchmark datasets for hand pose estimation: the NYU dataset [1] and the ICVL dataset [2]. There are different evaluation metrics for hand pose estimation following the literatures, and we report the numbers stated in the papers or measured from the graphs if provided, and plot the relevant

graphs for comparison.

We use two different metrics to evaluate the accuracy: First, we evaluate the accuracy of the 3D hand pose estimation as average 3D joint error. This is established as the most commonly used metric in literature, and allows comparison with many other works due to simplicity of evaluation. Second, we plot the fraction of frames where all predicted joints are below a given average Euclidean distance from the ground truth [20].

We train and test the network using pytorch [21]. We first segment the foreground and extract a cube from the depth image centered in the centroid of hand region [22]. Then resize the cube into 128x128 patch of depth values normalized to [?1, 1] as input for ConvNet. We use Adaptive Moment Estimation (Adam) with a mini-batch size of 16. The learning rate starts from 0.00005 and the model is trained for up to 50 epochs.

### A. Comparison With State-of-Arts

*1) ICVL Dataset:* The ICVL dataset [2] contains a training set of over 180000 depth frames of single-channel depth hand pose images. The test set includes two sequences, each of which has approximately 700 frames. The dataset contains 16 annotated joints and is recorded by a time-of-flight camera. The dataset has a high quality that it hardly has any missing depth values and has sharp outlines with little noise. Although the authors provide different artificially rotated training samples, we use the genuine 22000 frames only and apply the data augmentation proposed in [22]. However, compared to other datasets [1], [23], this dataset's pose variability is limited. As is discussed in [24], [3], annotations of hand joints are very inaccurate.

TABLE I
AVERAGE 3D ERROR ON ICVL DATASET [2].

| Method | Average 3D error |
|---|---|
| Deng et al. [25] (Hand3D) | 10.9mm |
| Tang et al. [2] (LRF) | 12.6mm |
| Wan et al. [26] | 8.2mm |
| Zhou et al. [27] (DeepModel) | 11.3mm |
| Sun et al. [23] (HPR) | 9.9mm |
| Wan et al. [28] (Crossing Nets) | 10.2mm |
| Fourure et al. [29] (JTSC) | 9.2mm |
| Krejov et al. [30] (CDO) | 10.5mm |
| Oberweger et al. [22] (DeepPrior++) | 8.1mm |
| **This work (Multiple transfer net)** | **7.3mm** |

We show a comparison to different state-of-the-art methods on ICVL dataset [2] in Table 1. Our method shows state-of-the-art accuracy. However, the gap to other methods is much smaller. This may be due to the fact that the dataset is much easier, with smaller pose variations [3], and attributed to deviation in the annotations for evaluating [24], [3].

In Fig. 4 we compare multiple transfer net to other methods on the ICVL dataset [2]. Our approach performs similar to the works of Oberweger et al. [22], Wan et al. [26], and Fourure et al. [29], all achieving state-of-the-art accuracies on this dataset. This might be an indication that the performance on the dataset is saturating, and the remaining error is due to the annotation uncertainty. This empirical finding is similar to the discussion in [3].

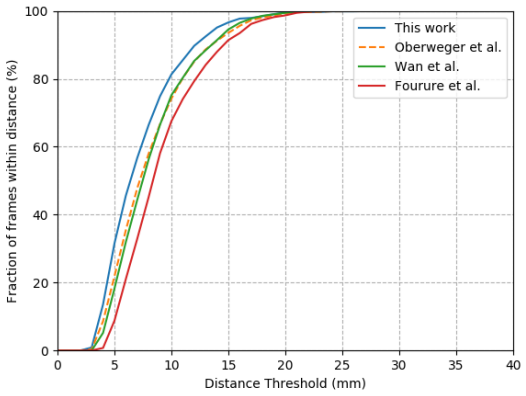


Fig. 4. Comparison with state-of-the-arts on ICVL [2] datasets: percentage of success frames.

*2) NYU Dataset:* The NYU dataset [1] comprises about 72000 training and 8000 test frames of multi-view RGB-D images. For the dataset was recorded using a structured light-based camera, the depth images have a series of missing values and noisy outlines, which makes the dataset very difficult to predict. We only use the single channel depth images captured from a single camera for our experiments. The training set has samples from a single person and the test set samples from two different persons. We follow the proposed evaluation metrics [34], [24] and choose 14 joints for evaluating. As is shown in Table 2 that our method outstands several current state-of-the-art methods.

TABLE II
AVERAGE 3D ERROR ON NYU DATASET [1].

| Method | Average 3D error |
|---|---|
| Oberweger et al. [31] (Feedback) | 16.2mm |
| Deng et al. [25] (Hand3D) | 17.6mm |
| Guo et al. [13] (REN) | 13.4mm |
| Zhou et al. [27] (DeepModel) | 16.9mm |
| Xu et al. [32] (Lie-X) | 14.5mm |
| Neverova et al. [33] | 14.9mm |
| Wan et al. [28] (Crossing Nets) | 15.5mm |
| Fourure et al. [29] (JTSC) | 16.8mm |
| Madadi et al. [21] | 15.6mm |
| **This work (Multiple transfer net)** | **13.1mm** |

In Fig. 5 we compare our method with other discriminative approaches on NYU dataset [1]. Our predictions significantly perform better for the majority of the frames.

