# PC-SuperPoint: interest point detection and descriptor extraction using pyramid convolution and circle loss

**Yu-Jie Xiong**[a,b,*] **Shuo Ma**[a] **Yongbin Gao**[a] **and Zhijun Fang**[a]

[a]Shanghai University of Engineering Science, School of Electronic and Electrical Engineering, Shanghai, China

[b]East China Normal University, Shanghai Key Laboratory of Multidimensional Information Processing, Shanghai, China

**Abstract.** Nowadays, deep learning is widely used to detect interest points and extract the corresponding descriptors and achieved suitable results for many applications of computer vision, such as image matching, three-dimensional reconstruction, simultaneous localization, and mapping. We propose an approach for interest point detection and descriptor extraction using pyramid convolution and circle loss, which is named as PC-SuperPoint. We utilize pyramid convolutions in the backbone network, which includes convolution kernels of different scales for multiscale feature extraction. The following well-designed networks are able to capture the local and global information from the obtained backbone feature maps. In addition, circle loss, which enhances weight attributes for each pair of descriptors, is also applied to improve the convergence speed in the training phase. Experiments on the HPatches dataset and KITTI dataset achieve promising results, which reveal the effectiveness of the proposed method.
 [DOI: 10.1117/1.JEI.30.3.033024]

**Keywords:** interest point detection; descriptor extraction; pyramid convolution; circle loss.

Paper 200735 received Oct. 28, 2020; accepted for publication Jun. 8, 2021; published online Jun. 26, 2021.

## 1 Introduction

Various computer vision applications rely on the accurate detection of interest points, such as image retrieval, simultaneous localization and mapping, and structure from motion (SfM). As a result, it is very important to calculate the correspondence between different images. Since the varieties of viewpoint and illumination, it is necessary to design a robust interest point detection and descriptor extraction model. The current researches can be divided into traditional methods and deep learning-based methods. Early study focused on handcrafted detector and descriptors, such as SIFT,[1] SURF,[2] and ORB.[3] These handcrafted descriptors are convenient but are easily affected by improper hyperparameters and unexpected image transformations.

Since the emergence of AlexNet[4] in 2012, deep learning-based methods are widely used in image recognition tasks. It is well known that the classification and recognition rates of deep learning have reached to human-level performance.[5–7] Deep learning methods also improved the performance of traditional methods in geometric computer vision,[8] such as pose estimation,[9,10] homography estimation,[11] stereo matching,[12] and visual ranging.[13] The study of interest point detection and descriptor extraction based on deep learning, especially the end-to-end method, gradually replaced the traditional methods, and achieved great results in some general applications.

The motivation of the proposed method is that we find pyramid convolutions are able to capture the local and global information at the same time. On the one hand, this characteristic is useful for fast and accurate point detection. On the other hand, it is also meaningful to describe the texture features of interest regions. An approach for interest point detection and descriptor extraction using pyramid convolution and circle loss is proposed in this paper. We employ pyramid convolution[14] with multiscale convolution kernels in the backbone. Pyramid convolution is

*Address all correspondence to Yu-Jie Xiong, xiong@sues.edu.cn

1017-9909/2021/$28.00 

used to extract global and local information. Meanwhile, multiscale convolution kernels expand the scope of receptive field, as to capture the different levels of detailed information. Spatial relationships of detected points are explored to avoid ambiguity of local information. We also design a flexible optimization approach using circle loss[15] to shorten the training time. The proposed method is extensively evaluated and achieved promising results. Comparing with the several present point detectors, the proposed method detects more significant points using less time in most cases. Moreover, the proposed descriptors are trainable. We can obtain suitable descriptors for different tasks using diverse training data. The three contributions of this paper are

- Multiscale convolution kernels are used to replace the standard single convolution kernel for obtaining multiscale feature maps of the image.
- Pyramid convolution that effectively captures global and local information are employed.
- Circle loss is employed to add weight attributes to each pair of descriptors for fast convergence during the training phase.

## 2 Related Work

In recent years, a lot of methods[16–22] are presented to extract the distinctive descriptors of the given regions of interest using convolutional neural networks. They use VGG,[23] GoogleNet,[24] and ResNet[25] as the backbone network to extract image features but pay little attention to detection of interest points. MatchNet[16] and DeepCompare[17] train the distance metric using a Siamese networks architecture and use it for descriptor extraction. DeepDesc[18] explores discriminative features from hard positive and negative samples. L2-Net[19] utilizes the basic concept of matching to calculate the descriptors. For a local patch, nearest-neighbor search is performed in the feature space to find the matching elements. At the same time, a progressive sampling is used to improve the training efficiency. Thus, L2-Net ensures that the extracted descriptors of the same patch are the closest neighbors of each other but does not care about the characteristics of the local details. GeoDesc[20] proposes a local descriptor learning approach that integrates geometry constraints from multiview reconstructions, which benefits the learning process in terms of data generation, data sampling, and loss computation. GeoDesc demonstrates its superior performance on various large-scale benchmarks and, in particular, shows its great success on challenging reconstruction tasks. ContextDesc[21] proposes a unified learning framework that leverages and aggregates the cross-modality contextual information and focuses on the spatial relationship of keypoints by introducing context awareness into local descriptors. LogPol[22] is a good work about local descriptors and achieves SOTA performance. However, interest point detection is not included in LogPol, and it uses SIFT to find keypoints and determine correspondences among them using the ground truth homographies.

Inspired by the great success of the descriptor extraction algorithms, there are several studies[26–29] that attempt to exploit the potential applications of keypoint detection. When interest point detection and descriptors extraction could share most of the computational resource, calculation efficiency will be improved immediately. SuperPoint[26] can predict points of interest and extract the corresponding descriptors at the same time. SuperPoint is trained using synthetic images of simple generated geometric shapes. Angles, intersections, spots, and line segments of the synthetic data are defined as pseudo ground truth of keypoints. The model is first trained on the synthetic data, then tested on the real images to generate the pseudo ground truth of keypoints by summarizing the predictions of 100 different homography transformations for each image. Key.net[27] is the research of detecting interest points, which combines handcrafted and learned convolutional neural network filters. Handcrafted filters provide anchor structures for learned filters, which localize, score, and rank repeatable features. Scale-space representation is used within the networks to extract keypoints at different levels. In the stage of descriptor extraction, the loss function plays an important role. SuperPoint adopts hinge loss for training and adds some hyperparameters to solve the problem of sample unbalance. Key.net designs a loss function to detect robust features across a range of scales and to maximize the repeatability score. Vassileios et al.[28,29] compared the effects of various samples with triplet loss.

