Nonparametric Structure Regularization Machine for 2D Hand Pose Estimation

Yifei Chen[*1], Haoyu Ma[*2], Deying Kong[2], Xiangyi Yan[2], Jianbao Wu[1], Wei Fan[†1], and Xiaohui Xie[†2]

[1]Tencent Hippocrates Research Lab
[2]Department of Computer Science, University of California at Irvine
{dolphinchen,jannwu,davidwfan}@tencent.com, {haoyum3,deyingk,x.yan,xhx}@uci.edu

Abstract

Hand pose estimation is more challenging than body pose estimation due to severe articulation, self-occlusion and high dexterity of the hand. Current approaches often rely on a popular body pose algorithm, such as the Convolutional Pose Machine (CPM), to learn 2D keypoint features. These algorithms cannot adequately address the unique challenges of hand pose estimation, because they are trained solely based on keypoint positions without seeking to explicitly model structural relationship between them. We propose a novel Nonparametric Structure Regularization Machine (NSRM) for 2D hand pose estimation, adopting a cascade multi-task architecture to learn hand structure and keypoint representations jointly. The structure learning is guided by synthetic hand mask representations, which are directly computed from keypoint positions, and is further strengthened by a novel probabilistic representation of hand limbs and an anatomically inspired composition strategy of mask synthesis. We conduct extensive studies on two public datasets - OneHand 10k and CMU Panoptic Hand. Experimental results demonstrate that explicitly enforcing structure learning consistently improves pose estimation accuracy of CPM baseline models, by 1.17% on the first dataset and 4.01% on the second one. The implementation and experiment code is freely available online[1]*. Our proposal of incorporating structural learning to hand pose estimation requires no additional training information, and can be a generic add-on module to other pose estimation models.*

1. Introduction

Hand pose understanding is an important task for many real world AI applications, such as human-computer interaction, augmented reality and virtual reality. However, hand pose estimation remains challenging because the hand is highly articulated and dexterous, and suffers severely from self-occlusion. Recently a significant amount of efforts have been dedicated to improving the accuracy of hand pose estimation from different perspectives, including 1) multi-view RGB systems [23, 11], 2) depth-based solutions [7, 28, 32], and 3) monocular RGB solutions [35, 19, 2]. Although some of these efforts focus on 3D hand pose or shape estimation, 2D hand pose estimation remains an essential component as it often constitutes a sub-module of 3D estimation problems and as such directly impacts the performance of downstream 3D pose or shape estimation.

Meanwhile, human pose estimation has advanced significantly since the advent of *Deep Convolutional Neural Network* (DCNN). Successful DCNN architectures typically have large receptive fields and strong representation power, such as the *Convolutional Pose Machine* (CPM) [30], the *Stacked Hourglass* (SHG) [18], and the *Residual Network* [10]. They are deployed by popular human pose estimation systems [3, 6, 9] to implicitly capture structure information of body parts. They are also utilized by many hand pose estimation algorithms to perform the 2D pose estimation subtask [23, 35, 2, 19, 28]. However, DCNNs only capture structure information implicitly and may not be adequately equipped to capture complex structure relationship between hand keypoints to handle severe articulation and self-occlusion of the hand [12, 14].

Recently, there is a trend to unify pose estimation and instance segmentation in a multi-task learning paradigm, and it is observed that the latter helps to improve the performance of the former [9, 29]. Unfortunately, this direction requires a large amount of manually labelled segmentation masks, which is costly to obtain. Hand mask datasets are even rarer than body mask datasets, making the multi-task approach less applicable to hand pose estimation.

In this paper, we propose the Nonparametric Structure Regularization Machine (NSRM) for 2D hand pose esti-

[*] These authors contribute equally to this work.
[†] These authors are co-corresponding authors of this work.
[1] https://github.com/HowieMa/NSRMhand

mation from a monocular RGB image. NSRM incorporates a nonparametric structure model and a pose model in a cascade multi-task framework. The structure learning is supervised by synthetic hand mask representations directly computed from keypoint positions, and is strengthened by a probabilistic representation of hand limbs and an anatomically inspired composition strategy of mask synthesis. The pose model utilizes the composite structure representation to learn robust hand pose representation. We comprehensively evaluate the performance of NSRM on two public datasets, *i.e.*, OneHand 10k [29], and the more challenging CMU Panoptic Hand [11]. Quantitative results demonstrate that NSRM consistently improves the prediction accuracy of the CPM baseline, by *1.17%* on the first dataset and *4.01%* on the second one, and that NSRM renders competitive performance compared to utilizing manually-labeled masks. Qualitative results show that NSRM effectively reinforces structure consistency to predicted hand pose especially when severe occlusion exists, and the learned structure representations highly resemble real segmentation masks.

The main contributions of this paper are as follows:

- We propose a novel cascade structure regularization methodology for 2D hand pose estimation, which utilizes synthetic hand masks to guide keypoints structure learning. The synthetic masks are derived directly from keypoint positions requiring no extra data annotations, making the method applicable to any existing pose estimation model.

- We propose a novel probabilistic representation of hand limbs and an anatomically inspired composition strategy for hand mask synthesis.

