Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 4, No. 2, 2025

Research on Quantitative Detection Algorithm Based on Hrnet

Zhuohui Li*, Yanfang Fu

School of Computer Science and Engineering, Xi'an Technological University, Xi'an, China
* **Corresponding author**: Zhuohui Li (Email: lizhuohui@st.xatu.edu.cn)

Abstract: Aiming at the insufficient localization accuracy of traditional algorithms due to complex texture interference, diverse fabric deformations and sensitivity to small size errors in garment size detection, this paper proposes an improved HRNet-cloth key point detection model. By introducing full-dimensional dynamic convolution (ODConv) in the HRNet backbone network, we enhance the feature adaptation ability of the network to nonlinear deformation such as garment folds and draping, and effectively reduce the key point coordinate offset error; we design the EMA cross-dimensional attention mechanism module, fusing the channel and spatial dimensional feature responses to improve the localization robustness of the neckline, sleeve holes, and other detail regions; for the sub-pixel level regression requirements, we Construct an adaptive focus loss function to optimize the heat map peak distribution by dynamically adjusting the weights of difficult samples. Experiments show that the PR of HRNet-cloth on the self-built dataset ClothData reaches 100%, which is 11.6% higher than that of the benchmark model, and the absolute measurement error (AKE) of the dimensions is stabilized within ±1cm.

Keywords: Garment Size Detection, Key Point Detection, Deep Learning, Hrnet, Attention Mechanism.

1. Introduction

Accurate measurement of apparel size information is of great significance in improving production efficiency and consumer experience. As the core parameters of garment specifications, size data such as garment length, shoulder width, chest circumference, etc. directly affect the rationality of production layout, size matching for online sales and reliability of quality inspection. Especially in the field of FMCG, the accuracy of size labeling is a key basis for consumers to make purchases, which is directly related to brand reputation and return rate. However, the current clothing size detection still generally relies on manual operation, which has the problems of low efficiency, large error and high cost. Therefore, it is of great significance to study the realization of automated clothing size detection. In recent years, deep neural network-based key point detection methods have been widely used in various industries in daily life, but they are mostly used for behavioral prediction and analysis without using them in the field of quantitative measurement. Compared with manual inspection, the deep learning-based inspection method greatly improves the efficiency of size inspection, realizes real-time automated inspection in all-weather, and effectively reduces the labor cost.

2. Related Work

Deep learning-based keypoint detection algorithms can be mainly categorized into coordinate regression-based methods and heatmap regression-based methods. Coordinate regression-based methods directly predict the coordinate location of keypoints, typical algorithms such as Regression Networks; while heat map regression-based methods locate keypoints by generating probabilistic heat maps, represented by Stacked Hourglass Networks (SHN), HRNet and HigherHRNet. . In recent years, HRNet, with its parallel multi-resolution subnet and progressive feature fusion mechanism, has shown significant advantages in maintaining high-resolution feature representation, and has gradually become a benchmark model in the field of keypoint detection.

In 2018, Jian Sun's[1] team proposed HRNet network, which effectively improves human pose estimation accuracy through cross-resolution feature interaction. Zhang [2] et al. designed a lightweight branching structure based on HRNet, but the model was less robust to occlusion scenes.In 2020, Wang Chen[3] et al. introduced deformable convolution in HRNet, which enhanced the model's ability to adapt to target deformation. For the problem of complex background interference, Li[4] team proposed a hybrid attention module integrating spatial attention and channel attention, but it increased the computational complexity of the network.In 2022, Hao-Ran Liu[5] et al. proposed a dynamic feature selection mechanism, which improves the detection of dense keypoints by adaptively adjusting the weights of multi-scale features. On the other hand, Zhao[6] et al. constructed a cascade optimization architecture based on HRNet, and gradually improved the quality of heatmap by using an iterative refinement strategy, but the speed of model inference decreased significantly.