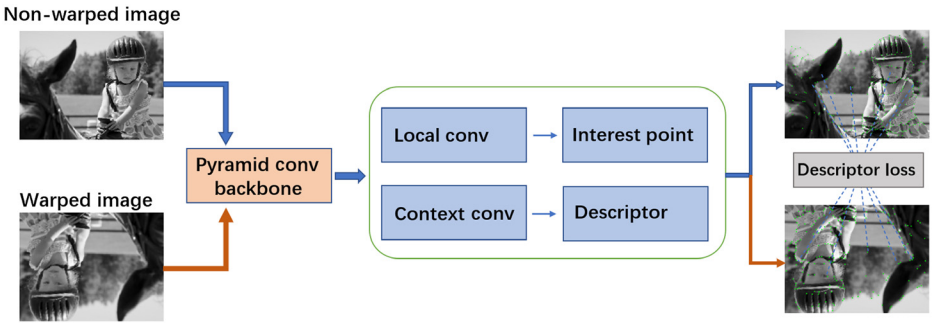



**Fig. 1** The architecture of the proposed method.

## 3 Proposed Method

We build a multitask network architecture using the convolutional neural networks, which detects the interest points and extracts fixed-length descriptors to complete the description of interest points. The architecture is shown in Fig. 1. The model employs a convolutional neural networks as an encoder to reduce the dimension and extract feature maps of the image. After that, the feature maps are used to process the detection of interest points and the description of interest points, respectively.

### 3.1 *Pyramid Convolution Backbone*

Our pyramid convolution backbone network is similar to VGG,[23] which includes four sets of convolutional layers. Instead of the original 3*3 convolution kernels, pyramid convolutions are used to adapt to multiscale feature information extraction. Our proposed PC-SuperPoint uses 3*3, 5*5, 7*7, and 9*9 convolution kernels of different sizes between convolutional layers to replace the original single-size convolution kernels and keep their total number of channels unchanged. The combination of different scale convolution kernels is called pyramid convolution. Figure 2 shows the structure of pyramid convolution. The pyramid convolution kernel inserts each set of convolutions. The convolutional layers perform spatial downsampling via pooling and activation functions. Each pooling operation reduces the width and height by one time. After three pooling operations, the image is reduced by eight times. For the image size $H * W$, the networks will return $H/8 * W/8$ pixel cells and define that $H_s = H/8$ and $W_s = W/8$. The next network will then proceed to each cell to obtain the position and descriptors of interest points.

### 3.2 *Interest Point Detection*

The network of interest point detection is shown in Fig. 3. The goal of interest point detection network is to calculate the probability of interest point for each pixel in the image. The primary network for extracting feature maps is a shared encoder that performs downsampling by max pooling operations to reduce spatial dimension. After local pyramid convolution module (LocalPyConv), which is composed of pyramid convolution with 256 channels takes the output feature maps from the backbone. Then, it applies a 1*1 convolution kernel to combine the

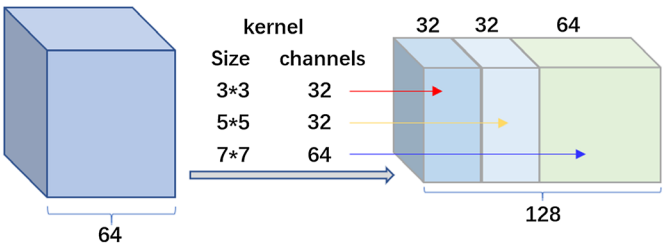



**Fig. 2** Pyramid convolution structure.

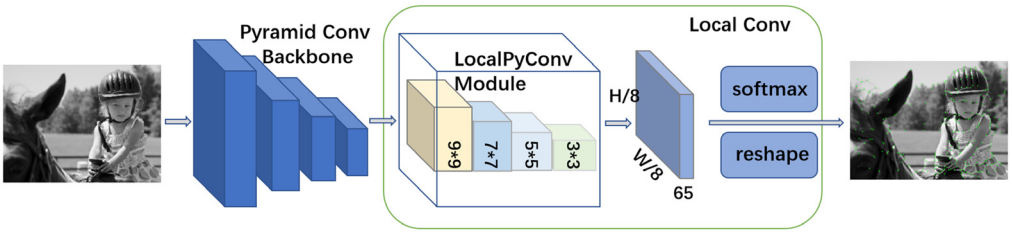



**Fig. 3** Interest point detection networks.


**Fig. 4** Local pyramid convolution module.

information extracted at different kernel sizes. The detailed information about each component of LocalPyConv is shown in Fig. 4. The feature size calculated by the interest point detection is $R^{H_s*W_s*65}$, and the output tensor is $R^{H*W}$. The 65 channels correspond to local and nonrepetitive 8*8 grid areas of the pixel and an additional placeholder for uninteresting points. After a softmax function, the irrelevant placeholder in the memory is removed, resulting in a feature of $R^{H_s*W_s*64}$ dimension. After the reshaping, an $R^{H*W}$ image with the annotation of interest points is obtained. Then, nonmaximum suppression is used for local optimal search to avoid local dense set.

## 3.3 *Descriptor Extraction*

The descriptor extraction network as shown in Fig. 5 calculates the feature of $R^{H_s*W_s*D}$. First, pyramid convolution is used to extract multiscale features. Then, extract the local feature information and global feature information and merge them together. The final output descriptor is a semidense grid[30] with a tensor of size $R^{H*W*D}$. In this module, local and global feature information can be extracted through specially designed network.

The first part is the LocalPyConv, which is mainly responsible for the extraction of detailed information by applying different sizes of convolution kernel. LocalPyConv obtains the output-image from the backbone and then performs four levels of pyramid convolutions to obtain different local details in four scales of kernel 3*3, 5*5, 7*7, and 9*9. Finally, a 1*1 convolution is performed to combine the information extracted under different kernel sizes.

The second part is global pyramid convolution module (GlobalPyConv), which is responsible for capturing global information. The components of GlobalPyConv are represented in Fig. 6. As the size of the input image is variable, to ensure that the full global information can be captured, feature maps keep the largest spatial size dimension as 9. To this end, the adaptive pooling operation on average was used to decrease the space size of the feature map to 9*9,

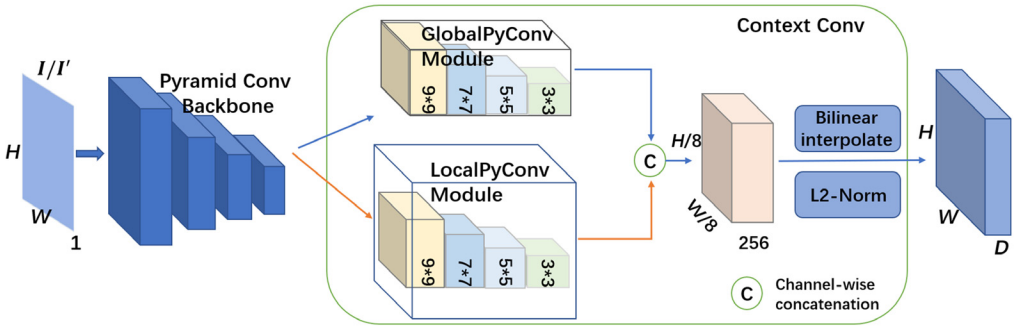



**Fig. 5** Descriptor extraction networks.

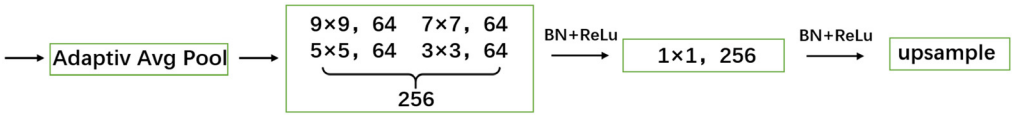



**Fig. 6** Global pyramid convolution module.

which still maintains a reasonable spatial resolution. Then, we use the pyramid convolution kernel to perform convolution operations similar to LocalPyConv. Since the space size of the feature map was transformed to 9*9, the four-layer pyramid convolutions with a maximum convolution kernel of 9*9 can cover all output feature maps of the backbone to capture global information. A 1*1 transformation as well is needed to combine the different information extracted at different kernel sizes. Finally, bilinear interpolation is used to upsample the feature map to the initial size.

