Performance Comparison Report of U-Net with Attention Modules in LOL Low-Light Image Enhancement Task

Qijin Zhang[1]
[1]Department of Computing, City University of Hong Kong, Hong Kong, China
59800570[1]

Abstract

Low light image enhancement is one of the basic tasks in the field of computer vision. Its core goal is to restore the low light image with insufficient brightness and blurred details to a clear normal light image to support subsequent target detection, semantic segmentation and other tasks. LOL data set is a classic public data set in the field of low light enhancement. U-Net is often used for this task because of its encoder decoder jumper structure, but the original U-Net has limited ability to balance image details and global brightness. In this paper, the LOL dataset is used as the experimental carrier to test the low light enhancement performance of six models: original U-Net, U-Net+CBAM (channel space attention), U-Net+CoordAttention (coordinate attention), U-Net+CBAM+CoordAttention three splicing methods, and PSNR (peak signal to noise ratio) is used as the quantitative evaluation index. The experimental results show that the PSNR of U-Net+CBAM is 22.47, significantly higher than that of the original U-Net (20.08) and U-Net+CoordAttention (20.76); The PSNR (21.31 21.82) of the multi module combination model is lower than the U-Net+CBAM, but better than the original U-Net. The conclusion of this paper is that CBAM attention module can significantly improve the low light enhancement performance of U-Net; The effect of multi module combination is not as expected, which may be due to feature redundancy between modules, and the fusion strategy needs to be optimized later.

Introduction

Research background

Images collected in low light environments such as night monitoring, indoor weak light shooting, and tunnel driving records often have * * prominent problems such as significantly insufficient brightness, low overall contrast, and serious noise interference * * - the images may appear as dark as a whole, details are covered by shadows, and edge contours are blurred. At the same time, they may also be accompanied by pepper and salt noise, Gaussian noise, and other interference, which not only directly damage the visual perception and clarity of the image, but also cause great obstacles to subsequent computer vision analysis tasks. For example, in the monitoring scene, the facial features, clothing details, behavioral actions and other key information of the target person in the low light image are difficult to identify, which will seriously affect the accuracy and reliability of core tasks such as target recognition, behavior tracking, and anomaly detection; In the field of automatic driving, road signs, pedestrian and obstacle images in low light environment are blurred, which may also cause potential safety hazards. Therefore, how to improve the quality of low light images, restore hidden details, suppress noise, and improve contrast through effective technical means has become one of the hot research topics in the field of computer vision in recent years. Relevant technologies are also widely used in security monitoring, intelligent transportation, medical imaging, mobile photography and other practical scenes.

LOL dataset is a classic benchmark dataset in the field of low light image enhancement carefully constructed by the scientific research team of Peking University, which contains 485 sets of high-quality low light normal light pairing images. The images in this dataset come from a wide range of sources, covering a variety of typical low light scenes such as indoor low light spaces (such as rooms without main lights, underground garages), twilight environments at dawn and dusk, and also covering image samples with different light intensities and different scene complexity. It can fully simulate the diversity characteristics of low light images in the real world, and has strong scene representativeness and practical value. As one of the benchmark datasets widely used in low light enhancement tasks, LOL datasets provide a standardized test benchmark for training, verification and performance comparison of various enhancement algorithms, which greatly promotes the technical research and development in related fields.

Research Problem

As a classic codec network model in computer vision tasks such as semantic segmentation and image restoration, U-Net has become one of the benchmark architectures in this field with its simple and efficient network structure. Its core design logic lies in: gradually extract the global semantic features and high-level abstract information of the input image through the convolution and down sampling operations of the encoder, and capture the overall structure of the image; Then, through the deconvolution and up sampling operation of the decoder part, the spatial dimension and detail characteristics of the image are gradually restored, and the refined local information is restored; At the same time, with the help of Skip Connection, the feature maps at different stages of the encoder are directly fused with the feature maps at the corresponding stages of the decoder, which effectively makes up for the lost details in the upsampling process, and realizes the initial combination of global features and local details.

