IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 19, 2019, accepted October 7, 2019, date of publication October 24, 2019, date of current version December 11, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2949269

An Improved Skeleton Extraction Method via Multi-Task and Variable Coefficient Loss Function in Natural Images

YOUQING XIAO, ZHANCHUAN CAI, (Senior Member, IEEE), AND XIXI YUAN
Faculty of Information Technology, Macau University of Science and Technology, Macau 999078, China

Corresponding author: Zhanchuan Cai (zccai@must.edu.mo)

This work was supported in part by the National Basic Research Program of China (973 Program) under Grant 2011CB302400, in part by the Science and Technology Development Fund of Macau under Grant 0012/2018/A1 and Grant 0069/2018/A2, in part by the Open Project Program of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, under Grant VRLAB2019C02, in part by the Open Project Program of the State Key Lab of CAD&CG of Zhejiang University under Grant A1910, and in part by the Open Fund of State Key Laboratory of Remote Sensing Science under Grant OFSLRSS201901.

ABSTRACT Extracting object skeleton and its scale features in natural images is helpful for object detection and recognition in computer vision. In order to advance the location accuracy of object skeleton pixels, a new method via multi-task and variable coefficient loss function is proposed in this paper. Adopting the hierarchical integration mechanism to mutually refine captured features at different network layers; a specific variable coefficient loss function is designed for multi-class imbalanced data handling problem, such as the skeleton pixels in natural images are always far less than the non-skeleton pixels; the regression algorithm is an added deep learning branch in the skeleton extraction network assisting the improvement of recognition accuracy. Besides, not only the skeleton pixels and its classification can be obtained, but also its scales are predicted without disturbing skeleton acquisition process. The experimental results verify that both the skeleton accuracy and the generalization abilities are promoted benefiting from the regression task and the new loss function in the new method, as satisfactory results are achieved on three public datasets, i.e., SK-LARGE, SK-SMALL, and WH-SYMMAX, which are indicated by F-measures and precision/recall curves. The results further demonstrate that the proposed method is superior to the best skeleton extraction method available currently.

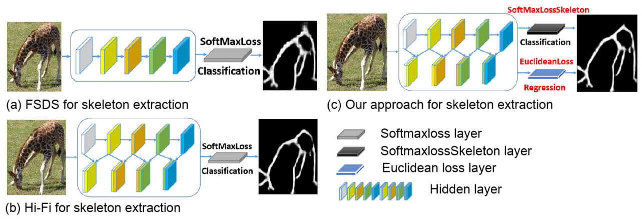
INDEX TERMS Skeleton extraction, Softmax function, regression algorithm, skeleton scale prediction, deep learning.

I. INTRODUCTION

As a high-level feature, skeleton has the characteristic of compactly representing the shape and structure of objects. With the development of artificial intelligence and deep learning techniques, identifying the object skeleton automatically from natural images is becoming achievable [1]–[3]. At present, image semantic segmentation is deep researched, meanwhile the machine learned features are more and more fine benefitting the skeleton extraction in natural images. Importantly, the skeleton extraction methods are successfully applied to many fields [4]–[10], such as image retrieval, scene text detection, human action recognition, gesture recognition, and medical examination.

The object skeleton is actually the *symmetry axis* treated as one shape descriptor, which possesses geometric characteristics (e.g., the size and angle metrics) and topology structures (e.g., the connection among different components) [14]. Object skeleton extraction from natural images means collecting all image pixels at the central of object contours without segmentation in advance. There are a mass of algorithms published recent years concerning the skeleton extraction technique [11], [12], [15], [16], and they are roughly sorted into three categories: (a) traditional image processing methods, that detect the skeleton according to the relative position of contours and the symmetry axis; (b) early machine learning based methods, which use hand-crafted

The associate editor coordinating the review of this manuscript and approving it for publication was Utku Kose.



(a) FSDS for skeleton extraction

(b) Hi-Fi for skeleton extraction

(c) Our approach for skeleton extraction

FIGURE 1. Brief illustration of our proposed method and some other skeleton extraction methods. (a) based on VGG16 network, FSDS method [11] fuses side-output features at different stages to extract object skeleton; (b) the Hi-Fi method [12] combines FSDS and RCF [13] methods to refine the skeleton; (c) in our proposed approach, we take hierarchical integration, regression algorithm, and new Softmax function to advance skeleton accuracy.

and multi-scale features of every pixel with surrounding pixels to train the classifier for selecting skeleton pixels; and (c) recently, the fully convolutional neural (FCN) network is widely used for image feature recognition, also for skeleton extraction.

We make a survey on the development of different skeleton extraction methods. To be specific, many early image processing methods [17]–[22] depend on the hypothesis that the skeleton stands between two parallel edges. The intensity gradient is used to calculate the edge pixels, and the skeletons are locked through the parallel edges. While, the limitation of this method is that the foreground and the background of test images should be easily distinguished. Besides, for complex scene of images, the computation complexity is huge [23], [24]. Afterwards, with the raising of machine learning techniques, researchers use hand-crafted features to train the classifier and calculate the predicted regression of skeleton pixels [25]–[28]. Limited by the pretreatment requirement, these traditional machine learning methods can not work with complex scenes, and they are also time consuming in pixel-by-pixel class prediction [29], [30]. Nowadays, the convolutional neural network (CNN) is extensively used to classify and predict features with automatic learning methods [31]–[33]. Long et al. [34] propose fully convolutional network (FCN) to solve pixel classification end to end and implement image semantic segmentation more efficiently. The HED in [35] pioneers side-output on the inner convolutional layers of FCN structure for multi-scale edge detection. Inspired by HED, the FSDS [11] (as shown in Fig. 1 (a)) uses scale-associated side-outputs (SSOs) to solve the scale unknown problem in skeleton detection. The side-output residual network (SRN) [36] based on HED as well, which employs deep to shallow residual connections to catch rich semantic features, so as to strengthen the shallower layers to distinguish real skeleton from local reflection structure. Liu et al. [13] propose a precise edge detector to catch rich convolutional features (RCF). In view of the merits of FSDS and RCF methods, Zhao et al. [12] present the hierarchical feature integration (Hi-Fi) mechanism (as shown in Fig. 1 (b)). Based on VGG16 network, we design new regression prediction method and Softmax loss function for skeleton extraction referring to the hierarchical structure raised in COB [37] and Hi-Fi, as shown in Fig. 1 (c).

Based on the research, we conclude the tasks of object skeleton extraction into three procedures: (1) judging and selecting skeleton pixels in the image; (2) quantizing the skeleton pixels output at every stage into scale-associated categories; (3) predicting specific scale of each skeleton pixel. The hierarchical feature integration method mainly solves the first two problems, which is adopted and advanced in our method to achieve better performance on these three problems. The most important contributions of the new proposed method as follows:

- The regression task added to the hierarchical convolutional network facilities the accuracy of skeleton pixel classification and prediction;
- A new variable coefficient loss function is put forward to deal with the multi-class imbalanced data handling problem;
- The performance of different hierarchical level networks for skeleton extraction is discussed comprehensively in the paper.

Further, we conduct sufficient experiments on SK-SMALL [11], SK-LARGE [14], and WH-SYMMAX [15] datasets to verify the skeleton accuracy and the generalization abilities of the proposed method with better precision/recall (PR) curves and F-measures. The rest arrangement of this paper is listed: section II provides a review of related methods. Section III includes some methodologies that are essential to skeleton feature extraction. The fourth section gives experiment results of the proposed method and other related methods on three datasets. The fifth part briefly summarizes the contribution of this paper and the possible further works.

II. RELATED WORKS

Since the CNN was pioneered in machine learning in 2012, it has boomed in image semantic recognition field, and successfully applied to many different applications [38]–[40]. The most famous peculiarity of CNN is that it has the ability to automatically learn multi-level features through multi-layer structure: the shallower convolutional layers have smaller receptive fields capturing local features; deep layers use larger receptive fields to learn senior semantic features. The deep abstract features neglect size, position, and orientation properties of objects that facilitate image recognition. However, the contour information is too small to be lost, which ruins the precise of image segmentation. In order to eliminate the deficiency of CNNs, the FCN proposed in [34] converts all fully connected layers of CNN into convolutional layers. The salient distinction of pixel-level classification achieves exact segmentation.

