International Journal of Educational Sciences and Arts (IJESA)
المجلة الدولية للعلوم التربوية والآداب
Vol. (5), No. (10) October 2026 الإصدار (5)، العدد (10)

HS-Former V3: An Uncertainty-Aware Hybrid Transformer–CNN Framework with Adaptive Frequency-Domain Learning and Edge-Aware Fusion for Low-Light Image Enhancement

Zainab Mohammed Essa
Mustansiriyah University, Iraq
zainab_dvam@uomustansiriyah.edu.iq

Abstract

improving low-light images is a very challenging vision task, since degradation in the low-light image is usually caused under complex low-light conditions that destroys the visibility of the image, sources of noise are often present, and the high-frequency structure information is lost. To address these problems, in this paper, we propose HS-Former V3: a hybrid Transformer-CNN setup, which aims to learn the modeling of global illumination together with the restoration of local structures. The setup contains a Swin Transformer V2 branch, a lightweight convolution branch, an Edge-Aware Cross Fusion (EACF) module, and a Dynamic Frequency-Selective Loss (DFL). Besides, we propose a masked-frequency pre-training strategy to further explore the effectiveness of unpaired low-light images and design an uncertainty estimation head to locate uncertain enhancement regions. The overall system is inspired by two deficiencies, i.e. deficiency of the traditional reconstruction objectives and the deficiency of features and efficiency of models. The evaluation protocol includes full-reference image-quality metrics, no-reference quality evaluation, computational complexity, corruption robustness, and downstream object detection. This structure will pave the way towards reproducible research on adaptive spectral supervision and edge-guided fusion, to be validated in a controlled successful implementation.

Keywords: Low-Light Images Enhancement, Swin Transformer V2, CNN, Learning in Frequency Domain, Edge Aware Fusion, Uncertainty.

1. Introduction

Low-light image enhancement (LLIE) is a very meaningful problem in computer vision. It has various applications like nighttime security, intelligent transportation, autonomous driving, and security monitoring. In most conditions, the images obtained in poor lighting conditions exhibit several degradations such as low contrast, incorrect colors, sensor noise and missing details that pose visual quality issues and affect high-level computer vision tasks.

The biggest problem is gathering sufficient light at the sensor, at low light levels the ratio of the signal of the object being observed to the signals of the noise is too low and without further enhancement to the image the entire scene structure cannot be differentiated from the noise at the sensor.

This is the issue that early low-light enhancement deep-learning approaches, for instance, Retinex

Net learned to estimate the decomposition of the observed image into illumination and reflectance [1], and Kind learned to estimate the illumination adjustment and the reflectance restoration [2]. Zero-DCE used a zero-reference based on image-specific enhancement curve [3]. In the meantime, Enlighten GAN studied the unpaired adversarial learning [4].

In addition, some methods have used the Transformer structure to make full use of their ability to model long-range dependency and context information. Restormer designed a new Transformer-based module to perform high-quality high-resolution image restoration [5]. Swin Transformer V2 used shifted-window attention and scaled cosine attention to construct a hierarchical visual feature representation [6]. Retinex former adopted an illumination-guided processing with the Transformer-based feature modeling [7].

Yet, there are still some drawbacks. The existing methods in global illumination correction cannot preserve local textures, and the process of recovering subtle details is also flawed.

Also, increase the noise. Another limitation is that traditional reconstruction loss functions may not provide the correct assessment of image features at different spatial frequencies.

In this paper, we present the third version of HS-Former, HS-Former V3, which addresses the limitation of existing methods through a comprehensive approach to include global context modeling, local feature restoration, edge-guided feature fusion, adaptive supervision in the frequency domain, masked frequency pretraining, and uncertainty estimation.

The main research objectives are:

1. To study on the complementarities of Transformer and CNN representations in boosting low-light images.
2. To build an adaptive spectral domain loss for spectral reconstruction.
3. Consolidate the edge information together into a feature fusion, to keep the structure.
4. Experimenting with masked-frequency pretraining for better data efficiency.
5. To provide an estimate of the spatially varying uncertainty, and of its correlation with the enhancement errors.
6. Develop and validate reproducible test protocol for evaluation including image-quality based metrics, robustness testing, computational measurements and downstream object detection.

The significant contribution is a uniform methodological basis to cover each of these components. They and their synergistic effects will be proved in systematic tests.

2. Related Work

2.1. Retinex-Based Low-Light Enhancement:

The Retinex-based methods treat the observed image as the product of a corresponding light source and a reflectance:

I=RL, (I is the original image, R the reflectance, L the illumination, and the asterisk represents the element-by-element product)

RetinexNet employs a learned decomposition for reflectance and illumination [1]. Kind also makes an additional step in which the illumination adjustment is separated from the reflectance restoration [2]. Both these methods have a scheme for an explainable brightness enhancement, but still poor decomposition results in color shifts, noise amplification, and structural artifacts.