- We carry out extensive experiments on two public datasets, and demonstrate that NSRM consistently outperforms baseline models.

2. Related work

2.1. Human pose estimation

DCNN has been massively applied to 2D human pose estimation since the seminal work of *DeepPose* [27]. As the human body naturally manifests an articulated graph structure, researchers have explored the combination of DCNN and graphical models (GM) for pose estimation [26, 4, 31, 25]. However, the GM component often suffers from two practical limitations: 1) the pairwise term typically takes some parametric form, which may not be true in reality; 2) belief propagation inference is performed frequently during training and computational intensive. As a result, mainstream algorithms [3, 6, 9] still rely on delicate DCNN architectures, such as CPM [30], SHG [18], and the Residual Network [10] to implicitly capture structure information, deploying their large receptive fields and strong representation power. To further improve the effect of structure regularization, some approaches attempt to modify the output of DCNN, for example, to introduce extra branches of the *offset field* [20] or the *structural-aware loss* [12]. Unfortunately, they still cannot fully characterize the structure conditions of limbs, especially their poses and interactions, resulting in very limited effect.

Meanwhile, 3D human pose estimation from monocular RGB, a challenger problem due to depth ambiguity, also advances significantly. Some researches explicitly infer 3D coordinates from the 2D pose [1, 15, 22, 21], while others incorporate 2D pose estimation networks into the whole architecture [16, 17, 33]. In both cases, DCNN-based 2D pose estimators are intensively utilized, such as the Mask RCNN [9] and SHG. To enforce structure constraints in 3D, a kinematic layer can be added on top of the network [34]. But this relies on known bone length and may suffer from the optimization difficulty, which limit its practical application.

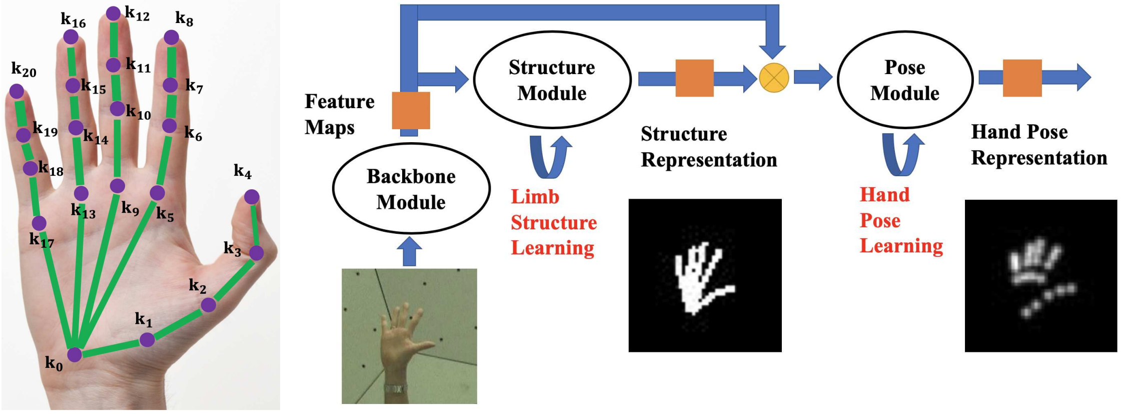
2.2. Hand pose estimation

Hand pose estimation is more difficult than body pose estimation, as the hand is highly articulated, dexterous, and suffers severely from self-occlusion. Although multi-view camera systems can solve the task reliably [23, 11], it is usually very costly to build such a system, involving numerous optical devices and complicated configuration. Therefore, their applications are mostly restricted to the laboratory scenarios. To circumvent this limitation, researchers also devote much effort to depth-based solutions [7, 28, 32]. However, depth devices have limited resolution and range, and are sensitive to lighting conditions [17]. And after all, they are still less ubiquitous than RGB cameras.

Due to the drawbacks of multi-view and depth-based solutions, monocular RGB approaches are drawing much more attention in recent years. Like in the case of 3D human pose estimation, many algorithms adopt a *two-stage* framework, i.e., first performing 2D hand pose estimation and then lifting from 2D to 3D [35, 19, 2]. The 2D subtasks commonly utilize prevalent 2D human pose algorithms, in particular CPM and SHG, which are also frequently used in the multi-view or depth-based solutions [23, 28]. Given the critical role of 2D hand pose algorithms in solving the complete 3D task, we focus on imposing novel structure regularization to 2D hand pose estimation in this paper.

3. The model

Nonparametric Structure Regularization Machine (NSRM) learns the hand's structure and keypoint representations in a cascade multi-task framework. A high-level illustration of the hand representation and our overall architecture is shown in Figure 1. The hand is modeled as



(a) the hand model

(b) the high-level architecture of NSRM

Figure 1: Illustration of the problem setting and our proposed framework. The hand is modeled as 21 keypoints, and 20 limbs that interconnect them anatomically. NSRM features a hierarchical multi-task architecture that learns structure representation and keypoint representation sequentially, which is generic for multi-stage pose estimation models such as CPM [30] and SHG [18]. Structure learning is guided by our novel synthetic hand mask representations (see 3.1, 3.2 for details).

