

Deformable Pose Network: A Multi-Stage Deformable Convolutional Network for 2D Hand Pose Estimation

Sartaj Ahmed Salman¹^a, Ali Zakir¹^b and Hiroki Takahashi^{1,2}

¹Department of Informatics, Graduate School of Informatics and Engineering, The University of Electro-Communications, Tokyo, Japan

²Artificial Intelligence Exploration/Meta-Networking Research Center, The University of Electro-Communications, Tokyo, Japan

Keywords: Deformable Convolution, Multi-Stage DC, EfficientNet, 2D HPE.


Abstract: Hand pose estimation undergoes a significant advancement with the evolution of Convolutional Neural Networks (CNNs) in the field of computer vision. However, existing CNNs fail in many scenarios in learning the unknown transformations and geometrical constraints along with the other existing challenges for accurate estimation of hand keypoints. To tackle these issues we proposed a multi-stage deformable convolutional network for accurate 2D hand pose estimation from monocular RGB images while considering the computational complexity. We utilized EfficientNet as a backbone due to its powerful feature extraction capability, and deformable convolution to learn about the geometrical constraints. Our proposed model called Deformable Pose Network (DPN) outperforms in predicting the 2D keypoints in complex scenarios. Our analysis on the Panoptic studio hand dataset shows that our proposed model improves the accuracy by 2.36% and 7.29% as compared to existing methods i.e., OCPM and CPM respectively.


1 INTRODUCTION

Convolutional Neural Networks (CNNs) have undergone considerable advancements and achieved substantial success in several applications such as visual recognition tasks such as pose estimation (Salman et al., 2023c; Salman et al., 2023b; Simon et al., 2017a; Kong et al., 2020; Zakir et al., 2024), object detection (Girshick et al., 2014) semantic segmentation (Long et al., 2015), and image classification (Krizhevsky et al., 2017). Their capability of modeling geometric transformation comes from extensive data augmentation, the large model capacity, and some hand-crafted modules (e.g., max pooling (Boureau et al., 2010)). Despite the merits, CNNs underperform in terms of modeling geometric transformations in object pose, viewpoint, scale, and part deformation. First, they are assumed to be known and fixed the data augmentation, features, and algorithms were designed on these assumptions which prevent generalization of a new task processing the unknown geometric transformation, which is not properly modeled. Second, even when the transformations are un-

known, hand-crafted designs of invariant features and algorithms are not feasible and it's difficult to overlay these transformations.

However, CNNs are limited to unknown transformations and large models and the origination of these limitations is from the fixed geometric structures of the CNN modules. Specifically, the convolution unit samples information from distinct points in the input feature maps, while reducing spatial resolution by a fixed ratio using pooling layers. Similarly, a RoI (region-of-interest) pooling layer segments a RoI into a set of spatial bins. There the model fails to handle the geometric transformations causing a noticeable problem i.e., the field sizes of the activation units of the same CNN layers are the same, which is quite undesirable for the high-level layers that encode the semantic over spatial locations. These different locations may correspond to the object with different scales or deformations, for visual recognition with fine localization adaptive determination of scales is favorable (Long et al., 2015). In object detection (Girshick et al., 2014), they rely on the features that are extracted based on the primitive bounding boxes, and in pose estimation (Zakir et al., 2024; Zakir et al., 2023) the geometric constraints of the keypoints.

^a <https://orcid.org/0000-0001-9344-6658>

^b <https://orcid.org/0000-0002-3187-9551>

In pose estimation, Hand Pose Estimation (HPE) is one of the prominent areas of CV with several real-world applications such as Virtual/Augmented Reality (VR/AR), sign language recognition, remote surgery, and so on. In addition to the aforementioned challenges of CNNs, HPE poses some new challenges such as self/object occlusion, size variability, high dexterity, and depth ambiguity. As a result, researchers turned their attention to resolving the above-mentioned issues, the model complexity in 2D HPE is also one of the issues causing trouble in making it more applicable in the real world. Despite these, numerous HPE approaches were proposed, including 2D and 3D HPE based on RGB (Wang et al., 2018; Chen et al., 2020; Pan et al., 2022), video (Khaleghi et al., 2022; Ren et al., 2022), and depth (Ren et al., 2022; Cheng et al., 2021) but still struggling to overcome these issues.

