Attention-Capsule Network for Low-Light Image Recognition

Shiqi Shen*, Zetao Jiang†, Xiaochun Lei, Xu Wu, Yuting He,
School of Computer Science and Information Security,
Guilin University of Electronic Technology, Guilin, Guangxi, China, 541004
College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060,
Email: *sq.shen.work@gmail.com, †zetaojiang@guet.edu.cn,
lxc8125@guet.edu.cn, csxuwu@163.com, yuting.h.work@gmail.com,

***Abstract*—Most of the existing deep learning algorithms are used for image recognition under sufficient light conditions. However, the critical feature information in the low-light image is hard to learn. Therefore, recognizing low-light images with insufficient light is a challenging task. We propose a novel end-to-end model named attention-capsule Networks (ACNet) for low-light image recognition tasks to overcome this drawback. The proposed model is based on the capsule structure, and its key component is the Global-Local Attention module (GLA). To effectively and comprehensively obtain important information, we design the global block with larger receptive fields to obtain global perception information and the local block to obtain local detail information. In addition, this paper proposes a directional learning loss, which guides the model to extract key features of the image by optimizing the error between the reconstructed image and the normal-light image. The experimental results demonstrate the effectiveness of the ACNet model for low-light images recognition.**

I. Introduction

Deep learning has made significant breakthroughs in image recognition of normal-light conditions. However, due to the exposure of the shooting equipment and the insufficient light of the real scene, the obtained images may have low brightness, blurred content, and loss of a lot of crucial information [1]. This leads to poor visual effects and makes subsequent image processing such as image recognition, detection, and segmentation work poorly. There are fewer recognition algorithms for low-light scenes. Previous low-light image recognition [2]–[4] methods use image preprocessing, such as image enhancement or screening, to improve the quality of input images. These methods have achieved certain success, but they have high complexity and low efficiency. Concretely, image enhancement operations improve the contrast and brightness of the image. However, the noise in the image is often amplified, which brings negative effects for the recognition model [5]. There are also methods [6]–[9] that focus on increasing the brightness of images, but these methods are not aimed explicitly at low-light recognition. Therefore, there are still many challenges in low-light image recognition.

This paper proposes an attention-capsule model (ACNet) for low-light image recognition to address the aforementioned problems. The ACNet performs the low-light image recognition end-to-end, and it consists of a capsule structure(CapsNet) [10] and an attention module. The capsule structure is selected to make up for the lack of spatial perception and spatial reasoning capabilities in CNNs. It can better obtain semantic information and spatial structure information in low-light images. The representation captured by CapsNet usually corresponds to the visual attributes of human-understandable objects. Since it is difficult to excavate information in extreme environments with low illumination, we designed a module to improve feature extraction ability. In addition, the reconstruction module is proposed to generate images with enhanced brightness, which makes the entire model learn more helpful information for image recognition. To train the reconstruction module, we design a directional learning loss which is performed by computing the error between normal-light images and the reconstructed image. Our contributions can be summarized as follows:

- We proposed an end-end model, ACNet, for low-light image recognition.
- We proposed a new global-local attention module, which simultaneously obtains global features and local detail features of low-light images and strengthens the model's ability to extract salient features in the early stage.
- We proposed a directional learning loss to guide the reconstruction module to generate normal-light images.
- Experiments show that our method is effective on low-light image recognition. Meanwhile, the proposed model has achieved better results than the state-of-the-art conventional recognition model.

II. Related Work

Convolution-based models for recognition Nowadays, deep learning has developed more maturely in image recognition, especially CNNs have achieved remarkable results. Since AlexNet [11] was proposed, CNN has ushered in the era of its explosion. GoogLeNet [12] proposed the concept of Inception, which uses convolution kernels of different sizes to extract features from images, and then perform feature fusion. On this basis, some improved versions [13], [14] further improve performance. Since then, as the depth of CNN continues to stack, many researchers find that the problem of model degradation serious follows. To this end, K He et al. [15] proposed ResNet. The model directly propagates

Manuscript 396 submitted to 26th International Conference
on Pattern Recognition. Received January 23, 2022.