21 keypoints and 20 limbs. The former are interconnected via the latter anatomically. First, the backbone module, usually a DCNN, *e.g.*, the VGG [24], processes an input image and generates a set of feature maps. Second, the structure module takes in image feature maps to learn the mask representation of the hand. Third, the pose module takes in both the mask representation and the feature maps to learn the pose representation of the hand, i.e., *keypoint confidence maps* (KCM). Both the structure and the pose modules are multi-stage convolutional neural networks.

3.1. Limb mask representation

Consider any particular limb L between Keypoint i and j as defined in Figure 1a. We define our basic mask representation as follows,

Limb Deterministic Mask (LDM). Pixels that belong to L are defined as those which fall inside a fixed-width rectangle centering around line segment $\overline{p_i p_j}$, *i.e.*,

$$\begin{cases} 0 \leq (p - p_j)^T (p_i - p_j) \leq \|p_i - p_j\|_2^2, \\ \left|(p - p_j)^T \mathbf{u}^{\perp}\right| \leq \sigma_{LDM} \end{cases} \tag{1}$$

where $\mathbf{u}^{\perp}$ is a unit vector perpendicular to $\overline{p_i p_j}$, and σ_{LDM} is a hyper parameter to control the width of the limb. The ground truth of LDM is a simple 0/1-mask defined outside/inside the rectangle, *i.e.*,

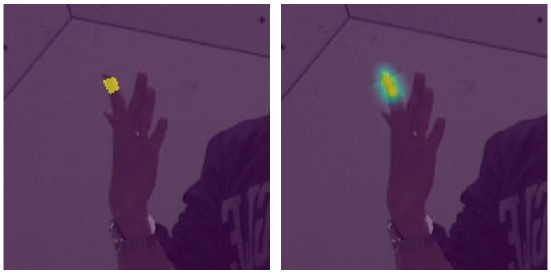
$$S_{LDM}(p|L) = \begin{cases} 1 & \text{if } p \in L \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $p \in I$ is an arbitrary pixel in the image. See Figure 2a for an illustration.

Limb Probabilistic Mask (LPM). LDM assigns 0/1 value to each pixel, depending on its belong to the rectangular mask. This crude treatment may not be optimal in practice. We further propose the novel LPM representation. Each pixel belongs to L with a Gaussian-alike confidence value as defined bellow,

$$S_{LPM}(p|L) = \exp\left(-\frac{\mathcal{D}(p, \overline{p_i p_j})}{2\sigma_{LPM}^2}\right) \tag{3}$$

where $\mathcal{D}(p, \overline{p_i p_j})$ is the distance between the pixel p and the line segment $\overline{p_i p_j}$, and σ_{LPM} is a hyper parameter to control the spread of the Gaussian. See Figure 2b for an illustration. LPM is a smoothed expansion of LDM.



(a) LDM

(b) LPM

Figure 2: Two limb mask representations (taking the index finger tip as an example). LDM: Limb Deterministic Mask (Equation 2); LPM, Limb Probabilistic Mask (Equation 3).

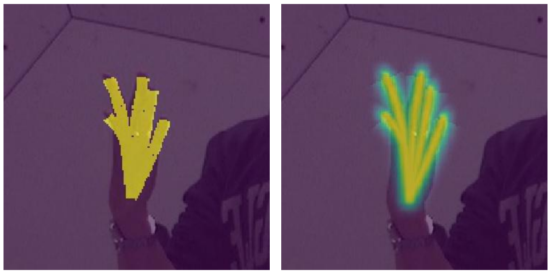
3.2. Limb composition

Given mask representations of single limbs, we further coalesce them into anatomically legitimate groups. Our basic strategy is to coalesce all the 20 limbs together, which renders one single mask representing the whole hand (denoted as *G1*). Alternatively, we also consider coalescing limbs separately into six groups, one representing each finger and one representing the palm (denoted as *G6*). G1 captures the overall structure of the hand, while G6 concerns more about detailed structure regarding local areas of the hand. See Figure 3 and 4 for an illustration. Formally, consider any particular limb group g containing $|g|$ limbs, $\{L_1, L_2, ..., L_{|g|}\}$. Using limb composition, the coalesced mask is defined as,

$$S^*(p|g) = \max\left(S(p|L_1), S(p|L_2), ..., S(p|L_{|g|})\right) \quad (4)$$

where $S(p|L)$ is computed using the basic representation of S_{LDM} (Equation 2) or S_{LPM} (Equation 3).

In practice, we mainly focus on utilizing G1 and *G1&6* (the combination of G1 and G6). We note that although G1 resembles the hand segmentation mask, it is much more efficient because it is readily obtained from keypoins without the extra work of mask annotation. In Section 4, we will compare the performance of utilizing our LDM/LPM representations against the real segmentation mask.



(a) LDM-G1

(b) LPM-G1

Figure 3: The G1 composition strategy: coalescing all the 20 limbs together to get one single mask representing the whole hand.