### B. Self-Comparison

We implement four baseline for comparison: (a) VGG-transfer has the same convolution structure with VGG-11 pre-trained network and a fully connected layer created by ourself. (b) Res-transfer has the same convolution structure

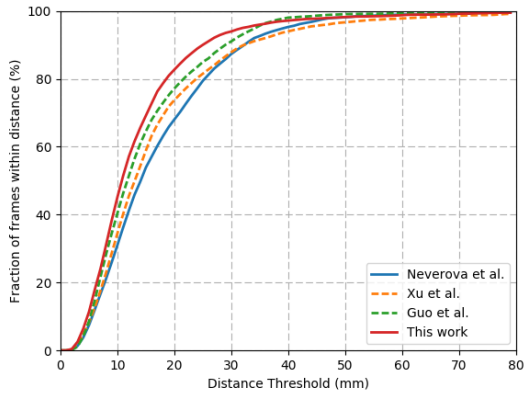


Fig. 5. Comparison with state-of-the-arts on NYU [1] datasets: percentage of success frames.

with ResNet-13 pre-trained network and a fully connected layer created by ourself. (c) Multiple-transfer-ensemble net has the same convolution structure with Fig. 6 without region ensemble step. (d) Multiple-transfer-region-ensemble net has the same convolution structure with Fig. 6. As is shown in Fig. 6, the results of VGG-transfer and Res-transfer are close and multiple-transfer-ensemble outperforms both. And Multiple-transfer-region-ensemble net achieves the best result.

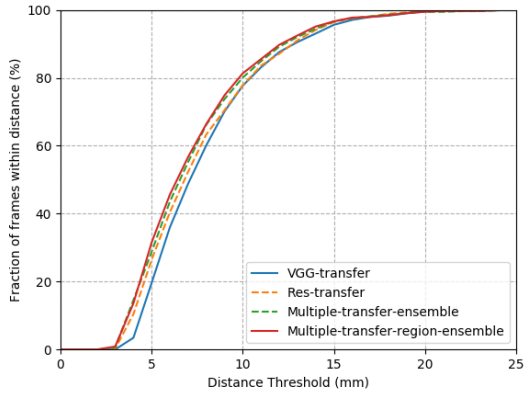


Fig. 6. Self-comparison on ICVL [2] dataset: percentage of success frames.

## IV. Conclusions

In this paper, an accurate multiple transfer region ensemble CNN based method for hand pose estimation is proposed. It provides a new channel extending method to extend hand and hand palm regions and match pre-trained networks' input format to fully utilize the pre-trained networks. And multiple transfer net reaches a better accuracy. Furthermore, a region ensemble for all transfer-net's output feature maps achieves further improvement of prediction. The experimental results demonstrate that proposed method outperforms state-of-art results by a considerable accuracy on NYU [1] and ICVL [2] datasets. Improving real-time performance of the system is our future work.

**Acknowledgment.**This work is supported by the Shenzhen Science and Technology Project (GGFW2017040714161462, JCYJ20170817161916238)

## REFERENCES


[1] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," *ACM Transactions on Graphics(TOG)*, vol. 33, pp. 1935–1946, August 2014.

[2] D. Tang, H. J. Chang, A. Tejani, and T. K. Kim, "Latent regression forest: Structured estimation of 3d articulated hand posture," in *Conference on Computer Vision and Pattern Recognition(CVPR)*. IEEE, 2014, pp. 3786–3793.

[3] J. Supancic, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan, "Depth-based hand pose estimation: Data, methods, and challenges," in *International Conference on Computer Vision(ICCV)*. IEEE, 2015.

[4] Y. Zhang, C. Xu, and L. Cheng, "Learning to search on manifolds for 3d pose estimation of articulated objects," in *arXiv preprint arXiv*, 2016.

[5] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun, "Realtime and robust hand tracking from depth," in *Computer Vision and Pattern Recognition(CVPR)*. IEEE, 2014, pp. 1106–1113.

[6] A. Makris, N. Kyriazis, and A. A. Argyros, "Hierarchical particle filtering for 3d hand tracking," in *Computer Vision and Pattern Recognition Workshops(CVPRW)*. IEEE, 2015, pp. 8–17.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Regionbased convolutional networks for accurate object detection and segmentation," in *IEEE transactions on pattern analysis and machine intelligence*, 2016, vol. 38(1), p. 142158.

[8] L. Chen, G. Papandreou, I. Kokkinos, and A. Murphy, K. andYuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," in *arXiv preprint arXiv*, 2016.

[9] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, p. 10971105.