After the above operations, we concatenate the output feature maps of the LocalPyConv and GlobalPyConv. Then, a convolution kernel with the size of 1*1 and 256 channels is operated on the generated feature maps. A single convolution kernel is used here, and since all levels of information have been obtained previously, the main task is to merge the feature information. After the convolution of LocalPyConv and GlobalPyConv, the standard relu-activation function is used. The operation that extracts both local and global information is named as ContextConv. The fixed length descriptor is generated by bicubic interpolation and $L2$-normalization.

## 4 Loss Function

This part introduces the loss function for training phase. The loss $L$ [shown as in Eq. (1)] consists of three terms: the first two terms $L_{\text{point}}$ and $L'_{\text{point}}$ are used for interest point detector, and the last one $L_{\text{desc}}$ is used for learning descriptor. $\mu$ is used as a weighting factor to normalize $L_{\text{desc}}$. It is assumed that the contributions of point's location and descriptor should be similar (with the same magnitude). But $L_{\text{desc}}$ is much larger than $L_{\text{point}}$. Thus, $\mu$ used to balance the contributions of point's location and descriptor. According to the experiments, the value of $\mu$ depends on the size of input images. The larger the size of input image is, the smaller the value of $\mu$ is. In our experiments, the value of $\mu$ is set to 0.0001, empirically. As done in SuperPoint,[26] the results of MagicPoint[31] are used as the ground truth for the first iteration:

$$L = L_{\text{point}} + L'_{\text{point}} + \mu L_{\text{desc}}. \tag{1}$$

### 4.1 *Interest Point Loss*

The loss function of interest point detector is the fully convolution cross-entropy loss. Softmax function is used to obtain the normalized value of each output interest point. The original image size is $H * W$. After going through the backbone, it becomes that $H_s = H/8$ and $W_s = W/8$. So $x_{hw} \in R^{H_s * W_s * 65}$ represents each unit and $y_{hw}$ is the ground-truth labels of interest points. The loss function of interest point can be calculated as

$$L_{\text{point}} = \frac{1}{H_s W_s} \sum_{h=1,w=1}^{H_s,W_s} -\log \frac{e^{y_{hw}}}{\sum_{k=1}^{65} e^{x_{hw}^k}}, \tag{2}$$

where $e$ is the natural constant, 65 channels correspond to local 8*8 grid regions and an extra dustbin.

### 4.2 *Descriptor Loss*

The descriptor loss function uses circle loss[15] to describe the similarity between positive and negative samples. The interest point is expressed as $i \in P$ in the image $I$, and the interest point

is expressed as $i \in P'$ in the image $I'$. The corresponding matrix of $P * P'$ is denoted as $S$, which contains the value of 0 or 1. Each entry $s_{ij}$ specifies the corresponding relationship of the descriptor in each pair of images. This means that the pixel distance between the corresponding points in each pair of images should be less than or equal to 8 pixels. Equation (3) determines whether the descriptors of the points on the original image correspond to the descriptors of the points on the homography transformed image:

$$s_{ij} = \begin{cases} 1 & \text{distance}(i,j) \leq 8, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

We define that $des_p$ is the descriptor corresponding to the interest point $P$ in the first image, and $des'_{p'}$ is the descriptor corresponding to the interest point $P'$ in the second image, which undergoes homography transformation. Then, the $des_p^T * des'_{p'}$ matrix is multiplied to intuitively indicate the similarity between descriptors. So the loss function is expressed as

$$L_{\text{desc}} = \log\left(1 + \sum_{i=1}^{P}\sum_{j=1}^{P'} \exp((1 - s_{ij})\theta_n^h(f_n - m_n) - \gamma_p s_{ij}\theta_p^k(f_p - m_p))\right), \tag{4}$$

where

$$\begin{cases} \theta_n^h = \max(0, f_n^h - U_n) & h \in (1,2,3\ \cdots\ h), \\ \theta_p^k = \max(0, U_p - f_p^i) & k \in (1,2,3\ \cdots\ k). \end{cases} \tag{5}$$

The similarity between each group of descriptors is divided into two cases. If a group of descriptor is noncorresponding, we will express it as $f_n$. On the contrary, it will be expressed as $f_p$. All we need is to minimize $f_p$ and maximize $f_n$. Since $f_n$ and $f_p$ are asymmetrical, negative margins $m_n$ and positive margin $m_p$ are added. $\theta$ is the weighting factor, the factor is not the same between different groups, $h$ is the number of negative samples, and $k$ is the number of positive samples. Because there are more negative samples, we increase the weight $\gamma_p$ to achieve a balance. The optimal value of $f_p$ is defined as $U_p$, and the optimal value of $f_n$ is defined as $U_n$. A larger weight is assigned when the similarity value deviates far from the optimal value, and a big step update is carried out. The same as circle loss,[15] the decision boundary of the Eq. (4) is achieved at $\theta_n(f_n - m_n) - \theta_p(f_p - m_p) = 0$. Combining with Eq. (5), the decision boundary is given as

$$\left(f_n - \frac{U_n + m_n}{2}\right)^2 + \left(f_p - \frac{U_p + m_p}{2}\right)^2 = \frac{(U_n - m_n)^2 + (U_p - m_p)^2}{4}. \tag{6}$$

There are several hyperparameters $U_p$, $U_n$, $\theta$, $m_n$, and $m_p$, by defining that $U_p = 1 + m$, $U_n = -m$, $m_p = 1 - m$, and $m_n = m$, Eq. (6) can be replaced by

$$(f_n - 0)^2 + (f_p - 1)^2 = 2m^2. \tag{7}$$

After the redefinition of hyperparameters, the optimization goal is changed to $f_p \to 1$ and $f_p \to 0$, and the parameter $m$ is the decision factor of the optimization radius.