In this research, we proposed a multi-stage deformable convolution network named Deformable Pose Network (DPN) for 2D HPE keeping in mind the above challenges, the deformable convolution (Dai et al., 2017; Chen et al., 2021) especially focuses on incorporating the geometrical constraints into the convolutional operation and the backbone deals with the hidden information overcoming the other issues. This approach consists of two modules one is the backbone and the other is the Deformable Convolution Block (DCB), we utilized the EfficientNet (EN) B0 as a backbone for feature extraction, to strike the balance between the computational cost and the model efficiency. As a DCB, we used the concept of Convolutional Pose Machine (CPM) (Wei et al., 2016) that utilizes a six-stage Convolutional Block (CB) for information processing, instead of the CB to deal with the geometrical constraints we replaced the six-stage CB with a four-stage DCB. These changes make our proposed model computationally efficient and enhance the model's capability to learn the unknown hidden information including the geometrical constraints, resulting in accurate 2D HPE.

The proposed approach is summarized below:

- We utilized the customized EfficientNet B0 version as a backbone by removing the fully connected layer for feature extraction, which is one of the best models striking the balance between computation efficiency and accuracy.
- The multi-stage deformable convolution network deals with the geometrical constraints and helps the model to be more generalized to learn the geometrical transformations.

The article consists of the following sections, Section 2 includes the related work on 2D HPE, the detailed network flow is explained in Section 3, exper-

imental setups are explained in Section 4, Section 5 presents the experimental results and analysis, and the conclusion and the future work are summarized in Section 6.

2 RELATED WORK

Hand Pose Estimation (HPE) is a CV task that involves localizing and identifying the hand keypoints (joints) of a hand in a video or an image. As CNNs (Schnürer et al., 2019; Charco et al., 2022) play a crucial role in CV, researchers have actively proposed different approaches to tackle the challenges in HPE, to address the problem of self/object occlusion multi-view RGB models (Simon et al., 2017a; Joo et al., 2015; Panteleris and Argyros, 2017) were proposed, but still constrained with a requirement of specific camera setups. On the other hand, depth-based pose estimation models (Schnürer et al., 2019; Cheng et al., 2021) achieve better accuracy based on depth values, resulting in a fast process. However, these models can be sensitive to the environment (i.e., noise, lightning conditions, and so on). Widespread adoption of RGB cameras in recent years for HPE tasks due to their affordability, anti-inference capabilities, and portability many approaches were proposed based on CNNs using RGB images. CPM (Wei et al., 2016), enforces CNNs to generate heatmaps indicating the location of each keypoints. Although CNNs tackle some of the key challenges but still struggle to deal with the geometrical constraints, self/object occlusion, and high-dexterity, to resolve these we utilized the idea of deformable convolutional (Chen et al., 2021) in our network to make it more generalized.

In recent days, researchers tried to reduce the computational complexity of 2D HPE models, while striking the balance between accuracy and computational cost (Salman et al., 2023a). CPM (Wei et al., 2016) was one of the state-of-the-art lightweight base models a few years back. Yifei Chen et al. (Chen et al., 2020) proposed an architecture based on cascade structure regularization, consisting of two lightweight modules Limb Deterministic Mask (LDM) and Limb Probabilistic Mask (LPM), and each module can be utilized separately for 2D HPE. Hinqing Yang et al. tried to improve those modules in terms of accuracy and computational efficiency and somehow succeeded in this. In (Pan et al., 2022) Tianhong Pan et al. optimized the CPM reducing the complexity of the models and improving the accuracy. However, the above-mentioned methods are state-of-the-art lightweight models but still not applicable in many cases because of the computational complex-