Although existing methods enhance the performance of keypoint detection through architectural improvements, the following challenges still exist: dense keypoints are susceptible to mutual occlusion interference leading to localization bias; feature abstraction in the deep network results in the loss of detail information; the traditional mean-square error loss is sensitive to low-quality labeled samples, and the accuracy of heatmap regression is limited. In this paper, HRNet-w48 is used as the baseline network, and the EMAttention[8] module is embedded in the feature aggregation stage, which enhances the feature discriminative property by establishing the attention correlation across channel-space dimensions; full-dimensional dynamic convolution[7] (ODConv) is introduced into the deep network, which improves the model's adaptive ability to clothing deformation and pose changes; and a new type of loss

function based on adaptive focus adjustment is also designed. optimize the heat map regression process by dynamically adjusting the weights of difficult samples. The final validation is carried out on a garment keypoint detection dataset containing complex folds. Experiments show that the method improves the localization accuracy of the garment keypoints while maintaining the real-time inference speed.

3. Proposed methodology

3.1. Full-dimensional dynamic convolution

Since the convolutional kernels of ordinary convolutional neural networks are static, recent dynamic convolution has shown that linear combinations of convolutional kernel weights to achieve conv's attentional weighting of the input data can significantly improve the accuracy of lightweight CNNs while maintaining high speed inference.ODConv (Omni-Dimensional Dynamic Convolution) argues that Existing dynamic convolution (e.g., CondConv and CyConv) only focuses on the dynamics of the number of convolution kernels, while ignoring the dynamics of spatial dimensions, input channels, and output channels. Based on this, full-dimensional dynamic convolution is realized using SE attention. The module structure is shown in Fig. 1:

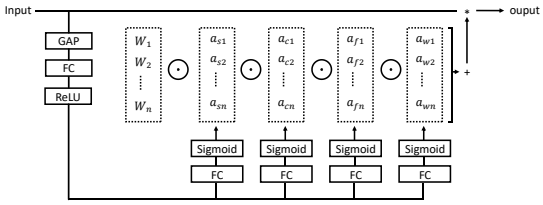


Figure 1. Omni-Dimensional Dynamic Convolution Module

ODConv is composed of convolution kernel parameters and attention mechanisms, and in ODConv, given n convolution kernels, its convolution kernel space contains four dimensions: the size of the convolution kernel ($k \times k$), the number of input channels (c_{in}), the number of output channels (c_{out}), and the number of convolution kernels (n). Among them, the convolution kernel size and the number of channels are inherent properties of the convolution operation, while the number of convolution kernels, n, introduces dynamism and represents the n sets of weighted parameters generated by the network based on the input data, which are ultimately weighted and summed to obtain the dynamic convolution kernel parameters.

$$y = \left(\alpha_{w1} \odot \alpha_{f1} \odot \alpha_{c1} \odot \alpha_{s1} \odot W_1 + \cdots + \alpha_{wn} \odot \alpha_{fn} \odot \alpha_{cn} \odot \alpha_{sn} \odot W_n\right) \quad (1)$$

Where α denotes the attention factor in different dimensions,W denotes the static convolution kernel weights, and α_{wi}、α_{fi}、α_{ci}、α_{si} denotes the attention weights associated with the spatial dimension of the convolution kernel, the input channel dimension, the output channel dimension, and the number of convolution kernels dimension, respectively, which are used for dynamically adjusting the convolution kernel weights in the corresponding dimensions. W_iThen it represents the base weights involved in dynamic weighted summation. After weighted summation of the two, the final convolution operation (*) is performed with the input feature map (x).

3.2. Attention Mechanism EMA Improvement

In deep learning-based keypoint detection, the attention mechanism enhances the model's ability to perceive critical regions by dynamically adjusting the feature weights.EMAttention (Efficient Multi-head Attention) is a lightweight multi-head attention module, whose core idea is to establish the channel and spatial dimensions separately through a decoupled multi-head interaction mechanism with the local-global feature associations. Different from the traditional multi-head attention, EMAttention adopts a grouped channel reorganization strategy, dividing the input feature map into multiple groups of sub-channels, and reconstructing the channel relationship through a dynamic weight fusion mechanism after each group of channels performs adaptive spatial attention computation respectively. This design retains the ability of multi-head attention to model long-distance dependence and reduces computational complexity through parameter sharing.

