BAYESIAN ENHANCEMENT MODELS FOR ONE-TO-MANY MAPPING IN IMAGE ENHANCEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Image enhancement is considered an ill-posed inverse problem due to its tendency to have multiple solutions. The loss of information makes accurately reconstructing the original image from observed data challenging. Also, the quality of the result is often subjective to individual preferences. This obviously poses a one-to-many mapping challenge. To address this, we propose a Bayesian Enhancement Model (BEM) that leverages Bayesian estimation to capture inherent uncertainty and accommodate diverse outputs. Our approach, integrated within a two-stage framework, first employs a Bayesian Neural Network (BNN) to model reduced-dimensional image representations, followed by a deterministic network for refinement. We further introduce a dynamic *Momentum Prior* to overcome convergence issues typically faced by BNNs in high-dimensional spaces. Extensive experiments across multiple low-light and underwater image enhancement benchmarks demonstrate the superiority of our method over traditional deterministic models, particularly in real-world applications lacking reference images, highlighting the potential of Bayesian models in handling one-to-many mapping problems.

1 INTRODUCTION

In computer vision, image enhancement refers to the process of enhancing the perceptual quality, visibility, and overall appearance of an image, which can involve reducing noise, increasing contrast, sharpening details, or correcting colour imbalances. In image enhancement tasks such as low-light image enhancement (LLIE) and underwater image enhancement (UIE), a common challenge arises from dynamic photography conditions, where a single degraded input image can correspond to multiple plausible target images. This phenomenon, known as the *one-to-many mapping* problem, arises because multiple valid outputs can be generated depending on varying conditions during image capture, such as changes in lighting, exposure, or other factors.

Recent advances in deep learning have shifted image enhancement towards data-driven approaches. Several deep learning-based models (Zamir et al., 2022; Cai et al., 2023) have achieved advanced results by learning mappings between low-quality inputs and their high-quality counterparts using paired datasets. However, we observe that existing datasets exhibit the one-to-many relationship between their input and target domains. Specifically, we observe cases where there exist at least two image pairs with input images that are either identical or visually indistinguishable, yet their corresponding targets exhibit notable variations. When such discrepancies arise due to ambiguity in the target domain, a traditional deep neural network—being a deterministic function—struggles to effectively model these one-to-many image pairs. Previous methods employing deterministic neural networks (DNNs) for image enhancement often overlook this class of one-to-many samples, leading to sub-optimal solutions. Figure 1 (middle) demonstrates how a deterministic neural network trained on one-to-many mapping data struggles to predict any specific target, instead producing an averaged output due to "regression toward the mean".

To tackle the inherent ambiguity in image enhancement tasks caused by one-to-many mappings, we adopt a Bayesian framework that models these mappings probabilistically. Rather than relying on a sub-optimal deterministic approach, our method leverages Bayesian inference to sample multiple sets of network weights from a learned distribution, effectively creating a diverse ensemble of deep networks. Each sampled network captures a distinct plausible solution, allowing our model to map a single input to a distribution of possible target outputs. This approach theoretically enables the

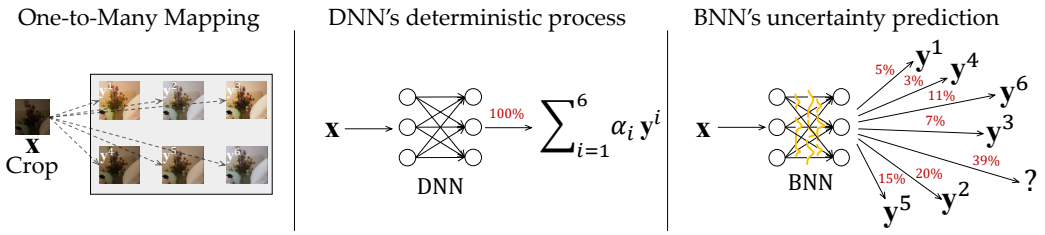


Figure 1: One-to-Many Mapping. The left panel shows an image crop $\mathbf{x}$ associated with multiple targets $\{\mathbf{y}^1, \ldots, \mathbf{y}^6\}$. A DNN (middle) trained on such data tends to predict the weighted average of all targets. In contrast, a BNN (right) models the one-to-many relation by producing different outputs according to a learned probability distribution.

mapping of all plausible variations, effectively modelling the complex one-to-many relationships present in real-world scenarios.

While Bayesian Neural Networks (BNNs) have shown promise in capturing uncertainty in various tasks (Kendall & Cipolla, 2016; Kendall et al., 2015; 2018), their potential in addressing the one-to-many mapping problem for image enhancement remains largely under-explored. By incorporating Bayesian inference into the enhancement process, our approach captures uncertainty in dynamic, uncontrolled environments, providing a more flexible and robust solution than traditional deterministic models. However, applying BNNs to these tasks presents significant challenges due to the high dimensionality of image data and the strong 2D spatial correlations between pixels. For example, the weight uncertainty in BNNs often leads to noisy image outputs, while models with high-dimensional weight spaces are prone to underfitting. Following our approach, we systematically address these challenges, unleashing the potential of BNNs in image-related tasks by overcoming their limitations and improving their performance in high-dimensional settings.

As the first work to explore the feasibility of BNNs for image enhancement, we selected tasks where the *one-to-many mapping* problem is particularly pronounced, such as LLIE and UIE, to effectively validate our theoretical framework. The main contributions of this paper are summarised as follows:

- We identify the one-to-many mapping issue between inputs and outputs as a primary bottleneck in image enhancement models, and propose the first Bayesian-based Enhancement Model (BEM) to learn this mapping relation.
- We introduce a dynamic prior called the *Momentum Prior* to mitigate the convergence difficulties typically encountered by BNNs in high-dimensional weight spaces.
- To reduce the complexity of BEM in modelling high-dimensional image data, we propose an innovative two-stage approach that combines the strengths of Bayesian Neural Networks (BNNs) and Deterministic Neural Networks (DNNs).

2 Related Work

Bayesian Deep Learning. BNNs quantify uncertainty by learning distributions over network weights, offering robust predictions (Neal, 2012). Variational Inference (VI) is a common method for approximating these distributions (Graves, 2011; Blundell et al., 2015). Gal & Ghahramani (2016) simplify the implementation of BNNs by interpreting dropout as an approximate Bayesian inference method. Recent advancements show that adding uncertainty only to the final layer can efficiently approximate a full BNN (Harrison et al., 2024). Another line of approaches, such as Krishnan et al. (2020), explored the use of empirical Bayes to specify weight priors in BNNs to enhance the model's adaptability to diverse datasets. These BNN approaches have shown promise across a range of vision applications, including camera relocalisation (Kendall & Cipolla, 2016), semantic and instance segmentation (Kendall et al., 2015; 2018). Despite these advances, BNNs remain underutilised in image enhancement tasks.

Probabilistic Models in Image Enhancement. Several works have utilised probabilistic models to address different aspects of image enhancement. Jiang et al. (2021) employed GANs to capture features for LLIE, while Fabbri et al. (2018) leveraged CycleGAN (Zhu et al., 2017) to generate

synthetic paired datasets, addressing data scarcity in UIE. FUnIE-GAN (Islam et al., 2020) further demonstrated effectiveness in both paired and unpaired UIE training. Anantrasirichai & Bull (2021) applied unpaired learning for LLIE when the scene conditions are known. Wang et al. (2022) applied normalising flow-based methods to reduce residual noise in LLIE predictions. However, its invertibility constraint limits model complexity. Zhou et al. (2024) mitigated this by integrating normalising flows with codebook techniques, introducing latent normalising flows. Diffusion Models (DMs) have been widely adopted for enhancement tasks (Hou et al., 2024; Tang et al., 2023). While DMs inherently address one-to-many mappings, their high latency for generating a single sample makes producing hundreds of candidates impractical due to prohibitive delays. Due to the practical limitations in generating multiple candidates, DM-based methods often prefer to produce an average of multiple targets, as this helps reduce the quality fluctuations within a single sampling process, as suggested by Jiang et al. (2023a).

2.1 PRELIMINARIES

In image enhancement, the output of a neural network can be interpreted as the conditional probability distribution of the target image, $\mathbf{y} \in \mathcal{Y}$, given the degraded input image $\mathbf{x} \in \mathcal{X}$, and the network's weights $\mathbf{w}$: $P(\mathbf{y}|\mathbf{x}, \mathbf{w})$. Assuming the prediction errors follow a Gaussian distribution, the conditional probability density function (PDF) of the target image $\mathbf{y}$ can be modeled as a multivariate Gaussian, where the mean is given by the neural network output $F(\mathbf{x}; \mathbf{w})$:

$$P(\mathbf{y}|\mathbf{x}, \mathbf{w}) = \mathcal{N}(\mathbf{y}|F(\mathbf{x}; \mathbf{w}), \mathrm{diag}(\boldsymbol{\sigma}^2)). \tag{1}$$

The network weights $\mathbf{w}$ can be learned through maximum likelihood estimation (MLE). Given a dataset of image pairs $\{\mathbf{x}^i, \mathbf{y}^i\}_{i=1}^N$, the MLE estimate $\mathbf{w}^{\mathrm{MLE}}$ is computed by maximising the log-likelihood of the observed data:

$$\mathbf{w}^{\mathrm{MLE}} = \underset{\mathbf{w}}{\mathrm{argmax}} \sum_{i=1}^{N} \log P(\mathbf{y}^i|\mathbf{x}^i, \mathbf{w}), \tag{2}$$

By optimising such an objective function in Eq. (2), the network $F_{\mathbf{w}}$ learns an injective function, $F_{\mathbf{w}} : \mathcal{X} \to \mathcal{Y}$. The deterministic nature of such a mapping implies that when $\mathbf{y}^i \neq \mathbf{y}^j$, the condition $\mathbf{x}^i \neq \mathbf{x}^j$ must hold. We argue that this deterministic process is inadequate in cases where one input corresponds to multiple plausible targets. In Sec. 3, we delve into methods for addressing this issue.