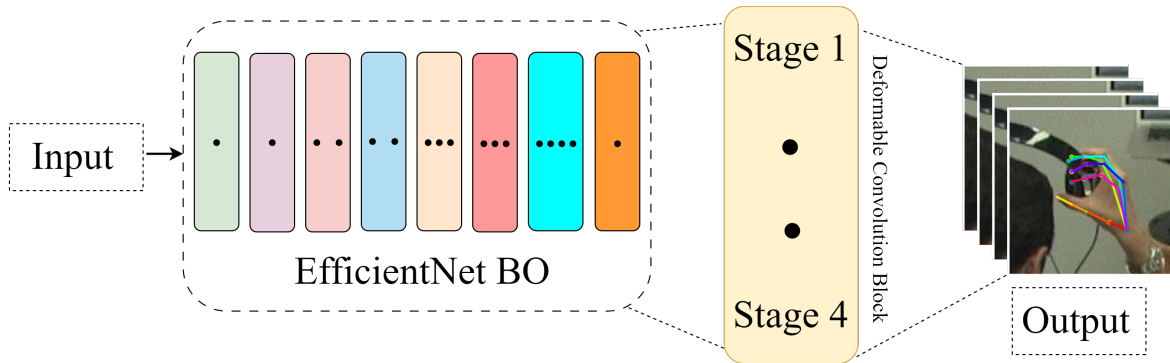


Figure 1: Detailed overview of Deformable Pose Network.

ity and high energy consumption. To tackle this we utilize the stages idea of CPM and reduce the number of stages to reduce the model complexity without affecting the accuracy (balancing the computational cost and accuracy).

3 DEFORMABLE POSE NETWORK

Generally, 2D HPE using heatmaps involves the keypoints detection to get the actual hand pose P , from an RGB image or a video frame I . Consider, K as a set of keypoints, wherein each keypoint k_i represents a distinct region on the hand such as joints or fingertips. These keypoint k_i are symbolized by individual heatmaps H , forming the objective to predict the heatmaps of each keypoint $\{H_1, \dots, H_i\}$. Consequently, the pose $P = \{(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k)\}$ denotes the coordinates with the highest probability in each heatmap. The count of keypoints K varies across the datasets, commonly comprising 21 keypoints. Therefore, for the given input we seek to estimate the pose P , expressed as the set of keypoints as described in Algorithm 1.

We proposed a new approach named Deformable Pose Network (DPN) for efficient and accurate 2D HPE. A multi-stage deformable convolution is utilized in our work inspired by the workflow of CPM stages, combining the power of EN as a backbone for feature extraction. Figure 1 shows the detailed architecture of our proposed method.

3.1 Modified EfficientNet Version B0 for Enhanced Feature Extraction

EN, a state-of-the-art network is utilized as a backbone of our proposed approach for feature extraction. EN is known for its ability to balance the model accu-

Data: RGB image or video frame I

Result: Estimated hand pose P represented as keypoints

Initialize $P = \emptyset$ (Set to store keypoints);

Detect keypoints K representing distinct hand regions in I ;

for $i = 1$ to K **do**

 Generate heatmap H_i for k_i in I ;

 Extract coordinates (x_i, y_i) with highest probability from H_i ;

 Add (x_i, y_i) to P as a keypoint;

end

Return P as the estimated hand pose;

Algorithm 1: 2D Hand Pose Estimation using Heatmaps.

racy and computational cost, based on this there are many versions of EN (B0-B7) each version varies in depth, and B0 the lightest version is utilized in our framework. Figure 2 shows the architecture of the modified EN acting as a backbone in our network. We employed the B0 version of EN to reduce the complexity in comparison with its variants and the other feature extraction networks (i.e. RestNet, VGG, and more). The modified B0 consists of seven blocks, containing varying numbers of MBConvs which are the structure of MobileNetv2, it further includes the removal of final fully connected convolution layers, reducing the model’s parameters and enabling it for feature extraction. The input data goes through several layers in a sequential process, the input is subjected to a 3×3 Conv, followed by the MBConvs operations. The final layer of the EN outputs 64 feature maps and passes to the deformable convolution block for further processing as shown in Figure 2.