## 5 Experiments

### 5.1 *Training Details*

In our experiments, we apply not only pyramid convolutions to the network structure but also pyramid convolutions to the backbone network structure. Different from the previous 3*3 convolution kernel, we use different scale convolution kernel this time to ensure the extraction of multiscale features. The original backbone network is a network architecture similar to VGG. It is divided into four parts, each of which contains two layers. In our pyramid convolution

backbone, pyramid convolution kernels of 3*3, 5*5, 7*7, and 9*9 are used between layers, and the number of each convolution kernel is 16. To replace the original 3*3 convolution kernel with 128 channels, we used 3*3 and 5*5 convolution kernels with 32 channels, and 7*7 convolution kernel with 64 channels. Our backbone is a shared encoder, and the features extracted from the backbone are output to the interest point detection and descriptor extraction, respectively. In the head network of the interest point detection, we use the pyramid convolution kernel with a total number of 256 and then obtained 65 units of the interest point detection through the 1*1 convolution kernel. In the part of the descriptor extraction, the head network uses pyramid convolutions to extract global and local information with the local pyramid convolution module and global pyramid convolution module. To extract global feature information, the feature information obtained from the backbone network is transformed into a matrix of $9 * 9 * D$. The spatial resolution of the feature stays reasonable during the compression process. Then, we use a pyramid convolution kernel with a maximum of 9*9 to perform convolution operations. After that, we merge the information of the local and the global feature maps. The convolutional layer with the kernel size of 1*1 and the number of channels is 256 to generate interest point descriptor. All of our convolution kernels are followed by the relu-activation and BatchNorm normalization.

In our experiments, the size of the descriptor is $D = 256$. The descriptor loss uses a positive margin $m_p = 0.75$ and a negative margin $m_n = 0.25$. To balance the difference in the number of positive and negative samples, we use a weighting term of $\gamma_p = 256$. In addition, the batch size of training is 32 and the learning rate of ADAM optimization parameter is 0.001.

## 5.2 *Training Data*

To train the PC-SuperPoint, we use images in COCO2014[32] to generate pair images by Homographic Adaptation in SuperPoint.[26] To make our model more robust in the real environment, we also employ standard data augmentation to the training images. As a result, 80,106 grayscale images are generated for training.

## 5.3 *Evaluation Data*

The results obtained by our network structure are evaluated in the HPatches dataset.[33] As the benchmark data set of image descriptors, HPatches dataset collects images from various sources, including existing datasets and camera capture sequences. HPatches dataset is divided into two scene changes, one is the difference in illumination and the other is the change of viewpoint. In 57 scenes, the main nuisance factors are photometric changes and the remaining 59 sequences show significant geometric deformations due to viewpoint change. In brief, "viewpoint" represents the images in HPatches that contain varying viewpoints, and "illumination" represents the images in HPatches that contain varying illuminations. To further investigate the effectiveness of the proposed method, we also applied it to visual odometry and tested on the KITTI dataset.[34]

## 5.4 *Ablation Study*

### 5.4.1 *Repeatability evaluation*

To measure the performance of our method, we use $240 \times 320$ images to calculate the repeatability, when the threshold pixel correct distance $e = 3$. Two kinds of image sequences are used for test. Viewpoint changes represent the images that contain varying viewpoints, whereas illumination changes represent the images that contain varying illuminations. Nonmaximum suppression (NMS = 4) is applied to make the points evenly distributed calculate 300 points. Compared with Shi,[35] FAST,[36] Harris,[37] BRISK,[38] SuperPoint[26] and Key. Net,[27] the experimental results are shown in Table 1 (the italicized numbers in the table represent the best performance among different methods). It can find that the performance of Key. Net is better than the proposed method in terms of illumination changes, whereas the proposed method has better repeatability when facing illumination changes. A potential reason is that pyramid convolutions provide a preferable scale-invariance and rotation-invariance for the model but neglect to

**Table 1** Repeatability of different methods on HPatches is the research of detecting interest points, whichdataset.

| Methods | Illumination changes | Viewpoint changes |
|---|---|---|
| Shi[35] | 0.584 | 0.629 |
| FAST[36] | 0.575 | 0.625 |
| Harris[37] | 0.630 | *0.755* |
| BRISK[38] | 0.567 | 0.610 |
| SuperPoint[26] | 0.641 | 0.621 |
| Key. Net[27] | *0.661* | 0.627 |
| PC-SuperPoint (ours) | 0.656 | 0.645 |

improve the networ's robustness against illumination changes. Taken together, our proposed method is still very competitive with state-of-the-art methods.

### 5.4.2 *Homography estimation*

Homography matching is an important indicator of our detection performance. The error of homography estimation (as shown in Fig. 7) is defined as the mean distance between target image corners transformed by the ground truth homography $H$ and the estimated homography $\hat{H}$. The matching score (M.S.) is the ratio of the correct matching occupancy of all points in the shared view, which can be used to describe the performance of interest point detector and descriptor. The correct matching is the two points that are the nearest neighbors in the descriptor space. After the true homography, the pair points have been transformed to the same view, and the distance of the separated pixels is less than correctness threshold (e.g., three pixels).

We create multiple sets of comparative experiments among SIFT,[1] ORB,[3] BRISK,[38] LF-net,[39] and SuperPoint.[26] The detector selects $N$ points from two images of the same scene, which satisfying interest points and descriptor matching, performs nearest-neighbor matching, and calculate the homography matrix at the same time. To estimate the homography, we perform nearest-neighbor matching from all interest points detected in the scene sequences. An OpenCV implementation with RANSAC is applied with all the matches to compute the final homography estimation as the ground truth. The correctness threshold $e$ is set up to three pixels and the NMS is set up to 8. In this experiment, we computed with a maximum of 1000 points between pairs of images with $480 \times 640$. The results are shown in Table 2. Two kinds of image sequences are used for test. Viewpoint changes represent the images that contain varying viewpoints, whereas illumination changes represent the images that contain varying illuminations. It is demonstrated that the proposed method has good homography estimation when facing illumination changes.

Table 3 shows the homography estimation results under different correctness thresholds and the M.S.s of all kinds of changes (all scene sequences in HPatches dataset are used for each test). For ORB, SIFT, and BRISK, we used the default OpenCV implementations. For Superpoint and

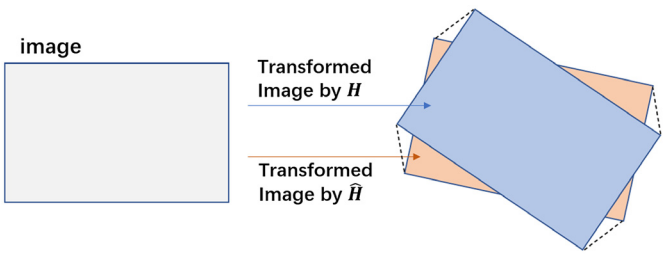



**Fig. 7** The expression of homography estimation (dotted lines represent the error of homography estimation).

**Table 2** Homography estimation of different methods on HPatches dataset.

| Methods | Illumination | Viewpoint |
|---|---|---|
| SIFT[1] | 0.807 | *0.766* |
| ORB[3] | 0.523 | 0.414 |
| BRISK[38] | 0.692 | 0.563 |
| LF-net[39](indoor) | 0.762 | 0.679 |
| LF-net[39](outdoor) | 0.820 | 0.712 |
| SuperPoint[26](SP) | 0.923 | 0.742 |
| SuperPoint+Pyconv (ours) | 0.941 | 0.749 |
| SuperPoint+circle loss (ours) | 0.932 | 0.746 |
| PC-SuperPoint (ours) | *0.945* | 0.753 |