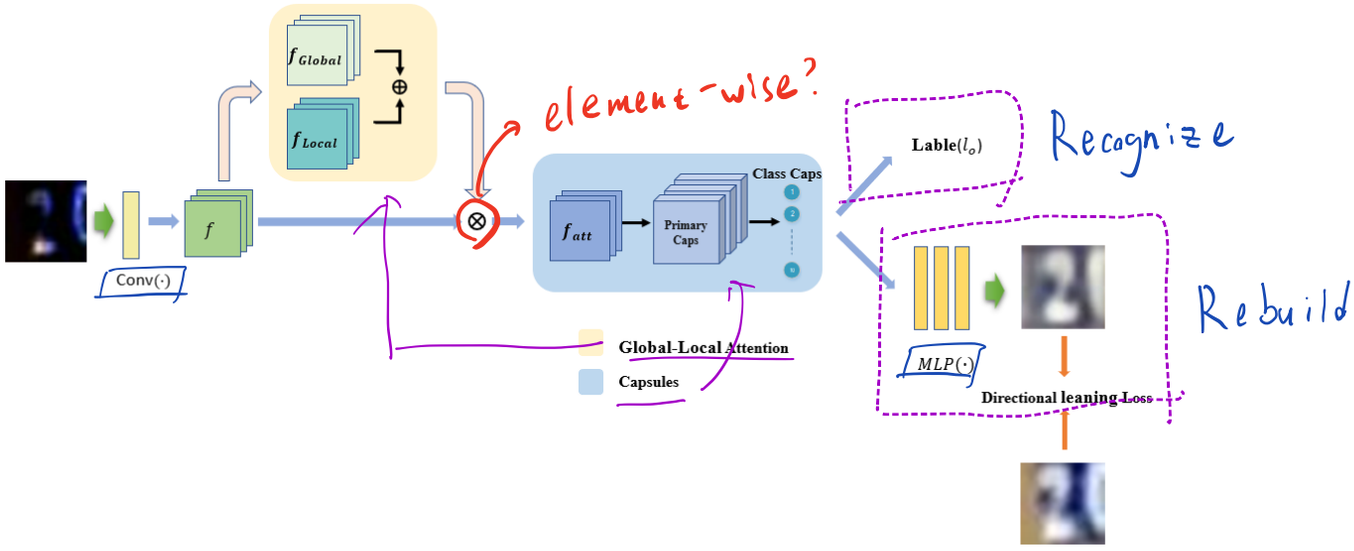

Fig. 1. Overview of our proposed model. Given arbitrary input low-light image I_x, we first apply a simple CNN layer to pre-process I_x. And then, the global-local attention module is applied to extract global and local representations and assign weights to fuse original features. Next, we feed the features to the capsule-based model for classification and reconstruction.

information from any lower layer to higher layers through residual learning, which alleviates the gradient vanishing and gradient exploding caused by network deepening. G Huang et al. [16] proposed DenseNet, which realizes feature reuse and reduces the amount of computation through the connection of features on channels. DPN [17] combines the advantages of ResNet and DenseNet to construct a new connection path topology internally. J Hu et al. proposed the channel attention module SENet [18], which explicitly models the relationship of each channel of the feature map. Subsequently, CBAM [19] adds spatial attention to channel attention to enhance the representation of features. In SKNet [20], the attention mechanism is proposed to be applied to the convolution branches of different scales to improve the adaptive ability of the model to the receptive field. T et al. [21] propose a new classification loss to match the proportion of known labels. FcaNet [22] rethinks channel attention using frequency analysis and generalizes the channel attention preprocessing mechanism in the frequency domain. ResNeSt [23] is based on ResNet and combines the idea of multiple models, such as multi-path and feature-map attention, to build the Split-Attention module. It improves the performance significantly while keeping the number of parameters not increasing dramatically.