3.2 Information Processing DCB

The CPM is one of the baseline CNN-based pose estimation models, which deals with the complexities involved in HPE. However, it encounters lim-

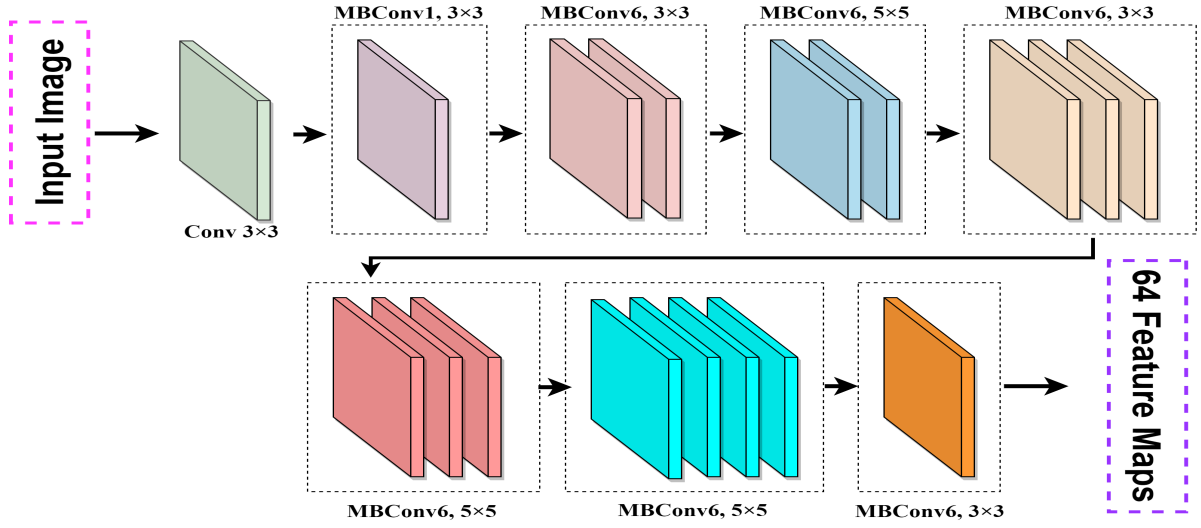


Figure 2: Overall architecture of modified EfficientNet BO.

itations in HPE due to unknown geometrical constraints and other mentioned challenges. To address this issue within the CNN-based models we integrated the DC, which focus on managing geometrical constraints and enhancing the model’s adaptability in learning the unknown features during the information processing. Our proposed approach is a four-stage network, The initial stage consists of two 3×3 DCBs with a channel count of 256. Subsequent stages consist of seven 3×3 DCBs, each with 128 channels. The detailed overview of this information processing DCB is shown in Figure 3.

The output feature maps generated by the backbone are directed to the DCB initial stage of our network for subsequent information processing. Within each stage, the DC, comprising two Convolutional Layers (CL) layers offset CL and a modulator CL, and a DC operation which is discussed in detail below:

3.2.1 Offset CL

It computes spatial offsets through learnable parameterization from the input feature map x which is the output feature map of the backbone, denoted by OF. It can be mathematically represented as Eq 1.

$$OF = OFC(x) \quad (1)$$

Where OFC denotes the convolutional operation on the input x for the computation of the offsets. Which helps to determine the sampling location in x , making it flexible to receptive fields.

3.2.2 Modulator CL

It governs the significance or modulation of sampled regions generating modulation weights by leveraging

the output of a sigmoid function as shown in Eq 2:

$$M = 2 \times \sigma(MC(x)) \quad (2)$$

Here, M represents the modulator, σ , and MC denotes the sigmoid function and the convolutional operation respectively. This factor helps in adaptive feature adjustments according to their importance.

3.2.3 DC Operation

After the first two CLs, DC operations play the role that is the core of DC, integrating the offset and modulator with the regular CL. Mathematically this operation can be expressed as in Eq 3:

$$x = deform2d(x, OF, w, b, M) \quad (3)$$

Here, x denotes the input feature map, OF signifies the spatial offsets, w , b , and M represents the convolutional weights, bias, and the modulating factor respectively. The incorporation dynamically adjusts the receptive fields, enabling the model’s capabilities to learn adaptive features and geometrical constraints, and the final output from the initial stage progresses to the second stage.

The sequence iterates across all four stages and in the final stage we got the 21 final keypoints. Along with the DC, we reduced the number of stages and channels in contrast to CPM, enhancing the overall adaptability and computational efficiency of our model.