3.3. Loss function and training

Intermediate supervision is applied to each stage of the structure module and the pose module. Following the common practice of instance segmentation [9], we apply cross-entropy loss to the output of our structure module, *i.e.*,

$$\mathcal{L}_S = \sum_{t=1}^{T_S} \sum_{g \in G} \sum_{p \in I} S^*(p|g) \log \hat{S}_t(p|g) \quad (5)$$
$$+ (1 - S^*(p|g)) \log \left(1 - \hat{S}_t(p|g)\right)$$

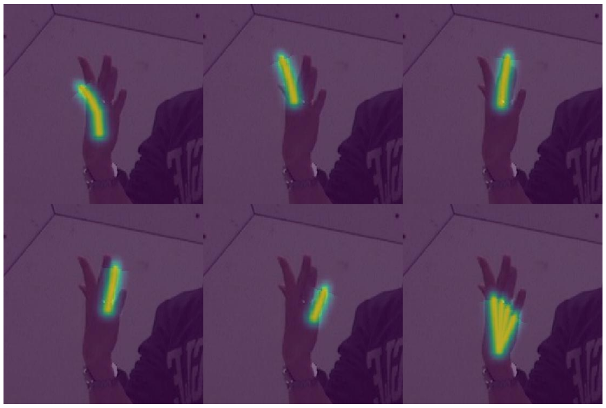


Figure 4: The G6 composition strategy (using LPM representation, *e.g.*, LPM-G6): coalescing the 20 limbs into 6 groups, representing five fingers and the palm separately.

where T_S is the number of stages of structure learning, and $\hat{S}_t(p|g)$ is the prediction of the structure module at Pixel p, Group g of Stage t.

Following the common practice of pose estimation [18, 3], we define the ground truth KCM of Keypoint k as a 2D Gaussian centering around the labelled keypoint with standard deviation σ_{KCM}, *i.e.*,

$$C^*(p|k) = \exp\left\{-\frac{\|p - p_k^*\|_2^2}{2\sigma_{KCM}^2}\right\} \quad (6)$$

We apply the sum-of-squared-error loss to the output of our pose module, *i.e.*,

$$\mathcal{L}_K = \sum_{t=1}^{T_K} \sum_{k=1}^{K} \sum_{p \in I} \left\| C^*(p|k) - \hat{C}_t(p|k) \right\|_2^2 \quad (7)$$

where T_K is the number of stages of pose learning, and $\hat{C}_t(p|k)$ is the prediction of the pose module at Pixel p, Keypoint k of Stage t.

The overall loss function is thus a weighted sum of the structure loss and the pose loss, *i.e.*,

$$\mathcal{L} = \begin{cases} \mathcal{L}_K + \lambda_1 \mathcal{L}_S^{G1}, & \text{for G1} \\ \mathcal{L}_K + \lambda_1 \mathcal{L}_S^{G1} + \lambda_2 \mathcal{L}_S^{G6}, & \text{for G1\&6} \end{cases} \quad (8)$$

where λ_1, λ_2 are hyper-parameters to control the relative weight of the structural regularization. The whole system is trained end-to-end.

4. Experiments

4.1. Datasets

We evaluate NSRM on two public hand pose datasets: the OneHand 10k dataset [29] (OneHand 10k), and the

CMU Panoptic Hand dataset [11] (Panoptic). Their overall statistics are summarized in Table 1. More descriptions are as follows.

OneHand 10k contains 11,703 in-the-wild hand images annotated with both segmentation masks and keypoints. Being collected online, the images often have cluttered background and cover various hand poses. Invisible keypoints are often left unannotated. Ground truth limb representations related to these missing keypoints are set to zero maps. We don't do any hand cropping, as most of the images are occupied by one hand. The dataset is already partitioned into training and testing subsets by Wang *et al*. [29].

Panoptic contains 14,817 images of persons from the Panoptic Studio, each with 21 annotated keypoints of the right hand. Since we focus on hand pose estimation instead of hand detection, we directly crop hands based upon their ground truth keypoints. Specifically, we crop a square patch of size 2.2B, where B is the maximum dimension of the tightest bounding box enclosing all hand keypoints. Then the cropped hand dataset is randomly divided into three subsets for training (80%), validation (10%) and testing (10%).

Table 1: Overall statistics of datasets used in this paper.

dataset	training	validation	testing
OneHand 10k	10,000	-	1,703
Panoptic	11,853	1,482	1,482

4.2. Experimental settings

Implementation details

We implement NSRM in Pytorch. To be compatible with and comparable to the CPM hand pose model [23], we adopt VGG-19 [24] (up to Conv 4_4) as our backbone network, which is pretrained on ImageNet [5] and generates 128-channel feature maps. The following architecture has 6 stages, each of which contains 5 convolution layers with 7x7 kernel and 2 convolution layers with 1x1 kernel (except the first stage). The first 3 stages are allocated to learn composite mask representations, and the last 3 stages are for pose representation learning. All hand image patches are resized to 368×368 before fed into our model, yielding 46×46 representation maps for both LDM/LPM and KCM. The detail network architecture is shown in Figure 5.