Capsule-based models for recognition Compared with traditional CNNs, CapsNet has excellent semantic information acquisition capabilities and can perceive subtle information changes in images and spatial reasoning capabilities. Meanwhile, it can overcome the shortcomings of CNNs that require a large amount of data to learn features. After that, EM Routing [24] was proposed, using the expectation-maximization algorithm to update iteratively. In order to further improve the capsule, SelfRouting [25] is proposed as a supervised non-iterative routing algorithm, which reduces the amount of calculation. In the latest research of CapsNet, some people also proposed more effective routing algorithms [26]–[28] and analyzed the contribution of routing algorithms [29], [30]. Scholars have also used the capsule network for very-low resolution image recognition [31], image segmentation [32], video processing [33], and so on, all of which have achieved outstanding results.

Low-Light Image Recognition Low-light image recognition has received relatively little attention, and the current practice in this field is mainly through low-light enhancement and then as a common recognition task. Several works have been proposed in recent years for low-light enhancement methods [6]–[9]. These methods achieve image enhancement by finding the mapping relationship from low-light to normal-light image distribution. More specifically, some works [2]–[4] have applied this idea to various classification tasks in recent years. Ren et al. [4] proposed a low-light image processing method for face recognition. The face image is preprocessed by the image multiplication method, which can improve the recognition rate of the face in the subsequent recognition algorithm. Zhang et al. [2] proposed an image light classification method under different illumination conditions, which trains the classification of image light according to the characteristics of different illumination images. Sharma et al. [3] is a method of supervised action recognition in the dark. It classifies the actions in the video by training the dynamic enhancement of low-light images and the sampling of the Delta strategy. HLA-Face [34] proposed a novel method to enhance low-light face images and detect the low-light face representation. In practice, our proposed method can also perform image augmentation to effectively learn helpful information for classification from the image distribution map and embed it well in the overall

network.

III. Attention-Capsule Network

A. Overview.

Low-light image recognition tasks have great potential for application in real scenes. Generally, low-light images are challenging to recognize because it always contains interference information, such as low brightness, uneven illumination, unobvious image features, and noise. Therefore, most recognition algorithms have poor performance in low-light images. We proposed an attention-capsule network(ACNet) for this shortage, which can well mine high-level semantic information in low-light images. The overview of our proposed framework as shown in Fig. 1.

We propose an end-to-end model mainly composed of two parts: feature enhancement and classification modules. Given a low-light image I_x, we first feed it to a CNN layer to get features f. And then use this feature enhancement module, namely the global-local attention module(GLA), to handle complex information in the low-light image. The formula is described as:

$$f_{att} = \mathcal{F}(GB(f), LB(f)) \tag{1}$$

where GB, LB denote global and local blocks, respectively. $\mathcal{F}$ is a fusion function to combine the information from GB and LB, and f_{att} is the enhanced feature.

Next, we apply capsule network [10], which is proven to have powerful feature mining and classification, as our backbone. Due to the features of low-light images are more difficult to mine than normal-light features. It makes it difficult for common backbones, such as VGG [35] and ResNet [15], to apply this task effectively. For stronger fitting ability, we adopt CapsNet as our backbone to process the enhanced feature f_{att}. Meanwhile, we proposed a novel directional learning loss to reconstruct input low-light image to normal-image(ground truth), which could improve the mining ability of our model. More formally, this process can be described as follows:

$$I_x^*, l_o = \mathcal{C}(f_{att}) \tag{2}$$

where I_x^* denotes the reconstructed image, l_o is the final predicted label of our model, and $\mathcal{C}$ is the capsule-based module in our model.

B. Global-Local Attention Module.

As shown in Fig. 2, the global-local attention module is mainly composed of two parts, namely the global block and the local block, so that the model can obtain global perception information and local detail information at the same time.

Global Block. This structure is used to deal with the influence of low brightness in low-light images. Since the low-light image's limited quality and identifiable information, we designed a global block with the idea of expanding the receptive field size. The input with a shape of $H \times W \times C_{in}$ is fed to the global block. We first fed an input to global average pooling to obtain global information and get an output of shape $1 \times 1 \times C_{in}$.