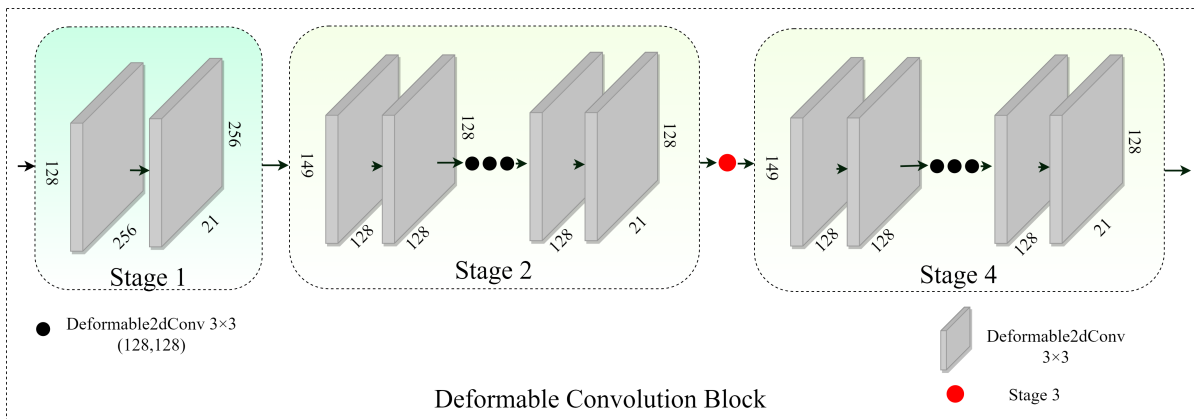


Figure 3: Detailed overview of stages of deformable convolution block.

4 EXPERIMENTAL SETUP

4.1 Dataset

In our research, we utilized a publicly available dataset The Carnegie Mellon University Panoptic Hand Dataset (CMU) (Simon et al., 2017b) from Panoptic Studio to evaluate our proposed model. The dataset includes 14,817 annotations of the right hand of individuals captured at the studio, the distribution is shown in Table 1. As our research is HPE to achieve this objective the annotated image patches were extracted from the full image using a box size of 2.2 times larger than the hand. The dataset is randomly divided into three subgroups by a random sampling technique for training, validation, and testing comprised of 80%, 10%, and 10% of the dataset respectively.

Table 1: CMU panoptic hand dataset distribution.

Dataset	Training	Validation	Test
CMU Panoptic	11,853	1482	1482

4.2 Implementation Details

We implemented our model using the PyTorch framework, with a batch size of 64 and a learning rate of 0.0001. The model is trained up to 100 epochs. The input images were scaled to $[0, 1]$ and normalized using a mean and standard deviation of $(0.485, 0.456, 0.406)$ and $(0.229, 0.224, 0.225)$ respectively. The Mean Squared Error (MSE) is utilized as a loss function. To prevent the loss from decreasing to an extremely low value, the loss function is adjusted using a scaling factor of 35.

4.3 Activation Function and Model Optimizer

To incorporate nonlinear aspects into the network, various activation functions were proposed such as ReLU (Banerjee et al., 2019), Softmax (Sharma et al., 2017), and, Mish (Misra, 2020). However, Mish outperforms others notably, due to its nonlinear nature, its mathematical representation is as follows:

$$f(x) = x \tanh(\ln(1 + e^x)) \quad (4)$$

Experimental results highlight Mish’s superior efficiency over other activation functions.

The model optimizers aim to decrease the loss function and enhance network performance by finding the best parameter values. We adopted a newly derived version of the Adam optimizer called AdamW can significantly bolster model optimization techniques. In contrast to the Adam optimizer, the AdamW algorithm separates the weight decay component from the learning rate, enabling individualized optimization of each component. This feature effectively addresses the issue of excessive overfitting. The results indicate that the model optimized with AdamW demonstrates better generalization performance. The AdamW optimizer was employed in the training of our proposed approach.

4.4 Evaluation Metric

As an evaluation metric commonly used for pose estimation Percentage of Correct Keypoints (PCK) was utilized in this study. It measures the probability that the predicted keypoints fall in a specified threshold distance, represented as σ from the ground truth. σ was uniformly distributed in a range of 0.04 to 0.12,

Table 2: Numerical comparison of DPM with other models on CMU panoptic hand dataset.