HS-Former V3 is a feature-learning based method that doesn't explicitly perform the Retinex decomposition.

2.2. Curve-Based and Unpaired Methods:

Zero-DCE learns the nonlinear curve of enhancement without normal-light references in training [3]. Enlighten GAN learns the enhancement from unpaired data with an adversarial learning method [4].

These kind of techniques will have applications where there is very little training data available for pairing. But, scene distribution, illumination variation and camera configuration will also be considerations. There will be the added difficulty of testing on new camera setups and dark scenes.

2.3. Transformer-Based Image Restoration:

Restormer relies on high powerful attention modules to restore image quality [5]. Swin Transformer V2 is built on hierarchical attention with scaled cosine similarity [6]. Retinex former combines illumination guided information and transformer representation for low light enhancement [7].

These are just some examples of how modeling contextual features can be applied. In our case, we'd like to explore if a global branch of the Transformer can be integrated with a light-weight local CNN branch to take advantage of their complementarity.

2.4. Frequency-Domain Learning:

Fourier analysis breaks down the image into spatial-frequency components. The lower spatial-frequency components include gradual changes in the intensity and the high spatial-frequency components include the edges and details. Unfortunately the higher spatial-frequency components also contain high levels of sensor noise.

The Parseval's theorem shows the energy comparison between the spatial view and Fourier view with equal normalization. So it does not mean the spatial L1 loss contains no high frequency.

The proposed Dynamic Frequency-Selective Loss assesses this as it tests if a learned, dynamic

weighting of the Fourier domain reconstruction errors can then contribute to the spatial domain losses.

2.5. Uncertainty Estimation:

Uncertainty estimation can help quantify the level of confidence in the prediction of a model. Spatial uncertainty maps of the enhancement algorithms in low-light enhancement shows, which parts of the image might have less reliable information.

Still this does have to be taken into account for the errors actually made for the reconstruction. These calibration tests need to be performed to determine whether or not the failure maps can be calibrated to be used as indicators of failure.

3. Proposed Method

3.1. Problem Formulation:

Now, let us consider the case where the input image is taken in low-light conditions.

Il [0, 1] HW3.

The module learns to map:

I^=f (Il),

In this final stage, I^ is the generated picture, and the parameters need to be trained.

If normal-light references can be obtained, supervised reconstruction loss could be used. For unpaired images, assessment of the quality is possible using no-reference metrics and visually.

Goals for more exposure and to keep the colour, structure details and the feeling of space of the original image.

3.2. Overall Architecture:

What are the parts of the HS-Former V3? It consists of these parts: global transformer branch, local CNN branch, the edge-aware fusion module and the reconstruction decoder.

The entire branch is feeding the context information:

Fg =T (Il).

Data provided by the local branch.

Fl=C (Il).

The fusion modules make use of these representations:

Ff=EACF (Fg Fl E),

Where E represents the edge representation of input image.

And the best image is reconstituted from:

I^=D (Ff).

Figure 1. The global Transformer branch, local CNN branch, edge-prior extraction, EACF module, decoder, and auxiliary training targets in general of HS-Former V3 architecture.

3.3. Global Transformer Branch:

The global branch encodes context and global brightness changes with the Swin Transformer V2 [6]. It uses scaled cosine attention, which can be written as: Attn (Q K V) = Softmax (cos(Q, K) + B)V where Q, K, and V are the query, key, and value features, respectively, and a learnable parameter is introduced for the temperature and B is the relative positional bias. The hierarchichal design encode relations between far away regions and a hierarchy of structured features. Number of stages, feature dimensions, depth of each block, patch window size, and normalization should be set for training.

3.4. Local CNN Branch:

The locally aware branch works with a light-weighted convolutional block to recover spatial details. Depth-wise separable convolution largely decreases the parameters compared to similar-sized convolution. Channel attention is used to reweight features to boost features. The trainable noise aware gate can be written as: Gn = g (Fl), Fl' = Fl (1-Gn), where g is a trainable convolutional mapping with the corresponding parameters, and the sigmoid function is applied. Gates are also used to suppress less certain parts of the network. Whether this is in fact detrimental or beneficial remains to be seen, as both true and noised information are at similar frequencies, and we will need to perform ablation and robustness tests to verify.