In recent years, some excellent skeleton extraction methods have been put forward. The FSDS method in [11] bases on FCN and HED to propose SSOs added network structure. More precisely, every SSO outputs several scale-associated skeleton maps under the supervision of ground-truths; the skeleton map of the same scale generated by different SSOs will be weighted average with a scale related weight, here the operation of weighted average utilizes an 1 × 1 convolutional layer to accomplish the task, this method can be seen from Fig. 2(c), because we believe that different scale maps have different weights. Finally, the object skeletons are extracted by fusing among multi-scale side-outputs. Afterwards, the LMSDS method [14] which is the progression of FSDS uses SSOs for skeleton localization and prediction. The added SSO branch as a regression task applied to skeleton prediction improves the skeleton detection accuracy compared with FSDS.

In addition, the Hi-Fi method [12] presents hierarchical feature integration and bidirectional mutual refinement strategies for skeleton detection, that enhances the ability of capturing enrich features of objects. The most vital point is that, it has verified that hierarchical feature fusion in convolutional network structure is better than non hierarchical feature fusion. A quantity of experiments of Hi-Fi manifest that it has better performance than FSDS in terms of skeleton extraction accuracy.

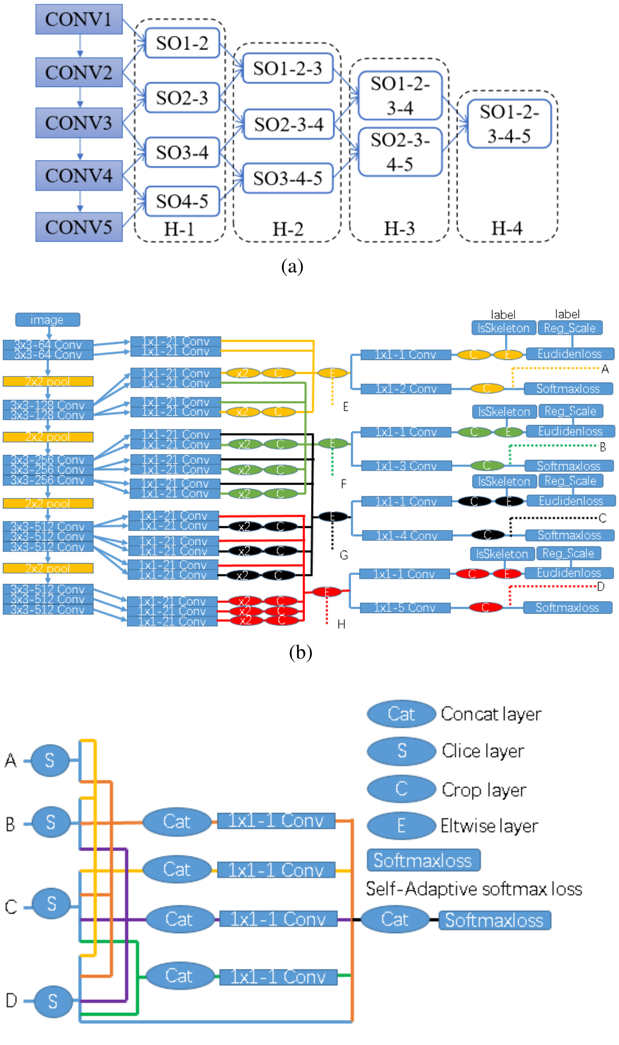
Above all, we have identified the advantages of different algorithms: (1) SSO plays an important role in skeleton extraction tasks based on deep learning; (2) multi-task (i.e., skeleton detection and skeleton scale prediction) can improve skeleton detection accuracy generally; and (3) the hierarchical feature integration network optimizes other skeleton extraction network structures. Therefore, we adopt these advantages in different algorithms to improve our method.

III. METHODOLOGY

In order to improve the accuracy of object skeleton extraction more than the current methods in natural images, we propose a new network architecture based on the state-of-the-art network. Considering the distribution of skeleton pixels in images is relatively less, resulting in learning sensitivity reduction of feature, we design a new variable coefficient loss function to handle this problem. Besides, we systematically summarize the classification of skeleton pixels and generation solutions of skeleton maps using deep learning, together with the skeleton scale prediction using regression algorithm.

TABLE 1. The sizes of the receptive fields (RF) at different stages based on VGG16 network [41] and the scale classification sets.

Stage(k)	Stage(1)	Stage(2)	Stage(3)	Stage(4)	Stage(5)
RF	5	14	40	92	196
Scale sets	$\varnothing$	$\{0,1\}$	$\{0,1,2\}$	$\{0,1,2,3\}$	$\{0,1,2,3,4\}$



(a)

(b)

(c)

FIGURE 2. The brief network architecture of proposed skeleton extraction method. (a) represents the network with four hierarchies. It includes five stages of convolutional layers (marked as CONV(x)), and the side-outputs (SOs) of feature maps connected to every convolutional layer are fused in four-level hierarchies. (b) and (c) are particular structure of H1, and the outputs of A to D of (b) are the inputs of (c). Besides, the E to H of (b) are the inputs of H2.

A. NETWORK ARCHITECTURE

The fundamental structure adopts VGG16 network [41], which has been successfully applied to object classification and detection in computer vision [42–44]. The network basically consists of 13 convolutional layers and three fully connected layers. The whole conv-layers are divided into five stages connecting to pooling layers (typically a 2 × 2 window with stride 2), which are used to change the sizes of receptive fields. In fact, with the increasing of the receptive field sizes, useful skeleton information captured at each stage are becoming increasingly rough. Table 1 lists the sizes of each receptive field at different stages. Based on VGG16 structure, we utilize some modified strategies to lift the skeleton extraction accuracy: on each layer of the network, the scale-associated side-output is added to the network to capture more edge and skeleton information; the hierarchical network structure is good at learning richer context information and achieving mutual refinement among different layers; the regression task is used for scale prediction which plays an important role in skeleton extraction.

Concretely speaking, the new network architecture has four hierarchies as shown in Fig.2. In Fig. 2(a), the leftmost column are five stages of convolutional layers: each of the

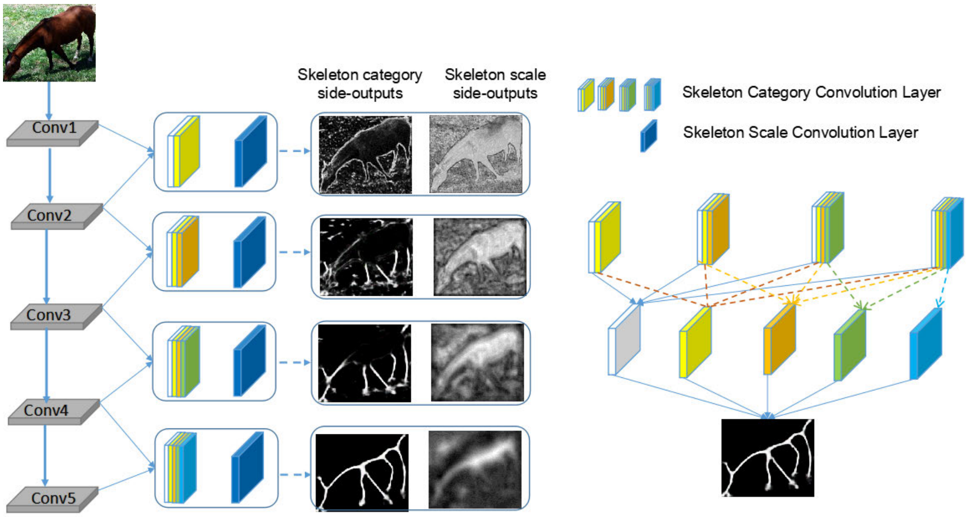

Conv1
Conv2
Conv3
Conv4
Conv5
Skeleton category side-outputs
Skeleton scale side-outputs
Skeleton Category Convolution Layer
Skeleton Scale Convolution Layer