Threshold σ	0.04	0.06	0.08	0.10	0.12	<i>Average</i>	<i>Improvement</i>
CPM(Wei et al., 2016)	56.76	74.66	82.50	86.67	89.45	78.01	–
LDM-6(Chen et al., 2020)	59.51	76.19	83.77	87.83	90.27	79.51	1.50
LPM-6(Chen et al., 2020)	60.71	77.60	84.93	88.76	91.10	80.62	2.61
OCPM(Pan et al., 2022)	63.67	80.26	87.10	90.65	93.01	82.94	4.93
DPN	67.19	82.81	89.27	92.63	94.48	85.30	7.29

it is formulated as:

$$PCK_{\sigma}^k = \frac{1}{|D|} = \sum_D 1 \left(\frac{\|p_k^{pt} - p_k^{gd}\|_2}{\max(w, h)} \leq \sigma \right) \quad (5)$$

Here p_k^{gd} represents the keypoints ground truth, 1 is the indicator function, p_k^{pt} denotes the predicted keypoints, k for the number of keypoints, D refers to the number of test or validation sample, and w and h indicates the height and width of the input image respectively.

5 RESULTS AND ANALYSIS

In this section, we discuss the performance analysis of our proposed network and compare it with various HPE methodologies.

5.1 Quantitative Results

The results presented in Table 2 are the quantitative analysis of our proposed model numerically and graphically. The results indicate that our proposed model achieves an improvement of 5.13 % at σ 0.12 and an average improvement of 7.29 % in comparison to CPM (Wei et al., 2016). Against OCPM (Pan et al., 2022) it achieves 1.57 % at σ 0.12 and 2.36 % on an average. Figure 4 depicts a PCK comparison of DPM with CPM, LDM-6, LPM-6, and OCPM, demonstrating its superior performance over existing lightweight methods.

To compare the computational complexity we did a parameter comparison, excluding LDM-6 and LPM-6 due to the absence of parameters. Table 3 indicates that our proposed architecture has fewer parameters in comparison with CPM and OCPM, signifying the reduction of the computational complexity of our methodology.

Table 3: Parameters comparison.

Model	<i>Parameters(M)</i>	<i>Flops(G)</i>
CPM(Wei et al., 2016)	36.80	103.23
OCPM(Pan et al., 2022)	29.28	80.53
DPN	8.55	16.38

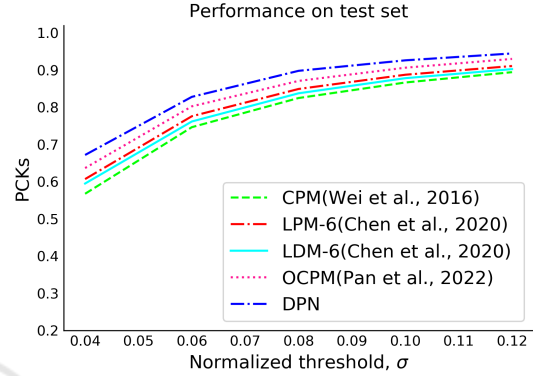


Figure 4: PCK comparison with other lightweight 2D HPE models.

5.2 Qualitative Results

To evaluate the effectiveness of DPN visually, we randomly select the images from the test set as input for the visualization. Figure 5 illustrates that our proposed network shows effective results the model’s efficiency on low light and blurred images is noteworthy. The findings suggested that our proposed DPM tends to perform better than the other lightweight state-of-the-art models.

5.3 Ablation Study

To demonstrate the effectiveness of DC in the stages we perform an ablation study by training the network without DC, the results reveal that DC performs better in comparison with the convolution even the model we trained without is a six-stage network with more parameters. The numerical results in Table 4 show the incorporation of DC promisingly improves the network performance in terms of accuracy.

Table 4: Comparison of six-stages without DC and four-stage with DC.