3 MODELLING THE ONE-TO-MANY MAPPING

During inference, the one-to-many mapping relation can be viewed as stemming from predictive uncertainty. To model this uncertainty, we can train multiple sets of network weights or even multiple networks, where each set is capable of predicting one of the potential targets. To train such diverse sets of weights, we adopt a Bayesian Probabilistic Model (Neal, 2012), assuming that the weights are drawn from an unknown distribution. By repeatedly sampling from this distribution, we obtain multiple sets of weights, which the network then maps to potential targets.

3.1 BAYESIAN ENHANCEMENT MODELS

We introduce uncertainty into the network weights $\mathbf{w}$ through Bayesian estimation, thus obtaining a posterior distribution over the weight, $\mathbf{w} \sim P(\mathbf{w}|\mathbf{y}, \mathbf{x})$. During inference, weights are sampled from this distribution. The posterior distribution over the weights is expressed as:

$$P(\mathbf{w}|\mathbf{y}, \mathbf{x}) = \frac{P(\mathbf{y}|\mathbf{x}, \mathbf{w})P(\mathbf{w})}{P(\mathbf{y}|\mathbf{x})} \tag{3}$$

where $P(\mathbf{y} \mid \mathbf{x}, \mathbf{w})$ represents the likelihood of observing $\mathbf{y}$ given the input $\mathbf{x}$ and weights $\mathbf{w}$, $P(\mathbf{w})$ denotes the prior distribution of the weights, and $P(\mathbf{y} \mid \mathbf{x})$ is the marginal likelihood.

Unfortunately, for any neural networks the posterior in Eq. (3) cannot be calculated analytically. This makes it impractical to directly sample weights from the true posterior distribution. Instead, we can leverage variational inference (VI) to approximate $P(\mathbf{w}|\mathbf{y}, \mathbf{x})$ with a more tractable distribution $q(\mathbf{w}|\boldsymbol{\theta})$. Such that, we can draw samples of weights $\mathbf{w}$ from the distribution $q(\mathbf{w}|\boldsymbol{\theta})$. As suggested

by (Hinton & Van Camp, 1993; Graves, 2011; Blundell et al., 2015), the variational approximation is fitted by minimising their Kullback-Leibler (KL) divergence:

$$\begin{aligned}\boldsymbol{\theta}^\star &= \arg\min_{\boldsymbol{\theta}} \mathrm{KL}\left[q(\mathbf{w}|\boldsymbol{\theta})\|P(\mathbf{w}|\mathbf{y},\mathbf{x})\right] \\ &= \arg\min_{\boldsymbol{\theta}} \int q(\mathbf{w}|\boldsymbol{\theta}) \log \frac{q(\mathbf{w}|\boldsymbol{\theta})}{P(\mathbf{w})P(\mathbf{y}|\mathbf{x},\mathbf{w})}\,\mathrm{d}\mathbf{w} \qquad \text{(Apply Equation 3)} \\ &= \arg\min_{\boldsymbol{\theta}} -\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\theta})}\left[\log P(\mathbf{y}|\mathbf{x},\mathbf{w})\right] + \mathrm{KL}\left[q(\mathbf{w}|\boldsymbol{\theta})\|P(\mathbf{w})\right].\end{aligned} \tag{4}$$

We define the resulting cost function from Eq. (4) as:

$$\mathcal{L}(\mathbf{x},\mathbf{y}) = \underbrace{-\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\theta})}\left[\log P(\mathbf{y}|\mathbf{x},\mathbf{w})\right]}_{\text{data-dependent term}} + \underbrace{\mathrm{KL}\left[q(\mathbf{w}|\boldsymbol{\theta})\|P(\mathbf{w})\right]}_{\text{prior matching term}}. \tag{5}$$

The loss function $\mathcal{L}(\mathbf{x},\mathbf{y})$ in Eq. (5), also known as the variational free energy, consists of two components: the prior matching term and the data-dependent term. The prior matching term can be approximated using the Monte Carlo method or computed analytically if a closed-form solution exists. The data-dependent term is equivalent to minimising the mean squared error between the input-output pairs in the training data. To optimise $\mathcal{L}(\mathbf{x},\mathbf{y})$, the prior distribution $P(\mathbf{w})$ must be defined. In Sec. 3.2, we define a dynamic prior that accelerates convergence and better models complex one-to-many mappings in the data.

3.2 Momentum Prior with Exponential Moving Average

BNNs with high-dimensional weight spaces often encounter challenges such as underfitting or even non-convergence. This limitation is a significant factor hindering their performance in low-level vision tasks. To address this, we propose the concept of *Momentum Prior*, which leverages an exponential moving average strategy to stabilise the training process and improve convergence.

Suppose that the variational posterior is a diagonal Gaussian, then the variational posterior parameters are $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma})$. A posterior sample of the weights $\mathbf{w}$ can be obtained via the reparameterisation trick (Kingma, 2014).

$$\mathbf{w} = \boldsymbol{\mu} + \boldsymbol{\sigma} \circ \boldsymbol{\epsilon} \quad \text{with } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{6}$$

Having liberated our algorithm from the confines of fixed priors, we propose a dynamic prior by updating the prior's parameters to the exponential moving average (EMA) of the variational posterior parameters. Specifically, for the prior $P(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \boldsymbol{\mu}_t^{\mathrm{EMA}}, {\boldsymbol{\sigma}_t^{\mathrm{EMA}}}^2\mathbf{I})$, the parameters are updated at each minibatch training step t over the training period $[0, 1, 2, \ldots, T]$ as follows:

$$\begin{aligned}&\boldsymbol{\mu}_0^{\mathrm{EMA}} = \mathbf{0}, \;\; \boldsymbol{\sigma}_0^{\mathrm{EMA}} = \sigma^{\mathrm{o}}\mathbf{1}, \\ &\boldsymbol{\mu}_t^{\mathrm{EMA}} = \beta\boldsymbol{\mu}_{t-1}^{\mathrm{EMA}} + (1-\beta)\boldsymbol{\mu}_t, \quad t = 1...T, \\ &\boldsymbol{\sigma}_t^{\mathrm{EMA}} = \beta\boldsymbol{\sigma}_{t-1}^{\mathrm{EMA}} + (1-\beta)\boldsymbol{\sigma}_t, \quad t = 1...T,\end{aligned} \tag{7}$$

where $\boldsymbol{\mu}_t$ and $\boldsymbol{\sigma}_t$ represent the mean and variance from the variational posterior $q(\mathbf{w}|\boldsymbol{\theta})$ at training step t, σ^{o} is a scalar controlling the magnitude of initial variance in the prior distribution, and β denotes the EMA decay rate. Thereafter, for minibatch optimisation with M image pairs, we update $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma})$ at step t by minimising minibatch loss $\mathcal{L}^{\mathrm{mini}}(\mathbf{x},\mathbf{y})$, reformulated from Eq. (5) as:

$$\begin{aligned}\mathcal{L}^{\mathrm{mini}}(\mathbf{x},\mathbf{y}) &= \underbrace{-\mathbb{E}_{q(\mathbf{w}|\boldsymbol{\theta})}\left[\log P(\mathbf{y}|\mathbf{x},\mathbf{w})\right]}_{\text{data-dependent term}} + \underbrace{\frac{1}{M}\mathrm{KL}\left[q(\mathbf{w}|\boldsymbol{\theta})\|P(\mathbf{w})\right]}_{\text{prior matching term}}, \\ &= \frac{1}{M}\Big[\underbrace{\sum_i^M \mathbb{E}_{\mathbf{w}\sim q(\mathbf{w}|\boldsymbol{\theta})}\|F(\mathbf{x}^i;\mathbf{w}) - \mathbf{y}^i\|_2^2}_{\text{data-dependent term}} + \underbrace{\log\frac{\boldsymbol{\sigma}_t^{\mathrm{EMA}}}{\boldsymbol{\sigma}} + \frac{\boldsymbol{\sigma}^2 + (\boldsymbol{\mu} - \boldsymbol{\mu}_t^{\mathrm{EMA}})^2}{2{\boldsymbol{\sigma}_t^{\mathrm{EMA}}}^2} - \frac{1}{2}}_{\text{prior matching term}}\Big],\end{aligned} \tag{8}$$

where the prior matching term is expressed as the analytical solution of $\mathrm{KL}\left[q(\mathbf{w}|\boldsymbol{\theta})\|P(\mathbf{w})\right]$.

