Journal of Computers Vol. 33 No. 6, December 2022, pp. 167-180
doi: 10.53106/199115992022123306014

Feature Fusion Method for Low-Illumination Images

Ya-Nan Li, Zhen-Feng Zhang, Yi-Fan Chen, Chu-Hua Huang*

State Key Laboratory of Public Big Data, College of Computer Science and Technology, Guizhou University,
Guiyang 550025, China
{gs.ynli20, gs.zfzhang20, gs.yfchen20, chhuang}@gzu.edu.cn

Received 10 March 2022; Revised 20 July 2022; Accepted 10 August 2022

Abstract. Aiming at the problem of inaccurate feature extraction of low illumination images, a method is proposed that fuses Scale Invariant Feature Transform into SuperPoint. Firstly, the low illumination image is light-enhanced. Secondly, SuperPoint and SIFT features are fused at feature map level, changing the deep neural network weight by labeling the prob output of the network with the SIFT of input image as the maximum of the current prob at pixel-level. Finally, the loss function is constructed based on homography transformation, its principle between image pairs is used to realize the constraint on network parameters. The training and evaluation are conducted on ExDark dataset, tests and comparisons are conducted on multiple indicators of SOTA on HPatches common dataset. The experimental results show that our method improves the precision and recall than SuperPoint, and performs well in multiple evaluation indicators.

Keywords: low-illumination image, feature fusion, feature map, SIFT, SuperPoint

1 Introduction

Image feature extraction is an essential part of computer vision tasks, especially visual localization [1-4], Structure from Motion (SFM) [5], and Simultaneous Localization and Mapping (SLAM) [6]. Meanwhile, it plays a vital role in pose tracking [7] and 3D point cloud reconstruction [8]. The sparse keypoints extracted from the image are sufficient for the aforementioned visual tasks. In addition, the descriptor obtained by learning and calculating is the crucial element of image matching, which has a direct impact on the accuracy of certain task results. However, the performance of feature extraction in low-illumination scenes is inaccurate, resulting in poor performance of many computer vision tasks. Therefore, this paper proposes a feature fusion method with light-enhanced for low-illumination images, laying a good foundation for feature matching, pose estimation and other tasks.

Early works on local feature extraction rely primarily on traditional hand-crafted rules. The representative methods include Scale Invariant Feature Transform (SIFT) [9], Speeded Up Robust Feature (SURF) [10], Binary Robust Invariant Scalable Keypoints (BRISK) [11], Oriented FAST and Rotated BRIEF (ORB) [12] and so on. These methods follow manually designed computational steps to produce fixed features, which are scale-consistent and robust in the image transformation scenes. Although they are widely used in vision applications, they are not scalable and cannot be further improved. So it is difficult to cope with vision tasks in constantly changing external surroundings.

With the recent development of deep learning, some researchers have applied it to feature extraction, which extracts features directly through deep neural networks and outperforms traditional methods in keypoint evaluation indexes such as precision and repeatability. Furthermore, unlike traditional methods, changing the network module can improve the effectiveness of these methods. On the basis of their structures, these methods are classified into two categories: one-stage and two-stage network methods. The former obtains local keypoints and descriptors independently via a single deep network. The latter receives targets through two sequentially connected modules or shared backbone branches. To some extent, the robustness of these methods to changes in the external environment is improved. However, the repeatability of the detected keypoints is poor for image transformation scenes, indicating that the keypoints at the same place in the two images with different viewpoints cannot be detected repeatedly, resulting in the failure of subsequent operations such as feature matching. Among these methods, DeTone et al. [13] proposed the SuperPoint deep learning method, which achieves state-of-the-art on the homography estimation task, and the feature map of keypoints output by this network has the possibility to further manipulate with the keypoints extracted by traditional methods. In addition, it can obtain pseudo-ground

* Corresponding Author

truths without manual labeling and exhibits excellent performance in illumination change scenes. Thus, this paper is based on this research. However, it does not make full use of the constraints between the homography transform image pairs, and there is a common problem with the traditional methods. That is, it is less effective at extracting features in low-illumination scenes.

As a result, we are motivated to incorporate an illumination-enhanced network to ensure that the method can extract better features even in dimly illuminated environments. Furthermore, inspired by these two methods, we hope to combine their benefits to design a network that can provide high-quality keypoints and descriptors for homography estimation and camera localization tasks. Moreover, the repeatability and reliability of the developed method for the scene based on image and external environment transformation, such as perspective and illumination transformation, are as high as possible.

The contributions are as follows.

1. In this paper, a novel feature fusion module is proposed, which is performed at the level of the feature map output from the deep learning network with SIFT keypoint labeling. The fused feature map then optimizes model parameters through network back propagation and improves the performance of features in illumination and perspective transformation scenes.

2. To address the issue of insufficient feature extraction in low-illumination scenes, the input image pairs are light-enhanced. In addition, the homography transform loss function is introduced in accordance with the homography transform principle of the image pairs. The network constraints are then strengthened to provide more precise feature extraction in scenarios with low illumination.

3. For training and testing, the ExDark low-illumination dataset and the popular HPatches dataset are utilized. On keypoint and multiple vision tasks evaluation metrics, the performance of two types of feature extraction methods is evaluated. The results reveal the efficacy of our suggested strategy by demonstrating a particular improvement in numerous indexes and achieving state-of-the-art performance in homography estimation task.

2 Related Work

2.1 Manual Feature Extraction

Traditional feature extraction methods follow fixed steps to extract image features. The typical manual feature extraction methods are SIFT, ORB, Shi-Tomas [14], etc. SIFT extracts features at various scales and calculates feature directions by constructing a Gaussian pyramid. The pyramid is constructed using the original image and Gaussian kernel functions with different parameters. The potential spatial interest points are identified by comparing the differences between the pyramid's layers, and the SIFT feature points are obtained after filtering. This method yields highly accurate features. In addition to rotation and scale consistency, it can maintain stability in the presence of noise and viewpoint transformation scenes, so this method is still widely used. However, its performance is general in external condition transformation scenarios such as illumination change scenes. In addition, this method cannot meet the needs of real-time operation due to its expensive computational cost. Then ORB detects feature points using the Features from Accelerated Segment Test (FAST) algorithm [15]. A FAST corner is identified based on the slight similarity between itself and its neighboring pixels. So far, this method is still utilized in the ORB-SLAM series of simultaneous localization and mapping systems that require real-time conditions, which has fast extraction speed and good performance in sequences with many rotations. However, the method is not scale-invariant and generates many false matches in low-texture or no-texture scenes owing to the tiny number of extracted features, resulting in initialization failure or tracking loss in the system. Shi-Tomas is a speed and performance enhancement of the Harris corner detection algorithm, which takes the difference between sliding windows composed of image pixel blocks as its detection criterion and is rotationally invariant. It is frequently employed to detect corners and can run in real-time. However, it is sensitive to scale transformation and has poor feature tracking performance in scenes with strong illumination changes [16].