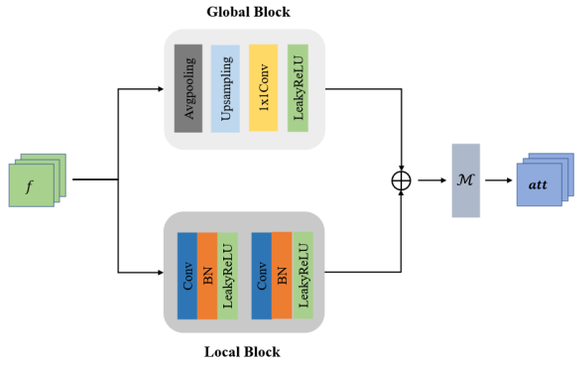

Fig. 2. The architecture of our global-local attention module(GLA). Our GLA module contains two blocks, global block and local block, which are responsible for global features and local detail features respectively. The $\mathcal{M}$ is a fusion function to combine features from the global block and the local block.

Then, we adopted an up-sampling layer to upscale the down-sampled global information and produced a feature map with a size of $H \times W \times C_{global}$. Next, we apply convolution operations in the channel dimension to further process the information. Our designed global block can handle global information and enables the model to enhance the ability to low-light image recognition. In addition, the extracted global context information provides relevant features required by subsequent capsule instances. This block can help the entire model have stronger feature resolution capabilities in the feature extraction process.

$$f_{\text{Global}} = F_G(f) = Conv_{1\times1}\left(F_{up}\left(F_{\text{avg}}\left(f\right)\right)\right) \tag{3}$$

where F_G denotes our global block. $F_{avg}(\cdot)$ is a global average pooling(GAP), which compresses the spatial dimensions of features through downsampling. Compared with global max-pooling(GMP), GAP pays more attention to the overall information of images. Then restore the size by Upsampling layer $F_{up}(\cdot)$. Moreover, the symbol $Conv_{1\times1}$ contains a 11 convolution and LeakyReLU activation function, which can bring more nonlinear to the model.

Local block. Compared with global features, tiny local features are also non-trivial. Benefiting from the excellent local response characteristics of convolution, we improve the local information extraction ability by increasing the fitting ability of the network in the local block. In practice, we build the local block with standard convolution blocks(Conv+BN+LeakyReLu) to achieve excellent results. It is because the local region we need to enhance can be well-mined by a convolution kernel of size 3×3. This block learns and pays attention to the detailed information in the low-light image, which is used as an essential feature in the subsequent recognition process.

$$f_{\text{Local}} = F_L(f) \tag{4}$$

where F_L denotes our local block, and f_{Local} is output features. The F_L is composed of two groups of convolution blocks. Each convolution block contains a convolu-

tion, batch normalization (BN) [13], and activation function (LeakyReLU).

Global-local attention module. We fuse the features (f_{Global} and f_{Local}) extracted from the global and local blocks. The attention module composed of a combination of global and local blocks can capture richer representational information in low-light images, which is beneficial to enhance feature information.

$$att = \mathcal{M}(f_{Global}, f_{Local})) \tag{5}$$

$$f_{att} = att \times f \tag{6}$$

where $\mathcal{M}$ denotes a fusion network to reintegrate the features. In this way, the weight att of the entire global-local attention module can be obtained. This weight will be adopted to the feature f of the input attention module to obtain a new attention feature map f_{att}.

C. Capsules Module.

As mentioned earlier, considering the advantages of CapsNet, such as stronger semantic information acquisition capabilities and equivariance characteristics, we select the capsule structure containing dynamic routing algorithms in CapsNet for image recognition. In our preliminary attempts, we found it can improve performance better to fuse the features from the previous module. As shown in Fig. 1, the capsule structure in the model consists of two capsule layers, namely the Primary Caps layer and the Class Caps layer.

$$c_i = F_{primary}(f_{att}) \tag{7}$$

$$v_j = Routing(c_i) \tag{8}$$

where $F_{primary}$ is the Primary Caps layer, which gets the feature maps through a convolution operation to obtain the capsule vector group $c_i \in \mathbb{R}^8 (i = 0, 1, \ldots, 2047)$. And $v_j \in \mathbb{R}^{16} (j = 0, 1, \ldots, 9)$ is the classification capsule of the Class Caps layer. This layer includes 10 capsules, each of which has a dimension of 16 and contains identifiable image information. $Routing$ is a dynamic routing algorithm [10] that maps low-level capsules to high-level capsules through three iterations.