FIGURE 3. The illustration of the new network structure. For the sake of simplicity, only one level (H-1) is taken as an example. On the left of the figure, the first column shows the network structure divided into five stages; the second column represents the H-1 level obtained by the integration of two adjacent stages (i.e., two adjacent stages), and the blue layer on the right represents the Euclidean loss layer; the third column is the outputs of the classification task and the regression task at different stages. The right side of the figure shows the final output of the integration at different stages.

first two stages (i.e., CONV1 and CONV2) contains two convolutional layers; the last three stages (i.e., CONV3, CONV4, and CONV5) contain three convolutional layers in several. Specifically, because of the pooling layers exist in the leftmost network, different scales feature maps will be obtained at different stages. In order to compensate the loss from shallow layers (early layers) to deep layers (later layers), at different levels, the feature maps generated by the convolutions of two adjacent phases are fused by eltwise operation. After feature information fusion, the obtained feature maps have two branches: Firstly, the side output includes skeleton classification and skeleton scale prediction, which is different from the Hi-Fi network, that only one task is included; Secondly, as the input of the next level, two adjacent stages produce one side output, therefore four different scale side outputs are generated at H1 level. At last, after the 'slice' and 'concat' operations of the feature maps, a final fused output will be generated. The processing approaches of other hierarchies, i.e., H2, H3, and H4 are similar. It should be noted that due to the pooling layers, the feature maps at adjacent deep convolution are smaller than the shallow layer (early layer), so it is necessary to fuse the feature maps of different scales by upsampling and cropping operations.

Generally, auxiliary task is helpful for skeleton extraction, so the logistic regression model is established and added to the new network architecture. Under the condition of linear correlation, the quantitative relationship between two or more independent variables and dependent variables is called multiple regression analysis. Linear regression can be used to fit a prediction model with the real data set Y and the variable data set X, and then for any new data of X, the corresponding value in Y can be predicted using this fitted model, usually Y is a real data instead of a category. For the scale prediction task of skeleton pixels, the scale is a specific value, then the task of skeleton scale prediction can be transformed to a regression problem. The specific strategy of this paper has two aspects: the Euclidean loss function added to SO(x)-x conducts a regression task, which is used to predict skeleton pixels; the variable coefficient Softmax loss function added to SO(x)-x is applied to classify pixels, as shown in Fig. 4.

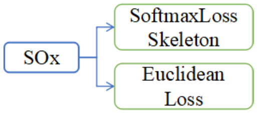


FIGURE 4. The structure of SO(x)-x with Euclidean loss and softmax loss operations, the details of the design can be found on the right side of Fig. 2(b).

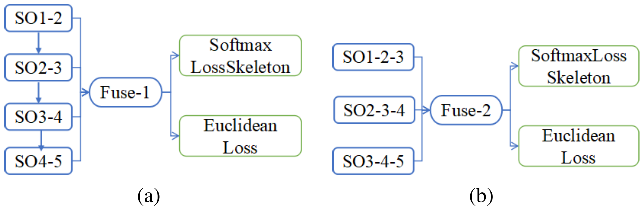


FIGURE 5. The fusion among different side-outputs in 1st-level hierarchy and 2nd-level hierarchy, respectively. The details of fusion operation can be seen from Fig. 2(C).

Because every hierarchy has multiple side-outputs, different SO(x)-xs at each hierarchy need to be fused. The fusion method on the first and second hierarchies (i.e., H-1 and H-2) are illustrated in Fig. 5. In order to capture the features preferably, we fuse the features of Fuse-1 and Fuse-2 outputted by adjacent hierarchies of H1 and H2 respectedly, it can be shown in Fig. 6.

As exhibition of the figures, there exists quite a few classification and regression tasks in the network architecture. In the specific calculation of pixel skeleton classification

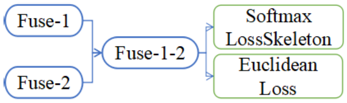


FIGURE 6. The fusion between multiple hierarchies.

and scale prediction, only the final feature fusion results are adopted. For instance, as for H-1 and H-2, the Fuse-1-2 is used for the final classification and regression prediction, while the Fuse-1 and Fuse-2 are only used as the interim parameters in training phase.

In the network structure, according to the change of receptive fields from small to large, for stage (k), the number of categories with scale of side output is k-1. On account of the sizes of receptive fields at different stages are various, the network can learn multi-scale information from low-level to object-level. For example, the convolutional layers of stage two can capture the skeleton pixels with scales equal to or less than 14, and the scale size of 196 are captured at stage five. Fig. 3 shows the intermediate results of all stages. In addition, based on the VGG16 structure, the hierarchical network structure can establish up to four levels, i.e., H-1, H-2, H-3, and H-4.

B. VARIABLE COEFFICIENT LOSS FUNCTION

The Softmax layer is connected to the last convolutional layer at each stage to deal with multi-class classification task in skeleton extraction. In fact, the Softmax is always used in multi-class classification problems, mapping outputs of multiple neurons into values at (0, 1) interval; the sum of all output values equals to 1; and the selected prediction corresponds to the largest value of Softmax. Then the parameters of feature model are trained by minimizing the Softmax loss function.

The Softmax loss function is defined as follows: Firstly, in equation (1), $J(\theta)$ is marked as the loss function, m is the number of samples in training dataset, k is the amount of categories. Specifically, a is the balance factor we defined, that's a new class-balancing weight of the loss function. The label y can take any k different values. As in (2), $1\{\cdot\}$ represents a indicator function, so that $1\{a\ true\ statement\}$ evaluates to 1, and $1\{a\ false\ statement\}$ evaluates to 0. So as for the training set $(x^{(1)}, y^{(1)}), \cdots, (x^{(m)}, y^{(m)})$, there is $y^{(i)} \in 1, 2, \cdots, k$. P represents the probability of classifying x into category j. Furthermore, the model parameter θ is trained to minimize the loss function $J(\theta)$.

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} a_j y_j log P_j\right] \tag{1}$$

$$y_i = 1\{y^{(i)} == j\} \tag{2}$$

$$P_j = \frac{e^{\theta_j^T x^{(i)}}}{\sum_{i=1}^{k} e^{\theta_1^T x^i}} \tag{3}$$

When training with deep learning method, the numbers of positive and negative samples always lack of balance. In order to overcome this problem, many scholars have adopted different schemes, such as the program in HED: since the numbers of edge pixels and non-edge pixels in images are extraordinary imbalanced, that means the non-edge pixels are much more than the edge pixels, the following balance coefficients are adopted when calculating the loss:

For the positive samples: $a_{pos} = 1 - \frac{Y_{neg}}{|Y|}$;

For the negative samples: $a_{neg} = 1 - \frac{Y_{pos}}{|Y|}$.

Pseudo Code 1 New Softmax Loss Funtion

Input: Original image

Output: Loss value

```
Calculate the number of skeleton pixels in each category ($i = 1, 2, \cdots, k$);
% e.g., the amount of skeleton pixels with category 1.
Calculate the loss of each category: $category\_loss_i$;
% e.g., the loss of skeleton scale with category 1.
for $i = 1$ to $k$ do
    Compute $|Y_1|, |Y_2|, \cdots, |Y_k|$ and $|Y|$;
    Compute $a_i$ with equation (4);
    % e.g., $a_1 = 1 - \frac{|Y_1|}{|Y|}$.
end for
for $i = 1$ to $k$ do
    $Loss+ = a_i \times category\_loss_i$;
end for
```

In skeleton extraction algorithm, the number of skeleton pixels is pretty less than the number of non-skeleton pixels, then the programs of [46], [47] use the same balance coefficient a of HED is used to handle this problem, while it is only effective for binary classification. However, the skeleton pixel belongs to multi-class problem, so it is necessary to propose a new appropriate balance coefficient for multi-class problem. Therefore, we define $a_i, (i \in 1, 2, \cdots, k)$ to adjust the weights of different skeleton pixel scales. The coefficient a_i is described in (4).

$$a_i = \frac{|Y| - \sum_{j=1, j\neq i}^{k} |Y_j|}{|Y|} = 1 - \frac{|Y_i|}{|Y|}, \quad (i \in 1, 2, \cdots, k) \tag{4}$$

where $|Y_1|, |Y_2|, \cdots, |Y_k|$ represent the amount of pixels with different scales, the category of which corresponding to the number in $\{0, 1, \cdots, k-1\}$, and $|Y|$ in (5) represents the amount of all skeleton pixels with different scales in the image, that's also regarded as skeleton map.

$$|Y| = |Y_1| + |Y_2| + \cdots + |Y_k| \tag{5}$$

For the minimization of $J(\theta)$, the iterative gradient descent method is applied to handle the task. After derivation, we get the gradient formula (6).

$$\nabla_{\theta_j} J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[a_j x^{(i)}(y_i - p_j)\right] \tag{6}$$

The particular calculation processes of the new loss function are presented in **Pseudo Code 1**.