Learning configuration

We use Adam [13] to train our model. The initial learning rate is set to 1e-4, and other parameters are set to default values. For G1, we set $\lambda_1 = 1$ for LDM, and $\lambda_1 = 0.5$ for LPM. For G1&6, we set $\lambda_1 = 0.2, \lambda_2 = 0.04$ for LDM, and $\lambda_1 = 0.1, \lambda_2 = 0.02$ for LPM. These configurations empirically make the structure loss and the pose loss on the same scale at the beginning of training. Further more, as structure learning is an auxiliary task and our ultimate goal is hand pose estimation, we propose to utilize a *decayed loss training schedule*. Specifically, we let λ_1 and λ_2 decay by a ratio of 0.1 every 20 epochs, so as to let training focus more on KCM in later iterations.

Evaluation metric

We adopt *Probability of Correct Keypoint within a Normalized Distance Threshold* [23] (shortly referred to as PCK) to perform model selection and evaluation. However, as the hand/head sizes are not explicitly provided by the datasets used in this paper, we resort to normalization with respect to the dimension of the tightest hand bounding box. The normalization threshold σ_{PCK} ranges from 0 to 1.

4.3. Quantitative results

OneHand 10k

As this dataset has a lot of missing values, G6 composition tends to generate incomplete masks. Therefore, we only consider about G1 composition. We retrain the Mask-pose Cascaded CNN [29], which utilize real segmentation masks (denoted as "Real Mask"). We also train the model with our proposed decayed loss training schedule (denoted as "Real Mask ++"). Figure 6a shows the performance comparison. Table 2 summarizes detailed numerical results. Our observations are as follows:

i) NSRM consistently improves the predictive accuracy of CPM, regardless of the choice of basic representation (LDM or LPM), and the value of the evaluation threshold σ_{PCK}. In particular, LPM-G1 achieves *0.0102* absolute improvement in average PCK, corresponding to *1.17%* relative improvement.

ii) LPM-G1 outperforms LDM-G1, and does slightly better than Real Mask. This result demonstrates the effectiveness of our proposed probabilistic mask, comparing to both the manually labeled mask and our proposed deterministic mask.

iii) Our proposed decayed loss training schedule, in combination with utilizing real masks, achieves the best performance (Real Mask ++).

To summarize, our proposed NSRM (along with its learning schedule) is both efficient and effective, as it avoids the overhead of mask labeling but still maintains competitive performance.

Panoptic

Figure 6b shows the performance of NSRM and the CPM baseline. Table 3 summarizes detailed numerical results. Our observations are as follows:

i) Like in the case of OneHand 10k, NSRM consistently outperforms the CPM baseline. In particular, the fully-fledged LPM-G1&6 achieves *0.0309* absolute improvement in average PCK, corresponding to *4.01%* relative improvement. The results suggest a systematic and significant ben-

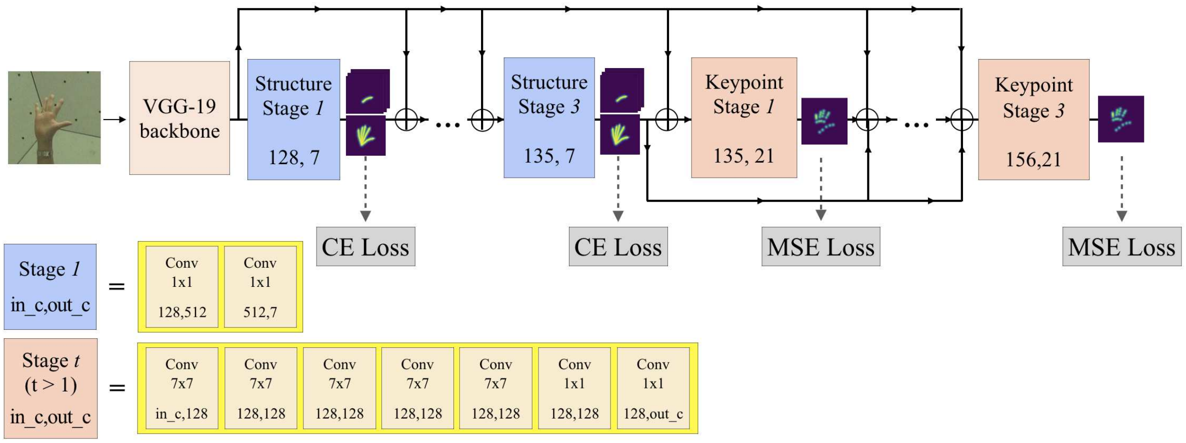


Figure 5: Detailed architecture of NSRM using G1&6 composition implemented based upon CPM. Each stage consists of a series of fully convolutional layers, whose kernel size is denoted as $k \times k$. The number of input feature maps and the number of output feature maps are shown lower insider each rectangle. The output of Structure Stage 3 is fed to each keypoint stage.

Table 2: Detailed numerical results of PCK (in %) evaluated at different thresholds on the OneHand 10k testing data. "ave" means the average PCK, whose absolute and relative improvement are shown in the right most column. The best improvement is highlighted in boldface.