Unlike empirical Bayes (Robbins, 1956; Krishnan et al., 2020), which defines a static prior based on MLE-optimised parameters, our momentum-based strategy incrementally refines the prior during training. This continuous adaptation prevents the model from exploiting shortcut learning when optimising the data-dependent term in Eq. (5), thereby avoiding sub-optimal solutions.

3.3 PREDICTIONS UNDER UNCERTAINTY

After optimising the variational posterior parameters $\boldsymbol{\theta}^\star$ via Eq. (4), predictions are made by sampling weights $\mathbf{w}$ from the variational posterior distribution $q(\mathbf{w}|\boldsymbol{\theta})$. As shown in Algorithm 1, we sample K sets of network weights $\{\mathbf{w}_k\}_{k=1}^K$, where each $\mathbf{w}_k$ is used to produce a corresponding output $\hat{\mathbf{y}}_k$ via $F(\mathbf{x}; \mathbf{w}_k)$. A quality metric D is then employed to rank the K candidates and select the most suitable output $\mathbf{y}^{\text{opt}}$, with higher D-values indicating better quality.

The prediction process is described for two cases depending on the availability of a reference:

i) With reference: When a reference image $\mathbf{y}$ is available, the quality metric D can be instantiated as the negative mean squared error (MSE) or other perceptual metrics to rank the K candidates, with the best score determining the final output.

ii) Without reference: in the absence of a reference image, the quality metric $D(\cdot)$ can be a no-reference image quality metric, such as NIQE (Mittal et al., 2012), UIQM (Panetta et al., 2015), or UCIQE (Yang & Sowmya, 2015). Alternatively, vision-language models like CLIP (Radford et al., 2021; Wang et al., 2023) can be used to find the best-matching image based on a given textual description. For instance, CLIP's encoders can extract features from a predicted image $\hat{\mathbf{y}}_k$ and a text prompt (e.g., "`A bright photo`"), denoted as $\mathbf{h}_k$ and $\mathbf{h}_{\text{text}}$, respectively. The quality metric D is then defined as their cosine similarity: $D(\hat{\mathbf{y}}_k) = \frac{\mathbf{h}_k^\top \mathbf{h}_{\text{text}}}{\|\mathbf{h}_k\| \|\mathbf{h}_{\text{text}}\|}$.

Algorithm 1: Prediction

Input: Input $\mathbf{x}$, network F
Initialisation: the best score $s^{\text{best}} \leftarrow 0$;
for $k \leftarrow 1$ **to** K **do**
  Sample $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$;
  $\mathbf{w}_k \leftarrow$ Calculate Eq. (6);
  $\hat{\mathbf{y}}_k = F(\mathbf{x}; \mathbf{w}_k)$;
  if reference $\mathbf{y}$ exists **then**
    $s_k = D(\hat{\mathbf{y}}_k, \mathbf{y})$; // reference
  else
    $s_k = D(\hat{\mathbf{y}}_k)$; // no-reference
  if $s_k > s^{\text{best}}$ **then**
    Update $s^{\text{best}} \leftarrow s_k$;
    Set $\mathbf{y}^{\text{opt}} \leftarrow \hat{\mathbf{y}}_k$;
Output: Optimal prediction $\mathbf{y}^{\text{opt}}$.

Meanwhile, our BEM can perform deterministic predictions (i.e., without requiring multiple weight samples) by simply setting $\mathbf{w} = \mathbf{u}$. We refer to this deterministic mode as BEM-DNN. However, due to its deterministic nature, BEM-DNN, like any deterministic model, is inherently sub-optimal for capturing complex one-to-many mappings.

4 A TWO-STAGE APPROACH

Image data is inherently high-dimensional. While BNN can be directly applied to model such data, it often compromises precision due to the complexity involved. To address this issue, we propose to use BEM to model the one-to-many mapping in a lower-dimensional feature representation of image. Then, we project the image features back to the original pixel space by a DNN.

4.1 THE FRAMEWORK

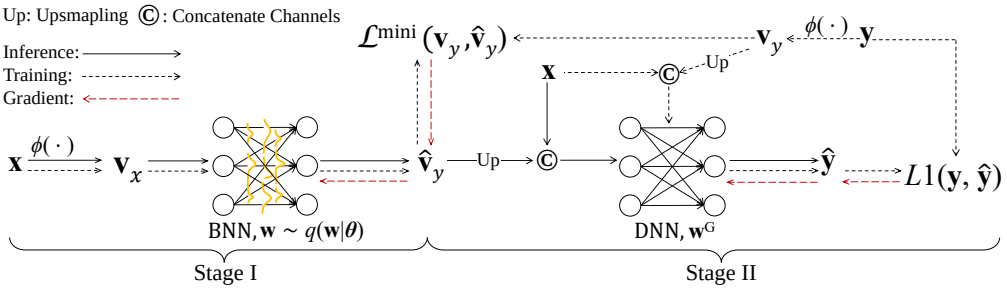


Figure 2: The two-stage pipeline. In Stage I, the BNN with weights $\mathbf{w} \sim q(\mathbf{w}|\boldsymbol{\theta})$ is trained by minimising the minibatch loss $\mathcal{L}^{\text{mini}}(\mathbf{v}_y, \hat{\mathbf{v}}_y)$ in Eq. (8). In Stage II, the DNN with weights $\mathbf{w}^{\text{G}}$ is trained by minimising the L1 loss, $L1(\mathbf{y}, \hat{\mathbf{y}})$. The inference process is denoted by $\rightarrow$, while the training process for each stage is indicated by $\dashrightarrow$. The gradient flow is shown with $\dashrightarrow$.

Figure 2 illustrates our proposed two-stage framework. We apply a reduction function ϕ to compress high-dimensional image data by either statistical summarisation or down-sampling, yielding compact representations $\mathbf{v}_x = \phi(\mathbf{x})$ and $\mathbf{v}_y = \phi(\mathbf{y})$ in a lower-dimensional space. In the first stage, the BEM

models the complex one-to-many mapping between $\mathbf{v}_x$ and $\mathbf{v}_y$. In the second stage, a DNN G refines the results by taking the first-stage low-dimensional output $\hat{\mathbf{v}}_y$ along with the original low-quality image $\mathbf{x}$ as inputs, producing a high-quality recovered image. The overall process is formulated as:

$$\hat{\mathbf{v}}_y = F(\phi(\mathbf{x}); \mathbf{w}), \quad \mathbf{w} \sim q(\mathbf{w} \mid \boldsymbol{\theta}), \tag{9}$$

$$\hat{\mathbf{y}} = G(\hat{\mathbf{v}}_y, \mathbf{x}; \mathbf{w}^{\mathrm{G}}), \tag{10}$$

where $\mathbf{w}^{\mathrm{G}}$ denotes the weights of the second-stage model. We explore two reduction functions: bilinear downsampling and local 2D histogram. Both methods are effective; however, bilinear downsampling provides higher measurement values on full-reference image quality assessment metrics. Additionally, considering bilinear downsampling offers a more efficient computation, we adopt it as the default setting. Further analysis of the reduction function ϕ is provided in Appendix A.

During the training phase of the second-stage model, we use the downsampled features of the target image $\mathbf{y}$ along with the low-quality image $\mathbf{x}$ as input to the DNN, instead of using the output from the first-stage model. This strategy removes constraints imposed by the first-stage model, thereby allowing the second stage to reach its full potential. Importantly, as illustrated in the inference flow in Figure 2, the inference process remains independent of the target image.

Backbone Network. For both the first and the second stage models, we adopt the same backbone network, but with different input and output layers. To enable weight uncertainty for the first stage model, we convert all the convolution and linear layers in the backbone network to their Bayesian counterparts, the weight parameters of which are obtained via Eq. (6). Inspired by Mamba (Gu & Dao, 2023) and VMamba (Liu et al., 2024b), featuring their linear computational complexity for long sequence modelling, we employ a Mamba as the backbone of our BEM. The overall framework is akin to a U-Net. We provide the details and experiment with the backbone in Appendix B.

4.2 Speeding up Inference

Similar to diffusion models, our BEM benefits from multiple inference passes to produce high-quality outputs. However, unlike the sequential denoising process of diffusion models, BEM allows parallel execution. We accelerate inference using two main strategies: I) Applying Algorithm 1 only to the first-stage model to generate a low-resolution output, $\mathbf{v}^{\mathrm{opt}}$. With a $16\times$ downsampling in function ϕ, this provides a theoretical $256\times$ speedup. II) Parallelising the K iterations along the batch dimension achieves a speedup proportional to the GPU's parallel computing capability. As illustrated in Figure 3, the accelerated inference speed for image resolutions of 512^2 and 1024^2, is in the same level of the single-pass inference. However, when the function D does not support parallel execution, the speed decreases proportionally to D's computational complexity. This acceleration strategy introduces a minor degradation in image quality: at $K = 100$, we observe an average drop of 3.2% in PSNR, while no decrease is noted in UIQM.