3.5 Edge-Aware Cross Fusion:

The EACF module makes use of the gradient information to fuse the global and local features. The gradient information is calculated using Sobel operators to determine horizontal and vertical gradients: Ex=SxIl, Ey=SyIl. The magnitude of the gradient is E=Ex2+Ey2+, where is a small constant for stability. The feature resolutions are first aligned and concatenated: Fc=concat(Fg, Fl', E). Then each learned fusion gate is obtained as A=h (Fc), where h is a learnable convolutional mapping. The fused representation then takes the form Ff=AFg + (1-A) Fl'+r (Fg), where r is a residual projection. This design allows the network to learn the interplay between global context features and local spatial features. Edge prior gives the gating block with structural hints, i.e. the decrease of halo artifacts should be visually examined and compared to the initial state and also quantitatively assessed with edge preservation index.

3.6. Dynamic Frequency-Selective Loss:

DFL is used to dynamically weight the spectral reconstruction errors. Let us take the Fourier transform of the predicted and reference images: Fp=F (I^), Ft=F (It), then the spectral-domain difference between them is: f=FpFt. We expect the trainable weighting network to learn the mask via the

frequency: W=M (f, c), where c is the optional extra conditioning features. The proposed loss reads: LDFL= 1 N k=1NWk fk, k+wR(W), where N is the number of coefficients, R(W) is the regularizer, w the weight of R(W). We now describe the implementation detail of this loss: what inputs would the weighting-network have, what mask size, what Fourier normalisation, and how the complex-valued coefficients are treated, with others providing the trivial solutions.

3.7. Masked-Frequency Pre-Training:

This is pre-training where we're trying to understand whether the unpaired low light image contains any spectral structural information.

Let:

Fl=F (Il),

And a binary mask M of the visibility of frequency coefficients. The network outputs:

This gives F^l = P (MFl).

A masked reconstruction objective is:

Lpre = (1M) (F^lFl) 11M1+.

The network is trained to complete the missing masked coefficients from the complete set of other coefficients. The amount of masked-out coefficients, the scheme to generate the mask, the adaptation to a complex number representation, and how the Fourier symmetry is processed should be explained.

Following pre-training, the model can be fine-tuned using coupled low light and normal light images. Data efficiency of the method can be tested through large number of experiments by using various ratios of coupled training data.

3.8. Uncertainty Estimation:

The other prediction head produces a spatial map for the log-variance s(x):

2(x) = exp (s(x)).

A Gaussian-style reconstruction objective is:

Lunc=1Nx [I^(x) It (x) 222exp (s(x)) + s(x) 2].

Because of this, the model can also be used as a proxy for the spatial distribution of the reconstruction uncertainty and should be verified with calibration and by comparing to the actual reconstruction errors.

This is a suggested guide only, not a safety guarantee.

3.9. Overall Training Objective:

The total goal is:

Ltotal=1L1+sLSSIM+fLDFL+eLledge+uLunc.

The reconstruction term is:

L1 = (1/N) ||I^ It||1.

The structural term is:

LSSIM = 1 - SSIM (I^, It).

The edge-preservation term is:

Ledge = I^ It1.

Coefficients 1sfeu are set by a written validation procedure. Uncertainty loss is used to train the uncertainty head.

4. Experimental Method

4.1. Datasets:

Benchmark datasets for low-light enhancement and night -time visual recognition are also included in the benchmark. Table 1. Datasets and evaluation objectives Dataset Characteristics Evaluation purpose LOL Dual light images this dataset includes low-light and normal-light image pairs for full-reference enhancement evaluation. SID observations from the low light sensor and the reference image for sensor-aware image restoration. ExDark contains 7,363 images captured in low-light conditions, covering 12 object categories. It is used for downstream object detection tasks. DICM Low-light sample without paired reference No-reference and qualitative evaluation. LOL gives paired data for measuring the quality of reconstructions [1]. SID helps low-light sensor imaging experiments [8] and ExDark gives annotations of low-light visual recognition [9]. Specify the versions of your datasets used, the nature of the data split, image formats and preprocessing details. Provide the preprocessing details for the RAW sensor data and for the RGB images. Documented collection location time coverage, camera capabilities annotation permissions, and sharing data must be provided for any Iraqi Night Surveillance Dataset. Size and experimental results will not be considered until verified.

5. Results Reporting Setup

5.1. Quantitative Comparison:

Table 2. Objectively evaluated algorithms for low-light enhancement. Dataset PSNR (dB) SSIM LPIPS Parameters FLOPs Retinex Net to be measured to be measured to be measured Verify. Zero-DCE Pending measurement Pending measurement Pending measurement To be verified. Restormer to be measured to be measured to be measured Verify. Retinex former measurement pending

measurement pending measurement pending Verification required. HS-Former V3 to be measured to be measured to be measured Measure. Note: This table is an example of an experimental reporting template. It is not a completed measurement.