σ_{PCK}	0.1	0.15	0.2	0.25	0.3	ave	improvement
CPM	78.48	84.73	88.54	90.89	92.64	87.06	-
LDM-G1	78.50	85.35	89.31	91.72	93.35	87.64	+0.59 (+0.67%)
LPM-G1	79.32	86.10	89.60	91.91	93.43	88.07	+1.02 (+1.17%)
Real Mask [29]	78.95	85.93	89.78	92.04	93.55	88.05	+0.99 (+1.14%)
Real Mask ++	79.62	86.38	90.05	92.34	93.92	88.46	**+1.41 (+1.62%)**

efit of utilizing our proposed structure regularization for 2D hand pose estimation.

ii) Like in the case of OneHand 10k, LPM consistently outperforms LDM, under both composition strategies (G1 and G1&6). The absolute improvement in average PCK is 0.64 under G1 and 0.71 under G1&6. This phenomenon consolidates that the probabilistic treatment in mask representation indeed benefits the performance of NSRM.

iii) Comparing the proposed composition strategies, we find that G1&6 moderately improves the performance of G1. We interpret this result from an anatomical perspective. As an overall representation, G1 covers important global structure information of the hand. However it cannot fully characterize local details, such as the shape of each finger, which are highly flexible and hard to distinguish due to self-occlusion. G6 is designed to cope with this problem. By combining G1 and G6, NSRM gets a representation that can cover both global and local structure information.

iv) The combination of centralized & distributed composition and the probabilistic representation (LPM-G1&6) renders the optimal performance, which correspond to 0.89 absolute improvement in average PCK, comparing to the basic version (LDM-G1).

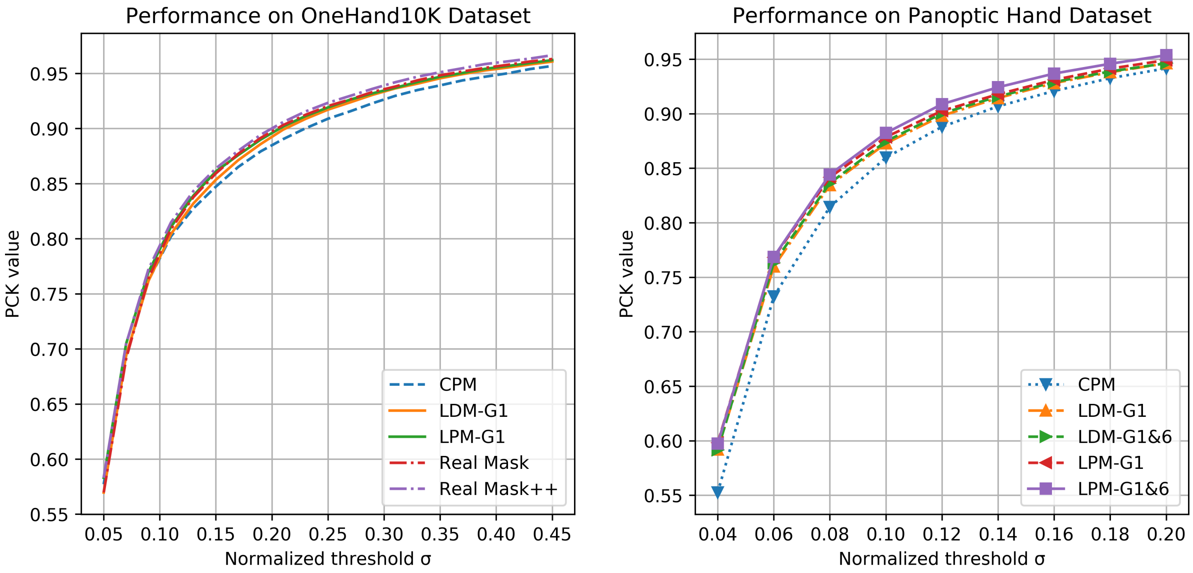
v) Although being consistent on both datasets, the improvement seems much more significant on Panoptic than on OneHand 10k. There are two potential reasons. First, Panoptic is much more challenging, with abundant hand gestures shot from different perspectives and complicated hand-hand & hand-object interactions. Second, OneHand 10K contains a lot of unlabelled keypoints, and therefore tends to make the mask learning signal too fragmented and noisy. Both factors suggest that Panoptic could benefit more from structure learning than OneHand 10k.

4.4. Qualitative results

Figure 7 visualizes the predictive results of CPM and NSRM on sample images from the Panoptic test data. We can clearly see that NSRM effectively reinforces structure

Table 3: Detailed numerical results of PCK (in %) evaluated at different thresholds on the Panoptic testing data. "ave" means the average PCK, whose absolute and relative improvement are shown in the right most column. The best improvement is highlighted in boldface.