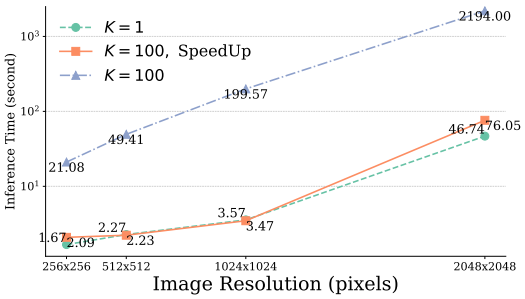


Figure 3: Inference speed before and after acceleration. A parallel implementation of D is employed. The model runs on an Nvidia RTX 4090.

5 Experiments

Datasets. We conduct experiments on several low-light image enhancement (LLIE) and underwater image enhancement (UIE) datasets. For LLIE, we evaluate our method on LOL-v1 (Wei et al., 2018) and LOL-v2 (real and synthetic subsets)(Yang et al., 2021), both of which have training and test splits, as well as the unpaired LIME(Guo et al., 2016), NPE (Wang et al., 2013), MEF (Ma et al., 2015), DICM (Lee et al., 2013), and VV (Vonikakis et al., 2018) datasets. For UIE, we use the UIEB (Li et al., 2019a), U45 (Li et al., 2019b), and UCCS (Liu et al., 2020) datasets. The UIEB dataset is further divided into training, validation (R90), and test (C60) subsets.

Metrics. For paired datasets, we evaluate pixel-level accuracy using PSNR and SSIM, and perceptual quality using LPIPS (Zhang et al., 2018). For real-world datasets, we use NIQE Mittal et al. (2012) as a no-reference metric. In UIE tasks, we additionally evaluate image quality using UIQM (Panetta et al., 2015) and UCIQE (Yang & Sowmya, 2015).

Settings. All models are trained with the Adam optimiser, starting at a learning rate of 2×10^{-4} and decaying to 10^{-6} using a cosine annealing schedule. The first-stage model is trained for 300K iterations on inputs reduced to a size of 24×24 through function ϕ, while the second-stage model is trained for 150K iterations on inputs of size 128×128. Batch size M is set to 8, and ϕ defaults to bilinear downsampling with a $\frac{1}{16}$ scaling factor. Unless stated otherwise, K is 100, D in Algorithm 1 is negative MSE, and σ^{o} in Eq. (7) is set to 0.05.

5.1 Quantitative Results

Table 1: Quantitative comparisons on the LOL-v1 and v2 datasets using PSNR, SSIM, and LPIPS. Models in grey adjust their output colour using the ground-truth mean (GT-Mean) value. For each section, the best results are in **bold**, and the second-best are underlined.

Method	GT Mean	LOL-v1			LOL-v2-real			LOL-v2-syn		
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
LLFlow (Wang et al., 2022)	✓	25.13	0.872	0.117	26.20	0.888	0.137	24.81	0.919	0.067
GlobalDiff (Hou et al., 2024)	✓	27.84	0.877	0.091	28.82	0.895	0.095	28.67	0.944	0.047
GLARE (Zhou et al., 2024)	✓	27.35	0.883	0.083	28.98	0.905	0.097	29.84	0.958	-
BEM-DNN (ours)	✓	28.30	0.881	0.072	31.41	0.912	0.064	30.58	0.958	0.033
BEM (ours)	✓	**28.80**	**0.884**	**0.069**	**32.66**	**0.915**	**0.060**	**32.95**	**0.964**	**0.026**
KinD (Zhang et al., 2019)	✗	19.66	0.820	0.156	18.06	0.825	0.151	17.41	0.806	0.255
Restormer (Zamir et al., 2022)	✗	22.43	0.823	0.147	18.60	0.789	0.232	21.41	0.830	0.144
SNR-Net (Xu et al., 2022)	✗	24.61	0.842	0.151	21.48	0.849	0.157	24.14	0.928	0.056
RetinexFormer Cai et al. (2023)	✗	25.16	0.845	0.131	22.80	0.840	0.171	25.67	0.930	0.059
RetinexMamba Bai et al. (2024)	✗	24.03	0.827	-	22.45	0.844	-	25.89	0.935	-
BEM (ours)	✗	**26.83**	**0.877**	**0.072**	**28.89**	**0.902**	**0.076**	**29.22**	**0.955**	**0.031**

Full-Reference Evaluation. For the LLIE tasks, we present quantitative comparisons with state-of-the-art methods on the LOL-v1 and LOL-v2 datasets, as detailed in Table 1. The table is divided into two sections: the first compares models that adjust their output colour using the ground-truth mean, while the second lists models that do not rely on this adjustment. Our BEM significantly outperforms all previous methods across all metrics. Notably, on LOL-v2-real, BEM achieves an exceptionally high PSNR of 28.89 dB, surpassing the second-best RetinexFormer by 6.09 dB. Although deterministic models are considered sub-optimal in the one-to-many mapping problem, our BEM-DNN (deterministic mode) still surpasses the previous methods across all benchmarks. We observed that previous methods often struggle to maintain high perceptual quality (measured by LPIPS) while ensuring pixel-level accuracy. However, our BEM excels in both, delivering the highest SSIM (0.877) and the lowest LPIPS (0.072). For the UIE tasks, we present quantitative comparisons

Table 2: Quantitative comparisons on the UIEB-R90, UIEB-C60, U45, and UCCS datasets in terms of PSNR, SSIM, UIQM, and UCIQE. Best results are in **bold**, second best are underlined.

Method	UIEB-R90		UIEB-C60		U45		UCCS	
	PSNR ↑	SSIM ↑	UIQM ↑	UCIQE ↑	UIQM ↑	UCIQE ↑	UIQM ↑	UCIQE ↑
WaterNet (Li et al., 2019a)	21.04	0.860	2.399	**0.591**	-	-	2.275	0.556
Ucolor (Li et al., 2021)	20.13	0.877	2.482	0.553	3.148	0.586	3.019	0.550
PUIE-MP (Fu et al., 2022)	21.05	0.854	2.524	0.561	3.169	0.569	2.758	0.489
Restormer (Zamir et al., 2022)	23.82	0.903	2.688	0.572	3.097	0.600	2.981	0.542
CECF (Cong et al., 2024)	21.82	0.894	-	-	-	-	-	-
FUnIEGAN (Islam et al., 2020)	19.12	0.832	2.867	0.556	2.495	0.545	3.095	0.529
PUGAN (Cong et al., 2023)	22.65	0.902	2.652	0.566	-	-	2.977	0.536
U-Shape (Peng et al., 2023)	20.39	0.803	2.730	0.560	3.151	0.592	-	-
Semi-UIR (Huang et al., 2023)	22.79	0.909	2.667	0.574	3.185	0.606	3.079	0.554
WFI2-Net (Zhao et al., 2024)	23.86	0.873	-	-	3.181	0.619	-	-
BEM (ours)	**25.62**	**0.940**	**2.931**	0.567	**3.406**	**0.620**	**3.224**	**0.561**

on the UIEB-R90 dataset, as shown in Table 2. Our BEM outperforms the second-best WFI2-Net by 1.76 dB in PSNR. This superior performance, observed consistently across both LLIE and UIE tasks, highlights the effectiveness and versatility of our method.

No-Reference Evaluation. For no-reference low-light images, we recover them using Algorithm 1 and D is instantiated as the NIQE metric. We then evaluate our method on five unpaired datasets as shown in Table 3, where we report the NIQE scores of SOTA methods. Our BEM consistently outperforms prior methods across all datasets. For enhancing no-reference underwater images, we instantiate D in Algorithm 1 as the UIQM and UCIQE metrics. We then evaluate our method on the C60, U45 and UCCS test sets. As shown in Table 2, BEM achieves the best UIQM scores across all test sets. With the UCIQE metric, we also achieve the best results in the U45 and UCCS test sets. These results, spanning different tasks and datasets, demonstrate the robustness and effectiveness of our method in real-world applications.

Table 3: No-reference evaluation on LIME, NPE, MEF, DICM and VV, in terms of NIQE↓. The best results are in **blodface**.

Method	DICM	LIME	MEF	NPE	VV
KinD (Zhang et al., 2019)	5.15	5.03	5.47	4.98	4.30
ZeroDCE (Guo et al., 2020)	4.58	5.82	4.93	4.53	4.81
RUAS (Liu et al., 2021)	5.21	4.26	3.83	5.53	4.29
LLFlow (Wang et al., 2022)	4.06	4.59	4.70	4.67	4.04
PairLIE (Fu et al., 2023b)	4.03	4.58	4.06	4.18	3.57
RFR (Fu et al., 2023a)	3.75	3.81	3.92	4.13	-
GLACE (Zhou et al., 2024)	3.61	4.52	3.66	4.19	-
CIDNet (Feng et al., 2024)	3.79	4.13	3.56	3.74	3.21
BEM (ours)	**3.55**	**3.56**	**3.14**	**3.72**	**2.91**

5.2 Visual Analysis

Predictions of One-to-Many. In Figure 4, we visualise the prediction process of BEM, where multiple plausible candidates are generated. As shown at the top of the figure, these candidates exhibit apparent visual differences. The best prediction candidate is selected using Algorithm 1, which is visually closer to the reference image. For no-reference prediction, we demonstrate that using the CLIP score with the text prompt, "`A bright photo`," results in the brightest image being outputted. By instantiating D as the NIQE metric, we can avoid generating overexposed predictions, as shown at the bottom right.