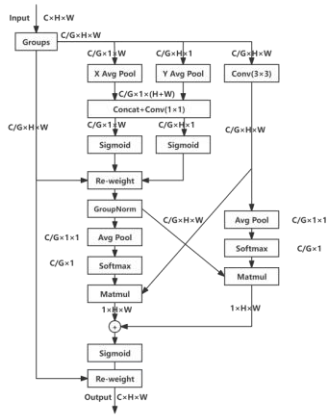


Figure 2. EMA Attention Mechanism Module

There are several reasons for incorporating the EMAttention wit module into HRnet Block: (1) Enhancing multi-scale feature fusion: HRNet maintains high-resolution feature streams through parallel multi-resolution subnets, but traditional convolution is difficult to capture cross-resolution semantic associations effectively. After embedding EMAttention at the end of branch Block, its group attention mechanism can adaptively establish spatial correspondences between feature maps of different resolutions. (2) Enhance deformation robustness: HRNet's fixed convolutional kernel is prone to losing local details when dealing with garment folds, and EMAttention's dynamic weighting mechanism enables the model to automatically adjust the attention distribution according to the input features: when deformation-prone regions such as cuffs and necklines are detected, the module enhances the local feature response through spatial attention, and cooperates with the channel reorganization strategy to retain the detailed texture. discriminative expression. (3) Optimize computational efficiency: Compared with the traditional multi-attention mechanism, EMA has lower computational requirements and higher efficiency.

As shown in Fig. 4, the EMA module is placed after the Block module in stage1 and stage2 of HRNet. This design enables the network to improve the localization accuracy of the key points of the garment through the attention-guided feature sufficiency mechanism while maintaining the original multi-scale feature extraction capability.

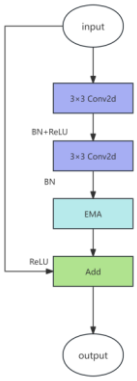


Figure 3. Improved Block module

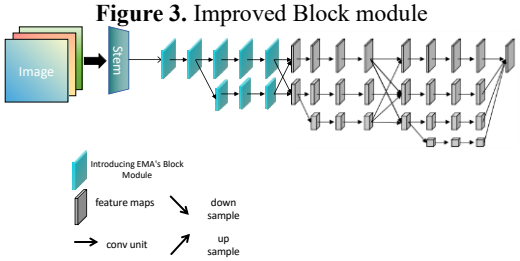


Figure 4. Improved HRnet network

3.3. Summary of the DGSBSA Algorithm Steps

Mean Squared Error Loss (MSE) is a loss function widely used in deep learning for regression tasks, which is mainly used to measure the difference between the predicted and true values of a model. The core idea is to quantify the prediction accuracy of the model by calculating the mean squared error between the predicted and true values. Specifically, for each prediction point, the difference between it and the corresponding true value is first calculated, and then the squares of all the differences are summed and averaged. The range of this loss value is $[0, +\infty)$, with smaller values indicating that the prediction is closer to the true value. It is defined as follows:

$$H_k^{gt}(x,y) = \exp\left(-\frac{(x-x_k)^2+(y-y_k)^2}{2\sigma^2}\right) \quad (2)$$

$$\mathcal{L}_{MSE} = \frac{1}{K}\sum_{k=1}^{K}\left\|H_k^{pred} - H_k^{gt}\right\|_2^2 \quad (3)$$

However, MSE Loss suffers from a major drawback.MSE Loss applies equal weights to all pixel points, but when the background region accounts for the majority of the pixels, the gradient contribution of the simple samples will dominate the process of parameter updating, resulting in an imbalance in the gradient distribution conflicting with the optimization of the difficult samples, a problem that can cause slow convergence of the difficult samples. Also, the model is more inclined to generate excessively smooth heatmaps to minimize the global error, resulting in a spread of response peak areas.

where a dynamic method is used in the loss to calculate the heat map regression loss, defined as follows:

$$\mathcal{L}_{Focal} = -\frac{1}{N}\sum_{i=1}^{N}\left[\alpha(1-\hat{p}_i)^{\gamma}\cdot y_i \log(\hat{p}_i) + \beta\hat{p}_i^{\gamma}\cdot (1-y_i)\log(1-\hat{p}_i)\right] \quad (4)$$

Where $\hat{p}_i$ is the prediction confidence andγ is the weight. In order to organize the simple background region to dominate the gradient update, this loss function distinguishes between difficult and easy samples by the prediction confidence, and applies a higher weight for difficult samples to suppress the contribution of simple samples.