D. Reconstruction Module.

In the original CapsNet, the reconstruction module fits the image generated by the digital capsule to the input image. This module has been proven to be an effective method to detect adversarial attacks to verify that the model can perceive the sufficient information we expect it to perceive [36]. In ACNet, to make the reconstruction module better learn the adequate category information in the image, we use this module to make the model learn to fit the normal-light image instead of the input low-light image. Therefore, we propose a directional learning loss, as described in III-E, to guide the reconstruction module, which forces the class capsule to contain more helpful information in the normal-light image. Experiments can prove that the reconstruction module in ACNet has additional contributions to low-light image recognition tasks.

$$I_x^* = MLP(v_j) \tag{9}$$

where v_j denotes that the predicted capsule output and I_x^* are reconstructed images from our model. v_j is fed to an MLP network, which contains three fully connected layers, to reconstruct images. Finally, we use the directional learning loss to constrain the model fitting to get the image I_x^* with normal illumination.

E. Losses

In this part, we summarize the training objectives applied in our method.

Margin Loss. The Class Caps layer predicts the existence probability of each class through the loss guidance. The length of the vector represents the probability of each class.

$$\begin{aligned} \mathcal{L}_M = T_j \max\left(0, m^+ - \|v_j\|\right)^2 \\ + \lambda \left(1 - T_j\right) \max\left(0, \|v_j\| - m^-\right)^2 \end{aligned} \tag{10}$$

where j is denoted as the class, and $T_j = 1$ when class j is present, otherwise $T_j = 0$. The upper and the lower margins are set to m^+=0.9 and m^-=0.1, respectively. In addition, $\|v_j\|$ is expressed as the L2-norm of the vector v_j.

Directional learning Loss In CapsNet, the ground truth image of reconstruction loss is the input image of the network. However, the reconstructed image expected in ACNet is not the low-light image. Since we hope that in training the model, we can learn identifiable and influential features, the normal-light image is used as the ground truth image, and the fitting model to learn more information related to the normal-light image. Through experiments, we have verified that the directional learning loss L_D can guide the class capsules to generate normal-light images. Furthermore, it prompted the model to have more differences in both the acquired features and the transmitted capsule information to strengthen the recognition ability of the entire model. The formula is as follows:

$$\mathcal{L}_D = ||I_t - I_x^*||_2^2 \tag{11}$$

where I_x^*, I_t is the restructured image of our model and ground truth image, respectively.

Total Loss. Combining the above two loss functions, the total loss is:

$$\mathcal{L}_{total} = \mathcal{L}_M + \lambda \mathcal{L}_D \tag{12}$$

To make $\mathcal{L}_M$ account for more proportion, we adopt $\lambda = 0.0005$ as the weight value of $\mathcal{L}_D$.

IV. Experiments

A. Dataset

Since the existing low-light image data sets are used for enhancement tasks, no class labels are annotated for recognition tasks. In this section, we mainly introduce how to build the low-light dataset for recognition.