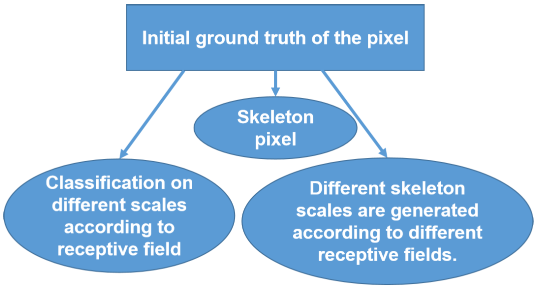

Initial ground truth of the pixel
Skeleton pixel
Classification on different scales according to receptive field
Different skeleton scales are generated according to different receptive fields.

FIGURE 7. Calculating the ground truth of classification and regression tasks.

C. GROUND TRUTH GENERATION DURING TRAINING

In our network structure, there are two types of tasks, one is skeleton classification, and the other is the regression skeleton scale of different network model stages. According to these tasks, in the training stage, the image and label data as the input, and the label data contains the initial skeleton scale. In the training phase, we need to generate three types of truth values (as shown in Fig. 7) according to the initial skeleton scale, which are skeleton category, regression scale and whether it is skeleton pixel or not. The specific algorithm is implemented as **Pseudo Code 2**.

Pseudo Code 2 Classification and Regression

```
Input: Initial scale: $init\_scale$;
       The size of current receptive field: $rf$;
Output: The true value of skeleton scale classification: $category$;
       The true value of skeleton scale regression: $scale$;
reg_scale: $reg\_scale = 2 \times \frac{init\_scale}{rf}$;
$IsSkeleton = True$;
if $1 \leq init\_scale < 10$ then
    $category = 1$;
else
    if $10 \leq init\_scale < 26$ then
        $category = 2$;
    else
        if $26 \leq init\_scale < 60$ then
            $category = 3$;
        else
            if $60 \leq init\_scale < 150$ then
                $category = 2$;
            else
                if $init\_scale < 1$ or $init\_scale \geq 150$ then
                    $category = 0$;
                    $IsSkeleton = False$
                end if
            end if
        end if
    end if
end if
return $category$, $reg\_scale$, $IsSkeleton$
```

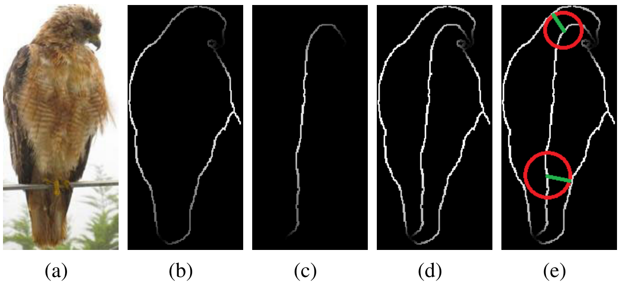


FIGURE 8. Skeleton scale illustration (skeleton is the collection of the largest inscribed circle center of the object contour). (a) is the original image; (b) is the object contour of the original image; (c) is the skeleton of the object; (d) combines the object skeleton and the contour; in (e), the red circle of the figure indicates that the largest inscribed circle of the skeleton pixel at object contour, and the green line represents the skeleton scale.

TABLE 2. The scale-associated skeleton pixels classification with piecewise quantization.

Scales	$\geq 150 \| < 1$	[1, 10)	[10, 26)	[26, 60)	[60, 150)
Y	0	1	2	3	4

In the regression task, IsSkeleton and feature map generated by convolution layer as the inputs, conduct the multi-operator of the eltwise layer, and get the result of the multi-operator serveing as the calculation basis for the regression loss, as shown in Fig. 2 (b) in the upper right corner.

D. SKELETON FEATURE MAP

The skeleton scale assists skeleton pixels localization is an important feature in skeleton extraction, so we make a clear classification of it in the paper. In addition, in order to make the method more effectively, we transform skeleton extraction into pixel classification task following piecewise quantization.

1) SKELETON PIXEL SCALE AND ITS CLASSIFICATION

The skeleton scale or thickness is the radius of the largest inscribed circle of the object contour centered on the skeleton point. In other words, it is the distance between the skeleton pixel and the closet object contour point. As shown in Fig. 8, the radii are colored in green. Obviously, the range of skeleton scale sizes is $[0, max(image_{\frac{width}{2}}, image_{\frac{height}{2}}]$.

The piecewise quantization inventory of skeleton scales is listed in Table 2. The scales are classified into five categories (marked as Y) $Y \in \{0, 1, 2, \cdots, k\}$. $Y = 0$ represents that's not a skeleton pixel; the categories of other values in $\{1, 2, \cdots, k\}$ have the corresponding scales in Table 2. For example, $Y = 2$ represents the scale of pixels lie in [10, 26).

The VGG16 network has five stages, and the receptive field is 5, 14, 40, 92, and 196 from the first stage to the fifth stage as shown in Table 1. At the first stage, the size of receptive field is too small to detect any skeleton scales, so skeleton pixels are captured from the second to the fifth stage. The output pixel scales are slightly smaller than the receptive field size in general. At different stages, the categories of the pixel scales are different, which can be seen in Table 1. Specifically, at Stage(k), $k \in \{2, 3, 4, 5\}$, the categories of the side-output pixels lie in $\{0, \cdots, k-1\}$. For example, the size of the receptive field at Stage(3) is 40, and the category of the side-output pixels lie in $\{0, 1, 2\}$, while the specific scale values are located at $[0, 10)$, $[10, 26)$, and $\{\geq 150 \| < 1\}$.

2) THE GENERATION OF THE SKELETON MAP

Assuming that $S_i, i \in \{0, 1, \cdots, k-1\}$ is the predicted pixel whose scales are corresponding to category i, e.g., S_2 represents the pixel set with scales belonging to category 2 (i.e., the pixel scales of S_2 lie in $[10, 26)$). Therefore, it can be found that the skeleton map is $S_1 \bigcup S_2 \bigcup S_3 \bigcup \cdots \bigcup S_{k-1}$. In an image, the total categories of pixels are marked as $\sum$, and S_0 is the non-skeleton pixel set. Then the formula (7) can be get.

$$\sum = S_0 \bigcup S_1 \bigcup S_2 \bigcup S_3 \bigcup \cdots \bigcup S_k \tag{7}$$

Hence, I-S_0 is the other form of skeleton in a binary image, where I is the identity matrix in a trainer.

E. REGRESSION PREDICTION METHOD OF SKELETON SCALE

Skeleton scale prediction is helpful for skeleton pixel position. The logical regression method is used to predict the skeleton scale. The definition and calculation of the involved functions are described in this section.

