Embedding Fourier Features for Low-light Enhancement

Jinyu Liu

Abstract—Low-light enhancement is a complex and challenging task due to the variety of noise, inconsistent color mapping, and detail loss in the low light images. It is a computer vision task currently dominated by deep learning models, such as HVI-CIDNet, which adapt to the changes in illumination ranges and complex interactions between brightness and color. However, deep learning approaches tend to be biased toward performing well on low-frequency components of an image. For this work, our goal is to encode frequency domain information into a state of the art low-light enhancement model like HVI-CIDNet. Our experiments find the last encoding layer of a U-Net model is the optimal layer to encode this frequency information for gains in low-light enhancement performance and visual quality.

Index Terms—Computational Photography, Low-Light Imaging, Deep Learning, Multi-Modal Networks

1 Introduction

Low-light imaging refers to the capture and processing of scenes under conditions with limited illumination. The lack of sufficient lighting can significantly degrade the quality of the captured images. At capture time, some operations can be applied to mitigate the effect of limited lighting, but all have their own drawbacks. For example, increasing the ISO increases the sensitivity of the image sensors to light, but also amplifies the amount of noise captured by the sensor. Increasing the exposure time will only work if the scene is static, otherwise motion blur artifacts are introduced. Finally, using flash to artificially brighten the environment may introduce undesirable highlights and unbalanced lighting.

Low-light enhancement is the computer vision task that involves improving the perceptive quality of images captured under low-light conditions. The aim of low-light enhancement is to improve brightness and clarity, while minimizing noise and distortion artifacts. Low-light enhancement has a wide range of applications in numerous areas, including surveillance, autonomous driving, and computational photography. As a result, low-light enhancement has emerged as an exciting research area.

Recently, advances in this area have been dominated by deep learning based solutions [1]. These solutions typically have higher accuracy, more robustness to noise, and are faster than conventional algorithmic methods. However, an interesting phenomenon observed in neural networks is that they are biased towards learning less complex functions [2]. For computational imaging tasks, this means that neural networks tend to perform well in the low-frequency regions of an image and poorly on the high-frequency regions of the image.

This project focuses primarily on exploring if embedding frequency features into a deep learning network can help reduce the effect of this low-frequency bias. More specifically, it examines if we can effectively implement cross-attention to extract meaningful information from the Fourier domain to enhance the performance of a low-light enhancement model.

• *J. Liu is a student under the Department of Computer Science, University of Toronto, Toronto, ON.*
E-mail: alexliu@cs.toronto.edu

2 Related Work

2.1 Fourier Features in Computational Imaging

Fourier features are representations of signals or images in the frequency domain, typically obtained by applying a Fourier transform. The frequency information in an image helps to identify fine details and edges, which are typically present in high frequency regions of the image, and overall structure and smooth textures, which are typically present in low frequency regions of the image.

The properties of Fourier features have been explored in deep neural networks extensively. Works such as [3] have explored the use of Fourier features in convolution blocks to increase the receptive field of a convolution kernel to be global. This ensures that the earlier layers in the convolutional neural network have vision of the entire input image, and have shown improved performance in image and facial recognition tasks. Earlier works such as [4] have analyzed the behavior of convolutional neural networks when the layers are put in the frequency domain, introducing a variety of tools such as spectral pooling and spectral re-parametrization of convolution filters. These techniques been have shown to improve translation invariance and converge faster than spatial domain layers in some tasks.

2.2 Neural Networks for Low-Light Image Enhancement

The first machine learning approach for low-light enhancement was LLNet [5], which employed a stack of denoising autoencoders. These denoising autoencoders were responsible for learning to adaptively enhance and denoise synthetically darkened images, and generalized well onto naturally low-light images.

Instead of directly computing loss from the RGB image channels, other works in low-light enhancement used

Retinex theory [6] to separate illuminance and reflectance maps from the input image and used these maps as input. Retinex-Net [7] was a pioneer in using this method and had better image enhancement performance than other models at this time. This paper also introduced the LOw-Light dataset (LOL), which comprised of low/normal light image pairs captured by varying exposure time and ISO of real scenes. This dataset became a popular benchmark in the low-light enhancement space. Unfortunately, Retinex-based models tended to have issues with color shifts and contain color bias. These issues were explored in later works.