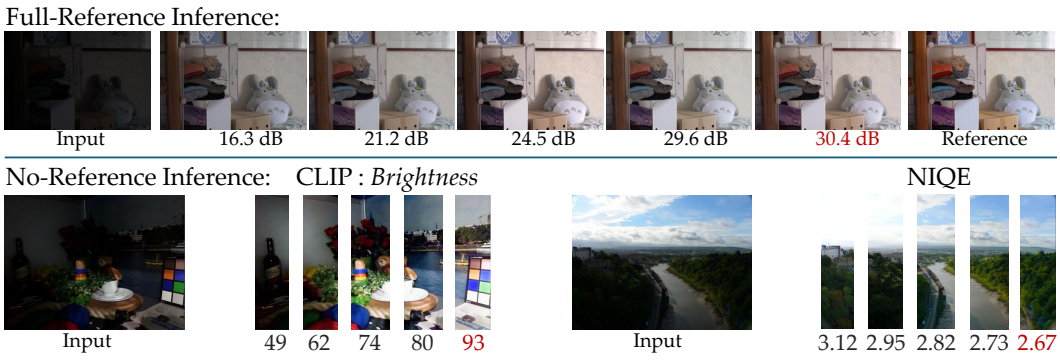


Figure 4: Visualisation of the predicting process of BEM in both cases with reference (top) and without reference (bottom). Zoom in for more details.

Qualitative Comparisons. We visually compare our BEM with twelve state-of-the-art UIE methods, including WaterNet (Li et al., 2019a), PRWNet (Huo et al., 2021), FUnIEGAN (Islam et al., 2020), PUGAN Cong et al. (2023), MMLE (Zhang et al., 2022), PUIE-MP (Fu et al., 2022), FiveA+(Jiang et al., 2023b), CLUIE (Li et al., 2023), Semi-UIR (Huang et al., 2023), UColor (Li et al., 2021), DM-Underwater (Tang et al., 2023), and CLIP-UIE (Liu et al., 2024a). As depicted in the first and second rows of Figure 5, our BEM achieves superior removal of underwater turbidity compared to other methods. In deeper ocean images with dominant blueish effects (last row in Figure 5), BEM can better enhance visual clarity. Visual comparisons on five unpaired LLIE test sets are shown in Figure 6, where our restored images offer better perceptual improvement. For example, in DICM, our method enhances brightness while effectively avoiding overexposure. These visual improvements align with the superior quantitative results presented in Sec. 5.1. HD visual results are included in Appendix E.

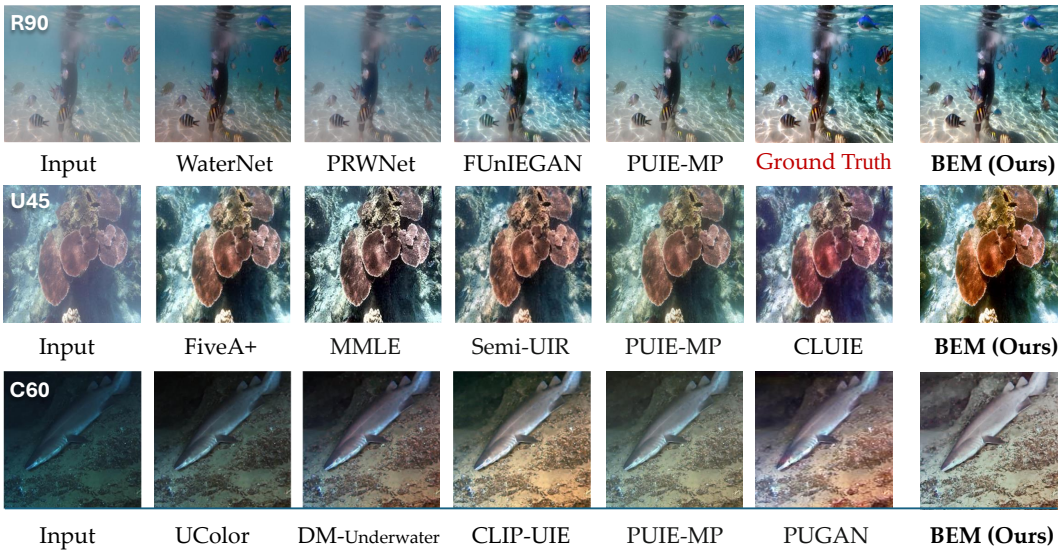


Figure 5: Visual comparisons on the R90, C60 and U45 datasets. Best viewed when zoomed in.

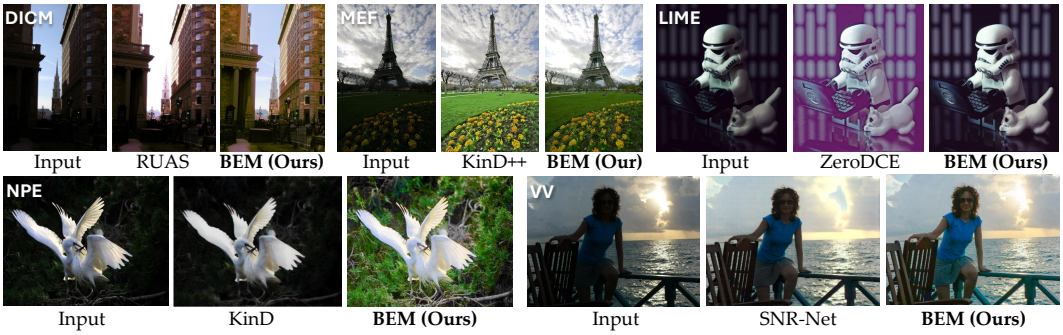


Figure 6: Visual comparisons on the DICM, LIME, MEF, NPE and VV datasets.

5.3 Ablation Studies

Single-Stage *vs*. Two-Stage Approaches. We assess the performance of our two-stage approach by comparing it against a single-stage variant. As discussed in Sec. 4, directly converting a DNN into a BNN typically results in noisy predictions. To generate smooth outputs, our single-stage model retains the last layer in the network as a deterministic layer, the entire process of which is opposite to the Bayesian last layer method (Harrison et al., 2024). While the two-stage approach introduces only marginal additional computational overhead, its performance significantly surpasses that of the single-stage model, as shown in Table 4. This highlights the efficiency and effectiveness of our two-stage approach.

Table 4: Single-stage *vs*. two-stage approaches on LOL-v1. FLOPs are calculated in an input size of 256×256 pixels.

Model	FLOPs	PSNR ↑	SSIM ↑
Single Stage	20.41G	24.78	0.852
Two Stages	20.49G	26.83	0.877

Magnitude of Uncertainty. The performance improvements of our BEM primarily stem from its ability to effectively model the one-to-many mapping using BNNs. To support this claim, we evaluate the influence of the variance in the variational posterior on model performance. As shown in Figure 7, except for BEM with $\sigma^{\circ} = 0.0001$, all other BEM instances outperform the DNN. This indicates that by setting a moderate variance in the momentum prior, BEM can significantly surpass its DNN counterpart.

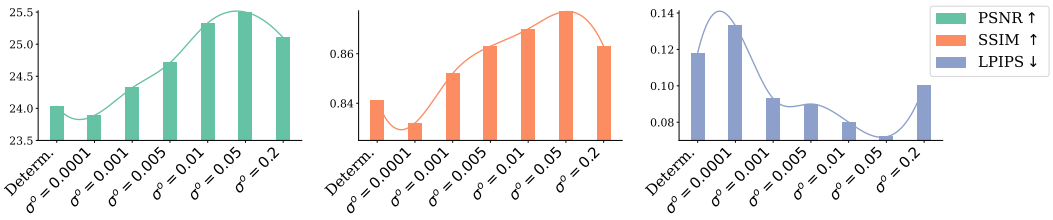


Figure 7: Effect of initial variance values (*i.e.*, σ^{o} in Eq. 7) on model performance. The results are obtained by evaluating single-stage models on the LOL-v1 dataset. "Determ." denotes the deterministic baseline model.