2.2 Deep Learning Feature Extraction

Early researchers rely primarily on manual feature extraction methods. Nevertheless, these features do not fully represent the essential characteristics of the target in the image and the requirements for accurate image classification. At the same time, their manually designed computational steps limit the enhancement possibilities. With the development of deep learning, feature extraction methods are gradually changing from manual extraction to deep learning research. The convolution features extracted by deep neural networks are more representative of the basic qualities of the target, and the performance in scenarios with changing illumination and viewpoint has significantly improved. Simo-Serra et al. [17] proposed a twin network method to extract descriptors consistent with SIFT dimension, which shows good performance in scale transformation scenes and lays the foundation for subsequent feature extraction networks. While the training procedure and sampling rule are relatively complex. Learned Invariant Feature Transform (LIFT) [18] employed a detector-based method to the keypoints, directions, and descriptors in order, replacing the hand-crafted implementation steps with convolution, showing good lighting and perspective change scene results. However, the serial phased network structure makes the network slow to get all features and unable to operate in real-time. MagicPoint, on the other hand, gets the feature extraction model by training a synthetic object dataset with labels and achieves more accuracy in testing on synthetic datasets but inferior results in real-world practical applications. Another two-stage deep learning network, SuperPoint, is a further extension of MagicPoint for natural scenes, which achieves state-of-the-art performance on the homography estimation and considerably improves the robustness of scenes with changing illumination and perspective. SuperPoint first extracts keypoints of natural scenes using the trained MagicPoint network as a pseudo-truth, taking an image pair after homography transformation as the network input, then outputs the probability maps and descriptors of the image pair in parallel. This method innovatively exploits the relationship between image pairs as descriptor constraints in a self-supervised manner, which satisfies the real-time operating requirements but does not completely utilize the relationship to constrain the keypoints. In particular, this method only performs well in specific scenes, and the effect needs to be improved for night scenes. Similar methods include [19-21] et al.

In this paper, based on the above method, we make full use of the constraints between the input image pairs of the deep learning network to add a homography transform loss to this feature, which enhances the constraints of the network and further improves the accuracy of the original network. For the low-illumination scenes with poor visual effects, we apply illumination enhancement to the input image to ensure that the network can extract better features even in low-illumination conditions.

2.3 Feature Fusion Methods

Since traditional and deep learning features have pros and cons and each plays a vital role in different scenarios, many scholars [22-23] fused the features extracted by both methods at different levels, hoping to combine the advantages of both feature extraction methods and apply them to the fields of face recognition, image segmentation and achieve better results. For example, Wang et al. [24] fused Histogram of Oriented Gradient (HOG) [25] features with convolution features for adaptive weight. The experimental results were improved compared with deep learning features for target recognition. However, this method directly expanded the extracted HOG feature dimension to the convolution feature dimension without considering the information mismatch problem. Moreover, the fusion in this method was a direct weighted superposition of the processed HOG features with the convolution features, making it challenging to ensure that the fused features will not combine their disadvantages simultaneously based on the advantages of the two methods. Zhang et al. [26] sent the SIFT features extracted from face images to a deep neural network to learn an optimal set of discriminative feature vectors. They used it to classify expressions with a high degree of accuracy. However, this method used the SIFT-processed features as the network input, which changed the information of the original image and may affect the convolution network's extraction ability.

These methods demonstrate the feasibility of combining two types of feature extraction methods as well as their effectiveness when applied to various vision tasks. Nonetheless, their implementations are relatively simple summations, which cannot capture the essential sense of feature fusion. Therefore, this paper explores a novel form of feature fusion, starting from fused features influencing the network weights by changing the feature map of the same size as the input image, allowing the network adaptively to learn the most optimal training parameters without dimensional transformation. It enables us to achieve a true sense of fusion, extracting more precise features in changing illumination and viewpoint scenes.

3 Methods

3.1 Overall Network Framework

As shown in Fig. 1, the overall network framework is improved based on SuperPoint. With the virtual frames as the primary improvement. The network consists of the illumination enhancement network and the feature extraction network. The feature extraction network consists of an encoder-decoder framework. The encoder is similar to the VGG structure, consisting mainly of convolution and pooling layers for extracting features of different dimensions. The decoder is divided into two branches, the keypoint decoder and the descriptor decoder, which output the keypoint probability map and the descriptor vector, respectively. Firstly, the original image is processed by the illumination enhancement network. Then the random homography transformation is performed on the high-illumination processed results to generate the image pairs as the network input. Following that, SIFT features are fused into the original network, combining the advantages of SuperPoint's high repeatability in illumination and perspective change scenes with the rotation and scale consistency of SIFT. Moreover, the descriptor constraint between image pairs and the detector constraint between keypoints and ground truths are used to realize self-supervised parameter training. Finally, more robust features are extracted.