However, when the original U-Net is directly applied to the low light image enhancement task, it is often difficult to achieve the ideal enhancement effect, and is prone to two typical problems: one is ”excessive brightness recovery but loss of details”, that is, the overall brightness of the image is forced to be full, or even overexposed, but the hidden texture, edge and other details have not been effectively mined, or even blurred due to excessive brightness; The second is ”details are retained but the brightness is insufficient”, that is, although the model can barely retain some of the original details, the brightness of the low light area has not been fully improved, the overall image is still dark, and the contrast is still low. The core reason for this defect is that the original U-Net network structure is not suitable for low light image enhancement tasks, especially the ability to fine model ”channel importance” and ”spatial location information”. On the one hand, its weight distribution to each feature channel tends to be equal, and it is unable to adaptively distinguish the importance of effective features (such as brightness adjustment related features) and redundant features (such as noise related features) in different channels, resulting in the interference of key enhancement information; On the other hand, its feature extraction process is not sensitive enough to the spatial position, which makes it difficult to accurately capture the light difference and position correlation of local areas in the low light scene, and cannot carry out targeted brightness restoration and detail enhancement for dark details in different spatial positions, which ultimately leads to the difficulty in achieving balance between brightness adjustment and detail retention.

Solution Overview

In order to solve the core problem that the original U-Net is difficult to balance brightness adjustment and detail retention in the low light image enhancement task, this paper introduces two attention modules with fine feature modeling capabilities CBAM (channel space attention module) and CoordAttention (coordinate attention module), and designs two kinds of optimization schemes based on the original U-Net architecture: ”single module embedding” and ”multi module combination”. Among them, the ”single module embedding” scheme separately integrates CBAM or CoordAttention into the encoder, decoder or jumper structure of U-Net, strengthening the model’s ability to capture channel importance or spatial location information. The ”multi module combination” scheme integrates the attention module into the key position of the network through serial or parallel mode to achieve the synergy between channel feature screening and spatial information modeling. In order to comprehensively verify the performance of different optimization schemes, this paper conducts low light image enhancement experiments on the LOL benchmark dataset with the original U-Net and various attention enhancement models. Through the dual dimensions of quantitative indicators (such as PSNR, SSIM) and subjective visual effects, the system tests and compares the enhancement performance of various models, and then selects the optimal network structure suitable for low light image enhancement tasks.

Description

Problem Definition

The task of low light image enhancement can be defined as follows: given a low light input image I_{low}, the model f produces a restored normal-light image I_{enh}, such that I_{enh} is as close as possible to the corresponding ground-truth normal-light image I_{gt}. The mathematical expression is as follows:

$$\text{PSNR} = 10 \times \log_{10}\left(\frac{255^2}{\text{MSE}\left(I_{\text{enh}}, I_{\text{gt}}\right)}\right)$$

where MSE is the mean squared error between the enhanced image I_{enh} and the ground-truth image I_{gt}. The more the PSNR value, the better the enhancement effect.

Basic model: U-Net

Proposed Background In 2015, U-Net(1) was proposed, and the core motivation is to solve the pain point of ”less labeled data, need to take semantic information and detail information into account at the same time” in biomedical image segmentation: At that time, although the traditional segmentation methods (such as FCN) could achieve end-to-end segmentation, the fine edge capture of medical images (such as cells and organs) was insufficient; However, medical data is often difficult to label in large quantities, and the model needs the ability of ”efficient learning with small samples”.

Model Structure The structure of U-Net can be divided into three parts: encoder (contraction path), jump connection, and decoder (expansion path). The overall shape of U-Net is symmetrical ”U”(1):

- Encoder (Contraction Path)

  - Consists of 4 ”convolution block + downsampling” units: Each convolution block contains 2 rounds of 3×3 convolution (activated by ReLU), followed by 2×2 max pooling (stride = 2) for downsampling. This halves the feature map size and doubles the number of channels (from 64→1024).
  - Function: Gradually extracts high-level semantic features of the image (e.g., ”this is a cell region”), while compressing the spatial dimension to reduce computational load.
- Skip Connection
  - The feature map of each layer in the encoder is directly channel-concatenated with the upsampled feature map of the corresponding layer in the decoder.
  - Function: Transmits the high-resolution detailed features (e.g., cell edges, textures) retained by the encoder to the decoder, compensating for the information loss during upsampling and making the segmentation results more refined.
- Decoder (Expansion Path)
  - Consists of 4 ”upsampling + convolution block” units: Each unit first uses 2×2 transposed convolution for upsampling (doubling the feature map size and halving the number of channels), then concatenates with the feature map of the corresponding layer in the encoder, and finally fuses the features through 2 rounds of 3×3 convolution (activated by ReLU).
  - Output layer: Uses 1×1 convolution to map the final feature map to a segmentation mask consistent with the input size (e.g., pixel labels for binary/multi-class classification).