1) THE NORMALIZATION AND REGRESSION WITH LOSS FUNCTION

The skeleton scale is acquainted referring to section III-D. In fact, the scale of skeleton pixel is a real number which can be forecasted by regression algorithm, after that the skeleton map is generated with normalization. At stage(i) $i \in \{2, 3, 4, 5\}$, the size of receptive field is represented as $Field\_Size_i$. After normalization the ground-truth of skeleton scale GT_Scale at Stage(i) is expressed in (8).

$$GT\_Scale = \begin{cases} \dfrac{Scale}{Field\_Size_i} \times rate, & Scale \leq Field\_Size_i \\ 0, & Scale > Field\_Size_i \end{cases} \tag{8}$$

The range of GT_Scale is $[0, rate]$, and $Scale$ is the truth value of skeleton pixel scale in the image. When the magnification of $rate$ sets as '1', the program achieves normalization, while we set it as '2' in experiment, which is the twice of the normalization and it is used to refine scale evaluation. $Pred\_Scale$ is the predicted scale in the image, and the standard Euclidean loss function of (9) is adopted to calculate the regression loss.

$$Loss = \sum_{i=1}^{n} (Pred\_Scale - GT\_Scale)^2 \tag{9}$$

2) SKELETON PIXEL SCALE PREDICTION

Assuming that i represents the category of skeleton pixel, and referring to Table 2 the size of receptive field is marked as $Field\_Size_i$. From the skeleton regression task, the scale of a certain skeleton pixel is predicted with (10).

$$\widetilde{Pred\_Scale} = Pred\_Scale \times \frac{Field\_Size_i}{rate}. \tag{10}$$

where $\widetilde{Pred\_Scale}$ is the predicted scale value and $rate$ is the magnification.

$$Final\_Scale = \begin{cases} \widetilde{Pred\_Scale}, \\ \qquad \widetilde{Pred\_Scale} \in Coarse\_Scale \\ \dfrac{Coarse\_Scale_{low} + Coarse\_Scale_{high}}{2}, \\ \qquad \widetilde{Pred\_Scale} \notin Coarse\_Scale. \end{cases} \tag{11}$$

The skeleton pixel classification method is introduced in section III-D, and the coarse range of pixel scales is $[Coarse\_Scale_{low}, Coarse\_Scale_{high})$. For instance, if a pixel belongs to category 2, that means the scale of this pixel ranges at $[10, 26)$, where the $Coarse\_Scale_{low}$ value is 10 and the $Coarse\_Scale_{high}$ value is 26. Then we can calculate the final skeleton scale $Final\_Scale$ according to (11).

IV. EXPERIMENTAL RESULTS

In order to verify the rationality of the new proposed method, sufficient experiments are carried out comparing with other related algorithms, such as FSDS [11] and Hi-Fi [12] methods. The procedures of skeleton extraction method based on neural network includes training and testing phases, and different datasets are chosen to make sufficient analyses. To be fair, the experimental results of all methods are produced under the same condition (i.e., the same server and the same type of GPU). In addition, the FSDS and Hi-Fi methods are reproduced using the source code provided by the original authors in experiment.

A. DATASETS AND IMPLEMENTATION DETAILS

The network architecture of the new method is important in experiment which has been introduced in the methodology section. Before training, the parameters of the network need to be initialized, and the hyper parameters in the training model are listed in Table 3. In the table, *mini_batch* represents the number of input images, and the default value of it sets 1; *Base_lr* represents basic learning rate; *weight_decay* represents weight attenuation; *momentum* is the momentum in the training model; *max_iter* is the maximum number of iterations; and *step_size* means that for every 20,000 times iteration, the learning rate is multiplied by 0.1. Besides, data augmentation is an important way to generate sufficient training samples and to improve the generalization ability of the model, while the data augmentation method of our method is the same as [14].

TABLE 3. Initial parameters in the training phase.

Properties	Values
$mini\_batch$	1
$Base\_lr$	1e-6
$weight\_decay$	0.0002
$momentum$	0.9
max_iter	40000
$step\_size$	20000

In aspect of experimental datasets, three public datasets are adopted, and they are SK-SMALL [11], SK-LARGE [14], and WH-SYMMAX [15]. The SK-LARGE is a public available skeleton detection dataset, which contains 746 training images, 745 test images, and their corresponding skeleton ground-truth. The SK-SMALL dataset contains 506 images, of which the first 300 are used for training and the last 206 are used for testing, and it is also a subset of SK-LARGE. The WH-SYMMAX dataset contains 328 horse images, and the first 228 are used for training while the last 100 are used for testing. Then, the GPU configuration is Tian XP in the operation condition.

B. EVALUATION PROTOCOL

There are two normal evaluations to measure the skeleton extraction performance, that's *F-measure* scores and the PR curves. This paper uses both of them to indicate the performance of different methods. The formulation of $F-measure$ are shown in (12).

$$F-measure = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{12}$$

where, the *Precision* and *Recall* are calculated with extracted skeleton features and the ground-truth. The PR curves are calculated in the following way: firstly, the extracted feature map is transformed into a binary map with a threshold, and then it is matched with the skeleton ground-truth allowing only small position errors during the matching procedure, that means rarely position offset is enabled between the extracted skeleton and the ground-truth. Secondly, setting the extracted skeleton pixel as a true-positive point when it matches at least one point of the ground-truth. Then if the extracted skeleton pixel can't match any ground-truth pixel, that will be signed as a false-positive point. At last, by using different thresholds, a series of *Precision* and *Recall* values can be got, and then the PR curve is generated by combining all of them.

C. OBJECT SKELETON EXTRACTION

Hi-Fi is the state-of-the-art skeleton extraction method that has been proposed at current, so it is an important target which is chosen to compare with our method. For the implementation of Hi-Fi and FSDS methods, the source programs and parameters provided by [11], [12] are used in training and testing. The tested images are processed with the standard non-maximum suppression (NMS) algorithm [48] to refine skeleton maps, which are evaluated at last. In addition, three datasets are selected to perform comparison between different methods, and the generalization ability of the new method is verified by across dataset experiments.

TABLE 4. The F-measures of different skeleton extraction methods on three datasets.

Datasets Methods	SK-LARGE	SK-SMALL	WH-SYMMAX
FSDS	0.6330	0.6230	0.7684
Hi-Fi-1	0.6565	0.6404	0.7882
Ours-1	0.6691	0.6396	0.7872
Ours-1-Reg	0.6724	0.6455	0.7923
Hi-Fi-2	0.6878	0.6519	0.8023
Ours-2	0.6869	0.6592	0.8056
Ours-2-Reg	**0.6922**	0.6604	0.8054
Ours-3	0.6861	0.6618	0.8063
Ours-3-Reg	0.6871	**0.6630**	0.8111
Ours-4	0.6775	0.6573	**0.8120**
Ours-4-Reg	-	0.6601	-

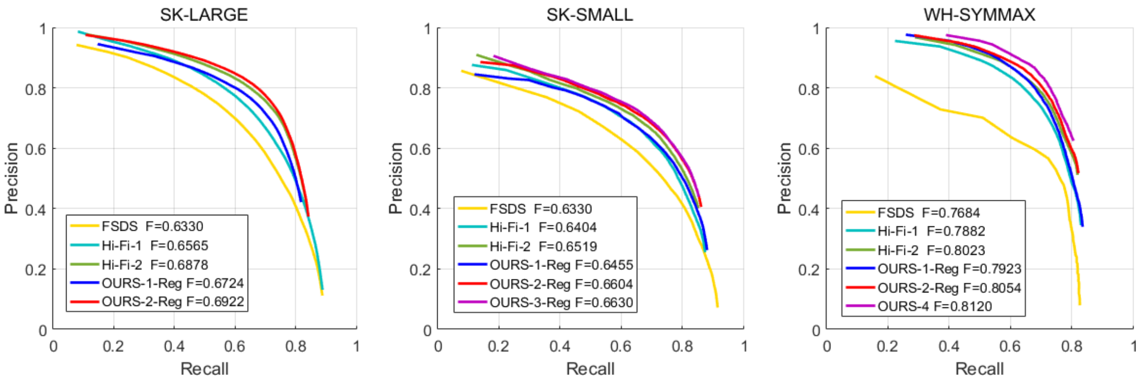
TABLE 5. The percentage of increase in the accuracy of skeleton localization of our method compared to Hi-Fi at the first two hierarchies (the negative sign indicates a decrease).

Datasets Methods	SK-LARGE	SK-SMALL	WH-SYMMAX
Ours-1(Hi-Fi-1)	+1.25%	-0.08%	-0.10%
Ours-1-Reg(Hi-Fi-1)	+1.60%	+0.50%	+0.40%
Ours-2(Hi-Fi-2)	-0.09%	+0.70%	+0.33%
Ours-2-Reg(Hi-Fi-2)	+0.44%	+0.85%	+0.30%

1) COMPARISION WITH COMMON DATASET

Although Hi-Fi network has four hierarchies (i.e., H-1, H-2, H-3, and H-4), limited by the memory of machine, in original paper Hi-Fi method only implements the first two hierarchies in experiments, that's Hi-Fi-1 and Hi-Fi-2 methods. In this paper we implement it on four hierarchies. Moreover, in order to verify the effect of the variable coefficient loss function of new proposed method, we conduct methods of Ours-1, Ours-2, Ours-3, and Ours-4 on four hierarchies independently. The Ours-1-Reg to Ours-4-Reg represent the methods with the new loss function and the regression task. While all the new methods have hierarchical network structure. The F-measures of skeleton extraction results in the experiments are shown in Table 4.