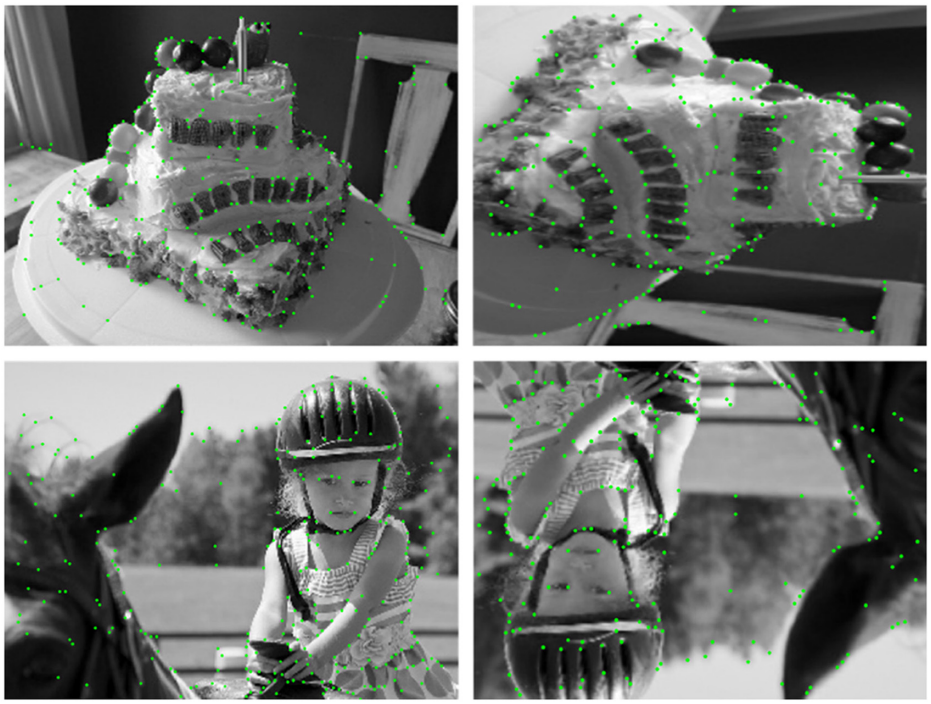
**Table 3** Homography estimation under different thresholds on HPatches dataset.

| Homography estimation | $e = 1$ | $e = 3$ | $e = 5$ | M.S. |
|---|---|---|---|---|
| SIFT[1] | *0.498* | 0.786 | 0.804 | 0.301 |
| ORB[3] | 0.162 | 0.467 | 0.564 | 0.259 |
| BRISK[38] | 0.300 | 0.653 | 0.746 | 0.211 |
| LF-net[39](indoor) | 0.216 | 0.659 | 0.815 | 0.301 |
| LF-net[39](outdoor) | 0.383 | 0.736 | 0.857 | 0.276 |
| SuperPoint[26](SP) | 0.438 | 0.833 | 0.914 | 0.375 |
| PC-SuperPoint (ours) | 0.442 | *0.847* | *0.921* | *0.384* |

LF-net, we used the trained models provided by the authors. SIFT is still a good interest point detector. It performs best in terms of viewpoint changes and can estimate low homography errors with a threshold of $e = 1$. The disadvantage of LF-net is that the detector and the descriptor do not share calculations, and the output generated by SFM is required for training. The work done by SuperPoint uses deep learning-based methods to make it powerful enough to learn a certain degree of invariance and achieve better results in homography estimation when large errors are tolerated ($e = 3$ and $e = 5$). The proposed method achieved better performance means that our innovations have positive impacts. Interest points detected in the corresponding scene images are shown in Fig. 8, which have high robustness and uniform distribution. As shown in Figs. 9 and 10, our method can produce dense and correct matching applying to image matching.

### 5.4.3 *Applications in visual odometry*

To further investigate the effectiveness of the proposed method, we applied it into visual odometry. KITTI dataset[34] is employed for experiments. In this experiment, we built three visual odometries for keypiont detectors, respectively. The absolute trajectory error (ATE) of each visual odometry reveals the position accuracy of keypoints extracted from different methods. As shown in Table 4, the ATE of the proposed method is much small than that of others. Figure 11 shows two visual results of trajectories using different visual odometries. The result shows that the our keypiont detection performances are improved. It demonstrates that the proposed method contributes to applications in visual odometry.



(a) (b)

**Fig. 8** Detected interest points in the images of various scenes. Each pair of images includes (a) the original image and (b) the warped image after homographic adaptation.