Threshold σ	0.04	0.06	0.08	0.10	0.12	<i>Average</i>
Six-stage without DC	62.09	78.82	85.64	90.27	92.42	81.85
Four-stage with DC	67.19	82.81	89.27	92.63	94.48	85.30

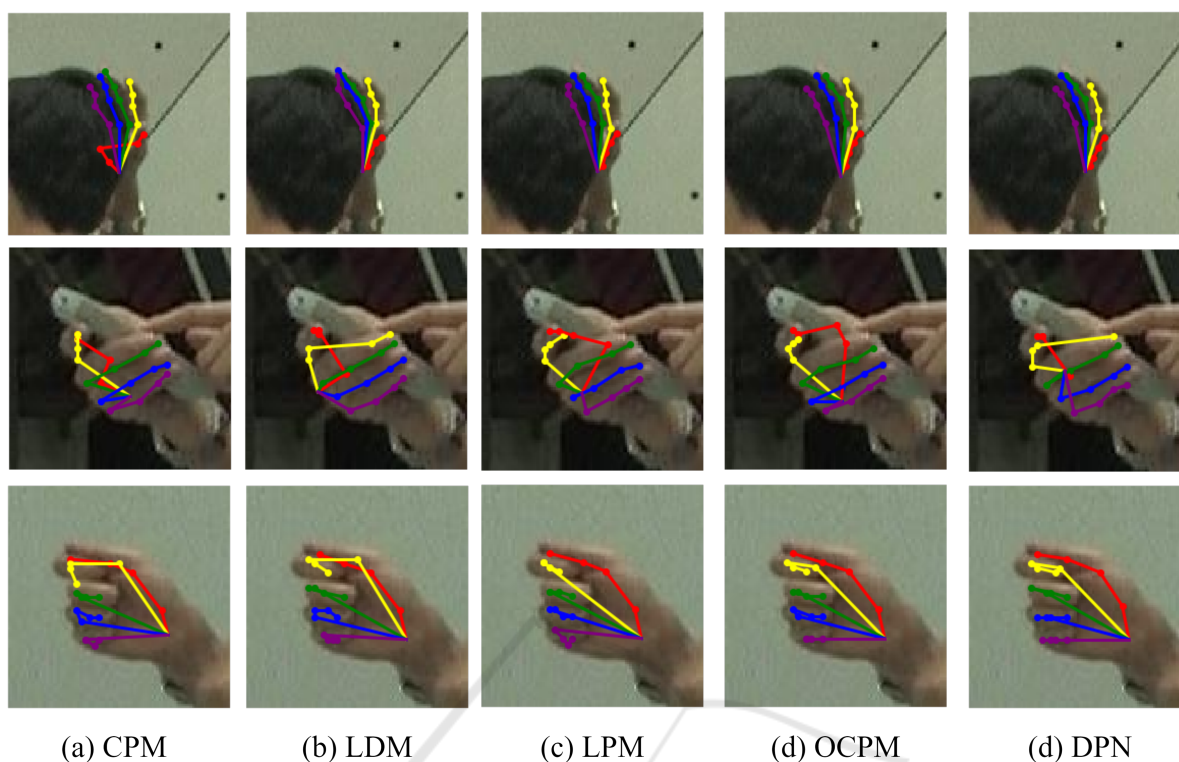


Figure 5: Visual illustration of predicted hand keypoints.

6 CONCLUSIONS

In this paper, we proposed a lightweight multi-stage deformable convolutional network for 2D hand pose estimation. To learn the hidden information EfficientNet was used as a backbone for enhanced feature extraction. To deal with the geometrical constraints we utilized deformable convolution in each stage instead of traditional convolutions. Evaluation on a publicly available CMU hand dataset, our proposed approach outperformed the state-of-the-art networks in terms of accuracy and computational complexity. With the potential of real-world application of hand pose estimation in AR, VR, HCI and so on we will extend our work to 3D HPE.