In view of this, a new criterion for loss function calculation is defined as follows:

$$\mathcal{L}_{FoL1} = \lambda_t\mathcal{L}_{Focal} + (1-\lambda_t)\left(\frac{1}{N}\sum_{i=1}^{N}|p_i - y_i|\right) \quad (5)$$

At the beginning of training (t/T< 0.3), the model needs to quickly locate the approximate region of the key point, at which time it strengthens the constraining effect of the L1 Loss ($\lambda_t = 0.2$), and utilizes its absolute error sensitivity to quickly reduce the overall deviation of the predicted heatmap from the true distribution. In the middle stage of training (0.3< t/T<0.7), the model gradually transitions to Focal Loss dominance (λ_t=0.2+0.6 ($\frac{t/T-0.3}{0.4}$)), optimizing the key point edge response through difficult sample reweighting and suppressing the overfitting of the background region. At the late stage of training (t/T> 0.7), it is completely dominated by Focal Loss ($\lambda_t = 0.8$), focusing on the peak fine tuning, and using its$(1-\hat{p}_i)$ term to enhance the high response gradient to make the heatmap peaks sharper and more accurately localized.

In short, in each scenario,$\mathcal{L}_{FoL1}$ exhibits better adaptability and robustness to more effective keypoint detection performance at $\mathcal{L}_{FoL1}$ compared to $\mathcal{L}_{MSE}$.

4. Experimentation

4.1. Introduction to the dataset

ClothData dataset is a self-constructed dataset based on the task requirements, using industrial ccd camera to capture the viewpoints, mainly used for detecting the key point detection of clothing. This collection comprehensively covers a variety of clothing types, including: round neck short sleeve, lapel short sleeve, pants, shorts, etc. For the dataset initial collection of images were screened layer by layer, eliminating the image data that do not match the standard clothing types, blurred, details are not clear enough, after Labelme tool for key point annotation, a total of 1,489 images were collected. In terms of dataset division, the ratio of training set, validation set and test set is 7:2:1. the results are as follows:

Figure 5. Example of clothing labeling

According to different kinds of clothes, the labeling made is also different. Take the short-sleeve front as an example, it defines 20 key point formats, in which the distance between key point 0 and key point 1 is the collar width, the distance between key point 4 and 5 is the collar height, the distance between key point 16 and 17 is the hem width, the distance between key point 18 and 19 is the hem height, the distance between key point 14 and 15 is the bust, the distance between key point 6 and 7 is the cuff width, the distance between key point 8 and 9 is the sleeve fat, and the length is calculated as:

the average of vertical coordinates 2 and 3 and the average of vertical coordinates 16 and 17 is the sleeve fat. The distance between No. 6 and No. 15 is the bust, the distance between No. 6 and No. 7 is the cuff width, the distance between No. 8 and No. 9 is the sleeve fat, and the length of the garment is calculated as the difference between the average of the vertical coordinates of No. 2 and No. 3 and the average of the vertical coordinates of No. 16 and No. 17.

4.2. Assessment of indicators

Since the task of this paper is for industrial level high precision keypoint detection task, the physical error that needs to be measured≤ 1cm. so, the model evaluation method used in this paper adopts AKE (Absolute Keypoint Error) as the core evaluation index to evaluate and quantify the actual spatial deviation directly, which is expressed as shown in Eq:

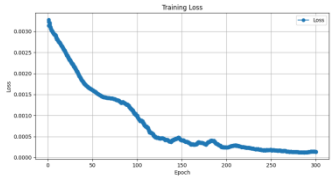
$$\mathrm{AKE}_i = \sqrt{\left(x_i^{\mathrm{pred}} - x_i^{\mathrm{gt}}\right)^2 + \left(y_i^{\mathrm{pred}} - y_i^{\mathrm{gt}}\right)^2} \times \rho \qquad (6)$$

Where, $(x_i^{\mathrm{pred}}, y_i^{\mathrm{pred}})$ are the i-th keypoint coordinates (pixel units) predicted by the model, $(x_i^{\mathrm{gt}}, y_i^{\mathrm{gt}})$ are the manually labeled true coordinates (pixel units), and ρ is the pixel resolution (cm/px), which is pre-determined by the calibration board.

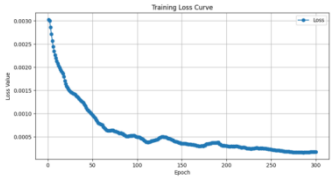
The pass rate (PR) was used to measure the probability of AKE≤ 1 in all detected images.