5.2. Ablation Study:

Table 3. Ablation study on the components we proposed. Set DFL EACF Noise gate Pre-training Uncertainty head. Base line NO Configuration A Yes No B Yes No Config C Yes No Full model Yes Test all configurations under the same test conditions. The conclusion should determine whether individual components have provided further benefit in combination.

5.3. Statistical Analysis:

Results can be reported as averages (means and standard deviations) over independent runs. For paired data similar tests can be used if their assumptions are met.

The statistical significance can be derived from the data from the test. Do not report p-values and CIs until calculations are done.

6. Discussion

Then, we ask whether global contextual modeling and local convolutional restoration can be integrated together. On top of modeling the global illumination correlation by the transformer branch, the CNN branch can be regarded as an auxiliary for local structure restoration.

~Inter Beam DFL builds on an additional reconstruction loss with adaptive spectral weights that does not have the trivial solution using a stable enough weighting network. Compare with fixed frequency weighting and regular reconstruction loss to illustrate its contribution.

EACF feature merges with an image-gradient image. In the very likely case of this aiding boundary preservation, the identical gradients will also aggravate halos and noise. Both subjective, visual, assessment, and measures of edge-preservation must be used to verify that this module continues to contribute a halo suppression.

The unpaired low-light images may be used to pre-train the masked frequency model. But, the effectiveness of spectral coefficient imputation will have to be validated by the data efficiency study because it is not the same as the real normal-light appearance recovery.

Uncertainty estimation could also be used to show areas where additional gains are not possible. Still, its effectiveness remains to be assessed given the calibration levels and how they correlate to the realized reconstructed errors. It must that means be regarded as a diagnostics aid, not a safety blanket.

7. Limitations and Future Work

Generalization to unseen artefacts is a matter of practical implementation and assessment. Reconstruction quality, cost of computation, robustness data efficiency, and downstream recognition performance are all unknowns from the network architecture.

And, future work should include network scaling and evaluation on the baseline benchmark so that is reproducible. Future experiments should also include generalisation over cameras, extremely low illumination, calibrating the uncertainty, and the relationship between denoising and fine details.

Before long, we can also extend this structure to video enhancement based on temporal consistency and the visible infrared fusion. These extensions need be experimentally tested first before any practical proposition.

8. Conclusion

Low Light Image Enhancement first, we designed a hybrid Transformer CNN architecture called HS-Former V3. Our model integrates the modeling of the global context and the local restoration of the convolution to the edge-aware fusion, adaptive supervision in the frequency domain, masked-frequency pre-training, and uncertainty estimation.

The motivation is from previous Retinex -based enhancement, non-linear curve estimation, Transformer-based restoration, and low-light image datasets. The structure is to give a principled exploration on their complementarity.

The empirical performance of HS-Former V3 is not known through implementation and controlled tests. When empirically evaluating HS-Former V3, other researchers would need to cite benchmark results, ablation studies, robustness tests and downstream detection results before claiming to outperform others.

References

1. Wei, C. Wang, W. Yang, W. & Liu, J. Deep retinex decomposition for low-light enhancement Proceedings of the British Machine Vision Conference (BMVC) 2018.
2. Y. Zhang, J. Zhang, and X. Guo, "Kindling the Darkness: A Practical Low-light Image Enhancer, " Proceedings of the ACM International Conference on Multimedia, vol. 2019-October, pp. 16321640, 2019. Doi: 10.1145/3343031.3350926.
3. C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, R. Cong, Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2020, pp. 17801789.
4. Y. Jiang, H. Liu, X. Wang, W. Qi, J. He. "Enlighten GAN: Deep Light Enhancement without Paired Supervision," IEEE Transactions on Image Processing, 2021 (open-access version) the full bibliographic record for this publication is also available in the publisher's record.
5. S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient Transformer for High-Resolution Image Restoration," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022 pp. 57285739.

6. Z. Liu et al. "Swin Transformer V2: Scaling Up Capacity and Resolution, " Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022 pp. 1200912019.
7. Y. Cai, H. Bian, J. Lin, H. Wang, R. Timofte, and Y. Zhang. Retinex former: One-Stage Retinex-Based Transformer for Low-Light Image Enhancement. Proceedings of the IEEE/CVF International Conference on Computer Vision 1250412513 2023.
8. C. Chen, Q. Chen, J. Xu, and V. Koltun Learning to see in the dark Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018 pp. 32913300. Doi: 10.1109/CVPR.2018.00347
9. Y. P. Loh and C. S. Chan, "Getting to Know Low-Light Images with the Exclusively Dark Dataset," Computer Vision and Image Understanding, vol. 178, pp. 3042, 2019. Https://doi.org/10.1016/j.cviu.2018.10.010.