

MASTER'S THESIS 2026

Exploiting Infrared Images for Hallucination Resistant Low-Light Image Enhancement

Samuel Naumann, Karl Hammar

Elektroteknik
Datateknik

ISSN 1650-2884

LU-CS-EX: 2026-19

DEPARTMENT OF COMPUTER SCIENCE

LTH | LUND UNIVERSITY



EXAMENSARBETE
Datavetenskap

LU-CS-EX: 2026-19

**Exploiting Infrared Images for
Hallucination Resistant Low-Light Image
Enhancement**

Hallucinationsresistent bildförbättring i
svagt ljus med hjälp av infraröda bilder

Samuel Naumann, Karl Hammar

Exploiting Infrared Images for Hallucination Resistant Low-Light Image Enhancement

Samuel Naumann
sa3266na-s@student.lu.se

Karl Hammar
ka4527ha-s@student.lu.se

June 1, 2026

Master's thesis work carried out at Saab.

Supervisors: Mårten Lager, Saab
Simon Kristoffersson Lind, simon.kristoffersson_lind@cs.lth.se

Examiner: Eren Aksoy, eren.aksoy@cs.lth.se

Abstract

Images captured in maritime environments often suffer from a low signal-to-noise ratio due to a lack of ambient light at night, which poses a challenge for autonomous systems relying on camera sensors. While precise object identification is vital for downstream tasks like navigation, the constant motion of sensors and subjects makes acquiring paired datasets for supervised learning unfeasible. Current unsupervised methods often fail to accurately recreate structural clarity necessary for reliable maritime perception. To address these limitations, this thesis proposes a semi-supervised enhancement framework that utilizes a novel structural loss function informed by infrared imagery. By applying this loss during training, the model learns to enhance low-light images using only RGB input at inference. Qualitative analysis indicates that this approach achieves superior structural recovery compared to existing unsupervised baselines, suggesting that IR-informed training is a viable solution for data-scarce environments.

Keywords: image enhancement, hallucination resistance, deep learning, IR supervision, autonomous systems

Acknowledgements

We would like to express our sincerest gratitude to Simon Kristoffersson Lind who through his guidance, support and continuous feedback helped elevate this thesis to its full potential. Through our regular meetings you kept us on track and on time.

We would also like to thank Mårten Lager, Johan Modin and the entire team at Saab. By being open to all our ideas and providing us with the resources and expertise to follow through on them, you created an inspiring environment to work in.

Lastly, we would like to thank Eren Aksoy for assuming the role of examiner for this thesis and providing feedback on the report.

Contents

1	Introduction	7
2	Related work	9
2.1	Low-Light Image Enhancement	9
2.2	Multi-modal IR-guided Enhancement and Fusion Models	10
2.3	Hallucinations	11
3	Background	13
3.1	Infrared Imaging and LWIR Sensors	13
3.2	Histogram Equalization	14
3.3	Zero-DCE	15
3.3.1	Light Enhancement Curves (LE-curves)	15
3.3.2	DCE-Net	16
3.3.3	Losses	16
3.4	U-Net	17
3.5	Attention Mechanisms	18
3.5.1	Squeeze-and-Excitation Attention	18
3.5.2	Coordinate Attention	19
3.6	DA-Zero-DCE	20
3.7	Retinex Theory	21
3.8	Inverse Problems	21
3.9	Hallucination	22
3.10	RAW Images	22
3.11	Sobel Filter	23
4	Method	25
4.1	System Overview	25
4.2	Dataset and Preprocessing	26
4.3	Network Architecture	27
4.3.1	U-Net Curve Estimator Backbone	28

4.3.2	Coordinate Attention	28
4.3.3	IR-based Enhancer	29
4.4	Loss Functions	30
4.4.1	Edge-Enhancing Structural Loss	30
4.4.2	Vertical Total Variation Loss	32
4.4.3	Total Loss	33
4.5	Baselines	33
4.6	Experiments	35
5	Results & Discussion	37
5.1	Evaluation Metrics	37
5.2	Ablation Study	37
5.3	Comparison with Current Methods	38
5.4	LE-curves & IRE Residual	41
5.5	Hallucination Resistance	42
5.6	Analyzing Histogram Equalization Performance	45
5.7	Limitations	45
6	Future Work	49
6.1	Improved Alignment	49
6.2	Quantitative Evaluation	50
6.3	Train on Larger Datasets	51
6.4	Reduce Noise in Edge Maps	51
6.5	Alternative architectures	51
7	Conclusion	53
	References	55

Chapter 1

Introduction

In recent years data driven deep learning methods have established themselves as the state-of-the-art for various image enhancement tasks, fundamentally shifting how computer vision systems process information. While these models have achieved great success in well-lit scenes, low-light conditions remain a challenge due to the physical limitations of cameras. Short-exposure images captured in poorly lit environments suffer from low Signal-to-Noise Ratio (SNR) due to the inherently low number of photons reaching the camera sensor. By capturing images at higher exposures brightness can be improved, but obtaining such data in domains with non-static scenes is difficult due to long exposures resulting in motion blur [6]. Traditional methods therefore often struggle to differentiate the true signal, corresponding to physical objects in the scene, from sensor noise in extremely low-light environments. Furthermore, aggressive neural network-based enhancement might lead to the generation of non-existent details and structures not present in the physical scene [27], often referred to as hallucinations [8], especially in regions with very limited RGB information.