4.3. Experimental Results

During the training phase, we used tensorboard to record the loss function of the model on the training set. As can be seen from the figure below, the training loss and validation loss of the model gradually decreases as the number of training sessions increases, which indicates that the model is continuously learning more accurate features. At the end of training, the model was evaluated on the dataset using the model and the following results were obtained, as shown in Figure 6.



(a) Before improvement



(b) After improvement

Figure 6. Comparison of loss function curves

The ablation experiments of the proposed modules are carried out on the ClothData validation set, and to ensure the fairness of the comparison of the experimental structure, the model input sizes are all 640×640, and the experimental configuration parameters such as the training batch, the training step size and the optimizer are kept consistent. The experimental results are shown in Table 1, where M1 and M2 denote the replacement of the full-dimensional dynamic deformation convolution and the joined EMAttention attention mechanism module, respectively, and M3 denotes the replacement of $\mathcal{L}_{\mathrm{FoL1}}$.

Table 1. the results of the ablation experiments performed on the model on the dataset

Baseline	M1	M2	M3	Params (M)	GFLOP (G)	PR (%)
HRNet-w48	-	-	-	63.8	154.3	88.4
HRNet-w48	√	×	×	70.5	172.0	92.5
HRNet-w48	√	√	×	76.9	180.3	98.1
HRNet-w48	√	√	√	77.0	180.3	100

As shown in the analysis of Table 1, the HRNet-cloth algorithm is significantly better than other HRNet improvement algorithms in terms of regression accuracy and pass rate for garment keypoint detection. From the three or four rows of the table, it can be seen that adding the EMA attention module or the ODConv module alone to HRNet can improve the detection accuracy, respectively, which can also indicate that the added EMA attention module and the ODConv module can help to improve the performance of the network, but the overall detection performance is compared with that of HRNet but neither of them has been able to fulfill the technical specifications. Finally, the new loss function calculation criterion is applied to the HRNet-ODC-EMA model and the HRNet-cloth model is able to fulfill the detection task well.

5. Conclusion

Using HRnet as the base model to realize the automated inspection of clothing size, products that do not meet the production specifications can be found and rejected in time, effectively reducing the probability of omission and misdiagnosis, and minimizing the labor cost and economic loss. This paper aims to construct a key point detection algorithm oriented to quantitative measurement, innovatively combines the key point detection algorithm in the field of human posture estimation and proposes the HRNet-cloth improvement algorithm. The key to this improvement is to add a realizable convolutional module to the backbone network to improve the accuracy of detecting keypoints on the feature map; this structural improvement injects a deeper understanding of the information into the whole algorithm and improves the model's perceptual ability, which in turn improves the detection of keypoints in clothing. The experimental results show that the improved HRNet-cloth model has a significant improvement in detection accuracy, which meets the actual engineering indexes, and provides a certain reference for the subsequent use of the key point detection network to detect the size information of other products.

References

[1] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5693-5703.

[2] Zhang X, Li H, Mo J, et al. Lightweight Human Pose Estimation with Hierarchical Feature Refinement[J]. IEEE Transactions on Image Processing, 2021, 30: 1234-1245.

[3] Wang C, Wang Y, Lin Z, et al. Deformable HRNet: A Deformable Convolution Enhanced Network for Human Pose Estimation[J]. IEEE Transactions on Multimedia, 2021, 23: 123-135.

[4] Li Y, Chen J, Zhang Z, et al. Mixed Attention Mechanism for Occluded Human Pose Estimation[C]//European Conference on Computer Vision. Springer, 2020: 456-472.

[5] Liu H, Fan Z, Wang T, et al. Dynamic Feature Selection for Dense Keypoint Detection[C]//AAAI Conference on Artificial Intelligence. 2022: 2145-2153.

[6] Zhao L, Li S, Wang Q, et al. Cascaded Refinement Network for High-Resolution Heatmap Regression[J]. International Journal of Computer Vision, 2022, 130(5): 1327-1345.

[7] Li X, Wang W, Hu X, et al. Dynamic Convolution: Attention over Convolution Kernels[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 11030-11039.

[8] Zhang Y, Li Q, Zhou B, et al. EMANet: Enhanced Multi-scale Attention for Keypoint Detection[C]//European Conference on Computer Vision. Springer, 2022: 678-694.