[10] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Computer Vision and Pattern Recognition(CVPR)*. IEEE, 2016, p. 47334742.

[11] A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung, and L. Fei-Fei, "Towards viewpoint invariant 3d human pose estimation," in *European Conference on Computer Vision(ECCV)*. Springer, 2016, p. 160177.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European Conference on Computer Vision(ECCV)*. Springer, 2014, p. 346361.

[13] G. Hengkai, W. Guijin, C. Xinghao, Z. Cairong, Q. Fei, and Y. Huazhong, "Region ensemble network: Improving convolutional network for hand pose estimation," in *International Conference on Image Processing(ICIP)*. IEEE, 2017.

[14] C. Xinghao, W. Guijin, G. Hengkai, and Z. Cairong, "Pose guided structured region ensemble network for cascaded hand pose estimation," in *arXiv preprint arXiv*, 2017.

[15] Y. Bar, I. Diamant, H. Greenspan, and L. Wolf, "chest pathology detection using deep learning with non-medical training," in *Biomedical Imaging (ISBI)*. IEEE, 2015, vol. 13.

[16] O. Maxime, B. Leon, L. Ivan, and S. Josef, "Learning and transferring mid-level image representations using convolutional neural networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1717–1724.

[17] K. Andrej, T. George, S. Sanketh, L. Thomas, S. Rahul, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1725–1732.

[18] L. Alexandre, "Networks with transfer learning between input channels," in *Intelligent Autonomous Systems*. Springer, 2015, vol. 13, pp. 889–898.

[19] P. Li, H. Ling, X. Li, and C. Liao, "3d hand pose estimation using randomized decision forest with segmentation index points," in *International Conference on Computer Vision(ICCV)*. IEEE, 2015, pp. 819–827.

[20] J. Taylor, J. Shotton, T. Sharp, and A. Fitzgibbon, "The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation," in *Computer Vision and Pattern Recognition(CVPR)*. IEEE, 2012.

[21] M. Madadi, S. Escalera, X. Baro, and J. Gonzalez, "End-to-end global to local cnn learning for hand pose recovery in depth data," in *arXiv Preprint*, 2017.

[22] O. Markus and L. Vincent, "Deepprior++: Improving fast and accurate 3d hand pose estimation," in *International Conference on Computer Vision(ICCV) Workshops*. IEEE, 2017.

[23] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun, "Cascaded hand pose regression," in *Conference on Computer Vision and Pattern Recognition(CVPR)*. IEEE, 2015, pp. 824–832.

[24] M. Oberweger, P. Wohlhart, and V. Lepetit, "Hands deep in deep learning for hand pose estimation," in *Computer Vision Winter Workshop(CVWW)*, 2015, pp. 21–30.

[25] X. Deng, S. Yang, Y. Zhang, P. Tan, L. Chang, and H. Wang, "Hand3d: Hand pose estimation using 3d neural network," in *arXiv Preprint*, 2017.

[26] C. Wan, A. Yao, and L. Van Gool, "Hand pose estimation from local surface normals," in *European Conference on Computer Vision(ECCV)*, 2016.

[27] X. Zhou, Q. Wan, W. Zhang, X. Xue, and Y. Wei, "Model-based deep hand pose estimation, international," in *International Joint Conference on Artificial Intelligence(IJCAI)*, 2016.

[28] C. Wan, T. Probst, L. Van Gool, and A. Yao, "Crossing nets: Dual generative models with a shared latent space for hand pose. estimation," in *Conference on Computer Vision and Pattern Recognition(CVPR)*. IEEE, 2017.

[29] D. Fourure, R. Emonet, E. Fromont, D. Muselet, N. Neverova, A. Tremeau, and C. Wolf, "Multi-task, multi-domain learning: Application to semantic segmentation and pose regression. neurocomputing," in *Neurocomputing*, 2017, vol. 1(251), p. 6880.

[30] P. Krejov, A. Gilbert, and R. Bowden, "Guided optimisation through classification and regression for hand pose estimation," in *Computer Vision and Image Understanding*, 2016, vol. 155(2), p. 124138.

[31] M. Oberweger, P. Wohlhart, and V. Lepetit, "Training a feedback loop for hand pose estimation," in *International Conference on Computer Vision(ICCV)*. IEEE, 2015.

[32] C. Xu, L. Govindarajan, Y. Zhang, and L. Cheng, "Lie-x: Depth image based articulated object pose estimation, tracking, and action recognition on lie groups," in *International Journal of Computer Vision(IJCV)*, 2016.

[33] N. Neverova, C. Wolf, F. Nebout, and G. Taylor, "Hand pose estimation through semi-supervised and weakly- supervised learning," in *arXiv Preprint*, 2015.

[34] J. Tompson, M. Stein, Y. Lecun, and K. Perlin, "Real-time continuous pose recovery of human hands using convolutional networks," in *ACM Transactions on Graphics*, 2014, vol. 33, p. 169.