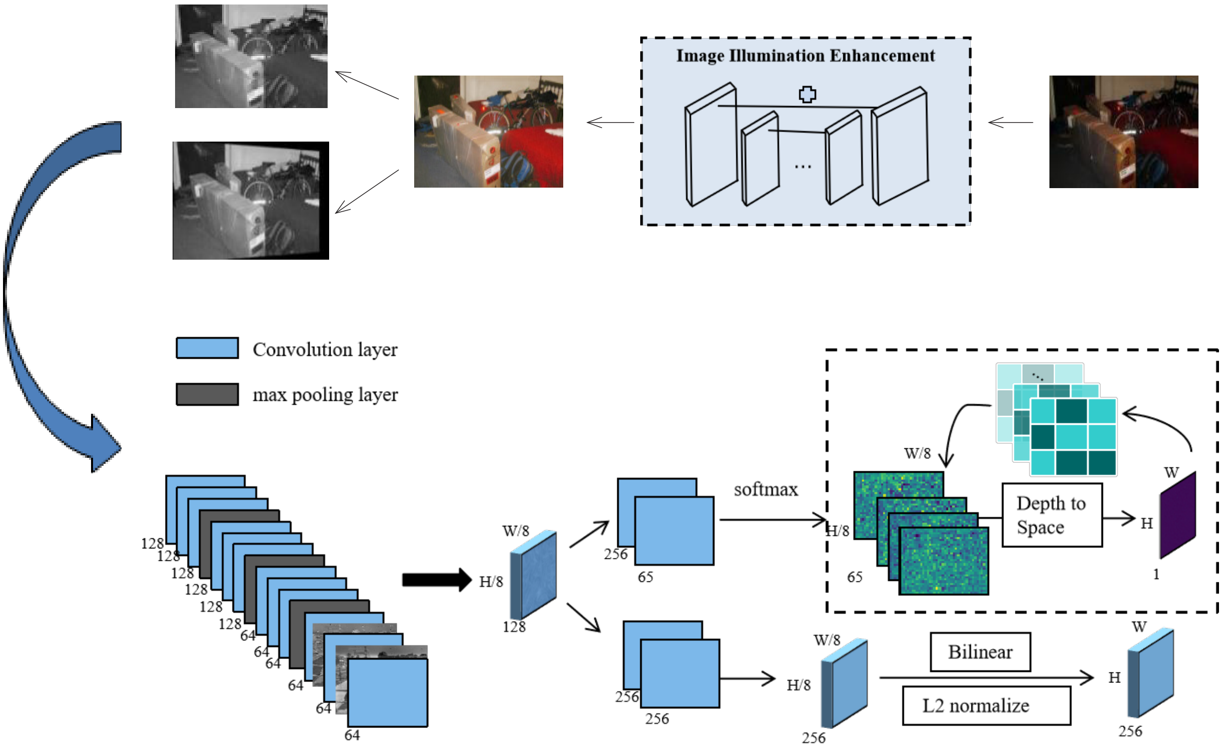


Fig. 1. Overall network structure diagram

3.2 Low-Illumination Image Enhancement Network

It is not conducive to feature extraction in weak light scenes because of its low visibility. In addition to random Gaussian noise and scale transformation, the paper preprocesses the training set for image illumination enhancement on the GLADNet [27] to improve the generalization of the feature extraction model for low-illumination images. The network structure of GLADNet is depicted in Fig. 2. The entire network consists of global illumination estimation and detail reconstruction modules. The former firstly downsamples the input image to fixed size by the nearest neighbor interpolation method. Then the cascade structure of convolution is used to realize the goal of global light prediction through feature mapping. Next, the results are upsampled to restore their size to the original image. By applying the jump connection structure, the coarse and fine features are combined to obtain more detailed information, which makes the network learning residual rather than the predicted image pixel value. Composed of convolution and activation layers, the latter is used to recover the lost image details during the sampling procedure. Consequently, illumination enhancement is accomplished by adjusting the illumination, and then the image details are restored to produce a high illumination image with the same quality as the original image.

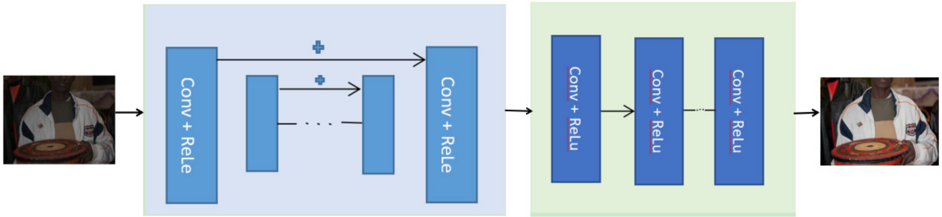


Fig. 2. Image illumination-enhanced network structure diagram

3.3 Feature Fusion Method Based on SIFT and SuperPoint

Corresponding to the operation module of the feature map in the virtual frame on the right side of Fig. 1, as shown in Fig. 3, we map SIFT coordinates of each image to a tensor of the input image's size, then obtain the SuperPoint network output logits and conduct dimensional transformation on it, the processed tensor is called Prob and represents each pixel's probability of being a keypoint. The probability values of the Prob at the SIFT coordinates are marked as the current maximum of the Prob. Last but not least, it is integrated into the new probability map Prob', which is used to calculate the detector loss and subsequent back propagation to change the network weight and improve prediction accuracy.

Since the network directly outputs logits and utilizes them to generate Prob by depth to space and other operations, the dimension of logits does not align with the SIFT. Consequently, we inverse logits by modifying Prob, then leverage the obtained new logits as the input of the detector's cross-entropy loss term and calculate it to complete the fusion.

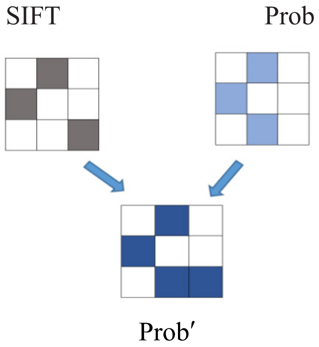


Fig. 3. Feature map fusion diagram

3.4 Loss Function

Based on the SuperPoint loss function, the paper constructs a new loss term of homography transformation. Wang et al. [28] propose an epipolar loss function. The epipolar constraint of two adjacent frames is used to construct the loss function so that the network can learn the descriptors only by the relative pose of the camera. Inspired by the idea, our loss item L_h is proposed to increase the constraint of the network by using the principle that one image's keypoints predicted by the feature extraction network after homography transformation should correspond to the homography transformed image's keypoints. The overall loss function defined as formula (1):

$$L = \alpha L_p(l', g) + \beta L_{wp}((wl)', wg) + \lambda L_d(s', (ws)') + L_h(wl', (wl)') . \quad \textbf{(1)}$$

L_p is the cross-entropy loss between l' that is logits obtained by the feature extraction network of the original image l and the ground truth g annotated by the MagicPoint of the original image. L_{wp} is the cross-entropy loss between the feature points $(wl)'$ output by the network of the warped image and the ground truth wg of the warped image wl. α and β are the parameters to adjust the proportion of detector loss in terms of the original and warped images. L_d is the hinge loss between the descriptors s' and $(ws)'$ obtained by the feature extraction network of the original image l and the warped image wl. λ is a coefficient used to balance the final loss. L_h is the new homography transformation loss item. That is, the constraint between wl' transformed by feature points l' and $(wl)'$.