Analyzing the experiment results from Table 4, the Hi-Fi method and ours method have higher F-measures than FSDS in experiments on three datasets. Specifically, for the methods with the 1st-level hierarchical feature integration, ours-1-reg method works best on SK-LARGE, SK-SMALL, and WH-SYMMAX datasets. In the case of second-level



(a) (b) (c)

FIGURE 9. Skeleton extraction results with PR curves in three datasets. (a) Skeleton localization evaluation on SK-LARGE [14] dataset, and the OURS-2-Reg achieves the best performance of all. (b) Skeleton localization evaluation on SK-SMALL [11] dataset, and the OURS-3-Reg achieves the best performance of all. (c) Skeleton localization evaluation on WH-SYMMAX [15] dataset, and the OURS-4-Reg achieves the best performance of all.

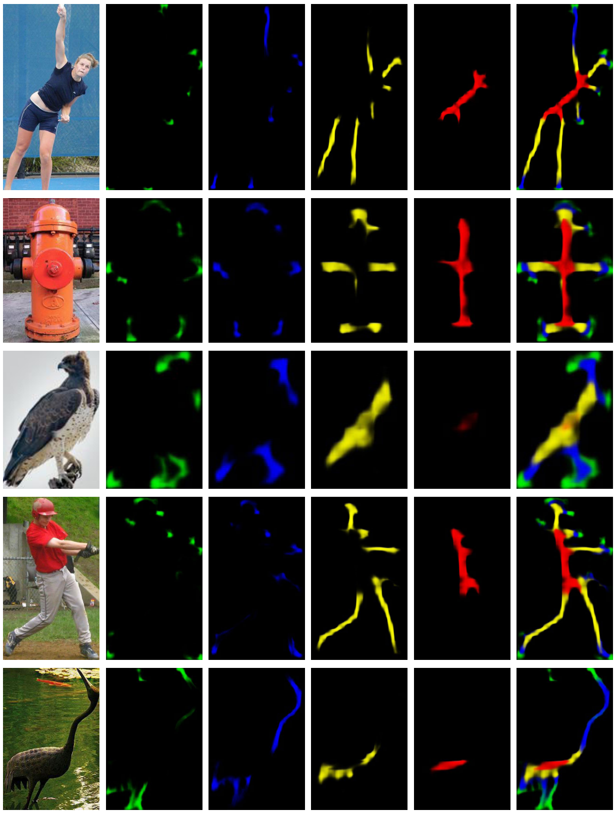


FIGURE 10. Illustration of ours-2-reg skeleton extraction method on SK-LARGE [14] with five selected images. In every row, the first is the original natural image, and the last is the object skeleton map with the classification of all scales, the middle four images are different scales marked with color green, blue, yellow, and red represent classification one, two, three, four, and five.

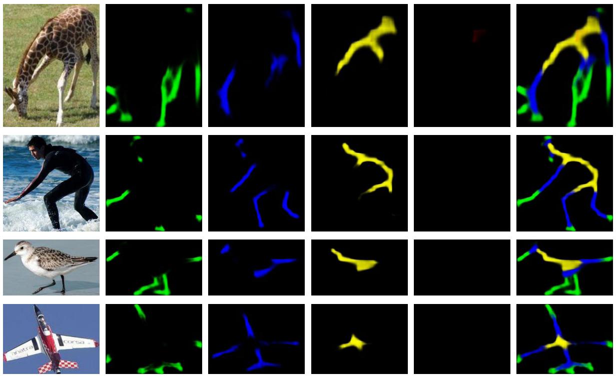


FIGURE 11. Illustration of ours-3-reg skeleton extraction method on SK-SMALL [11] with five selected images. In every row, the first is the original natural image, and the last is the object skeleton map with the classification of all scales, the middle four images are different scales marked with color green, blue, yellow, and red represent classification one, two, three, four, and five.

TABLE 6. The F-measures of different skeleton extraction methods on cross dataset generalization.

Datasets Methods	SK-LARGE /WH-SYMMAX	WH-SYMMAX /SK-LARGE
FSDS	0.6850	0.4400
Hi-Fi-1	0.7095	0.4949
Ours-1	**0.7235**	0.4892
Ours-1-Reg	0.7143	**0.4972**
Hi-Fi-2	0.7271	0.4915
Ours-2	**0.7324**	0.4936
Ours-2-Reg	0.7285	0.4937

of hierarchical feature integration, ours-2-reg method works mostly the best on those three datasets. Therefore, the method added with the new variable coefficient loss function and regression outperforms other compared methods. Besides, through the Tables 4 and 5, the new method with variable coefficient loss function performs slightly better than Hi-Fi, while the method with both new loss function and regression task improves the extraction performance of Hi-Fi entirety. Furthermore, the Hi-Fi does not test the feature integration mechanism on the third-level and the fourth-level hierarchies with memory limitation. While the experiments

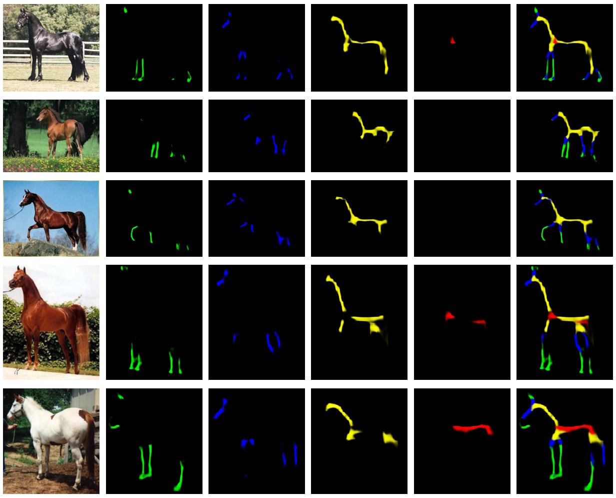


FIGURE 12. Illustration of ours-4 skeleton extraction method with WH-SYMMAX [15] for five selected images. In every row, the first is the original natural image, and the last is the object skeleton map with the classification of all scales, the middle four images are different scales marked with color green, blue, yellow, and red represent classification one, two, three, four, and five.

TABLE 7. The percentage of increase in the accuracy of skeleton localization of our method compared to Hi-Fi on cross dataset generalization (the negative sign indicates a decrease).

Datasets Methods	SK-LARGE /WH-SYMMAX	WH-SYMMAX /SK-LARGE
Ours-1(Hi-Fi-1)	+1.40%	-0.50%
Ours-1-Reg(Hi-Fi-1)	+0.53%	+0.20%
Ours-2(Hi-Fi-2)	+0.50%	+0.20%
Ours-2-Reg(Hi-Fi-2)	+0.13%	+0.22%

indicating that the higher the level, the better effect does not always occur, such as the experiments on SK-LARGE dataset, the performance of Ours-2-Reg is the best, while Ours-3, Ours-3-Reg, Ours-4, and Ours-4-Reg methods are not as good as Ours-2-Reg. The sign '-' in Table 4 represents that there is no way to verify the results due to the memory limitation of hardware.

Some experimental results of ours-2-reg method on SK-LARGE dataset, ours-3-reg method on SK-SMALL dataset, and ours-4 method on WH-SYMMAX dataset are illustrated in Figs. 10, 11 and 12 respectively. The methods selected are the best of all on different datasets, and the figures illustrated in the first row are the original images, the skeleton pixels with four kind of scales are drawing in different colors, and the final skeleton maps of objects have integrate scales. From the PR curves in Fig. 9, it can be found that the performance of the new methods on three datasets are better than Hi-Fi methods, although there is no unique method to achieve the best results on all three datasets simultaneously, the best performance can always be found in the new methods. The specific result is that the ours-1-reg method is better than the Hi-Fi-1 method, and the ours-2-reg method is better than the Hi-Fi-2 method, that means the new proposed method is better than Hi-Fi.