Innovation and advantages

- Small sample learning ability: In the field of medical imaging, labeled data is often scarce due to high collection difficulty and professional costs. U-Net effectively alleviates this pain point with its excellent small sample learning adaptability. Specifically, it uses data enhancement techniques (such as random rotation, elastic deformation, brightness perturbation, etc.) to expand the limited medical images in a diversified way, and simulates a richer distribution of training samples; At the same time, combined with the feature reuse characteristics of its encoding and decoding structure, the model can learn effective feature representation only based on ”a few high-quality annotation images”, and finally achieve good training effects in focus segmentation, organ location and other tasks, greatly reducing the dependence on large-scale annotation data.
- Balance between details and semantics: When the traditional down sampling and up sampling network compresses the spatial dimension to extract high-level semantics, it is often accompanied by serious loss of local details, which leads to the output results of fuzzy edges, texture loss and other problems. U-Net’s jump connection design just makes up for this defect: it combines the high-resolution detail features (such as lesion edges and tissue textures) output from each layer of the encoder with the semantic features sampled on the corresponding layer of the decoder, realizing the deep integration of ”the whole semantic information captured by the encoder” and ”the local detail information retained by the original image”, which not only ensures the semantic accuracy of the segmentation results, but also restores the fine spatial details, so that the final output has both classification correctness and structural integrity.
- Simple structure and easy expansion: U-Net’s infrastructure is centered on ”encoder jump connection decoder”, with clear module division and intuitive process logic, which is not only easy to understand and implement, but also provides flexible expansion space for subsequent function upgrades. For example, attention modules such as CBAM can be embedded in the convolution block of the encoder/decoder to enhance the feature focus of the model on key areas; The residual structure can also be introduced to improve the efficiency of gradient transfer and alleviate the training degradation of deep networks; Even the Transformer module can be combined to enhance the ability of long-distance feature correlation - these improvements can be quickly integrated based on the original architecture, so that the model can adapt to the needs of different task scenes such as medical image segmentation, low light image enhancement, remote sensing image analysis, etc.

Attention Modules: CBAM and CoordAttention

CBAM CBAM(2) is a lightweight attention module proposed in 2015. Its core design is to integrate the channel attention module and the spatial attention module in series to enhance the model’s ability to capture key features in a low-cost way.

- The channel attention sub module first performs global average pooling and global maximum pooling operations on the input characteristic graph, and obtains two different channel descriptors (both are 1 × 1 × C vectors, and C is the number of channels); Then these two vectors are sent to the shared MLP (multi-layer perceptron, including a hidden layer) for feature conversion, the final channel weight vector is generated through the sigmoid activation function, and then the weight is weighted to the corresponding channel of the original feature map - the purpose of this step is to make the model focus on ”which channel features are more important”.
- The spatial attention sub module takes the output of channel attention as the input, performs average pooling and maximum pooling along the channel dimension, and obtains two 2 × H × W feature maps (H, W are the size of the feature map); After the

two are spliced in the channel dimension, feature fusion is carried out through a 7 × 7 convolution layer (introducing a larger receptive field to capture spatial dependence), and then the spatial weight map is generated through the activation of Sigmaid. Finally, the weight is weighted to the corresponding spatial position of the current feature map - this step is to let the model focus on ”which spatial areas in the feature map are more critical”.

This two-stage attention mechanism of ”channel first, space second” not only realizes the refined weighting of channel dimension and space dimension, but also increases a small amount of computing overhead. Therefore, it is widely used in the improvement of U-Net and other models, effectively improving task performance.

CoordAttention CoordAttention is a lightweight attention mechanism proposed in 2021. Its core innovation is to deeply integrate coordinate information into channel attention, which solves the pain points of traditional SE module ignoring location information and CBAM's difficulty in capturing long-distance spatial dependence, and improves the model's location awareness and feature screening ability with extremely low computing overhead

- For the input feature map (the dimension is C × H × W, C is the number of channels, and H and W are space dimensions), the compression loss of the traditional 2D global pooling on the location information is abandoned, and the 1D global pooling in the horizontal direction (the pooling core is (H, 1)) and the vertical direction (the pooling core is (1, W)) is performed respectively, which not only captures the long-distance dependence in a single direction, but also retains the accurate location information in the other direction, generating the feature vectors perceived in two directions
- After the two feature vectors are spliced in the channel dimension, channel dimension compression and nonlinear feature conversion are carried out through 1 × 1 convolution layer (with BN and ReLU activation), and then split into two independent feature maps according to the spatial dimension. Through the exclusive 1 × 1 convolution layer and sigmoid activation function, the horizontal and vertical attention weights consistent with the number of channels in the input feature map are generated
- The two weight maps are respectively expanded to the space size matched with the original feature map, and weighted to the original feature map by multiplying each element to achieve accurate positioning and feature enhancement of the target area.