σ_{PCK}	0.04	0.06	0.08	0.10	0.12	ave	improvement
CPM	55.25	73.23	81.45	85.97	88.80	76.94	-
LDM-G1	59.20	75.98	83.45	87.28	89.81	79.14	+2.20 (+2.86%)
LDM-G1&6	59.16	76.32	83.63	87.46	90.03	79.32	+2.38 (+3.09%)
LPM-G1	59.81	76.82	84.16	87.86	90.26	79.78	+2.84 (+3.69%)
LPM-G1&6	59.73	76.86	84.43	88.23	90.87	80.03	**+3.09 (+4.01%)**



(a)

(b)

Figure 6: PCK curves on testing data of a) OneHand 10k, and b) Panoptic. Best viewed in color.

consistency and reduces inference ambiguity. Even during severe occlusion, LPM-G1&6 still makes anatomically legitimate prediction while CPM cannot. Moreover, Figure 8 visualizes the learned structure representations on sample images. We can see that they highly resemble hand segmentation masks. This indicates that our NSRM framework could be potentially applied to multi-task learning of both hand pose estimation and instance segmentation.

4.5. Discussion

NSRM learns compositional structure representation of the hand, and utilizes it to regularize the learning process of KCM in a nonparametric fashion. This implicit hierarchical treatment is fundamentally different to graphical-model-involved approaches [4, 31, 25, 14], and those which introduce simple keypoint-induced output/loss [20, 12] or kinematic constraints [34].

Our basic LDM representation takes a rectangular shape, similar to the Part Affinity Field (PAF) [3]. However, NSRM goes significantly beyond PAF in three core aspects. First, our segmentation-inspired representation and its probabilistic expansion are completely different from the vector field representation of PAF. Second, we propose structure composition to coalescing limbs into anatomically inspired groups, a key feature not considered by PAF. Last but not least, PAF is an auxiliary representation proposed for differentiating multiple human instances, not intended for hand structure regularization as in our case.

Previous researches have explored simultaneous pose estimation and instance segmentation [9, 29], but all require mask annotation. Our structure representation is automatically constructed from keypoints, avoiding laborious anno-

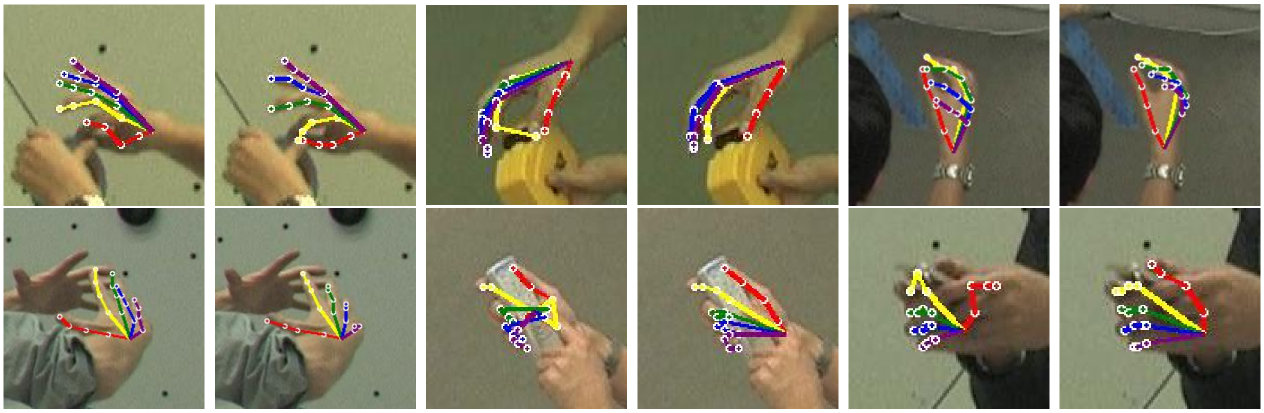


Figure 7: Visualization of predicted hand pose of samples from the Panoptic test data. For each pair of images, the left shows CPM's prediction and the right shows LPM-G1&6's prediction. Best viewed in color.

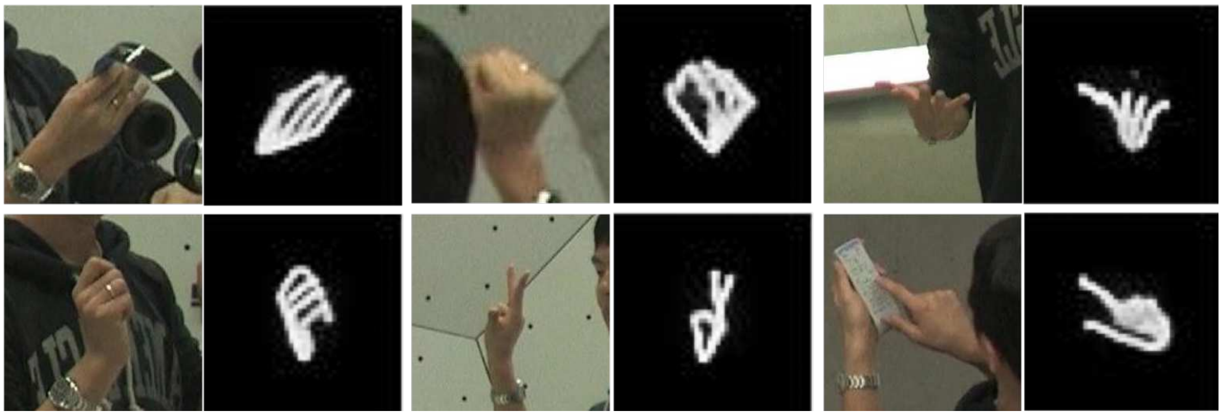


Figure 8: Visualization of the learned structure representation of samples from the Panoptic test data. For each pair of images, the left shows the original image and the right shows the learned mask representation of LPM-G1, *i.e.*, the output of Structure Stage 3 (see Figure 5 for an illustration).