Impact of Different Priors. We evaluate the effectiveness of our momentum prior against two baseline priors: a naive Gaussian prior and an empirical Bayes prior. The naive Gaussian prior is defined as $P(\mathbf{W}) = \mathcal{N}(\mathbf{0}, 0.1\mathbf{I})$. The empirical Bayes prior, MOPED (Krishnan et al., 2020), is defined as $P(\mathbf{W}) = \mathcal{N}(\mathbf{w}^{\text{MLE}}, 0.1\mathbf{I})$, where $\mathbf{w}^{\text{MLE}}$ represents the maximum likelihood estimate (MLE) of the weights learned by optimising a deterministic network. In the case of the empirical Bayes prior, the mean $\boldsymbol{\mu}$ of the variational posterior $q(\mathbf{w}|\boldsymbol{\theta})$ is initialised as the MLE of the weights, $\mathbf{w}^{\text{MLE}}$, and the posterior variance $\boldsymbol{\sigma}$ is set to $0.1\mathbf{w}^{\text{MLE}}$, as suggested by Krishnan et al. (2020). As shown in Figure 8, the momentum prior demonstrates a clear advantage over both baselines, providing faster convergence and superior performance. While the empirical Bayes prior accelerates training during early iterations, its performance degrades over time due to the fixed nature of the prior. The fixed prior, learned from the same data, can act as a shortcut during the optimisation of the variational posterior parameters, minimising the loss function in Eq. (5) predominantly by reducing the prior matching term $\text{KL}\left[q(\mathbf{w}|\boldsymbol{\theta})\|P(\mathbf{w})\right]$. This behaviour bypasses data-driven learning, ultimately resulting in sub-optimal solutions that do not fully capture the data's inherent uncertainty.

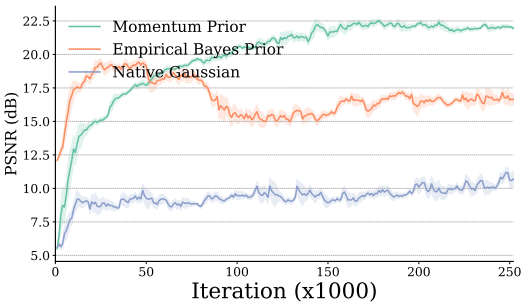


Figure 8: Training curves of one-stage BEMs with different priors. The PSNR for each iteration is calculated using the mean weight $\boldsymbol{\mu}$.

6 Discussion and Conclusion

Although BEM demonstrates stronger generalisation capability than DNN-based methods, fully realising its potential will require intentionally collecting target images under diverse capture settings to further increase label diversity. While using small image crops as training data can alleviate the label diversity problem to some extent, similar to conventional data augmentation strategies in DNNs, this approach has limitations. We leave these aspects for future work. Additionally, the distinction between image enhancement and image restoration is not always well-defined, as some restoration tasks (e.g., image colourisation and de-raining) may also present one-to-many mapping challenges. Consequently, our BEM could be extended to certain image restoration scenarios.

Overall, we identified the one-to-many mapping problem as a key limitation in existing image enhancement tasks and introduced the first Bayesian-based model to address this issue. To facilitate efficient training on high-dimensional data, we proposed a *Momentum Prior* that dynamically refines the prior distribution during training, enhancing convergence and performance. Our two-stage framework integrates the strengths of BNNs and DNNs, yielding a flexible yet computationally efficient model. Extensive experiments on various image enhancement benchmarks demonstrate significant performance gains over state-of-the-art models, showcasing the potential of Bayesian probabilistic models in handling the inherent ambiguities of image enhancement tasks, paving the way for future research in modelling complex one-to-many mappings in low-level vision tasks.

REFERENCES

Nantheera Anantrasirichai and David Bull. Contextual colorization and denoising for low-light ultra high resolution sequences. In *2021 IEEE International Conference on Image Processing (ICIP)*, pp. 1614–1618. IEEE, 2021.

Jiesong Bai, Yuhao Yin, and Qiyuan He. Retinexmamba: Retinex-based mamba for low-light image enhancement. *arXiv preprint arXiv:2405.03349*, 2024.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pp. 1613–1622. PMLR, 2015.

Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12504–12513, 2023.

Runmin Cong, Wenyu Yang, Wei Zhang, Chongyi Li, Chun-Le Guo, Qingming Huang, and Sam Kwong. Pugan: Physical model-guided underwater image enhancement using gan with dual-discriminators. *IEEE Transactions on Image Processing*, 32:4472–4485, 2023.

Xiaofeng Cong, Jie Gui, and Junming Hou. Underwater organism color fine-tuning via decomposition and guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1389–1398, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Cameron Fabbri, Md Jahidul Islam, and Junaed Sattar. Enhancing underwater imagery using generative adversarial networks. In *2018 IEEE international conference on robotics and automation (ICRA)*, pp. 7159–7165. IEEE, 2018.

Yixu Feng, Cheng Zhang, Pei Wang, Peng Wu, Qingsen Yan, and Yanning Zhang. You only need one color space: An efficient network for low-light image enhancement. *arXiv preprint arXiv:2402.05809*, 2024.

Huiyuan Fu, Wenkai Zheng, Xiangyu Meng, Xin Wang, Chuanming Wang, and Huadong Ma. You do not need additional priors or regularizers in retinex-based low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18125–18134, 2023a.

Zhenqi Fu, Wu Wang, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Uncertainty inspired underwater image enhancement. In *European conference on computer vision*, pp. 465–482. Springer, 2022.

Zhenqi Fu, Yan Yang, Xiaotong Tu, Yue Huang, Xinghao Ding, and Kai-Kuang Ma. Learning a simple low-light image enhancer from paired low-light instances. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22252–22261, 2023b.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1780–1789, 2020.

Xiaojie Guo, Yu Li, and Haibin Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Transactions on image processing*, 26(2):982–993, 2016.

James Harrison, John Willes, and Jasper Snoek. Variational bayesian last layers. In *International Conference on Learning Representations (ICLR)*, 2024.

Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pp. 5–13, 1993.

Jinhui Hou, Zhiyu Zhu, Junhui Hou, Hui Liu, Huanqiang Zeng, and Hui Yuan. Global structure-aware diffusion process for low-light image enhancement. *Advances in Neural Information Processing Systems*, 36, 2024.

Shirui Huang, Keyan Wang, Huan Liu, Jun Chen, and Yunsong Li. Contrastive semi-supervised learning for underwater image restoration via reliable bank. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18145–18155, 2023.

Fushuo Huo, Bingheng Li, and Xuegui Zhu. Efficient wavelet boost learning-based multi-stage progressive refinement network for underwater image enhancement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1944–1952, 2021.

Md Jahidul Islam, Youya Xia, and Junaed Sattar. Fast underwater image enhancement for improved visual perception. *IEEE Robotics and Automation Letters*, 5(2):3227–3234, 2020.

Hai Jiang et al. Low-light image enhancement with wavelet-based diffusion models. *ACM Transactions on Graphics (TOG)*, 42(6):1–14, 2023a.

Jingxia Jiang, Tian Ye, Jinbin Bai, Sixiang Chen, Wenhao Chai, Shi Jun, Yun Liu, and Erkang Chen. Five a^+ network: You only need 9k parameters for underwater image enhancement. *British Machine Vision Conference (BMVC)*, 2023b.

Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE transactions on image processing*, 30:2340–2349, 2021.

Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *2016 IEEE international conference on Robotics and Automation (ICRA)*, pp. 4762–4769. IEEE, 2016.

Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. *arXiv preprint arXiv:1511.02680*, 2015.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018.

Diederik P Kingma. Auto-encoding variational bayes. *International Conference on Learning Representations (ICLR)*, 2014. URL `http://arxiv.org/abs/1312.6114`.

Ranganath Krishnan, Mahesh Subedar, and Omesh Tickoo. Specifying weight priors in bayesian deep neural networks with empirical bayes. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 4477–4484, 2020.

Chulwoo Lee, Chul Lee, and Chang-Su Kim. Contrast enhancement based on layered difference representation of 2d histograms. *IEEE transactions on image processing*, 22(12):5372–5384, 2013.

Chongyi Li, Chunle Guo, Wenqi Ren, Runmin Cong, Junhui Hou, Sam Kwong, and Dacheng Tao. An underwater image enhancement benchmark dataset and beyond. *IEEE transactions on image processing*, 29:4376–4389, 2019a.

Chongyi Li, Saeed Anwar, Junhui Hou, Runmin Cong, Chunle Guo, and Wenqi Ren. Underwater image enhancement via medium transmission-guided multi-color space embedding. *IEEE Transactions on Image Processing*, 30:4985–5000, 2021.

Hanyu Li, Jingjing Li, and Wei Wang. A fusion adversarial underwater image enhancement network with a public test dataset. *arXiv preprint arXiv:1906.06819*, 2019b.

Kunqian Li, Li Wu, Qi Qi, Wenjie Liu, Xiang Gao, Liqin Zhou, and Dalei Song. Beyond single reference for training: Underwater image enhancement via comparative learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(6):2561–2576, 2023.

Risheng Liu, Xin Fan, Ming Zhu, Minjun Hou, and Zhongxuan Luo. Real-world underwater enhancement: Challenges, benchmarks, and solutions under natural light. *IEEE transactions on circuits and systems for video technology*, 30(12):4861–4875, 2020.

Risheng Liu, Long Ma, Jiaao Zhang, Xin Fan, and Zhongxuan Luo. Retinex-inspired unrolling with cooperative prior architecture search for low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10561–10570, 2021.

Shuaixin Liu, Kunqian Li, and Yilin Ding. Underwater image enhancement by diffusion model with customized clip-classifier. *arXiv preprint arXiv:2405.16214*, 2024a.