More recently, HVI-CIDNet [8] has been released to address color misalignment issues. HVI-CIDNet works by first transforming the input image into a trainable color space, Horizontal/Vertical Intensity (HVI). The ability for the color space to be trained reduces the color instability during enhancement. When released, HVI-CIDNet was state of the art on the LOL dataset.

2.3 Fusion

There are various way to combine information from multiple inputs in a neural network. Works like [9] found that multi-modal models that use RGB and depth inputs outperform single-modality RGB or depth models that have similar architecture. In addition, the work suggests that middle level information exchange produced the most increase in performance. Another work [10] found similar results in that dense fusion (analogous to middle fusion) outperformed both early and late fusion. This work confirmed that multi-modal models have better performance in middle fusion in various different tasks.

3 Proposed Method

In this paper, we will explore the various methods for multi-modal fusion for the computational imaging task of low-light image enhancement. We want to separate out the Fourier features from the input image by applying a fast Fourier transformation on the image. The goal of this work to see if a multi-modal network that takes the Fourier features and input image can have improved results over the single input network. We also aim to find the best method for fusing the two inputs together.

Past works on multi-modal fusion have dealt with synchronized inputs, such as RGB-depth pairs [9] or CT-PET scans [10]. Fourier features differ in that they operate globally, and are not synchronized locally. That is, any input at pixel (x, y) for the RGB input will not correspond to the same pixel location in the Fourier input.

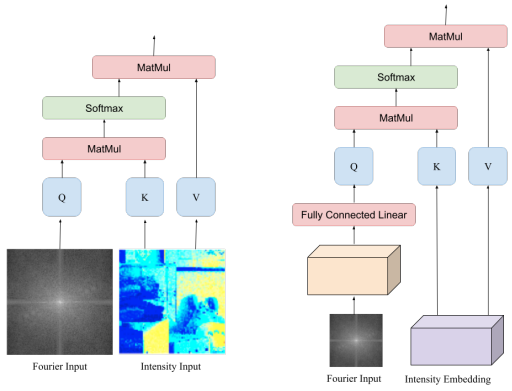
3.1 HVI-CIDNet Baseline Architecture

The baseline model we will be using for the pipeline is HVI-CIDNet [8]. As previously mentioned, this model has state of the art performance on the commonly used LOL dataset.

A diagram for the HVI-CIDNet framework can be found in Figure 1. To address the common color misalignment problems, the HVI color space has a trainable mapping from the RGB space. The first few layers transform the RGB image into the HVI color space. The intensity map represents the maximum value of the input image along any of the channels. As such, we propose it makes the most sense to encode Fourier features into the Intensity channel, as it would be the most sensitive to details and edge discontinuities. The image in the HVI color space is then fed into the enhancement network, which is a U-Net based network that performs the main processing. We will modify the network here, introducing early/middle/late fusion.

3.2 Encoding Fourier Features

The first fusion approach we will attempt is early fusion. At the beginning of the enhancement network, we will apply the fast Fourier transform on the RGB input image to get the Fourier representation of the image. Other works have used convolution layers to fuse the inputs together, however these convolution layers have limited receptive fields. Since the Fourier representation is not locally synchronized, the relevant Fourier information might not be present in the convolution's receptive field. Instead, to encode this Fourier representation into the intensity channel, we will employ cross-attention between the two inputs. The intensity channel will serve as the query and the keys and values will be obtained from the Fourier features. This is done before the intensity features are extracted from the input and before the enhancement network is run. A diagram of this is shown in Figure 2, on the left.



(a) Encoding the Fourier input into an Intensity Map

(b) Encoding the Fourier input into the embedding of the U-Net

Fig. 2: Cross-attention mechanism for encoding Fourier information into the Intensity information.

For late fusion, we will apply the same encoding method as for early fusion, except we will encode the Fourier features into the intensity channel output of the enhancement network instead of the input. Since the inputs' sizes are all identical, no modifications need to be made to the cross-attention mechanism.