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} h_1 & h_2 & h_3 \\ h_2 & h_2 & h_3 \\ h_3 & h_2 & h_3 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} . \quad \textbf{(2)}$$

As shown in formula (2), the hinge loss function is calculated according to the principle that the output predicted by the neural network after homography transformation is consistent with the output predicted by the neural network of warped input:

$$L_{\hbar} \ = \ \frac{1}{n}\sum_{i=1}^{n} max(0, 1 - (wl)' * wl') \cdot \quad \textbf{(3)}$$

4 Experimental Analysis

4.1 Experimental Environment and Dataset Description

The hardware configurations of the training experiment are as follows: Intel Core i7-10700F CPU 2.90GHz, NVIDIA GTX2080Ti 12GB. In terms of software, tensorFlow1.14.0 and OpenCV3.4.2 are selected.

The Exclusively Dark (ExDark) low-illumination image dataset [29], a benchmark test for low-light research, is used for training. The dataset covers 7363 low-light images in environments ranging from extremely weak to low-light, including 12 object classes such as chairs and buses. The total number of iterations is set to 600,000, and the precision and recall are evaluated once every 5000 iterations. The initial learning rate is set to 10^{-4}, default NMS (non-maximum suppression parameters) is 4, and the filter threshold percentage of the feature map is 10^{-3}, descriptor loss term weight λ is 10^{-4}, original and warped detector loss weight parameters α and β can be adjusted according to the actual situation. In addition, the HPatches dataset [30] and GladNet-Dataset are selected as the test dataset. HPatches is the benchmark dataset for evaluating repeatability and homography, which contains 116 scenes, divided into 57 illumination changes and 59 perspective changes. Each scene contains six images as well as a homography matrix document relative to the first image. Then quantitative experiments are carried out to evaluate and compare typical feature extraction algorithms with deep learning methods and ours on multiple indicators. Qualitative experiments are carried out on GladNet-Dataset for the visualization effect of keypoints before and after illumination and model improvement.

To enhance the applicability of the data, we firstly add random Gaussian and speckle noise, illumination, and contrast transformation to the training set. Then the low-illumination images of the dataset are enhanced and resized into 240 * 320 grayscale images for the feature extraction network's inputs. Next, we extract the image feature of the dataset using MagicPoint, which has been pretrained by the synthetic shape dataset as ground truth. In addition, SIFT feature points need to be extracted and preserved from the processed image, and each image's SIFT detector attribute is added to the preprocessing part of SuperPoint to provide data support for subsequent feature fusion.

4.2 Comparison Illumination Enhancement Network

As we have discussed, light enhancement is necessary for feature extraction in low-illumination scenes. To determine the effect of different light enhancement structures on feature extraction, in this section, we conduct comparative experiments on the effects of classical light enhancement networks while ensuring that other conditions are the same. Table 1 shows the precision and recall of the feature extraction network trained on the high-illumination dataset obtained from the GLADNet, MMLLEN [31], and RetinexNet [32] illumination-enhanced pretrained models, quantitatively demonstrating the impact of different illumination-enhanced networks on the feature extraction task. While RetinexNet obtains better results only with a specific dataset, so the effect decreases when applied to the ExDark dataset. In contrast, GLADNet shows a more extensive improvement. To gain a more intuitive understanding of the effect of different networks, we extract four sets of test images for display. As shown in

Fig. 4, we can see that RetinexNet corresponds to the effect of quantitative experiments and shows a tendency of over-sharpening on the ExDark dataset. From the qualitative and quantitative experiments, we can conclude that GLADNet obtains higher-quality images with high illumination and performs the feature extraction task with the best performance.

Table 1. ExDark dataset illumination enhancement comparison experiment

Method	Precision	Recall	Loss
GLADNet	0.266	0.479	2.793
MMLLEN	0.230	0.428	2.863
RetinexNet	0.149	0.291	3.005
None	0.200	0.384	1.825

4.3 Multi-index Comparison of Classical Feature Extraction Methods

To evaluate the effectiveness of our method, in this section, we compare it with several classical traditional and deep learning methods. In addition to the ExDark dataset, the HPatches common dataset is selected to compare feature extraction results. As an essential metric for feature extraction evaluation, we evaluate repeatability metrics with several representative methods in viewpoint and illumination transformation scenarios for quantitative comparison. As a second step, we select the two most classical manual and deep learning feature extraction methods as our baseline for two feature extraction evaluation metrics and two visual task metrics experiments. Our method achieves the best performance on the homography estimation task. Therefore, we further analyze in detail the impact of different parameters on the homography estimation task and compare it with the latest deep learning method LoFTR. Our method surpasses it and achieves the best performance. Last but not least, to visually validate the efficacy of our method, we perform qualitative comparison experiments with a baseline for the feature matching task.

First, we choose the traditional manual feature extraction methods Fast, Harris, Shi-Tomas, and the deep learning feature extraction method SuperPoint to evaluate the repeatability in the illumination and perspective change scenes, respectively. The parameters that affect the evaluation results are mainly non-maximum suppression (NMS), which is used to obtain feature local maxima. A larger NMS value can ensure that the points are evenly distributed in the image, which we set to 4 and 8 for comparison. The correction distance parameter ε is set to 3 pixels by default and the first 300 points are selected for testing in a pixel size of 480 * 640. The final test results are shown in Table 2. Our method has a certain degree of improvement compared with SuperPoint in illumination and perspective change scenes, so it can be inferred that our strategy of incorporating feature fusion and homography constraints improves the keypoints' repeatability in different scenes.