2) COMPARISION WITH CROSS DATASET

To further verify the generalization capabilities of the proposed model, we perform cross-validation on two different datasets (A/B indicates training on the dataset A and testing on the dataset B). It can be analyzed from the Tables 6 and 7 that the generalization ability of the new method is better than Hi-Fi.

D. FAILURE CASE EXPLORATION

The training dataset of SK-LARGE contains 745 images, which is more than SK-SMALL and WH-SYMMAX

TABLE 8. The top ten of F-measures in SK-LARGE dataset.

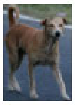
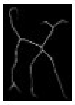
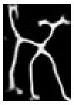
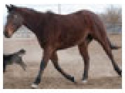
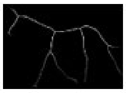
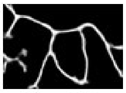
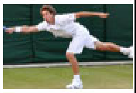
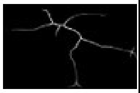
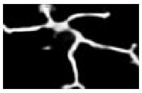
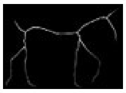
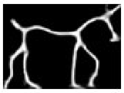
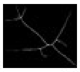
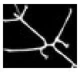
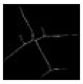
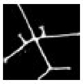
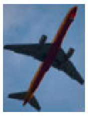
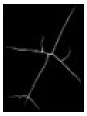
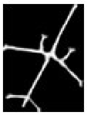
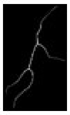
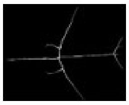
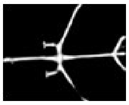
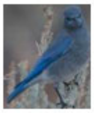
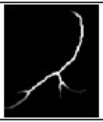
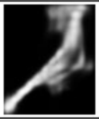
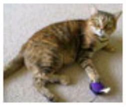
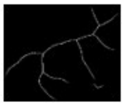
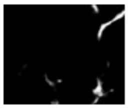
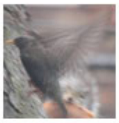
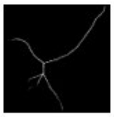
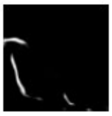
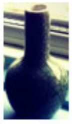
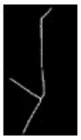
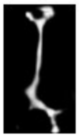
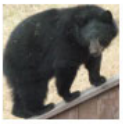
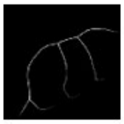
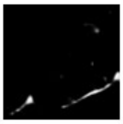
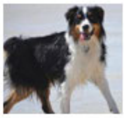
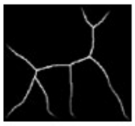
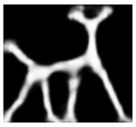
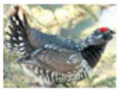
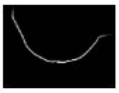
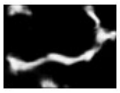
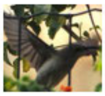
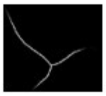
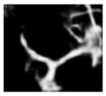
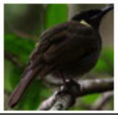
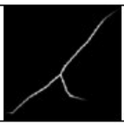
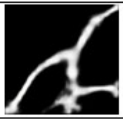
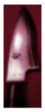
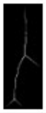
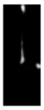
SN	Origin image	Ground Truth	Output skeleton	F-Measure	SN	Origin image	Ground Truth	Output skeleton	F-Measure
1				0.9419	6				0.9481
2				0.9425	7				0.9486
3				0.9444	8				0.9523
4				0.9444	9				0.9562
5				0.9457	10				0.9577

TABLE 9. The last ten of F-measures in SK-LARGE dataset.

SN	Origin image	Ground Truth	Output skeleton	F-Measure	SN	Origin image	Ground Truth	Output skeleton	F-Measure
1				0.1181	6				0.2685
2				0.2007	7				0.2727
3				0.2205	8				0.2738
4				0.2580	9				0.2796
5				0.2654	10				0.2828

datasets, so it is selected to discuss the failure cases. The top ten and the last ten of F-measures are selected out to analyze the extraction performance. When the background is relatively simple or the number of the same kind of images is large in the training set (such as 'aircraft', the proportion of it in the training set is 35/746), the skeleton extraction effect is better. Relatively, when the probability of such image is low in the training set (such as the 10*th* knife-shaped figure in Table 9), the skeleton extraction result is poor. In addition, when the background is complicated or the image is more ambiguous, the skeleton extraction effect is not good (such as the first image and the second image in Table 9). In summary, in order to get better skeleton extraction results, it is necessary to have a large enough training dataset, and the samples are accurately labeled.

V. CONCLUSION

This paper proposes an advanced skeleton extraction method with multi-task and variable coefficient loss function handling the multi-class imbalanced data problem. The significant contributions demonstrate in four aspects: proposing a new strategy for object skeleton extraction in natural images; a new Softmax function is presented to deal with multi-class problem; rearranging the skeleton pixel classification manner; putting forward a new regression prediction method for skeleton pixel. The experimental results show that the new method proposed in this paper is superior to the best Hi-Fi method in terms of skeleton extraction accuracy and generalization ability, and our method can accurately calculate the skeleton scales. In addition, we fully introduce the skeleton extraction method in natural images using deep learning, as well as the skeleton pixel and scale classification methods.

REFERENCES

[1] M. Ehatisham-Ul-Haq, A. Javed, M. A. Azam, H. M. A. Malik, A. Irtaza, I. H. Lee, and M. T. Mahmood, "Robust human activity recognition using multimodal feature-level fusion," *IEEE Access*, vol. 7, pp. 60736–60751, 2019.

[2] G. Batchuluun, D. T. Nguyen, T. D. Pham, C. Park, and K. R. Park, "Action recognition from thermal videos," *IEEE Access*, vol. 7, pp. 103893–103917, 2019.

[3] Y. Yang, C. Deng, D. Tao, S. Zhang, W. Liu, and X. Gao, "Latent max-margin multitask learning with skelets for 3-D action recognition," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 439–448, Feb. 2017.

[4] H. El-Ghaish, M. E. Hussein, A. Shoukry, and R. Onai, "Human action recognition based on integrating body pose, part shape, and motion," *IEEE Access*, vol. 6, pp. 49040–49055, 2018.

[5] L. Song, W. Lin, Y.-G. Yang, X. Zhu, Q. Guo, and J. Xi, "Weak micro-scratch detection based on deep convolutional neural network," *IEEE Access*, vol. 7, pp. 27547–27554, 2019.

[6] A. B. Spanier, N. Caplan, J. Sosna, B. Acar, and L. Joskowicz, "A fully automatic end-to-end method for content-based image retrieval of CT scans with similar liver lesion annotations," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 1, pp. 165–174, 2018.

[7] X. Zhang, S. Wang, Z. Li, and S. Ma, "Landmark image retrieval by jointing feature refinement and multimodal classifier learning," *IEEE Trans. Cybern.*, vol. 48, no. 6, pp. 1682–1695, Jun. 2018.

[8] J. Hou, G. Wang, X. Chen, J.-H. Xue, R. Zhu, and H. Yang, "Spatial-temporal attention res-TCN for skeleton-based dynamic hand gesture recognition," in *Lecture Notes in Computer Science* (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11134. Springer, 2019, pp. 273–286.

[9] A. Sain, A. K. Bhunia, P. P. Roy, and U. Pal, "Multi-oriented text detection and verification in video frames and scene images," *Neurocomputing*, vol. 275, pp. 1531–1549, Jan. 2018.

[10] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 3, pp. 807–811, Mar. 2018.

[11] W. Shen, K. Zhao, Y. Jiang, Y. Wang, Z. Zhang, and X. Bai, ''Object skeleton extraction in natural images by fusing scale-associated deep side outputs,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 222–230.

[12] K. Zhao, W. Shen, S. Gao, D. Li, and M.-M. Cheng, ''Hi-Fi: Hierarchical feature integration for skeleton detection,'' Jan. 2018, *arXiv:1801.01849*. [Online]. Available: https://arxiv.org/abs/1801.01849

[13] Y. Liu, M.-M. Cheng, X. Hu, K. Wang, and X. Bai, ''Richer convolutional features for edge detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5872–5881.