These physical limitations pose a major challenge for many current methods which often rely on fully supervised learning, necessitating large datasets of paired low-light and well-lit ground truth images [6, 30]. Acquiring pixel-aligned ground truth data in dynamic scenes is very difficult due to the challenges of long-exposure image capture. To address this some research, especially within medical imaging, has explored utilizing Long-Wave Infrared (LWIR) cameras in multi-modal fusion models [44]. As opposed to RGB sensors, LWIR cameras are illumination-independent and can thus capture image structure in low exposure and low illumination scenarios. Such cameras are however significantly more expensive and the captured data does not have as rich information as cameras operating in the visual spectrum. Moreover, fusion models often require IR images to be available as input during inference, and usually still rely on ground truth RGB images during training [36]. This makes them susceptible to similar data acquisition problems as the traditional models.

Saab Kockums is a leader within computer vision and autonomous systems for naval defense and security. The findings of this thesis are especially relevant in such maritime environments, where data collection is complicated as both subjects and the camera sensors

themselves are moving, motivating a semi-supervised approach. Furthermore, maritime environments often contain little to no ambient lighting at night, resulting in images with very low SNR and making a model which can reliably differentiate true signal from noise extremely valuable. Autonomous systems are heavily reliant on correctly identifying structures and objects in their surrounding to achieve robustness and safety on tasks such as navigation, object detection and decision making in general.

In this thesis we aim to address the challenges of extreme low-light image enhancement in dynamic environments by proposing a semi-supervised training framework, *IR-based Enhancer* (IRE). By utilizing IR imagery as privileged information during training, we intend to guide the structural enhancement and denoising of RGB images without relying on well-lit ground truth datasets.

Research Questions

- RQ1:** To what extent can IR images be used as privileged information during training to guide high-frequency enhancement of structure and contrast?
- RQ2:** How can an architecture be designed to provide both global illumination enhancement and local structural refinement in extreme low-light scenarios?
- RQ3:** Can loss functions and training constraints be formulated to punish hallucinated structure, while still encouraging strong contrast and detail recovery?

Chapter 2

Related work

Computer vision, like many other fields, has undergone a transformation in the past decade from classical signal processing algorithms to data-driven deep learning methods. This shift has fundamentally altered how low-light image enhancement is approached, moving from pixel wise transformations and fixed statistical models to learnable representations that can extract deeper features in latent space by training on large datasets [40]. However, while many deep learning models have achieved great success in visibility restoration tasks, they are limited by the reliance on paired data and struggle to differentiate signal from noise in scenes with little to no ambient light, leading to hallucination of nonexistent structure and texture.

In this chapter we review existing literature across three main areas to contextualize our work. First, we explore how low-light image enhancement has previously been approached, from traditional methods to both supervised and unsupervised deep learning models. We then explore how thermal imaging has been utilized to supplement RGB data in fusion architectures. Lastly, we discuss the phenomenon of hallucination in generative models, existing ways to quantify them, and suppression techniques.

2.1 Low-Light Image Enhancement

Early approaches to low-light image enhancement relied heavily on statistical and physical models, most notably Histogram Equalization (HE) and Retinex theory [37, 39]. HE usually operates on a pixel-level or patch-level basis and balances the histogram of the entire image by redistributing the pixels uniformly. While fast and mathematically interpretable, HE and its variations struggle on high dynamic range images with varying lighting and very dark regions, often introducing severe noise and washing out details [10]. Alternatively, Retinex theory assumes that images can be decomposed into illumination and reflectance components. This decomposition has inspired both traditional algorithms [39] and deep learning architectures [45] to enhance the illuminance of the image while retaining its intrinsic physical

qualities (reflectance). However, both traditional and deep-learning Retinex-based methods frequently fail in low SNR scenarios, such as low-light images, as heavy noise corrupts the reflectance estimation [35, 45].

The advent of Convolution Neural Networks (CNNs) and later Generative Adversarial Networks (GANs) shifted the paradigm from hand-crafted features to data-driven mapping [6]. Seminal works by Chen et al. [6] and Lore et al. [30] demonstrated how end-to-end convolutional networks could outperform the traditional image processing pipeline by learning direct mappings between raw noisy data to enhanced denoised RGB outputs. These models employ U-Net or Auto-Encoder (AE) architectures to learn latent features in low-light images and produce denoised enhanced outputs. Subsequent supervised methods utilize more complex architectures such as transformers to capture global context better through a broader receptive field [49]. The success of these models is bottlenecked by the need for paired datasets consisting of strictly aligned short-exposure images and corresponding long-exposure ground truth. The acquisition of such data is difficult, costly, and highly susceptible to misalignment and motion blur, especially in dynamic environments. While a widely utilized alternative is synthetic datasets where noise is artificially added and gamma lowered, models trained on such data often struggle to generalize to the complex noise profiles of real-world scenarios [1].