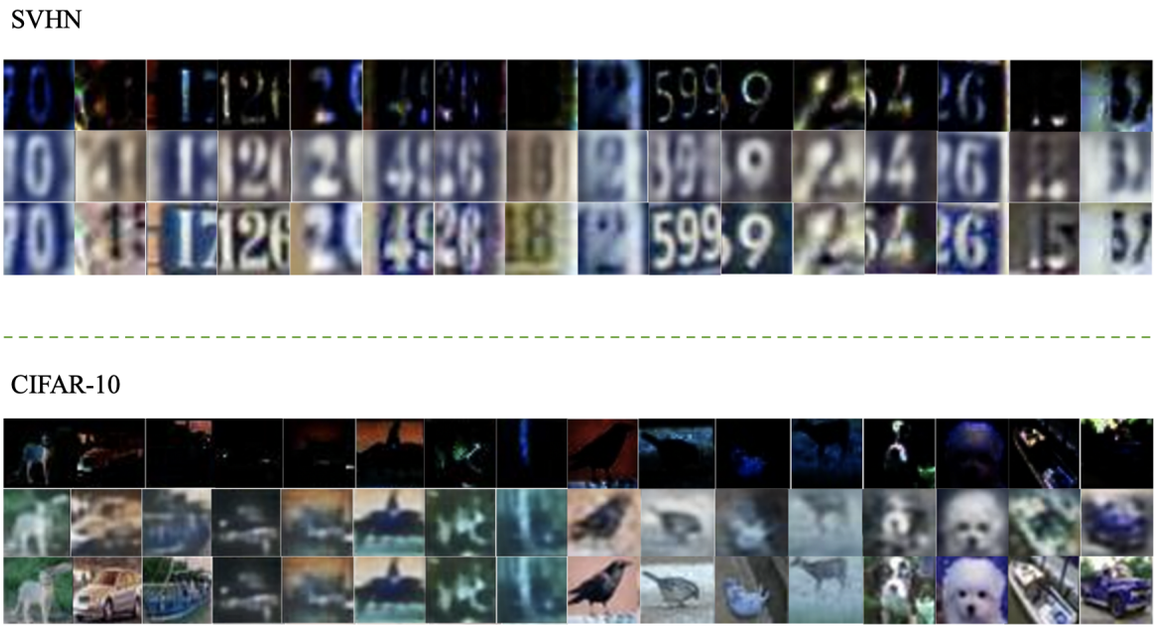

Fig. 3. Image visualization comparison of SVHN and CIFAR-10 datasets. The upper and lower parts represent the SVNH and CIFAR-10, respectively, and each group contains three sets of images. The first row is the shifted image by gamma correction and gaussian blur. The second row is the reconstructed image output by our model, and the third row is the normal-light image which is the ground truth in our experiment.

The built datasets are shown in Fig. 3. We use Gamma correction as the non-linear transformation to adjust the brightness degree of the image and turn the normal-light image into the low-light image. The mathematical formula of gamma correction is defined as follows:

$$I_{low} = I_{normal}^{\gamma} \tag{13}$$

where $\gamma \geqslant 1$ denotes the hyper-parameter of Gamma correction.

In addition,low-light images generally have blur characteristics in the wild. Thus, we also adopted Gaussian blur to make the image look more realistic and finally get a low-light image data set with low brightness and blurred imaging. The whole process is formulated as:

$$I_{low} = G(\mu_{blur}, \sigma_{blur}) \times I_{normal}^{\gamma} \tag{14}$$

In practice, we randomly choose the values of γ and σ in the range of $[4, 4.5]$ and $[0.1, 0.5]$, respectively. Meanwhile, we set μ equal to 0.

In order to prove the effectiveness of the proposed model, we create the $< I_{low\text{-}light}, I_{normal\text{-}light} >$ image pair from the SVNH dataset and CIFAR-10 dataset based on the method described above.

B. Implementation Details

After the process described by IV-A, we crop and resize all input images to the resolution of 32x32 for all training proceed. Our proposed model is implemented based on Tensorflow framework and trained with the widely-used Adam optimizer ($\beta_1 = 0.5$, and $\beta_2 = 0.999$). In the training process, we train our model for 50 epochs with a batch size of 128. The initial learning rate of our model is set to 1e-4. All experiments were run on two NVIDIA GTX 2080ti GPUs.

TABLE I
QUANTITATIVE COMPARISON IN LOW-LIGHT SVHN.

Algorithm	Accuracy
ResNet18 [15]	88.59
ResNet34	89.25
ResNet50	90.37
GoogleNet-v1 [12]	90.61
DenseNet101 [16]	90.62
Se-ResNet18 [18]	88.62
CBAM-Resnet50 [19]	88.86
Fcanet [22]	90.18
ResNeSt [23]	92.55
Ours	**94.30**

C. Comparison

In this part, we compare our model with other recognition networks. As far as we know, there is currently no widely recognized recognition network that focuses on low-light recognition. Thus, we compare our method with other widely-knows recognition networks with strong representation ability for various recognition tasks.