tation. Our experiments demonstrate that NSRM achieves comparable pose estimation performance to models trained with real masks [29]. More over, our learned structure representations highly resemble real masks, which indicates potential applications to hand instance segmentation.

5. Conclusion

In this paper, we have proposed a novel Nonparametric Structure Regularization Machine for 2D hand pose estimation. NSRM is a cascade architecture of structure learning and pose learning. The structure learning is guided by self-organized hand mask representations, and strengthened by a novel probabilistic representation of hand limbs and an anatomically inspired composition strategy of mask synthesis. The pose module utilizes the structure representation to learn robust hand pose representation. We comprehensively evaluate NSRM on two public datasets. Experimental results demonstrate that, 1) NSRM consistently improves the prediction accuracy of the baseline model; 2) NSRM renders comparable performance to utilizing manually labeled masks; 3) NSRM effectively reinforces structure consistency to predicted hand poses, especially during severe occlusion. We should note that although we used CPM as our baseline pose estimation model, the proposed method is generic, independent of any particular choice of the baseline. The structure learning module can be readily added to other prevalent models [18, 8] to improve pose estimation performance, which we intend to explore in future work.

References

[1] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016.

[2] Y. Cai, L. Ge, J. Cai, and J. Yuan. Weakly-supervised 3D hand pose estimation from monocular RGB images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 666–682, 2018.

[3] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. *arXiv preprint arXiv:1812.08008*, 2018.

[4] X. Chen and A. L. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems*, pages 1736–1744, 2014.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[6] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu. RMPE: regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017.

[7] L. Ge, Y. Cai, J. Weng, and J. Yuan. Hand PointNet: 3D hand pose estimation using point sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8417–8426, 2018.

[8] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran. Detect-and-track: Efficient pose estimation in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 350–359, 2018.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, I. Matthews, et al. Panoptic Studio: A massively multiview system for social interaction capture. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):190–204, 2019.

[12] L. Ke, M.-C. Chang, H. Qi, and S. Lyu. Multi-scale structure-aware network for human pose estimation. In *European Conference on Computer Vision*. Springer, 2018.

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] D. Kong, Y. Chen, H. Ma, X. Yan, and X. Xie. Adaptive graphical model network for 2d handpose estimation. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2019.

[15] K. Lee, I. Lee, and S. Lee. Propagating LSTM: 3D pose estimation based on joint interdependency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–135, 2018.

[16] M. Lin, L. Lin, X. Liang, K. Wang, and H. Cheng. Recurrent 3D pose sequence machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 810–819, 2017.

[17] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. VNect: Real-time 3D human pose estimation with a single RGB camera. *ACM Transactions on Graphics (TOG)*, 36(4):44, 2017.

[18] A. Newell, K. Yang, and J. Deng. Stacked Hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[19] P. Panteleris, I. Oikonomidis, and A. Argyros. Using a single RGB frame for real time 3D hand pose estimation in the wild. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 436–445. IEEE, 2018.

[20] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4903–4911, 2017.

[21] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019.

[22] M. Rayat Imtiaz Hossain and J. J. Little. Exploiting temporal information for 3D human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 68–84, 2018.

[23] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017.

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[25] J. Song, L. Wang, L. Van Gool, and O. Hilliges. Thin-slicing network: A deep structured model for pose estimation in videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017.

[26] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems*, pages 1799–1807, 2014.

[27] A. Toshev and C. Szegedy. DeepPose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, 2014.

[28] C. Wan, T. Probst, L. Van Gool, and A. Yao. Dense 3D regression for hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5147–5156, 2018.

[29] Y. Wang, C. Peng, and Y. Liu. Mask-pose cascaded cnn for 2D hand pose estimation from single color image. *IEEE*

Transactions on Circuits and Systems for Video Technology, 2018.

[30] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[31] W. Yang, W. Ouyang, H. Li, and X. Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3073–3082, 2016.

[32] S. Yuan, G. Garcia-Hernando, B. Stenger, G. Moon, J. Yong Chang, K. Mu Lee, P. Molchanov, J. Kautz, S. Honari, L. Ge, et al. Depth-based 3D hand pose estimation: From current achievements to future goals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2636–2645, 2018.

[33] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. Wei. Towards 3D human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 398–407, 2017.

[34] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. Deep kinematic pose regression. In *European Conference on Computer Vision*, pages 186–201. Springer, 2016.

[35] C. Zimmermann and T. Brox. Learning to estimate 3D hand pose from single RGB images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4903–4911, 2017.