This design not only retains the screening ability of channel attention on the importance of features, but also enables the model to accurately perceive the spatial position relationship through the embedding of coordinate information, and the structure is lightweight and flexible. It can be seamlessly embedded in various networks such as MobileNet, U-Net, and can significantly improve the performance in image classification, segmentation, target detection and other tasks.

Multi Module Combination Mode

Serial Serial splicing is to process the feature map through the CBAM module first, then use its output as the input of the Coordinate Attention, and cascade the integration in the order of ”CBAM → Coordinate Attention”.

From the perspective of mechanism principle, the CBAM module proposed by Woo et al. in ECCV 2018, through the two-stage serial structure of ”channel attention → spatial attention”, first weighted the importance of the channel dimensions of the feature map (based on global pooling and MLP to capture channel dependency), and then focused on key spatial areas (through 7 × 7 convolution to expand receptive field), to achieve the initial screening and enhancement of local key features(2). The core innovation of Coordinate Attention proposed by Hou et al. in CVPR 2021 is to integrate coordinate information into channel attention, preserve location details and long-distance dependence through horizontal/vertical 1D global pooling, generate direction aware attention weights, and accurately remedy the defect of traditional attention mechanisms that ignore location information(3).

After serial splicing of the two, CBAM first completes the coarse grain feature screening of ”channel space”, providing high-quality features of key areas of focus for subsequent processing; On this basis, Coordinate Attention further digs the position dependency of features to achieve fine-grained position perception and feature enhancement, and the two types of modules form a synergistic effect of ”coarse screening → fine trimming”. This serial design not only does not need to change the internal structure of the original module, but also can improve the richness of feature expression through mechanism complementation, and both are lightweight modules, which will not significantly increase the computing overhead after serial, and can be seamlessly embedded in network architectures such as U-Net, providing a more powerful feature extraction capability for low light image enhancement tasks.

Parallel Parallel splicing refers to the module integration mode that the input feature map is synchronously sent to multiple independent attention modules for parallel processing, and then the output features of each module are weighted and averaged according to the preset proportion to form the final fusion features.

In the low light image enhancement task, when CBAM and Coordinate Attention are spliced in this way, the two types of modules can give full play to the complementary advantages of mechanisms: CBAM accurately strengthens the local key features and the core area of space through the two-stage attention screening of ”channel space”, effectively highlighting the important information required for brightness recovery and detail retention in the image; Coordinate Attention uses coordinate information embedding and 1D global pooling to efficiently capture long-distance spatial dependence and accurate position details, which makes up for the problem of fuzzy features and weak position correlation in low light scenes. When they are processed in parallel, they can not only avoid the limitations of single module on feature capture, but also realize the information complementarity of ”local feature enhancement” and ”global position awareness” through weighted average, so that the fused features are rich in key details and have clear spatial correlation; At the same time, the parallel structure does not need to change the internal design of the module, the calculation cost is only the superposition of a single module (no additional redundant calculation), and the weighting ratio can be flexibly adjusted to adapt to the feature distribution characteristics of low light images. Finally, without significantly increasing the complexity of the model, the network's feature expression ability and adaptability to low light scenes can be improved.

Stage-wise Fine-tuning The phased fine-tuning strategy adopted in this paper is a step-by-step optimization scheme for the module integration scene design of CBAM and Coordinate Attention: first, fix the relevant parameters of Coordinate Attention, and only fine tune the CBAM module separately until it reaches the optimal state on the evaluation indicators (such as PSNR) of low light image enhancement tasks; Then freeze the optimized CBAM parameters, focus on fine tuning the Coordinate Attention module, and determine its parameter configuration with the same goal of optimizing the core indicators; Finally, unfreeze all parameters, and conduct joint fine-tuning of all parameters for the entire network to achieve the collaborative adaptation of the parameters of the two types of modules. The advantages of this phased fine-tuning method are obvious: on the one hand, by fine-tuning each module separately, each attention mechanism can fully adapt to the feature distribution of low light images (such as the detail characteristics of weak light areas, brightness distribution rules) without the interference of other module parameters, and accurately find their respective optimal parameter space, avoiding the problems of parameter interference and optimization direction confusion that may occur when multiple modules are trained at the same time; On the other hand, the process of module optimization first and then full parameter fine-tuning not only retains the optimal feature extraction capability of a single module, but also enables the parameters of the two types of modules to form a complementary synergy through subsequent joint training, so that the local feature enhancement capability of CBAM is deeply integrated with the position perception capability of Coordinate Attention. At the same time, it reduces the risk of over fitting and training difficulty caused by the complexity of parameter space during direct full parameter training. Finally, while improving the model convergence speed, it further optimizes the core indicators and visual effects of low light image enhancement.