Table I shows that the compares our best results to the state-of-the-art methods on the low-light SVNH dataset. We first contrast our method with general networks, such as ResNet-based models [15], DenseNet101 [16]. The recognition ability of our model is significantly improved than this type of method. And then, we compare our method with attention-based recognition networks. SENet [18] and CBAM [19]

TABLE II
QUANTITATIVE COMPARISON IN LOW-LIGHT CIFAR-10.

Algorithm	Accuracy
ResNet18 [15]	58.20
ResNet34	44.74
ResNet50	53.30
GoogleNet-v1 [12]	60.54
DenseNet101 [16]	63.01
Se-ResNet18 [18]	63.49
CBAM-Resnet50 [19]	54.64
FcaNet [22]	55.24
ResNeSt [23]	61.78
Ours	**64.24**

are two typical attention-based networks of channel attention and spatial attention. Our ACNet also adopts the attention mechanism, but our results are superior. Not only that, but ACNet also shows apparent advantages compared to new models such as FCANet and ResNeSt in the last two years.

Table II shows the compared results in the low-light CIFAR-10 dataset. Since low illumination and blur have a more significant impact on the retention of detailed information, we found that most methods cannot maintain good performance in the low-light cifar-10 dataset. Our method still has obvious advantages and strong competitiveness in such a complex situation.

D. Ablation Study.

This section conducts ablation experiments on the low-light SVHN and low-light CIFAR-10 datasets to analyze each component of the proposed ACNet. Our baseline strategy is to use native CapsNet to recognize low-light images. As observed from Table III, we train the baseline on the **low-light** datasets and attain the top-1 classification accuracy of 89.09% and 54.92%, respectively. For more convenience, we denoted the method trained on low-light datasets and validated on low-light datasets as ACNet-l. Inspired by [31], to verify the necessity of a recognition model for low-light, we train the native CapsNet on the **normal-light** dataset and evaluate it on the low-light dataset. Similarly, we define the method trained on normal-light datasets and validated on low-light datasets as ACNet-n. The table above shows that the ACNet-n achieves a classification accuracy of 83.38% and 47.23%, separately. It can be observed that a model trained with normal-light images cannot be applied for low-light image recognition well.

To verify the effectiveness of directional learning loss, we design two experiments. First, we combine our global-local attention module to the ACNet-l model and adopt the low-light image as the ground truth image to optimize the directional learning loss. Second, we turn low-light images to normal-light images as the objective. With low-light ground truth, the classification accuracies of 92.92% on SVHN and 62.22% on CIFAR-10 are obtained. Meanwhile, the normal-light ground truth attained 94.30% and 64.24%, separately. Referring to the results in the third and fourth rows in Table III, we can conclude that the directional learning loss function contributes helpfully to the recognition ability of the model.

TABLE III
ABLATION STUDY IN LOW-LIGHT CIFAR-10 AND LOW-LIGHT SVHN.

Algorithm	Accuracy	
	SVHN	CIFAR-10
ACNet-l(w/o GLA)	89.09	54.92
ACNet-n(w/o GLA)	83.38	47.23
ACNet(dl-loss w/ low-light)	92.92	62.22
ACNet(dl-loss w/ normal-light)	**94.30**	**64.24**

V. CONCLUSION

This paper proposes a novel attention-capsule network for low-light image recognition. To enable the model to obtain more recognizable and influential information from the low-light image, we first build a global-local attention module to enhance the features. Next, we propose a directional learning loss that enables the model to learn features that are difficult to capture from the low-light image to enhance recognition. Comprehensive experimental analysis shows that our proposed model effectively performs low-light image recognition.

Acknowledgments:

This work is supported by Nature Science Foundation of China (62172118,61876049) and and Nature Science key Foundation of Guangxi2021GXNSFDA196002in part by the Guangxi Key Laboratory of Image and Graphic Intelligent Processing under Grants (GIIP2006, GIIP2007, GIIP2008); and in part by the Innovation Project of Guangxi Graduate Education under Grants (YCB2021070 , YCBZ2018052, 2021YCXS071).

REFERENCES

[1] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8174–8182.