[14] W. Shen, K. Zhao, Y. Jiang, Y. Wang, X. Bai, and A. Yuille, ''Deepskeleton: Learning multi-task scale-associated deep side outputs for object skeleton extraction in natural images,'' *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5298–5311, Nov. 2017.

[15] W. Shen, X. Bai, Z. Hu, and Z. Zhang, ''Multiple instance subspace learning via partial random projection tree for local reflection symmetry in natural images,'' *Pattern Recognit.*, vol. 52, pp. 306–316, Apr. 2016.

[16] K. Saeed, M. Tabędzki, M. Rybnik, and M. Adamski, ''K3M: A universal algorithm for image skeletonization and a review of thinning techniques,'' *Int. J. Appl. Math. Comput. Sci.*, vol. 20, no. 2, pp. 317–335, 2010.

[17] Z. Yu and C. Bajaj, ''A segmentation-free approach for skeletonization of gray-scale images via anisotropic vector diffusion,'' in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun./Jul. 2004, pp. 1–6.

[18] S. Schlüter, A. Sheppard, K. Brown, and D. Wildenschild, ''Image processing of multiphase images obtained via X-ray microtomography: A review,'' *Water Resour. Res.*, vol. 50, no. 4, pp. 3615–3639, 2014.

[19] M. Hisada, A. G. Belyaev, and T. L. Kunii, ''A skeleton-based approach for detection of perceptually salient features on polygonal surfaces,'' *Comput. Graph. Forum*, vol. 21, no. 4, pp. 689–700, 2002.

[20] D. Ziou and S. Tabbone, ''Edge detection techniques—An overview,'' *Pattern Recognit. Image Anal. C/C Raspoznavaniye Obrazov I Analiz Izobrazhenii*, vol. 8, pp. 537–559, Dec. 1998.

[21] N. K. Ratha, S. Chen, and A. K. Jain, ''Adaptive flow orientation-based feature extraction in fingerprint images,'' *Pattern Recognit.*, vol. 28, no. 11, pp. 1657–1672, 1995.

[22] D. Maio and D. Maltoni, ''Direct gray-scale minutiae detection in fingerprints,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 1, pp. 27–40, Jan. 1997.

[23] W. Shen, X. Bai, X. Yang, and L. J. Latecki, ''Skeleton pruning as trade-off between skeleton simplicity and reconstruction error,'' *Sci. China Inf. Sci.*, vol. 56, no. 4, pp. 1–14, 2013.

[24] X. Bai, X. Wang, L. J. Latecki, W. Liu, and Z. Tu, ''Active skeleton for non-rigid object detection,'' in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 575–582.

[25] S. Tsogkas and I. Kokkinos, ''Learning-based symmetry detection in natural images,'' in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 41–54.

[26] N. Widynski, A. Moevus, and M. Mignotte, ''Local symmetry detection in natural images using a particle filtering approach,'' *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5309–5322, Dec. 2014.

[27] C. L. Teo, C. Fermüller, and Y. Aloimonos, ''Detection and segmentation of 2D curved reflection symmetric structures,'' in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1644–1652.

[28] F. Valeur, D. Mutz, and G. Vigna, ''A learning-based approach to the detection of SQL attacks,'' in *Proc. Int. Conf. Detection Intrusions Malware, Vulnerability Assessment*. Berlin, Germany: Springer, 2005, pp. 123–140.

[29] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, ''Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks,'' in *Proc. AAAI*, 2016, vol. 2, no. 5, pp. 3697–3703.

[30] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras, ''Two-person interaction detection using body-pose features and multiple instance learning,'' in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2012, pp. 28–35.

[31] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, ''A new representation of skeleton sequences for 3D action recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4570–4579.

[32] H. Noh, S. Hong, and B. Han, ''Learning deconvolution network for semantic segmentation,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[33] B. Li, M. He, Y. Dai, X. Cheng, and Y. Chen, ''3D skeleton based action recognition by video-domain translation-scale invariant mapping and multi-scale dilated CNN,'' *Multimedia Tools Appl.*, vol. 77, no. 17, pp. 22901–22921, 2018.

[34] J. Long, E. Shelhamer, and T. Darrell, ''Fully convolutional networks for semantic segmentation,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[35] S. Xie and Z. Tu, ''Holistically-nested edge detection,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.

[36] W. Ke, J. Chen, J. Jiao, G. Zhao, and Q. Ye, ''SRN: Side-output residual network for object symmetry detection in the wild,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 302–310.

[37] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool, ''Convolutional oriented boundaries: From image segmentation to high-level tasks,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 819–833, Jan. 2018.

[38] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, ''Beyond a Gaussian Denoiser: Residual learning of deep CNN for image denoising,'' *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142–3155, Jul. 2017.

[39] B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, ''Classification of mass and normal breast tissue: A convolution neural network classifier with spatial domain and texture images,'' *IEEE Trans. Med. Imag.*, vol. 15, no. 5, pp. 598–610, Oct. 1996.

[40] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, ''Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning,'' *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[41] K. Simonyan and A. Zisserman, ''Very deep convolutional networks for large-scale image recognition,'' Sep. 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[42] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, ''DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[43] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, ''TextBoxes: A fast text detector with a single deep neural network,'' in *Proc. AAAI*, 2017, pp. 4161–4167.

[44] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, ''HCP: A flexible CNN framework for multi-label image classification,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1901–1907, Sep. 2016.

[45] s9xie. *HED*. Accessed: Mar. 16, 2016. [Online]. Available: https://github.com/s9xie/hed/blob/master/src/caffe/layers/sigmoid_cross_entropy_loss_layer.cpp

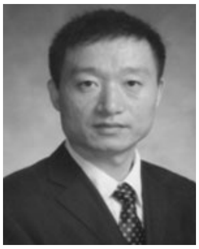
[46] K. Zhao. *DeepSkeleton*. Accessed: Oct. 9, 2017. [Online]. Available: https://github.com/zeakey/DeepSkeleton/blob/master/src/caffe/layers/softmax_loss_layer.cpp

[47] K. Zhao. *Skeleton*. Accessed: Oct. 9, 2017. [Online]. Available: https://github.com/zeakey/skeleton/blob/master/caffe/src/caffe/layers/softmax_loss_layer.cpp

[48] A. Neubeck and L. Van Gool, ''Efficient non-maximum suppression,'' in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 3. Aug. 2006, pp. 850–855.

YOUQING XIAO received the M.S. degree in computer software and theory from Sun Yat-sen University, China, in 2005. He is currently pursuing the Ph.D. degree with the Macau University of Science and Technology, Macau, China.

He worked at The Hong Kong Polytech University as a Research Assistant, from 2007 to 2008. He is currently working on the technology for autonomous vehicles with the Guangzhou Automobile Industry Research Institute (GAEI) and serves as a Lead Algorithm Engineer. His research interests include image processing, machine learning, and video-based perception problems.



ZHANCHUAN CAI (M'16–SM'19) received the Ph.D. degree from Sun Yat-sen University, Guangzhou, China, in 2007.

From 2007 to 2008, he was a Visiting Scholar with the University of Nevada at Las Vegas, Las Vegas, NV, USA. He is currently a Professor with the Faculty of Information Technology, Macau University of Science and Technology, Taipa, Macau, China, where he is also with the State Key Laboratory of Lunar and Planetary Sciences. His research interests include intelligent information processing, image processing, and computer graphics.

Prof. Cai is a member of the ACM and Chang'e-3 Scientific Data Research and Application Core Team and the Asia Graphics Association. He is also a Senior Member of the CCF. He was a recipient of the Third Prize of the Macau Science and Technology Award-Natural Science Award, in 2012, the BOC Excellent Research Award from the Macau University of Science and Technology, in 2016, and the Third Prize of the Macau Science and Technology Award-Technological Invention Award, in 2018.

XIXI YUAN received the B.S. degree in software engineering from the Beijing Information Science and Technology University, in 2016, and the M.S. degree in information technology from the Macau University of Science and Technology, in 2018, where she is currently pursuing the Ph.D. degree with the Faculty of Information Technology.

Her research interests include image processing and computer graphics.