In order to verify that our model performs better than other methods, we compare the baseline with our method on two keypoint evaluation metrics and two visual task metrics. The first two are classical metrics for evaluating the performance of keypoints. The visual localization is a valuable application area, and we calculate localization error by our feature detection method and nearest neighbor search. Then homography estimation refers to solving the homography transformation matrix by extracting keypoints and descriptors of known image pairs, which is widely used in image registration and panorama stitching. The methods provided by OpenCV accomplish the traditional feature extraction. LIFT is realized by the pretrained model supplied by the author. The correction distance parameter ε is set to three pixels, and the first 1000 points are selected from the pixel size of 480 * 640 for testing. As seen in Table 3, compared to baseline, our method achieves the best performance in all four metrics, especially in the homography estimation task, which improves by more than 10% compared to other methods, mainly thanks to the constraints between image pairs and feature fusion we implemented, which enhances relevance between warped image pairs and optimize network's parameters. The improvements in these metrics prove our method's effectiveness.

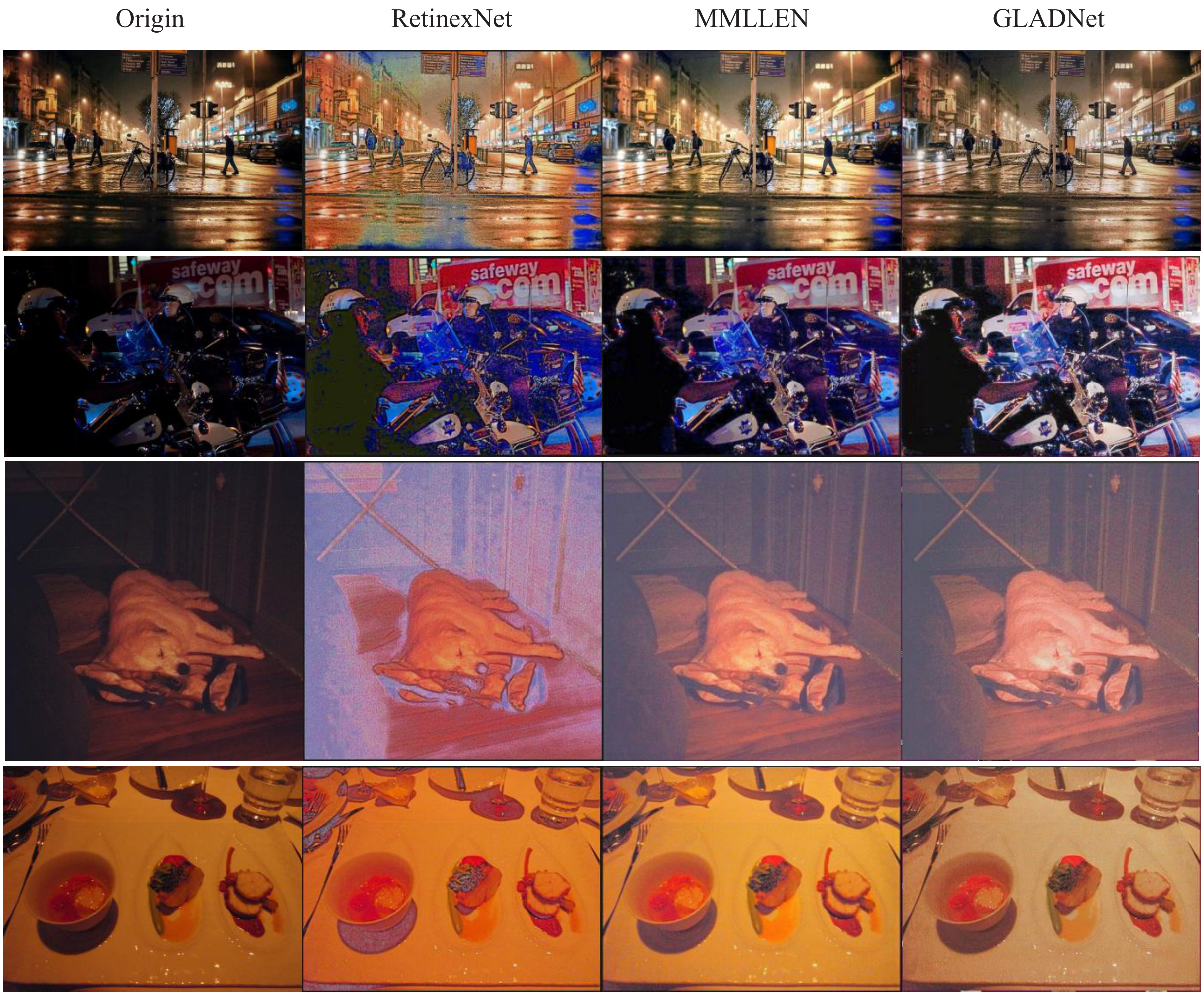


Fig. 4. Visualization comparison of different light enhancement methods

To further compare the impact of parameters on the homography estimation task, we choose the correction distance parameter ε, which greatly impacts the results, for the experiments. The ε is measured in pixels, meaning that as long as the predicted feature points are within ε pixels from the ground truth, it can be considered a correct prediction. Therefore, the larger the ε is, the higher the predicted feature points' accuracy. Here we set ε to 1,3,5 pixels for experiments, respectively. Besides, we compare the latest feature extraction methods based on the original baseline. The results are shown in Table 4, where our method still delivers the best performance and achieves state-of-the-art homography estimation.

Meanwhile, we randomly select several groups of images in the Hpatches dataset under different perspective transformation scenes to visualize the image matching and intuitively compare the model effects before and after the improvement. As shown in Fig. 5, the red points represent the unmatched points, whereas the green lines represent successfully matched pairs. We mark the differences in matching between the two methods with yellow boxes, and we can see that our model is able to find more accurate matching pairs than the original baseline method, which is more capable of meeting precise feature matching requirements.