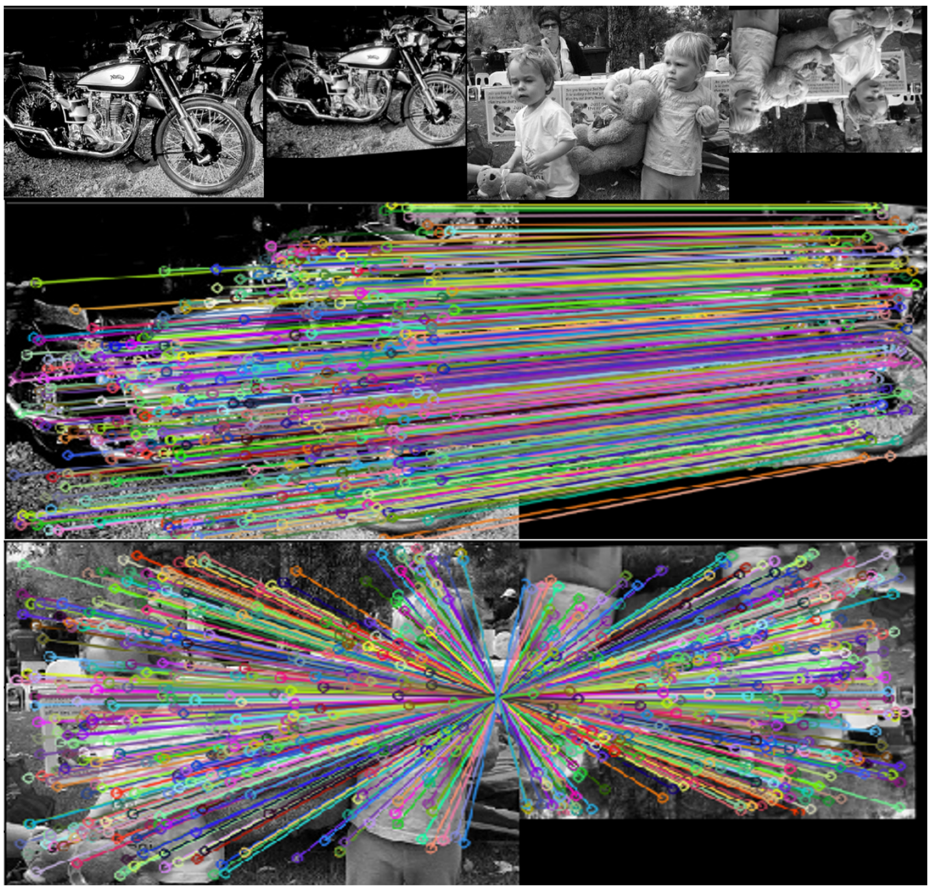


**Fig. 9** Matching results using our proposed method (a) (matching points are indicated by color lines).

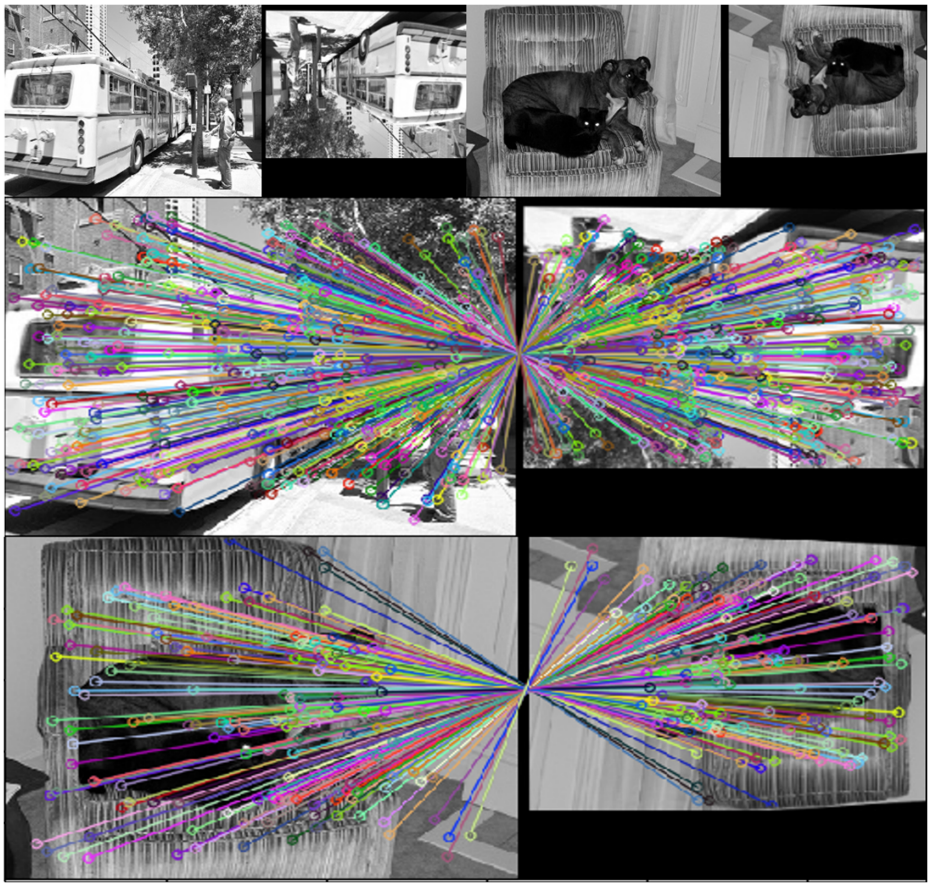


**Fig. 10** Matching results using our proposed method (b) (matching points are indicated by color lines).

**Table 4** ATE (m) of different visual odometries.

| Image sequence ID | SuperPoint | ORB | PC-SuperPoint (ours) |
|---|---|---|---|
| 00 | 21.671 | *13.579* | 18.325 |
| 01 | 148.536 | 463.745 | *63.743* |
| 02 | 27.423 | *24.961* | 34.829 |
| 03 | 8.769 | 10.031 | *7.257* |
| 04 | 4.983 | 2.279 | *1.967* |
| 05 | 30.394 | 64.757 | *21.698* |
| 06 | 11.486 | 12.038 | *9.577* |
| 07 | 13.988 | 38.392 | *8.072* |
| 08 | 43.674 | *29.574* | *33.347* |
| 09 | 20.873 | 21.351 | *14.703* |
| 10 | 13.423 | 94.456 | *11.057* |

### 5.4.4 *Calculation time*

Table 5 shows the comparison of the calculation time between our proposed method and state-of-the-art methods. All methods are performed with the same hardware platforms. The images in HPatches with two kinds of resolution ($240 \times 320$ and $480 \times 640$) are used for keypoint

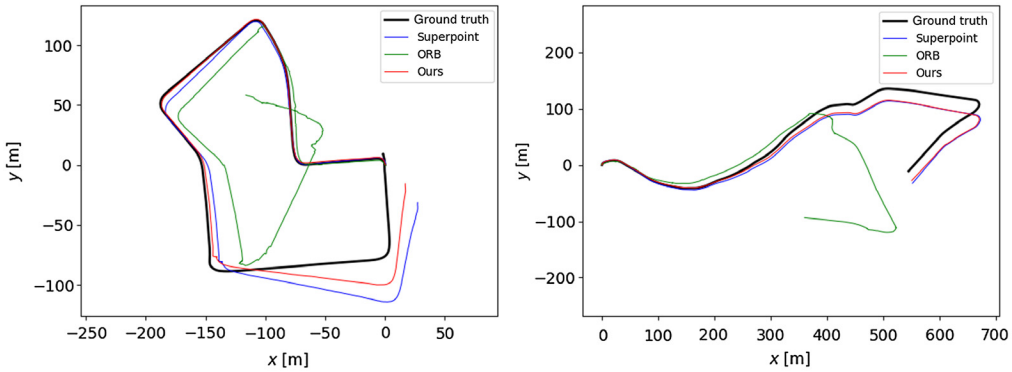



**Fig. 11** Motion trajectory of two image sequences (ID: 07 and 10).

**Table 5** Frames per second (FPS) for different keypoint detection methods.

| Methods | 240 × 320 | 480 × 640 |
|---|---|---|
| SIFT[1] | 42 | 9 |
| FAST[36] | 106 | 41 |
| SURF[2] | 40 | 10 |
| LF-net[39](indoor) | 62 | 25 |
| Key. Net[27] | 97 | 43 |
| SuperPoint[26] | *159* | *66* |
| PC-SuperPoint (Ours) | 138 | 54 |

detection. It is shown that the FPS of PC-SuperPoint is larger than that of most previous methods on the both cases. However, the runtime speed of SuperPoint is faster than the proposed method. Because PC-SuperPoint is based on SuperPoint and contains more complex network architecture. There is a trade-off between efficiency and performance.

## 6 Conclusions

In this paper, we aim to the issues on interest point detection and descriptor extraction. We propose a new well-designed networks named as PC-SuperPoint to achieve the goals. The proposed methodology is composed of interest point detection networks and descriptor extraction networks. Just as their names imply, the first part is used for interest point detection, and the last part is used for descriptor extraction. We carry out extensive simulation experiments to validate the proposed method by comparing them with other state-of-art approaches. Experimental results reveal that our proposed methodology is effective and accurate. In this work, we pay more attention to interest point detection but neglect to improve network's robustness against illumination changes. However, this in fact is not always applicable since generalization performance of feature descriptors have become the mainstream of research trends and illumination changes are also very common in many real-world scenarios. Therefore, we would like to investigate the availability of feature descriptors in the future.