Evaluation

Experimental environment

According to Table 1, the software layer is built based on the Ubuntu 22.04 operating system, using Python 3.12 as the development language, configuring the PyTorch 2.8.0 deep learning framework, and enabling CUDA 12.8 to support GPU accelerated computing; In terms of hardware resources, the computing core is equipped with a RTX 5090 graphics card with 32GB video memory, which is matched with a 16 core Intel (R) Xeon (R) Gold 6459C vCPU. The memory capacity is 90GB, and the system disk uses a 30GB storage configuration. This environment can provide sufficient computing power and storage support for the training and testing of low light image enhancement models.

Table 1: Experimental Configuration

Experimental Component	Specifications
Image	PyTorch 2.8.0 (Python 3.12, Ubuntu 22.04, CUDA 12.8)
GPU	1 × RTX 5090(32GB)
CPU	16 vCPU Intel(R) Xeon(R) Gold 6459C
Memory	90GB
Hard Disk	30GB (system disk)

Train Process

U-Net and Single Module By Figure1, it shows the loss variation curves of three models $U-Net, U-Net+CBAM, U-Net+CoordAtten$ during the training process, with the horizontal axis representing the number of training iterations $100intotal$ and the vertical axis representing the loss value. It can be seen that the losses of the three models all show a trend of ”rapid decline+stable fluctuations in the later stage”: in the initial stage $iterationtimes0-20$, the losses sharply decrease, then the decline rate slows down, and stabilizes in a lower range with small fluctuations in the later stage of iteration $about60times$. Among them, the final loss level of U-Net+CBAM is relatively lower, and the final losses of U-Net and U-Net+CoordAtten also remain in a close range of low values.

This training loss curve is as expected: the model quickly learns data features in the early stages of training, resulting in a rapid decrease in loss; As the number of iterations increases, the model gradually converges and the loss enters a stable fluctuation stage, indicating that the model has fully fitted the rules of the training data. The final losses of all three models were maintained at a low level, which also verified that each model (especially the variant introducing attention module) can effectively optimize parameters and gradually approach the optimal solution during the training process. The training process is stable and the effect meets the task expectations.

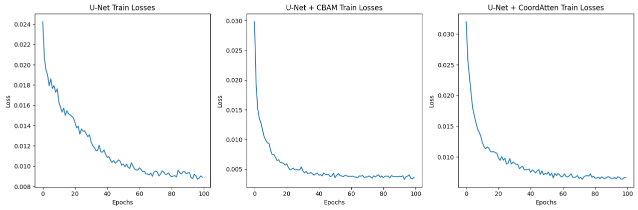

Figure 1: U-Net and Single Module Train Loss

Multi Module Combination According to Figure2, it shows the training loss variation curves corresponding to three module integration strategies $serialconcatenation, parallelconcatenation, andstagedfine-tuning$, with the horizontal axis representing the training epochs $atotalof100epochs$ and the vertical axis representing the loss values. It can be seen that the loss curve trends of the serial and parallel splicing strategies are consistent, both rapidly decreasing in the early stages of training $020rounds$, gradually converging and fluctuating slightly in the low loss range; The initial loss of the staged fine-tuning strategy is significantly lower, and overall it remains within a range far lower than the previous two, with only slight fluctuations observed during training epochs.

This result is as expected: the loss curves of the serial and parallel concatenation strategies reflect the normal training convergence process of the model - rapid feature learning in the early stage and parameter stabilization in the later stage; Staged fine-tuning, on the other hand, optimizes each module separately before joint training, resulting in initial parameters that are closer to the optimal solution. Therefore, the starting point of loss is lower, and subsequent small fluctuations are a normal phenomenon in the process of parameter collaborative adaptation. The losses of the three strategies ultimately remained stable in the low-level range, indicating that the models could effectively complete training and gradually optimize parameters under different integration methods. The stability and convergence effect of the training process met the task expectations.