Table 2. Repeatability evaluation of illumination and perspective change scenes on HPatches dataset

Methods	57 illumination transformation scenes		59 perspective transformation scenes	
	NMS=4	NMS=8	NMS=4	NMS=8
Fast	0.481	0.438	0.570	0.499
Harris	0.526	0.495	0.628	0.543
Shi-Tomas	0.486	0.452	0.585	0.530
Random	0.026	0.027	0.046	0.046
SuperPoint	0.538	0.537	0.547	0.546
Ours	0.550	0.540	0.564	0.551

Table 3. Evaluation metrics of the feature extraction algorithms on the HPatches dataset

Methods	Repeatability	mAP	Localization Error	Homography Estimation
SIFT	0.495	0.694	0.833	0.676
ORB	0.641	0.735	1.157	0.395
LIFT	0.449	0.664	1.102	0.598
SuperPoint	0.543	0.853	0.229	0.667
Ours	0.557	0.864	0.223	0.784

Table 4. Homography estimation results in different correction distance parameter

Methods	Homography estimation		
	ε=1	ε=3	ε=5
SIFT	0.424	0.676	0.759
ORB	0.150	0.395	0.538
LIFT	0.284	0.598	0.717
SuperPoint	0.310	0.684	0.829
LOFTR	--	0.659	0.756
Ours	0.424	0.784	0.871

4.4 Ablation Studies

We conduct ablation studies on each improvement to verify their effect on the feature extraction network. According to the default training and testing partitioning of ExDark, we evaluate features using precision and recall metrics. The ablation studies are conducted on the basis of SuperPoint from three aspects: image enhancement, feature fusion, and loss optimization. Results are shown in Table 5. Each of them is added based on the previous one.

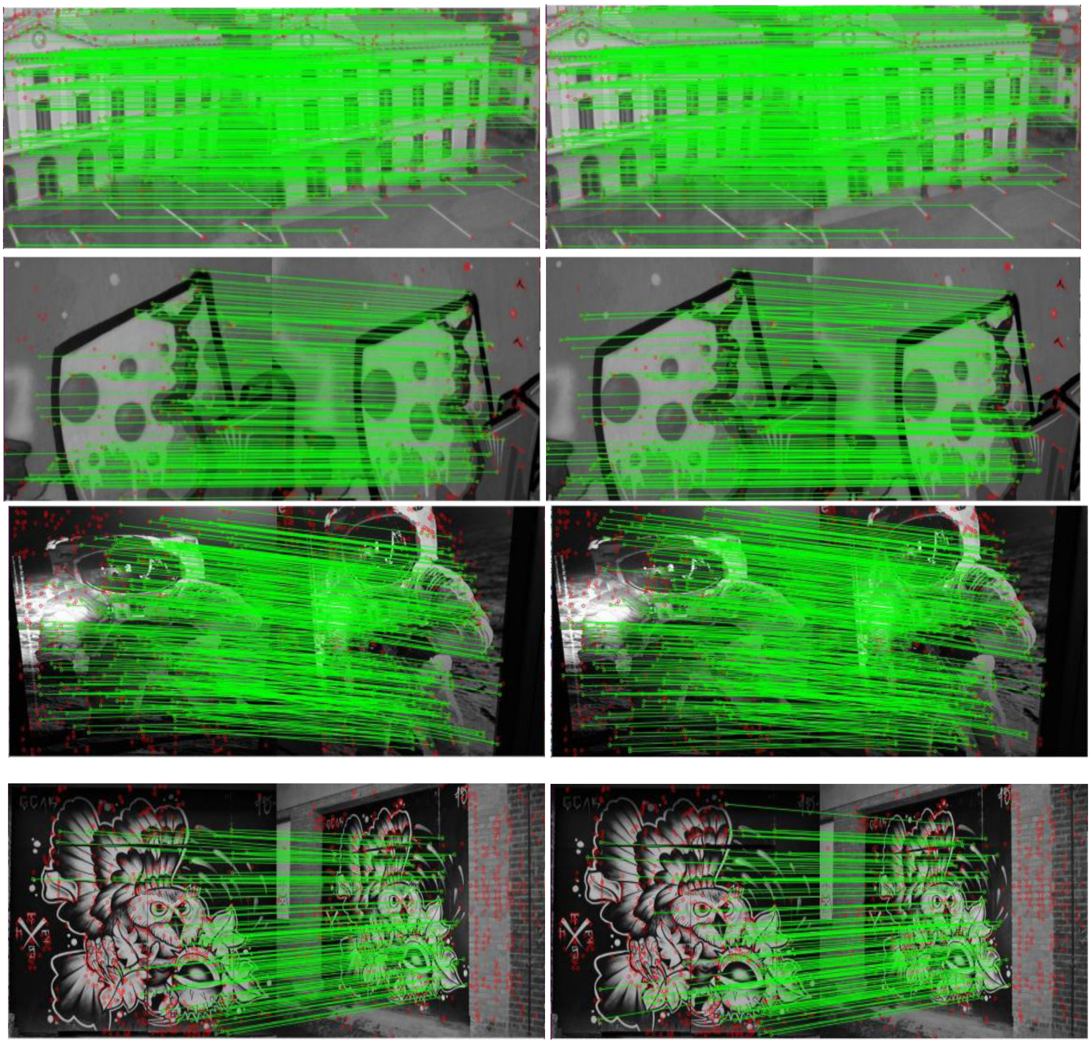
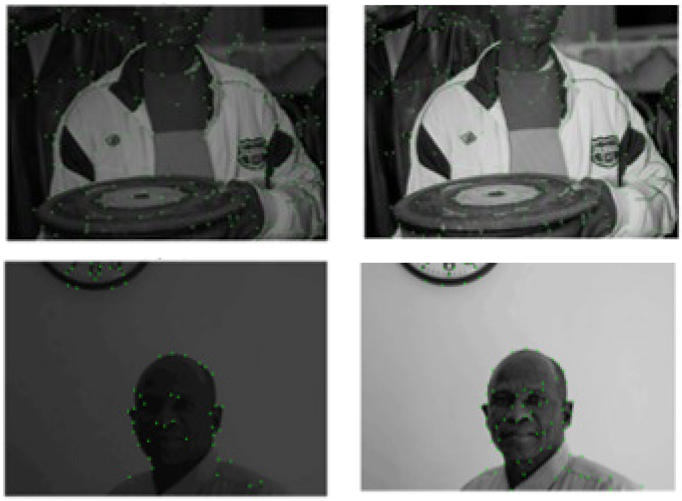


Fig. 5. Visualization of image matching results based on perspective transformation scenes

Table 5. Feature extraction network ablation studies on the ExDark dataset

Methods	Precision	Recall	Loss
SuperPoint	0.200	0.384	1.825
Low-illumination enhancement	0.230	0.436	1.899
Feature fusion	0.262	0.483	1.791
Homography loss	0.266	0.479	2.793

Image enhancement increases the visibility of low-illumination images and makes the keypoints clearer so that the model can show a better feature extraction performance in low-illumination scenes. As shown in Tab. 5, its effectiveness can be verified by the precision and recall of 3% and 5.2% increments, respectively. In addition, we conduct qualitative experiments on image feature extraction before and after light enhancement. We test the model trained with ExDark on the GladNet-Dataset low-illumination dataset by randomly selecting two images for keypoint extraction visualization and comparing them with the effect after light enhancement. As shown in Fig. 6, the number and quality of extracted keypoints are better in high-illumination images than in low-illumination images, demonstrating the necessity of light enhancement for feature extraction in low-illumination scenes.