## Acknowledgments

This work was jointly sponsored by the National Natural Science Foundation of China (Grant No. 62006150), Shanghai Young Science and Technology Talents Sailing Program (Grant No. 19YF1418400), Shanghai Key Laboratory of Multidimensional Information Processing (Grant No. 2020MIP001), and the Fundamental Research Funds for the Central Universities.

## References


1. D. G. Lowe, "Distinctive image features from scale invariant keypoints," *Int. J. Comput. Vision* **60**, 91–110 (2004).
2. H. Bay et al., "Speeded-up robust features," *Comput. Vision Image Understanding* **110**, 346–359 (2008).
3. E. Rublee et al., "ORB: an efficient alternative to SIFT or SURF," in *Proc. Int. Conf. Comput. Vision*, pp. 6–13 (2011).
4. K. Alex, S. Ilya, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, pp. 1097–1105 (2012).
5. K. He et al., "Delving deep into rectifiers: surpassing human-level performance on ImageNet classification," in *Proc. Int. Conf. Comput. Vision (ICCV)*, pp. 1026–1034 (2015).
6. A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature* **542**, 115–118 (2017).
7. S. C. Joon et al., "Lip reading sentences in the wild," in *Proc. IEEE Comput. Vision and Pattern Recognit. (CVPR)*, pp. 21–26 (2017).
8. H. Richard and Z. Andrew, *Multiple View Geometry in Computer Vision*, Cambridge University Press (2003).
9. A. Kendall, G. Matthew, and R. Cipolla, "PoseNet: a convolutional network for real-time 6-DOF camera relocalization," in *Proc. Int. Conf. Comput. Vision (ICCV)*, pp. 2938–2946 (2015).
10. A. Kendall and R. Cipolla, "Geometric loss functions for camera pose regression with deep learning," in *Proc. IEEE Comput. Vision and Pattern Recognit. (CVPR)*, pp. 5974–5983 (2017).
11. D. DeTone, T. Malisiewicz, and A. Rabinovich, "Deep image homography estimation," https://arxiv.org/abs/1606.03798 (2016).
12. W. Luo, A. G. Schwing, and R. Urtasun, "Efficient deep learning for stereo matching," in *Proc. IEEE Comput. Vision and Pattern Recognit. (CVPR)*, pp. 27–30 (2016).
13. W. Sen et al., "DeepVO: towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *Proc. Int. Conf. Rob. and Autom. (ICRA)*, pp. 2043–2050 (2017).
14. I. C. Duta et al., "Pyramidal convolution: rethinking convolutional neural networks for visual recognition," https://arxiv.org/abs/2006.11538 (2020).
15. Y. Sun et al., "Circle loss: a unified perspective of pair similarity optimization," in *Proc. IEEE Comput. Vision and Pattern Recognit. (CVPR)*, pp. 13–19 (2020).
16. X. Han et al., "MatchNet: unifying feature and metric learning for patch-based matching," in *Proc. IEEE Comput. Vision and Pattern Recognit. (CVPR)*, pp. 3279–3286 (2015).
17. S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Comput. Vision and Pattern Recognit. (CVPR)*, pp. 4353–4361 (2015).
18. S.-S. Edgar et al., "Discriminative learning of deep convolutional feature point descriptors," in *Proc. Int. Conf. Comput. Vision (ICCV)*, pp. 118–126 (2015).
19. T. Yurun, F. Bin, and W. Fuchao, "L2-Net: deep learning of discriminative patch descriptor in Euclidean space," in *Proc. IEEE Comput. Vision and Pattern Recognit. (CVPR)*, pp. 661–669 (2017).
20. Z. Luo et al., "GeoDesc: learning local descriptors by integrating geometry constraints," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, pp. 170–185 (2018).
21. Z. Luo, T. Shen, and L. Zhou, "ContextDesc: local descriptor augmentation with cross-modality context," in *Proc. IEEE Comput. Vision and Pattern Recognit. (CVPR)*, pp. 15–20 (2019).
22. P. Ebel et al., "Beyond Cartesian representations for local descriptors," in *Proc. Int. Conf. Comput. Vision (ICCV)*, pp. 253–262 (2019).
23. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," https://arxiv.org/abs/1409.1556 (2014).
24. C. Szegedy, W. Liu, and Y. Jia, "Going deeper with convolutions," in *Proc. IEEE Comput. Vision and Pattern Recognit. (CVPR)*, pp. 7–12 (2015).

25. K. He, X. Zhang, and S. Ren, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Vision and Pattern Recognit. (CVPR)*, pp. 27–30 (2016).
26. D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: self-supervised interest point detection and description," in *Proc. IEEE Comput. Vision and Pattern Recognit. Workshops (CVPR)*, pp. 224–236 (2018).
27. A. B. Laguna et al., "Key.net: keypoint detection by handcrafted and learned CNN filters," in *Proc. Int. Conf. Comput. Vision (ICCV)*, pp. 5835–5843 (2019).
28. B. Vassileios et al., "PN-Net: conjoined triple deep network for learning local image descriptors," https://arxiv.org/abs/1601.05030 (2016).
29. B. Vassileios et al., "Learning local feature descriptors with triplets and shallow convolutional neural networks," in *Proc. Br. Mach. Vision Conf. (BMVC)*, pp. 1–11 (2016).
30. C. B. Choy et al., "Universal correspondence network," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, pp. 449–460 (2016).
31. D. DeTone, T. Malisiewicz, and A. Rabinovich, "Toward geometric deep SLAM," https://arxiv.org/abs/1707.07410 (2017).
32. L. Tsung-Yi et al., "Microsoft COCO: common objects in context," *Lect. Notes Comput. Sci.* **8693**, 740–755 (2014).
33. V. Balntas et al., "HPatches: a benchmark and evaluation of handcrafted and learned local descriptors," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR)*, pp. 3852–3861 (2017).
34. A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 3354–3361 (2012).
35. J. Shi and T. Carlo, "Good features to track," in *Proc. IEEE Comput. Vision and Pattern Recognit. (CVPR)*, pp. 21–23 (1994).
36. R. Edward and D. Tom, "Machine learning for high-speed corner detection," in *Proc. Eur. Conf. Comput. Vision (ECCV)*, pp. 430–443 (2006).
37. C. G. Harris and M. J. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vision Conf.*, pp. 147–151 (1988).
38. L. Stefan, C. Margarita, and Y. S. Roland, "BRISK: binary robust invariant scalable keypoints," in *Proc. Int. Conf. Comput. Vision (ICCV)*, pp. 6–13 (2011).
39. Y. Ono et al., "LF-Net: learning local features from images," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, pp. 6234–6244 (2018).

**Yu-Jie Xiong** received his BE degree from Central South University in 2011 and his PhD in computer application technology from East China Normal University in 2018. He is currently a lecturer with Shanghai University of Engineering Science. His research interests include pattern recognition, writer identification, and biometrics.

**Shuo Ma** is a graduate student at Shanghai University of Engineering Science. He received his BS degree in communications engineering in 2018. His current research interests include deep learning and image processing.