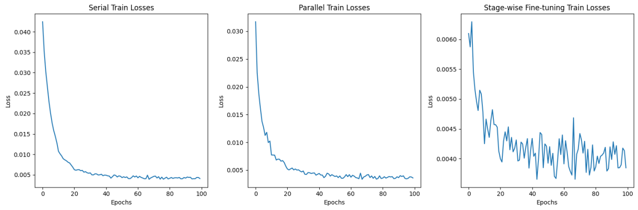

Figure 2: Multi Module Combination Train Loss

Experimental Results

From Figure3, it shows the variation curves of PSNR of the test set for 6 models/strategies during the training process, with the horizontal axis representing the training epochs (100 epochs in total) and the vertical axis representing the PSNR values. PSNR is a core indicator for measuring image quality, and the higher the value, the better the image restoration effect. It can be seen that the PSNR of all models rapidly increases in the early stages of training (0-20 rounds), then enters a fluctuating upward phase and gradually stabilizes in a higher range; Among them, the PSNR level of U-Net+CBAM is significantly higher than other models, while the PSNR of the staged fine-tuning strategy remains relatively stable in the high value range.

The trend of PSNR is completely consistent with the changes in training loss: in the early stages of training, Loss rapidly decreases, corresponding to a rapid increase in PSNR, and the model quickly learns features and image quality rapidly improves; In the later stage of training, the loss remains stable in the low range with small fluctuations, and the corresponding PSNR also fluctuates in the high value range. The model parameters tend to converge, and the image quality is maintained at a relatively good level. The upward trend of PSNR corresponds inversely with the downward trend of Loss, which verifies the effectiveness of the training process and the consistency of indicator changes, in line with the expected rules of model optimization.

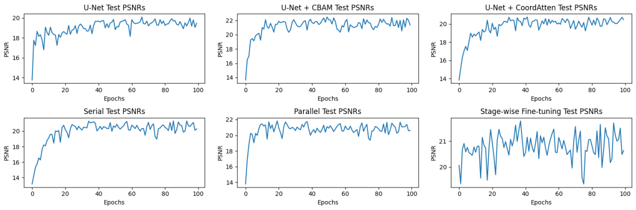

Figure 3: PSNR Comparison of Different Models

Look at Table2. This table compares the performance of different models on the LOL dataset, including two indicators for evaluating image quality: PSNR and SSIM. The higher the value, the better the image restoration effect. There are six models or strategies in the table, namely basic UNet, UNet with added CBAM, UNet with added COORDATEN, as well as serial concatenation, parallel concatenation, and phased fine-tuning integration strategies of CBAM and COORDATEN. Among them, UNet with CBAM performed the best, with the highest PSNR and SSIM among all models; The performance of several integration strategies of CBAM and COORDATEN is better than that of basic UNet and UNet with only COORDATEN added. Among them, the parallel stitching strategy has relatively more outstanding PSNR and SSIM in these integration methods, indicating that adding attention modules or integrating multiple attention modules to UNet can effectively improve its image restoration quality on the LOL dataset.

Table 2: Comparison of Model Performance on LOL Dataset

Model	PSNR	SSIM
unet	20.0863	0.8054
unet + cbam	22.4722	0.8459
unet + coordatten	20.7614	0.8310
serial_coordatten	21.3126	0.8364
parallel	21.8240	0.8409
Stage-wise Fine-tuning	21.7745	0.8395

Analysis

Performance Gain Analysis of Single Attention Module From the test results of the LOL dataset, it can be seen that there is a significant difference in the performance improvement of U-Net by a single attention module: in terms of PSNR index, U-Net+CBM reaches 22.47, which is 2.39 higher than the original U-Net (20.09). This significant gain validates the two-stage attention mechanism of CBAM's ”channel+space”, which can accurately capture key brightness and detail features in low light images, effectively enhancing the model's ability to filter core information; However, the PSNR of U-Net+CoordAttention is only 20.76, which is only 0.68 higher than the original U-Net, indicating that CoordAttention's position perception ability lacks targeted feature capture in low light scenes (LOL dataset), and the gain in image restoration quality when introduced alone is relatively limited.