(a) Keypoints in low illumination scenes (b) Keypoints in high illumination scenes

Fig. 6. Visualization results of low and high illumination images

In addition, feature fusion improves feature point accuracy by 3.2% and recall by 4.7%. This is primarily due to the fact that we combine the two methods on the feature map to capitalize on their strengths, allowing the network to learn the optimal weights to predict more accurate features in an adaptive manner. To visualize the changes brought by feature fusion, we perform qualitative experiments on the prediction points derived from the SuperPoint and the model after adding feature fusion. We randomly select four images from the test set, as shown in Fig. 7. Both models detect the white points before and after improvement. The SuperPoint detects yellow points, but not our model. Our model detects blue points, whereas the SuperPoint did not, and there are more yellow points than blue ones. Nonetheless, the majority of these extra points are of lower quality, indicating that our model screens out a large number of non-feature points and retains more meaningful feature points, thereby improving the accuracy of feature point prediction.

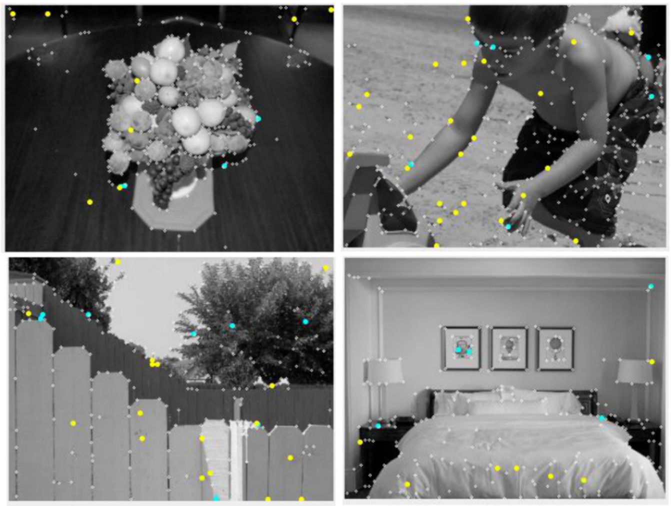
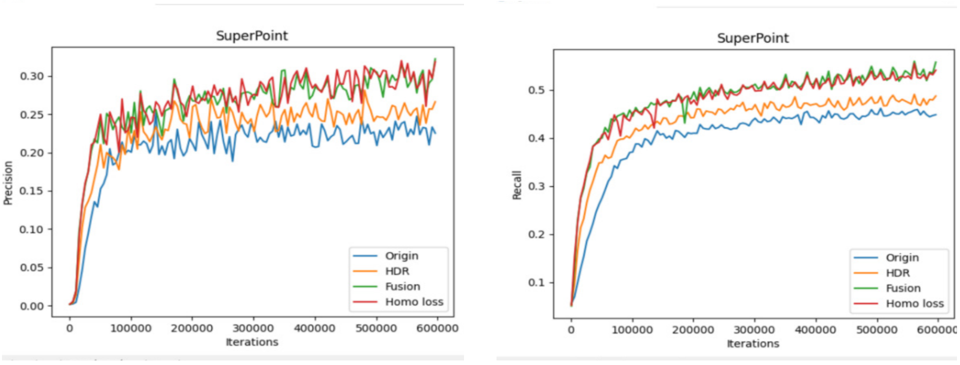


Fig. 7. The original and feature fusion models extracted the feature points

Last but not least, the loss term is added to fully use the homography transformation input image pairs and optimize the performance effect of the features extracted by the network on the visual tasks. The precision obtained by this term is similar to the previous one. However, because of the increased constraints between homography images, the probability of detecting the same points in images with scene changes has increased. The predicted feature points' repeatability is improved, making a more significant contribution to the homography estimation task. Besides, weighting the loss term can also achieve better results in practical applications.

We visualize the training process for each improvement of the ablation experiment, as shown in Fig. 8. Because the last change adds one item to the loss function, the loss value is similar before and after this improvement. The training speed was comparable before and after the addition of each improvement, whereas the differences are magnitudes of accuracy. The blue curve represents the trend of the original SuperPoint model, we take it as our baseline owing to its significant performance in deep learning methods, and conduct effect comparative experiments on the ExDark as above. While the orange curve is the trend of the enhanced low-illumination dataset, the green and red curves are the results of adding feature fusion and homography loss items to the image enhancement basis, respectively.



(a) Precision (b) Recall

Fig. 8. Visualization of precision and recall training results for ablation studies

In addition, we fine-tune the loss term scaling parameters α and β. In addition to $\alpha = 1.5$ and $\beta = 0.5$, we test two sets of parameters used in the aforementioned ablation studies and obtain the results shown in Table 6.

Table 6. Training effect comparison of different weighting parameters

Parameters	Precision	Recall	Loss
$\alpha = 1, \beta = 1$	0.248	0.459	2.841
$\alpha = 1.2, \beta = 0.8$	0.256	0.465	2.831

5 Conclusion

This paper proposes a feature fusion method between manual features and deep learning ones for low-illumination images. The image enhancement network is added to the original SuperPoint network. Then SIFT manual extraction features are fused into the SuperPoint framework at the feature map level. Combined with the advantages of the two methods, the network weight is affected by changing output value of the back propagation network. The fusion is realized in the true sense. The results show that the network performance and robustness of the feature extraction method are improved. On this basis, the homography transformation between image pairs is used to enhance the constraints of the network and optimize the network parameters, improve the repeatability of perspective transformation scenes and the accuracy of homography estimation task. From the results, our proposed method improves the precision and recall of the feature and performs well in multiple indicators.