Yue Liu et al. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024b.

Kede Ma, Kai Zeng, and Zhou Wang. Perceptual quality assessment for multi-exposure image fusion. *IEEE Transactions on Image Processing*, 24(11):3345–3356, 2015.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Karen Panetta, Chen Gao, and Sos Agaian. Human-visual-system-inspired underwater image quality measures. *IEEE Journal of Oceanic Engineering*, 41(3):541–551, 2015.

Lintao Peng, Chunli Zhu, and Liheng Bian. U-shape transformer for underwater image enhancement. *IEEE Transactions on Image Processing*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Ali M Reza. Realization of the contrast limited adaptive histogram equalization (clahe) for real-time image enhancement. *Journal of VLSI signal processing systems for signal, image and video technology*, 38:35–44, 2004.

Herbert Robbins. An empirical bayes approach to statistics. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1:157–163, 1956.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1874–1883, 2016.

Yi Tang, Hiroshi Kawasaki, and Takafumi Iwaguchi. Underwater image enhancement by transformer-based diffusion model with non-uniform sampling for skip strategy. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 5419–5427, 2023.

Vassilios Vonikakis, Rigas Kouskouridas, and Antonios Gasteratos. On the evaluation of illumination compensation algorithms. *Multimedia Tools and Applications*, 77:9211–9231, 2018.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023.

Shuhang Wang, Jin Zheng, Hai-Miao Hu, and Bo Li. Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE transactions on image processing*, 22(9):3538–3548, 2013.

Yufei Wang, Renjie Wan, Wenhan Yang, Haoliang Li, Lap-Pui Chau, and Alex Kot. Low-light image enhancement with normalizing flow. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2604–2612, 2022.

Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *British Machine Vision Conference (BMVC)*, 2018.

Xiaogang Xu, Ruixing Wang, Chi-Wing Fu, and Jiaya Jia. Snr-aware low-light image enhancement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17714–17724, 2022.

Miao Yang and Arcot Sowmya. An underwater color image quality evaluation metric. *IEEE Transactions on Image Processing*, 24(12):6062–6071, 2015.

Wenhan Yang, Wenjing Wang, Haofeng Huang, Shiqi Wang, and Jiaying Liu. Sparse gradient regularized deep retinex network for robust low-light image enhancement. *IEEE Transactions on Image Processing*, 30:2072–2086, 2021.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5728–5739, 2022.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Weidong Zhang, Peixian Zhuang, Hai-Han Sun, Guohou Li, Sam Kwong, and Chongyi Li. Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement. *IEEE Transactions on Image Processing*, 31:3997–4010, 2022.

Yonghua Zhang, Jiawan Zhang, and Xiaojie Guo. Kindling the darkness: A practical low-light image enhancer. In *Proceedings of the 27th ACM international conference on multimedia*, pp. 1632–1640, 2019.

Chen Zhao, Weiling Cai, Chenyu Dong, and Chengwei Hu. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8281–8291, 2024.

Han Zhou, Wei Dong, Xiaohong Liu, Shuaicheng Liu, Xiongkuo Min, Guangtao Zhai, and Jun Chen. Glare: Low light image enhancement via generative latent feature based codebook retrieval. *Proceedings of the European conference on computer vision (ECCV)*, 2024.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

A EXPERIMENTS ON REDUCTION FUNCTION ϕ

Regarding the form of reduction function ϕ in Eq. (9). we consider two instantiations: bilinear downsampling and local 2D histogram. As illustrated in Figure 9. with the local histogram, the recovered images preserve more details than that of bilinear downsampling, due to additional configuration for the histogram's bin number, avoiding losing much information when the downsample scale is larger.

(a) Bilinear Down (b) Local Histogram

Figure 9: With the same downsampling scale, the local histogram offers more precise control over the amount of retained information by adjusting the number of bins (corresponding to the number of channels). In contrast, bilinear downsampling tends to lose excessive details, especially when using larger downsampling strides.

The discrete nature of histograms poses challenges in both prediction accuracy and computational speed. To address this, we approximate the histogram calculation using Kernel Density Estimation (KDE), which significantly improves both computation efficiency and prediction accuracy. As shown in Table 5, while the pixel-level PSNR of local histogram-based ϕ is slightly lower than that of bilinear downsampling, we attribute this to the larger variance inherent in histogram values, which the model struggles to fit effectively.

Table 5: Comparisons of different instantiations of ϕ. The PSNR values on LOL-v1 are reported. K is set to 100.

Function ϕ	Down Scale	Bins	Channels	PSNR ↑
Bilinear Down	8	N/A	3	25.87
Local Histogram	8	3	9	25.29
Local Histogram	8	10	30	24.96
Local Histogram	8	16	48	24.80
Bilinear Down	16	N/A	3	26.83
Local Histogram	16	10	30	25.89
Local Histogram	16	16	48	25.83

Despite this, we observe that the local histogram approach exhibits slightly better colour representation compared to the bilinear instance. In Figure 10, we present a visual comparison between the two implementations, highlighting that the histogram-based model generates more vivid colours. However, the bilinear downsampling method performs better in restoring details in areas where significant information loss occurs.

B INVESTIGATION ON MAMBA BACKBONE

Considering Mamba's linear computational complexity for long sequence modelling, we adopt the VMmaba Liu et al. (2024b) to build the backbone of our BEM. The overall framework is akin to a U-Net, but we replace all the Transformer blocks Dosovitskiy et al. (2020) with the Visual State-Space (VSS) blocks, each of which is composed of a 2D Selective Scan (SS2D) module Liu et al. (2024b) and a feedforward network (FFN). The formulation of VSS block Liu et al. (2024b) in layer l can be expressed as

$$\begin{aligned} \mathbf{h}_l &= \text{SS2D}\left(\text{LN}\left(\mathbf{h}_{l-1}\right)\right) + \mathbf{h}_{l-1}, \\ \mathbf{h}_{l+1} &= \text{FFN}\left(\text{LN}\left(\mathbf{h}_l\right)\right) + \mathbf{h}_l, \end{aligned} \tag{11}$$

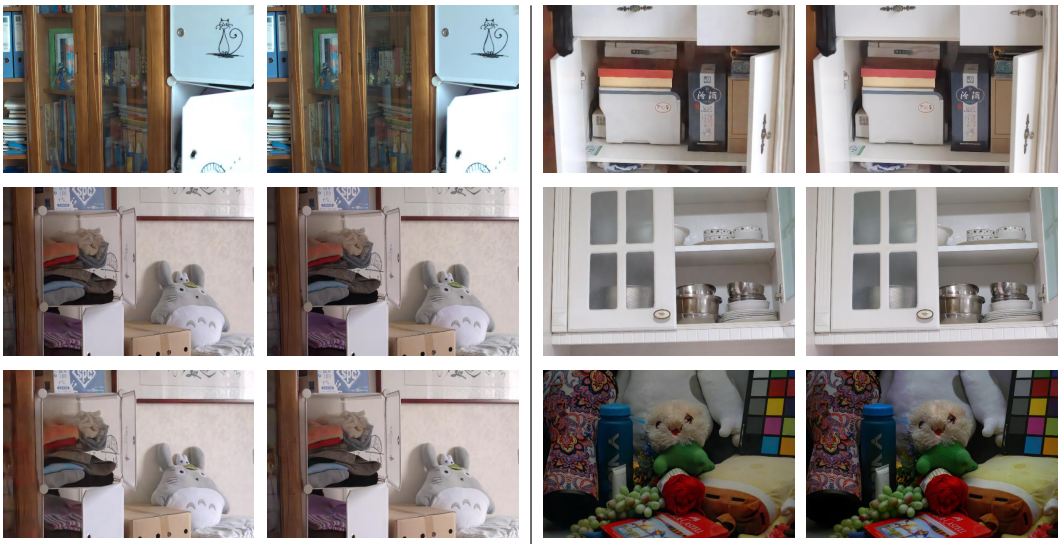


Figure 10: Visual comparison between the local histogram and bilinear downsampling implementations of the reduction function ϕ. The bilinear ϕ demonstrates better restoration capability compared to the histogram-based counterpart. However, the histogram-based ϕ shows better global colour representation. Best viewed when zoomed in.

where FFN denotes the feedforward network and LN denotes layer normalisation. $\mathbf{h}_{l-1}$ and $\mathbf{h}_l$ denote the input and output in the l-th layer, respectively. Our network backbone consists of an input convolutional layer, 12 VSS blocks, and an output layer. The first 6 VSS blocks form the encoder of a U-Net, where the spatial dimensions of the feature maps are halved every two blocks, while the number of channels is doubled. Specifically, given an input image with a shape of $H \times W \times 3$, the encoding blocks obtain hierarchical feature maps of sizes $H \times W \times C$, $\frac{H}{2} \times \frac{W}{2} \times 2C$ and $\frac{H}{4} \times \frac{W}{4} \times 4C$. The remaining 6 VSS blocks constitute the decoder, upsampling these encoding feature maps hierarchically with the `pixelshuffle` layers (Shi et al., 2016). At each scale level, lateral connections are built to link corresponding blocks in the encoder and decoder.