**Yongbin Gao** received his PhD from Chonbuk National University, South Korea. He is currently an associate professor of the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, China. He has published numerous SCI papers in prestigious journals such as TIP, Information science, and pattern recognition letters. His research interests are image processing, pattern recognition, and computer vision.

**Zhijun Fang** received his PhD from Shanghai Jiao Tong University, Shanghai, China. He is currently a professor and the Dean of the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science. His current research interests include image processing, video coding, and pattern recognition. He has received the Ganpo 555 Talents Program Award and the One-Hundred, the One-Thousand, and the Ten-Thousand Talent Project Award of Jiangxi Province. He was the general chair of the Joint Conference on Harmonious Human Machine Environment, in 2013, and the general co-chair of the International Symposium on Information Technology Convergence, from 2014 to 2017.

关闭 Web of Science 打印
第 1 页 (记录 1 -- 1)
[1]

第 1 条，共 1 条

标题: PC-SuperPoint: interest point detection and descriptor extraction using pyramid convolution and circle loss

作者: Xiong, YJ (Xiong, Yu-Jie); Ma, S (Ma, Shuo); Gao, YB (Gao, Yongbin); Fang, ZJ (Fang, Zhijun)

来源出版物: JOURNAL OF ELECTRONIC IMAGING 卷: 30 期: 3 文献号: 033024 DOI: 10.1117/1.JEI.30.3.033024 出版年: MAY 2021

Web of Science 核心合集中的 "被引频次": 0
被引频次合计: 0
使用次数 (最近 180 天): 0
使用次数 (2013 年至今): 0
引用的参考文献数: 39

摘要: Nowadays, deep learning is widely used to detect interest points and extract the corresponding descriptors and achieved suitable results for many applications of computer vision, such as image matching, three-dimensional reconstruction, simultaneous localization, and mapping. We propose an approach for interest point detection and descriptor extraction using pyramid convolution and circle loss, which is named as PC-SuperPoint. We utilize pyramid convolutions in the backbone network, which includes convolution kernels of different scales for multiscale feature extraction. The following well-designed networks are able to capture the local and global information from the obtained backbone feature maps. In addition, circle loss, which enhances weight attributes for each pair of descriptors, is also applied to improve the convergence speed in the training phase. Experiments on the HPatches dataset and KITTI dataset achieve promising results, which reveal the effectiveness of the proposed method. (C) 2021 SPIE and IS&T

入藏号: WOS:000705982600005
语言: English
文献类型: Article
作者关键词: interest point detection; descriptor extraction; pyramid convolution; circle loss

地址: [Xiong, Yu-Jie; Ma, Shuo; Gao, Yongbin; Fang, Zhijun] Shanghai Univ Engn Sci, Sch Elect & Elect Engn, Shanghai, Peoples R China.
[Xiong, Yu-Jie] East China Normal Univ, Shanghai Key Lab Multidimens Informat Proc, Shanghai, Peoples R China.
通讯作者地址: Xiong, YJ (通讯作者), Shanghai Univ Engn Sci, Sch Elect & Elect Engn, Shanghai, Peoples R China.
Xiong, YJ (通讯作者), East China Normal Univ, Shanghai Key Lab Multidimens Informat Proc, Shanghai, Peoples R China.
电子邮件地址: xiong@sues.edu.cn

出版商: IS&T & SPIE
出版商地址: 1000 20TH ST, BELLINGHAM, WA 98225 USA
Web of Science 类别: Engineering, Electrical & Electronic; Optics; Imaging Science & Photographic Technology
研究方向: Engineering; Optics; Imaging Science & Photographic Technology
IDS 号: WF0CB
ISSN: 1017-9909
eISSN: 1560-229X
29 字符的来源出版物名称缩写: J ELECTRON IMAGING
ISO 来源出版物缩写: J. Electron. Imaging
来源出版物页码计数: 14

基金资助致谢:

| 基金资助机构 | 授权号 |
|---|---|
| National Natural Science Foundation of China | 62006150 |
| Shanghai Young Science and Technology Talents Sailing Program | 19YF1418400 |
| Shanghai Key Laboratory of Multidimensional Information Processing | 2020MIP001 |
| Fundamental Research Funds for the Central Universities | |

This work was jointly sponsored by the National Natural Science Foundation of China (Grant No. 62006150), Shanghai Young Science and Technology Talents Sailing Program (Grant No. 19YF1418400), Shanghai Key Laboratory of Multidimensional Information Processing (Grant No. 2020MIP001), and the Fundamental Research Funds for the Central Universities.

输出日期: 2021-11-01

关闭 Web of Science 打印
第 1 页 (记录 1 -- 1)
[1]

## 1. PC-SuperPoint: Interest point detection and descriptor extraction using pyramid convolution and circle loss

**Accession number:** 20212710591019
**Authors:** Xiong, Yu-Jie (1, 2); Ma, Shuo (1); Gao, Yongbin (1); Fang, Zhijun (1)
**Author affiliation:** (1) Shanghai University of Engineering Science, School of Electronic and Electrical Engineering, Shanghai, China; (2) East China Normal University, Shanghai Key Laboratory of Multidimensional Information Processing, Shanghai, China
**Corresponding author:** Xiong, Yu-Jie(xiong@sues.edu.cn)
**Source title:** Journal of Electronic Imaging
**Abbreviated source title:** J. Electron. Imaging
**Volume:** 30
**Issue:** 3
**Issue date:** May 1, 2021
**Publication year:** 2021
**Article number:** 033024
**Language:** English
**ISSN:** 10179909
**E-ISSN:** 1560229X
**CODEN:** JEIME5
**Document type:** Journal article (JA)
**Publisher:** SPIE
**Abstract:** Nowadays, deep learning is widely used to detect interest points and extract the corresponding descriptors and achieved suitable results for many applications of computer vision, such as image matching, three-dimensional reconstruction, simultaneous localization, and mapping. We propose an approach for interest point detection and descriptor extraction using pyramid convolution and circle loss, which is named as PC-SuperPoint. We utilize pyramid convolutions in the backbone network, which includes convolution kernels of different scales for multiscale feature extraction. The following well-designed networks are able to capture the local and global information from the obtained backbone feature maps. In addition, circle loss, which enhances weight attributes for each pair of descriptors, is also applied to improve the convergence speed in the training phase. Experiments on the HPatches dataset and KITTI dataset achieve promising results, which reveal the effectiveness of the proposed method. © 2021 SPIE and IS&T.
**Number of references:** 39
**Main heading:** Convolution
**Controlled terms:** Deep learning - Extraction
**Uncontrolled terms:** Convergence speed - Convolution kernel - Descriptor extractions - Global informations - Interest point detections - Multi-scale features - Simultaneous localization , and mappings - Three-dimensional reconstruction
**Classification code:** 716.1 Information Theory and Signal Processing - 802.3 Chemical Operations
**DOI:** 10.1117/1.JEI.30.3.033024
**Compendex references:** YES
**Database:** Compendex

**Data Provider:** Engineering Village

EI收录
2021-755
林晨旭