However, the network requires more preparation for image keypoints extraction of SIFT and illumination enhancement processing. Therefore, our future research is to simplify the network steps and then apply them to augmented reality tasks in low-illumination scenes using the advantage that the method can extract more accurate features in low-illumination conditions and achieve better results in real-time.

Acknowledgements

This work is supported by National Natural Science Foundation (62162007) and Guizhou Provincial Natural Science Foundation (Joint Science Foundation [2019] 1088).

References

[1] X.Y. Jiang, Q.M. Chen, C.H. Huang, A dynamic-traversal-based Hierarchical Feature Network for visual location, Computer Engineering 47(9)(2021) 197-202.

[2] H. Taira, M. Okutomi, T. Sattler, M. Cimpoi, InLoc: Indoor visual localization with dense matching and view synthesis, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[3] P.E. Sarlin, F. Debraine, M. Dymczyk, R. Siegwart, C. Cadena, Leveraging Deep Visual Descriptors for Hierarchical Efficient Localization, in: Proc. Conference on Robot Learning (CoRL), 2018.

[4] P.E. Sarlin, A. Unagar, M. Larsson, H. Germain, Back to the Feature: Learning Robust Camera Localization from Pixels to Pose, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[5] S. Zhu, R. Zhang, L. Zhou, T. Shen, Very large-scale global sfm by distributed motion averaging, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[6] C. Campos, R. Elvira, J.J.G. Rodríguez, J.M.M. Montiel, J.D. Tardós, ORB-SLAM3: An Accurate Open-Source Library for Visual, Visual–Inertial, and Multimap SLAM, IEEE Transactions on Robotics 37(6)(2021) 1874-1890.

[7] C. Wang, R. Martín-Martín, D. Xu, J. Lv, 6-pack: Category-level 6d pose tracker with anchor-based keypoints, in: Proc. IEEE International Conference on Robotics and Automation (ICRA), 2020.

[8] C.R. Qi, H. Su, K. Mo, L.J. Guibas, Pointnet: Deep learning on point sets for 3d classification and segmentation, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[9] T. Lindeberg, Scale invariant feature transform, Scholarpedia 7(5)(2012) 10491.

[10]H. Bay, T. Tuytelaars, L.V. Gool, Surf: Speeded up robust features, in: Proc. European Conference on Computer Vision (ECCV), 2006.

[11]S. Leutenegger, M. Chli, R.Y. Siegwart, BRISK: Binary robust invariant scalable keypoints, in: Proc. IEEE International Conference on Computer Vision (ICCV), 2011.

[12]E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient alternative to SIFT or SURF, in: Proc. IEEE International Conference on Computer Vision (ICCV), 2011.

[13]D. DeTone, T. Malisiewicz, A. Rabinovich, Superpoint: Self-supervised interest point detection and description, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[14]J. Shi, Good features to track, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1994.

[15]E. Rosten, T. Drummond, Machine learning for high-speed corner detection, in: Proc. European Conference on Computer Vision (ECCV), 2006.

[16]H. Yu, F. Guo, J. Wang, Q. Fu, Robust monocular visual-inertial SLAM based on the improved SuperPoint network, Chinese Journal of Scientific Instrument 42(1)(2021) 116-126.

[17]E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, Discriminative Learning of Deep Convolutional Feature Point Descriptors, in: Proc. IEEE International Conference on Computer Vision (ICCV), 2016.

[18]K. Yi, E. Trulls, V. Lepetit, P. Fua, Lift: Learned invariant feature transform, in: Proc. European Conference on Computer Vision (ECCV), 2016.

[19]Y.F. Song, L. Cai, J. Li, Y. H. Tian, M.Y. Li, SEKD: Self-evolving keypoint detection and description. <https://arxiv.org/abs/2006.05077>, 2020 (accessed 09.06.2020).

[20]J. Sun, Z. Shen, Y. Wang, H. Bao, X. Zhou, LoFTR: Detector-free local feature matching with transformers, in: Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[21]J. Tang, J. Folkesson, P. Jensfelt, Geometric correspondence network for camera motion estimation, IEEE Robotics and Automation Letters 3(2)(2018) 1010-1017.

[22]Q. Jia, X. Gao, H. Guo, Z. Luo, Y. Wang, Multi-layer sparse representation for weighted LBP-patches based facial expression recognition, Sensors 15(3)(2015) 6719-6739.

[23]L. Wang, R.F. Li, K. Wang, J. Chen, Feature representation for facial expression recognition based on FACS and LBP, International Journal of Automation and Computing 11(5)(2014) 459-468.

[24]L. Wang, Z. Zhang, L. Su, W. Nie, Target classification with adaptive weights based on multi-feature fusion, Journal of Huazhong University of Science and Technology(Natural Science Edition) 48(9)(2020) 38-43.

[25]N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2005.

[26]Y. Zhang, N. He, R. Wei, Face Expression Recognition Based on Convolutional Neural Network Fusing SIFT Features, Computer Applications and Software 36(11)(2016) 161-167.

[27]W. Wang, C. Wei, W. Yang, J. Liu, GLADNet: Low-light enhancement network with global awareness, in: Proc. IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2018.

[28]Q. Wang, X. Zhou, B. Hariharan, N. Snavely, Learning feature descriptors using camera pose supervision, in: Proc.

European Conference on Computer Vision (ECCV), 2020.

[29]Y.P. Loh, C.S. Chan, Getting to know low-light images with the exclusively dark dataset, Computer Vision and Image Understanding 178(2019) 30-42.

[30]V. Balntas, K. Lenc, A. Vedaldi, K. Mikolajczyk, HPatches: A benchmark and evaluation of handcrafted and learned local descriptors, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[31]R.J. Huang, H. Cui, Q.M. Cheng, C.H. Huang, Low-light image enhancement based on multi-branch residual and affine transformation, Application Research of Computers 38(12)(2021) 3786-3790+3807.

[32]C. Wei, W. Wang, W. Yang, J. Liu, Deep retinex decomposition for low-light enhancement. <https://arxiv.org/abs/1808.04560>, 2018 (accessed 14.08.2018).