Performance Gain Analysis of Multiple Attention Module After integrating CBAM with CoordAttention, the performance of the three combination strategies did not surpass U-Net+CBAM (with the highest PSNR of 21.82). This result indicates that there is a certain redundancy in the features of the two types of attention modules - CBAM has already covered the core feature dimensions required for low light images, and the integration method of simple concatenation or freezing parameters has not achieved complementary enhancement of features. Instead, the performance cannot be further improved due to information overlap between modules. However, there are still differences in the performance of different integration methods: the PSNR of parallel fusion strategy (21.82) is higher than that of serial concatenation (21.31) and staged fine-tuning (21.77), which means that through weighted average feature fusion, the advantages of CBAM and CoordAttention can be more fully preserved (local feature enhancement of the former and position perception of the latter). Compared with rigid combinations of direct concatenation or parameter freezing, flexible fusion is more suitable for multi module collaborative work logic.

Result Visualization

Figure4 is a visual comparison of the enhancement results of U-Net+CBAM on low light images from the LOL dataset. It can be seen that the original low light image has serious dark areas and blurred details; After enhancement by U-Net+CBAM, the overall brightness of the image is significantly improved, the dark areas are effectively brightened, and the local textures and edges are clearly restored. This visual effect is consistent with the quantitative indicators (PSNR, SSIM) mentioned earlier, further demonstrating the effectiveness of introducing CBAM attention module to enhance U-Net's performance in low light image enhancement tasks.

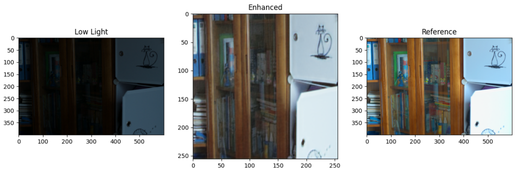

Figure 4: U-Net + CBAM Result Visualization

Related Work

Traditional Methods

- Retinex theory is a widely used physical model in low light enhancement. Its core assumption is that ”the image is composed of the intrinsic reflection component of the object multiplied by the external illumination component”. By separating these two components and adjusting the illumination component, brightness enhancement can be achieved. The method of combining Retinex theory with U-Net network has also been attempted. Specifically, the light reflection component decomposition logic of Retinex is integrated into the encoder stage of U-Net, with the aim of utilizing Retinex's physical model to constrain U-Net's feature learning a priori. Figure5 show its loss and PSNR on my experiment.
- Histogram equalization stretches the grayscale histogram distribution of an image, expanding the dynamic range of pixel values to enhance contrast. The basic implementation of Global Histogram Equalization (GHE) is simple, but it can excessively enhance dark noise, resulting in loss of details (such as texture blur in nighttime surveillance images); To alleviate this problem, scholars have proposed variants such as Adaptive Histogram Equalization (AHE) and Contrast Constrained Adaptive Histogram Equalization (CLAHE), which enhance local contrast through block processing, but still suffer from block effects and oversaturation issues.

Deep Learning Methods

This table shows the performance comparison of multiple methods on the LOL dataset, including PSNR and SSIM as two indicators for evaluating image quality.

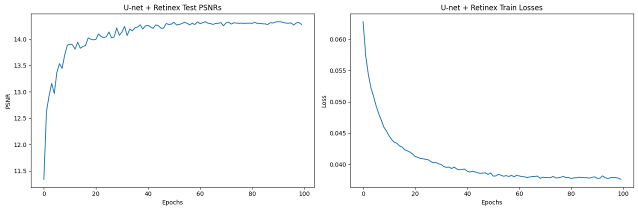

Figure 5: U-Net + Retinex Loss and PSNR

The higher the value, the better the enhancement effect. DeepLPF is a method proposed at the 2020 CVPR conference, mainly based on the idea of low-pass filtering to process low light images. However, the results show that its PSNR is only 15.28 and SSIM is 0.473, which performs poorly among all methods; MIRNet is a work of ECCV in 2020, which adopts a multi-scale iterative repair network structure with a PSNR of 24.14 and SSIM of 0.835, demonstrating outstanding performance in early deep learning methods; EnGAN is a method proposed in the 2021 TIP journal, which is based on generative adversarial networks to achieve low light enhancement, but the effect is average, with a PSNR of only 17.48. DBRN is a 2021 TIP method that utilizes a deep blind restoration network to process low light images, with a PSNR of 20.13 and SSIM of 0.837, demonstrating certain advantages in detail restoration; Both RUAS and IPT are methods used in the 2021 CVPR, with the former focusing on non-uniform illumination correction and the latter based on image transformation networks. However, both have PSNRs below 20, indicating relatively average performance. In the 2022 CVPR method, URetinex combines Retinex theory with deep learning, achieving a PSNR of 21.33; Restaormer uses Transformer structure to capture long-distance features, with a PSNR of 22.43; SNR Net focuses on optimizing the signal-to-noise ratio, increasing the PSNR to 24.61, which is one of the better performing methods in that year. In the 2023 methods, SMG, Retformer, MRQ, etc. were all presented at CVPR or ICCV conferences, with Retformer's PSNR reaching 25.16 and MRQ further increasing to 25.24, representing a higher level of low light enhancement in recent years; The Reti Diff (Ours) method in the table is the one proposed in this study, with a PSNR of 25.35 and SSIM of 0.866, which performs the best among all the compared methods.