[2] W. Zhang, H. Li, and Z. Wang, "Research on different illumination image classification method," in *2017 2nd International Conference on Automation, Mechanical Control and Computational Engineering (AMCCE 2017)*. Atlantis Press, 2017, pp. 574–581.

[3] V. Sharma, A. Diba, D. Neven, M. S. Brown, L. Van Gool, and R. Stiefelhagen, "Classification-driven dynamic image enhancement," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4033–4041.

[4] D. Ren, H. Ma, L. Sun, and T. Yan, "A novel approach of low-light image used for face recognition," in *2015 4th International Conference on Computer Science and Network Technology (ICCSNT)*, vol. 1. IEEE, 2015, pp. 790–793.

[5] V. Mnih, N. Heess, A. Graves *et al.*, "Recurrent models of visual attention," in *Advances in neural information processing systems*, 2014, pp. 2204–2212.

[6] Y. Zhang, X. Di, B. Zhang, and C. Wang, "Self-supervised image enhancement network: Training with low light images only," *arXiv preprint arXiv:2002.11300*, 2020.

[7] Y. Wang, Y. Cao, Z.-J. Zha, J. Zhang, Z. Xiong, W. Zhang, and F. Wu, "Progressive retinex: Mutually reinforced illumination-noise perception network for low-light image enhancement," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2015–2023.

[8] Y. Zhang, X. Di, B. Zhang, Q. Li, S. Yan, and C. Wang, "Self-supervised low light image enhancement and denoising," *arXiv preprint arXiv:2103.00832*, 2021.

[9] W. Wang, C. Wei, W. Yang, and J. Liu, "Gladnet: Low-light enhancement network with global awareness," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 751–755.

[10] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," *arXiv preprint arXiv:1710.09829*, 2017.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.
[13] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.
[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.
[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[16] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.
[17] Y. Chen, J. Li, H. Xiao, X. Jin, S. Yan, and J. Feng, "Dual path networks," *arXiv preprint arXiv:1707.01629*, 2017.
[18] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.
[19] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.
[20] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 510–519.
[21] T. Durand, N. Mehrasa, and G. Mori, "Learning a deep convnet for multi-label classification with partial labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 647–657.
[22] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 783–792.
[23] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha *et al.*, "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020.
[24] G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with em routing," in *International conference on learning representations*, 2018.
[25] T. Hahn, M. Pyeon, and G. Kim, "Self-routing capsule networks," *Advances in neural information processing systems*, vol. 32, pp. 7658–7667, 2019.
[26] L. Zhang, M. Edraki, and G.-J. Qi, "Cappronet: Deep feature learning via orthogonal projections onto capsule subspaces," *arXiv preprint arXiv:1805.07621*, 2018.
[27] Y.-H. H. Tsai, N. Srivastava, H. Goh, and R. Salakhutdinov, "Capsules with inverted dot-product attention routing," *arXiv preprint arXiv:2002.04764*, 2020.
[28] J. Gu and V. Tresp, "Improving the robustness of capsule networks to image affine transformations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7285–7293.
[29] J. Gu, B. Wu, and V. Tresp, "Effective and efficient vote attack on capsule networks," *arXiv preprint arXiv:2102.10055*, 2021.
[30] P. Afshar, A. Mohammadi, and K. N. Plataniotis, "Brain tumor type classification via capsule networks," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 3129–3133.
[31] M. Singh, S. Nagpal, R. Singh, and M. Vatsa, "Dual directed capsule network for very low resolution image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 340–349.
[32] R. LaLonde and U. Bagci, "Capsules for object segmentation," *arXiv preprint arXiv:1804.04241*, 2018.
[33] K. Duarte, Y. S. Rawat, and M. Shah, "Videocapsulenet: A simplified network for action detection," *arXiv preprint arXiv:1805.08162*, 2018.
[34] W. Wang, W. Yang, and J. Liu, "Hla-face: Joint high-low adaptation for low light face detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 195–16 204.
[35] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
[36] N. Frosst, S. Sabour, and G. Hinton, "Darccc: Detecting adversaries by reconstruction from class conditional capsules," *arXiv preprint arXiv:1811.06969*, 2018.