REFERENCES

- Banerjee, C., Mukherjee, T., and Pasiliao Jr, E. (2019). An empirical study on generalizations of the relu activation function. In *Proceedings of the 2019 ACM Southeast Conference*, pages 164–167.
- Boureau, Y.-L., Ponce, J., and LeCun, Y. (2010). A theoretical analysis of feature pooling in visual recognition. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 111–118.
- Charco, J. L., Sappa, A. D., and Vintimilla, B. X. (2022). Human pose estimation through a novel multi-view scheme. In *VISIGRAPP (5: VISAPP)*, pages 855–862.
- Chen, F., Wu, F., Xu, J., Gao, G., Ge, Q., and Jing, X.-Y. (2021). Adaptive deformable convolutional network. *Neurocomputing*, 453:853–864.
- Chen, Y., Ma, H., Kong, D., Yan, X., Wu, J., Fan, W., and Xie, X. (2020). Nonparametric structure regularization machine for 2D hand pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 381–390.
- Cheng, W., Park, J. H., and Ko, J. H. (2021). Handfold-net: A 3d hand pose estimation network using multiscale-feature guided folding of a 2d hand skeleton. In *Proceedings of the IEEE/CVF Conference on Computer Vision*, pages 11260–11269.
- Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., and Wei, Y. (2017). Deformable convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision*, pages 764–773.
- Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587.
- Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., and Sheikh, Y. (2015). Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE*

- International Conference on Computer Vision*, pages 3334–3342.
- Khaleghi, L., Sepas-Moghaddam, A., Marshall, J., and Etemad, A. (2022). Multi-view video-based 3d hand pose estimation. *IEEE Transactions on Artificial Intelligence*.
- Kong, D., Ma, H., Chen, Y., and Xie, X. (2020). Rotation-invariant mixed graphical model network for 2D hand pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1546–1555.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.
- Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440.
- Misra, D. (2020). Mish: A self regularized non-monotonic activation function. *BMVC*.
- Pan, T., Wang, Z., and Fan, Y. (2022). Optimized convolutional pose machine for 2D hand pose estimation. *Journal of Visual Communication and Image Representation*, 83:103461.
- Panteleris, P. and Argyros, A. (2017). Back to rgb: 3d tracking of hands and hand-object interactions based on short-baseline stereo. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 575–584.
- Ren, P., Sun, H., Hao, J., Wang, J., Qi, Q., and Liao, J. (2022). Mining multi-view information: a strong self-supervised framework for depth-based 3d hand pose and mesh estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20555–20565.
- Salman, S. A., Zakir, A., Benitez-Garcia, G., and Takahashi, H. (2023a). Acenet: Attention-driven contextual features-enhanced lightweight efficientnet for 2d hand pose estimation. In *2023 38th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6.
- Salman, S. A., Zakir, A., and Takahashi, H. (2023b). Cascaded deep graphical convolutional neural network for 2D hand pose estimation. In *International Workshop on Advanced Imaging Technology (IWAIT) 2023*, volume 12592, pages 227–232. SPIE.
- Salman, S. A., Zakir, A., and Takahashi, H. (2023c). SDF-PoseGraphNet: spatial deep feature pose graph network for 2d hand pose estimation. *Sensors*, 23(22).
- Schnürer, T., Fuchs, S., Eisenbach, M., and Groß, H.-M. (2019). Real-time 3d pose estimation from single depth images. In *VISIGRAPP (5: VISAPP)*, pages 716–724.
- Sharma, S., Sharma, S., and Athaiya, A. (2017). Activation functions in neural networks. *Towards Data Sci*, 6(12):310–316.
- Simon, T., Joo, H., Matthews, I., and Sheikh, Y. (2017a). Hand keypoint detection in single images using multi-view bootstrapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1145–1153.
- Simon, T., Joo, H., and Sheikh, Y. (2017b). Hand keypoint detection in single images using multiview bootstrapping. *CVPR*.
- Wang, Y., Peng, C., and Liu, Y. (2018). Mask-pose cascaded cnn for 2d hand pose estimation from single color image. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(11):3258–3268.
- Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732.
- Zakir, A., Salman, S. A., Benitez-Garcia, G., and Takahashi, H. (2023). Aeca-prnetcc: Adaptive efficient channel attention-based poseresnet for coordinate classification in 2d human pose. In *2023 38th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–6.
- Zakir, A., Salman, S. A., and Takahashi, H. (2024). Sahf-lightposeresnet: Spatially-aware attention-based hierarchical features enabled lightweight poseresnet for 2d human pose estimation. In Park, J. S., Takizawa, H., Shen, H., and Park, J. J., editors, *Parallel and Distributed Computing, Applications and Technologies*, pages 43–54, Singapore. Springer Nature Singapore.