Since the enhancement network employs a U-Net architecture, the middle embedding of the network is made up of more dimensions than the Fourier representation. To ensure the dimensions match, we apply a trainable fully-connected linear layer to transform the Fourier input, ensuring the inputs to the cross-attention mechanism are compatible. A diagram for this can be found in Figure 2, on the right.

Due to the increased number of channels, we also apply layer normalization to stabilize the attention scores during

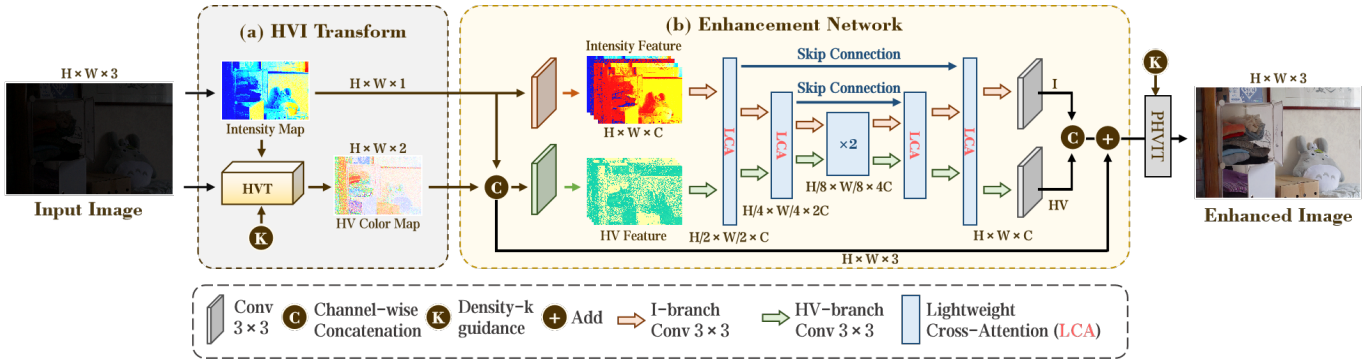


Fig. 1: Overview of the HVI-CIDNet architecture. Taken from [8].

training [11]. Since the dimensionality for the input for early and late fusion is 1, we did not use layer normalization for them.

3.3 Training Parameters

3.3.1 Dataset

We will be using LOLv1 for training and testing. LOLv1 is made up of 500 images, which we split 485:15 for training:test datasets. We will use a batch size of 2 for training. The images are originally of size 400x600 and in RGB format. We perform random crop on the images down to size 400x400 and normalize the inputs with the same mean and variance as the standard convention for images trained on the ImageNet dataset. We also augment the data by applying random horizontal and vertical flip.

3.3.2 Experiment Settings

We will be training the baseline model and fusion-modified models using the Adam optimizer [12] for 1500 epochs over all the training data. We use a learning rate of 1×10^{-5} initially and steadily decrease the learning rate to 1×10^{-8} using the cosine annealing scheme [13].

We use the same loss as the original HVI-CIDNet paper [8], which is a combination of L1 loss L_1, structural similarity (SSIM) loss L_s [14], edge loss L_e [15], and perceptual loss L_p [16] between the training and test images in HVI colour space. The total loss becomes:

$$\begin{aligned} l(\hat{X_{HVI}}, X_{HVI}) = \lambda_1 \cdot L_1(\hat{X_{HVI}}, H_{HVI}) \\ + \lambda_s \cdot L_s(\hat{X_{HVI}}, H_{HVI}) \\ + \lambda_e \cdot L_e(\hat{X_{HVI}}, H_{HVI}) \\ + \lambda_p \cdot L_p(\hat{X_{HVI}}, H_{HVI})) \end{aligned}$$

where $\lambda_1, \lambda_s, \lambda_e, \lambda_d$ are the weights for their respective loss components.