Construct the backbone. We build our backbone by gradually evaluating each configuration of a vanilla Mamaba-based UNet. We thoroughly investigate settings including `ssm-ratio`, block numbers, `n_feat` and `mlp-ratio`. The training strategies for all variants are identical. Setting `n_feat` denotes the number of feature maps in the first `conv3×3`'s output. Setting `d_state` denotes the state dimension of SSM. Note that the established baseline assures two things: 1) Further naively introducing additional parameters and FLOPs, e.g., scaling models with more blocks, will not help boost the performance. 2) A technique with additional parameters introduced to the baseline model can no doubt demonstrate its effectiveness if the modified model shows better results than the baseline.

To balance both speed and performance, we selected the model in the second row of Table 6 as the backbone for our BEM. The chosen backbone features a simple architecture with no task-specific modules, enhancing its generalisability and establishing a solid foundation for extending our method to other types of vision tasks.

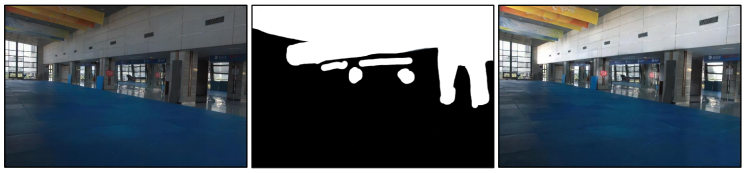
C Controllable Local Enhancement

Thanks to the interpretability of the lower-dimensional representations in both the spatial and channel dimensions, we can easily achieve local adjustment with a masking strategy. The local adjustment is particularly useful in the cases where the input images are unevenly distorted, and we want to retain the undistorted regions consistent before and after enhancement. The local adjustment process can be achieved by using a mask layer $\mathbf{M}$: $\mathbf{y}^{\text{local}} = G(\gamma\mathbf{M} \odot \mathbf{v}, \mathbf{x}; \mathbf{w}^{\text{G}})$, where $\mathbf{v}$ can be lower-dimensional

Table 6: The performance of deterministic Mamba UNet variants with different `d_state`, `ssm-ratio`, `mlp-ratio`, `n_feat` and `block numbers`. PSNR and SSIM on LOL-v1 are reported. Since the deterministic networks trained using minibatch optimisation are likely to fit very different targets each time, the results will fluctuate greatly. We train each model five times and report the average performance.

d_state	ssm-ratio	mlp-ratio	n_feat	block numbers	FLOPs (G)	Params (M)	TP img/s	PSNR (dB)	SSIM
1	1	2.66	40	[2,2,2]	14.25	1.23	125	22.45	0.828
1	1	4	40	[2,2,2]	20.41	1.52	78	23.76	0.842
16	1	2.66	40	[2,2,2]	25.50	1.37	84	23.83	0.840
32	1	2.66	40	[2,2,2]	37.49	1.52	61	21.93	0.812
16	2	4	40	[2,2,2]	44.36	2.08	58	23.67	0.830
16	2	4	52	[2,2,2]	65.10	3.37	40	23.21	0.833
16	2	4	40	[2,2,2,2]	54.82	7.77	51	23.44	0.838
1	2	4	40	[2,2,2]	21.87	1.79	82	22.73	0.834

features extracted from a real image or estimated by the first stage model via Eq. (9). We can use a scalar γ to control the strength of the enhancement effect. A demonstration of the local enchantment is shown in Figure 11.



Before Mask Layer After

Figure 11: The local brightness of an image before adjustment (left) can be edited locally by providing a mask layer (middle). The image after adjustment (right) shows improved brightness in the regions indicated by the mask.

Compared to directly applying the mask to the output, our local enhancement strategy not only reduces the dependency on mask accuracy but also results in smoother transitions at the mask boundaries. This mitigates issues such as excessive roughness or colour inconsistencies between processed and unprocessed regions.

D Label Diversity Augmentation

Theoretically, an infinite number of target images could correspond to a single input. However, current paired datasets often lack sufficient label diversity, which may become a bottleneck for BEM model performance.

Table 7: Evaluation of label augmentation strategies for enhancing label diversity. PSNR scores are obtained using single-stage models on LOL-v1.

Model	Gamma Correction	Saturation Shift	CLAHE	PSNR ↑
BEM				24.78
BEM	✓			24.89
BEM	✓	✓		24.93
BEM	✓	✓	✓	24.86
DNN				24.02
DNN	✓	✓	✓	21.58

Without relying on additional data collection to increase label diversity, we propose two strategies for augmenting label diversity within existing datasets:

i) When training a deep network, high-resolution images are often divided into smaller crops (e.g., 128×128). Many of these smaller image crops may represent the same scene, but due to various factors, such as being captured at different moments in a video or having different capture settings, the corresponding target crops show differences in colour or brightness. Thus, using these crops as input during training, the actual label diversity within the training data is naturally increased.

ii) Existing labels can be further enriched by applying data augmentation techniques such as random brightness adjustments, saturation shifts, changes in colour temperature, gamma corrections, and histogram equalisation.

Both strategies contribute to increasing label diversity to some extent. In Table 7, we evaluate whether expanding the number of target images using gamma correction, saturation shift, and CLAHE Reza (2004) can further improve the model's performance. Among these, saturation shift is a linear transformation, while gamma correction and CLAHE are nonlinear methods. We observed that deterministic networks showed a decline in performance after applying these label augmentation techniques. This can be attributed to DNNs overfitting to local solutions that deviate further from the inference image as uncertainty in the data increases. In contrast, BEM exhibited a slight increase in PSNR when using these augmented labels. For consistency, these augmentation strategies were not applied in other experiments.

E Supplementary Visualisations

HD Visualisation for LLIE. To facilitate a closer inspection of enhanced image details, we present high-resolution visual comparisons in Figure 12, where the predictions of state-of-the-art models are displayed at their original resolutions. The high-resolution visualisation reveals that previous state-of-the-art methods tend to exhibit varying degrees of noise artefacts in the enhanced results, significantly degrading perceptual quality. In contrast, our method effectively suppresses these noise

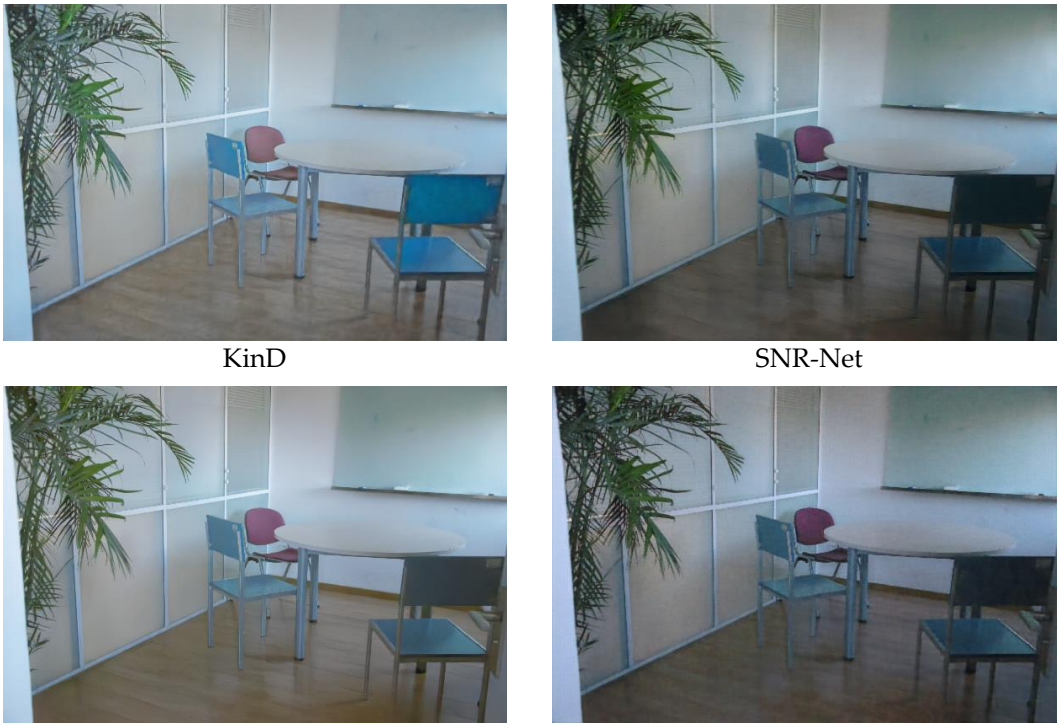


Figure 12: Visual comparisons with KinD, SNR-Net and RetinexFormer under images' original resolution. The sample is from the LOL-v2-real dataset.

artefacts, which are often introduced by low-light conditions. Furthermore, our approach achieves superior detail restoration, while other methods show signs of blurring and detail loss.

More Visualisations for UIE. In Figure 13, we present additional visual comparisons on the U45 and UCCS datasets, demonstrating that our method consistently outperforms PUGAN and PUIE-MP in enhancing various underwater scenes.

Figure 13: Visual comparisons with PUGAN and PUIE-MP on the U45 and UCCS test sets.