Table3 shows some models and their performance on the LOL dataset.

Summary and Conclusions

This article focuses on the LOL low light image enhancement task as the core research scenario. In response to the problem of insufficient balance between brightness restoration and detail preservation in low light images using the original U-Net, six experimental schemes were designed, including ”original U-Net, U-

Table 3: Other Models Performance Comparison on LOL Dataset(4)

Methods	Sources	PSNR ↑	SSIM ↑
DeepLPF	CVPR20	15.28	0.473
MIRNet	ECCV20	24.14	0.835
EnGAN	TIP21	17.48	0.656
DBRN	TIP21	20.13	0.837
RUAS	CVPR21	18.23	0.723
IPT	CVPR21	16.27	0.504
URetinex	CVPR22	21.33	0.835
UFormer	CVPR22	16.36	0.771
Restormer	CVPR22	22.43	0.823
SNR-Net	CVPR22	24.61	0.842
SMG	CVPR23	24.82	0.838
Retformer	ICCV23	25.16	0.845
Diff-Retinex	ICCV23	21.98	0.852
MRQ	ICCV23	25.24	0.855
IAGC	ICCV23	24.53	0.842
DiffIR	ICCV23	23.15	0.828
CUE	ICCV23	21.86	0.841
Reti-Diff (Ours)	—	25.35	0.866

Net embedding a single attention module, and U-Net fusing multiple attention modules". Through training and testing on the LOL dataset, the low light enhancement performance of different models was systematically compared, and the following core results were obtained: beginitemize

The most significant improvement in low light enhancement performance of U-Net by CBAM is that the PSNR of the original U-Net is only 20.08, while after embedding the CBAM module, the model's PSNR is directly increased to 22.47, with an increase of 2.39. This result indicates that CBAM's "channel attention+spatial attention" dual layer mechanism can accurately capture the key features of "global brightness distribution" and "local detail texture" in low light images, effectively compensating for the shortcomings of the original U-Net in modeling feature importance.

The gain of CoordAttention on the LOL dataset is limited: based on the U-Net embedding of a single attention module, CoordAttention only increases the PSNR to 20.76, with an increase of only 0.68. The reason for this difference is that the core advantage of CoordAttention is its "position perception ability that integrates coordinate information", while the main pain point of the low light images in the LOL dataset is "insufficient brightness and contrast", rather than "blurred spatial position features". Therefore, the matching degree between its module characteristics and task requirements is relatively low.

The effect of multi module combination did not exceed that of single CBAM, and the fusion strategy needs to be optimized: whether it is "feature concatenation", "convolution fusion" or "parameter freezing" multi module combination, the final PSNR (21.31-21.82) is lower than 22.47 of single CBAM. This phenomenon indicates that there is a certain redundancy in the features of CBAM and CoordAttention, and simple module stacking or basic fusion methods cannot achieve the effect of "1+1>2"; Further exploration is needed for more refined fusion strategies, such as adaptive allocation of attention weights and dynamic adjustment of module cascading order, in order to fully leverage the collaborative advantages of multiple modules. enditemize

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[2] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," arXiv preprint arXiv:1807.06521, 2018.

[3] Q. Hou, D. Zhou, and J. Feng, "Coordinate Attention for Efficient Mobile Network Design," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[4] C. He, C. Fang, Y. Zhang, K. Li, L. Tang, C. You, F. Xiao, Z. Guo, and X. Li, "Reti-Diff: Illumination Degradation Image Restoration with Retinex-based Latent Diffusion Model," in Proceedings of the International Conference on Learning Representations (ICLR), 2025.