This loss is then weighted by λ_{HVI}, and added to the same loss function but over the RGB color space. Therefore, the total loss function L becomes

$$L = \lambda_{HVI} \cdot l(\hat{X_{HVI}}, X_{HVI}) + l(\hat{X_{RGB}}, X_{RGB})$$

Unfortunately, the default values for the loss weights provided by the HVI-CIDNet authors did not produce satisfactory results. Instead, we performed grid search on the loss weights and found the following values performed the best: $\lambda_{HVI} = 1.5, \lambda_1 = 2.5, \lambda_s = 0.7, \lambda_e = 10, \lambda_p = 0.05$

4 Experimental Results

4.1 Qualitative Results

In Figure 3, we see that the middle fusion based model produces results with illumination closest to the ground truth. It also produces colors most similar to the ground truth. However, the bottom row of Figure 3 has some saturation loss in color. Even so, the perceptive quality is the highest for the middle fusion result, and it is the only fusion result that surpasses the baseline in perceptive quality.

Augmenting the model by adding early fusion seemed to reduce the saturation and created less vibrant colors, resulting in a muted visual effect. We hypothesize the reason for this behavior is that the Fourier transformation is very complex and difficult to encode correctly. When we are passing this information through the skip connections, the model is not disentangling the complex information properly. As a result, the model is tending toward more conservative predictions, resulting in the muted color.

On the other hand, late fusion result exhibits regions of excessive or unatural color saturation. It also seemed to lose some edge details in regions of extremely low light, seen in the top row of Figure 3. We suspect that encoding the Fourier information at the output intensity feature map was ineffective since the model had already formed a strong prediction on the output at this stage. Few layers follow this encoding layer and it is difficult for the model meaningfully incorporate the Fourier encoding in a beneficial way, and thus the additional information may have corrupted the output.

4.2 Quantitative Results

We compare our fusion-modified HVI-CIDNet architectures against the baseline and each other using various metrics. These metrics include the peak signal-to-noise ratio (PSNR), SSIM, and perceptive quality loss, shown in Table 1 and plotted in Figure 4.

Note these benchmarks confirm that middle fusion based fusion produced the best results, outperforming the baseline model in PSNR, SSIM, and perceptual quality. We also verify that early fusion and late fusion based models

(a) Low-light image example 1 (b) High-light image example 1 (c) Baseline output example 1 (d) Early fusion example 1 (e) Middle fusion example 1 (f) Late fusion example 1

(g) Low-light image example 2 (h) High-light image example 2 (i) Baseline output example 2 (j) Early fusion example 2 (k) Middle fusion example 2 (l) Late fusion example 2

Fig. 3: LOLv1 test set examples and output from the various models.

perform worse than the baseline in PSNR, SSIM, and perceptual quality. Early fusion works slightly better than late fusion in terms of SSIM and perceptual quality, but slightly worse than late fusion in terms of PSNR.

TABLE 1: Quantitative Benchmarks of the various model architectures

Model	PSNR	SSIM	Perceptual Loss
Baseline	19.8997	0.8182	0.1293
Early Fusion	15.3205	0.6869	0.3236
Middle Fusion	**22.1694**	**0.8382**	**0.1160**
Late Fusion	15.4061	0.6590	0.3168

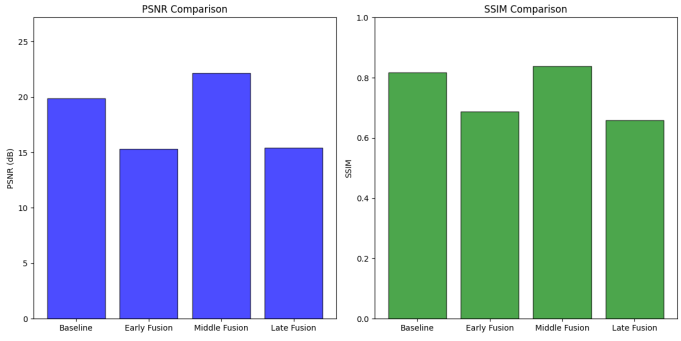


Fig. 4: PSNR and SSIM bar graphs of the various model architectures.

5 Conclusion

5.1 Summary

In this work, we evaluate the performance of a single modality low-light enhancement network and experiment to see if we get any performance gains from integrating another input in the Fourier space. We find that early and late fusion of these additional Fourier features do not improve the model performance in terms of PSNR, SSIM, and perceptual quality. However, when we embed these features in an embedding layer, we see improvements in all the above metrics over the baseline model.

5.2 Limitations

The baseline model and the fusion modifications have only been evaluated on the LOLv1 dataset. Numerous other low-light enhancement datasets are available, including See-in-the-Dark dataset [17] and LOLv2 [18]. It's possible there is a data sparsity problem and it would be interesting to see if there are any changes in the results if we increasing the dataset size and variety.

In addition, the HVI-CIDNet is unique in that it works in it's own color space. It is not immediately clear whether or not these results extend to other models working in the RGB color space. One potential interaction is that the HVI color space is trained to integrate the Fourier features better. However it could also be the case that the complexity results in training instabilities that make it difficult for the inputs to be fused. More experimentation is needed and a potential future work could be to evaluate a RGB-based model and see if augmenting it to encode Fourier space features would have the same effect as it did on HVI-CIDNet.

Lastly, we found during experimentation that HVI-CIDNet was very sensitive to hyperparameter changes. Small adjustments in learning rate and loss weights would sometimes produce an fully black image or a noisy image. Due to the large number of hyperparameters, we cannot make the conclusion that the results we obtained are the best possible results for the model architecture, just the best results we were able to find.

Acknowledgments

The authors would like to thank Parsa Mirdehghan for his invaluable mentorship and support throughout the duration of this project, particularly during its initial stages, where his guidance was instrumental in shaping and refining the core ideas. The authors would also like to thank Dr. David Lindell, Dr. Aviad Levis, and Shayan Shekarforoush for their instruction and support throughout the course.

References

[1] C. Li, C. Guo, L. Han, J. Jiang, M.-M. Cheng, J. Gu, and C. C. Loy, "Low-light image and video enhancement using deep learning: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 12, pp. 9396–9416, 2021.

[2] Y. Cao, Z. Fang, Y. Wu, D.-X. Zhou, and Q. Gu, "Towards understanding the spectral bias of deep learning," *arXiv preprint arXiv:1912.01198*, 2019.

[3] L. Chi, B. Jiang, and Y. Mu, "Fast fourier convolution," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4479–4488, 2020.

[4] O. Rippel, J. Snoek, and R. P. Adams, "Spectral representations for convolutional neural networks," *Advances in neural information processing systems*, vol. 28, 2015.

[5] K. G. Lore, A. Akintayo, and S. Sarkar, "Llnet: A deep autoencoder approach to natural low-light image enhancement," *Pattern Recognition*, vol. 61, pp. 650–662, 2017.

[6] E. H. Land, "The retinex theory of color vision," *Scientific american*, vol. 237, no. 6, pp. 108–129, 1977.

[7] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," *arXiv preprint arXiv:1808.04560*, 2018.

[8] Q. Yan, Y. Feng, C. Zhang, P. Wang, P. Wu, W. Dong, J. Sun, and Y. Zhang, "You only need one color space: An efficient network for low-light image enhancement," *arXiv preprint arXiv:2402.05809*, 2024.

[9] A. Roitberg, T. Pollert, M. Haurilet, M. Martin, and R. Stiefelhagen, "Analysis of deep fusion strategies for multi-modal gesture recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[10] G. M. Alshmrani, Q. Ni, R. Jiang, and N. Muhammed, "Hyper-dense_lung_seg: Multimodal-fusion-based modified u-net for lung tumour segmentation using multimodality of ct-pet scans," *Diagnostics*, vol. 13, no. 22, p. 3481, 2023.

[11] S. Brody, U. Alon, and E. Yahav, "On the expressivity role of layernorm in transformers' attention," *arXiv preprint arXiv:2305.02582*, 2023.

[12] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[13] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[14] B. A. WANGZ, H. Sheikh *et al.*, "Image qualityassessment: Fromerrorvisibilitytostructural similarity," *IEEETransactionsonImageProcessing*, vol. 13, no. 4, p. 600G612, 2004.

[15] G. Seif and D. Androutsos, "Edge-based loss function for single image super-resolution," in *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 1468–1472.

[16] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 694–711.

[17] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to see in the dark," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3291–3300.

[18] W. Yang, S. Wang, Y. Fang, Y. Wang, and J. Liu, "From fidelity to perceptual quality: A semi-supervised approach for low-light image enhancement," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3063–3072.