To address the scarcity of paired training data, unsupervised approaches such as Zero-DCE and EnlightenGAN, have been explored [14, 22]. Zero-DCE approaches the enhancement process as a curve estimation problem where zero-reference loss functions are used to eliminate the requirement of paired data. Conversely, EnlightenGAN utilizes an adversarial framework and unpaired low-light and normal-light data to guide the model towards generating a realistic light distribution. While these methods generalize better to unseen data and solve the data acquisition problem, the lack of structural ground-truth guidance means they struggle to differentiate noise from true signal in extreme low-light scenes, often resulting in lost details and suboptimal enhancement results [36].

2.2 Multi-modal IR-guided Enhancement and Fusion Models

Due to the degradation of images captured by RGB sensors in low-light and poor environmental conditions, research has increasingly looked towards using IR imaging. Since IR sensors capture thermal radiation, they are inherently robust against these illumination limitations [36]. The most wide-spread use of IR images in RGB image enhancement has been in multi-modal fusion networks where features are extracted from both modalities and fused in some stage of the processing pipeline to generate a composite image [36][29][32].

However, a persistent challenge across all multi-modal architectures is the strict requirement for cross-modal alignment. As noted by Ma et al. [31], because RGB and IR sensors have different fields of view and physical placements, achieving perfect pixel-level registration is difficult. This spatial misalignment of network inputs can introduce severe parallax artifacts during fusion or guided enhancement [48]. Several registration techniques have been explored in the literature to address this spatial misalignment [31]. These typically include area-based methods, which align images by minimizing pixel-wise differences under a defined metric, and feature-based methods, which measure the correspondence between

common structural elements, such as edge maps, extracted from both modalities.

Furthermore, existing multi-modal low-light image enhancement methods usually require both RGB and IR inputs during inference. One solution to this is the Learning Under Privileged Information (LUPI) paradigm [43] where the a model is provided auxiliary information that is only available during training to improve performance, which has been utilized in various cross-modal fusion networks [17, 2]. Within low-light image enhancement however, to the best of our knowledge, no research has yet explored using IR exclusively for structural supervision during training. By embedding IR-based structural understanding into an RGB-only network, this approach eliminates the need for an IR imagery at inference, thereby bypassing the alignment issues.

2.3 Hallucinations

As low-light image enhancement is fundamentally an ill-posed problem, algorithms are forced to infer severely degraded or entirely absent spatial and textural information [1]. Consequently, existing generative models often struggle to accurately reconstruct missing information, especially in extreme low-light scenarios [1, 8]. In particular, diffusion and GAN-based methods have been found to produce erroneous semantic hallucinations [7, 34], producing fictitious object contours and textures derived from learned training distributions rather than the information found in the actual scene [16]. The implications of these hallucinations are significant, particularly in domains where structural integrity is critical, such as autonomous navigation and nighttime surveillance [1].

To mitigate erroneous hallucinations different methods have been proposed, for example integrating semantic segmentation networks to preserve information useful for higher level tasks during the image restoration process [7]. Another strategy that is commonly utilized is enforcing structural losses, such as SSIM or edge-aware penalties during training [50]. Given the inherent deterioration of RGB data in extreme low-light conditions, this fundamental limitation motivates approaches incorporating IR information as guidance through specialized loss functions.

Chapter 3

Background

Enhancing the visibility of structures in extreme low-light environments requires a method which can learn to differentiate true signal from pure noise in very low SNR images. The current literature highlights how this task lends itself well to unsupervised or semi-supervised approaches due to the difficulties of data collection, as well as approaches incorporating cross-modal guidance. To contextualize our methodology and findings, this chapter establishes the theoretical foundations that underpin modern low-light enhancement networks. While the broader fields of computer vision and machine learning encompass a wide range of image processing and deep learning techniques, this chapter focuses primarily on unsupervised and cross-modal methods. We assume the reader has a basic familiarity with concepts such as neural networks, digital image representations and how deep networks learn through gradient descent and back-propagation.

3.1 Infrared Imaging and LWIR Sensors

Infrared (IR) imaging operates within the portion of the electromagnetic spectrum between visible light and microwaves. While visible light occupies the $0.4\mu\text{m}$ to $0.7\mu\text{m}$ range, the infrared spectrum extends from approximately $0.7\mu\text{m}$ to $1000\mu\text{m}$, though imaging applications typically focus on the region between $0.7\mu\text{m}$ and $14\mu\text{m}$ [11]. This spectrum is further categorized into distinct sub-bands based on their physical properties and interaction with matter, as detailed in Table 3.1.

Unlike the Near-Infrared (NIR) cameras, which rely heavily on reflected photon energy and often requires active illumination in total darkness, LWIR cameras capture emitted radiation, eliminating the need for external light sources [11].

This property makes LWIR a compelling tool for low-light enhancement tasks, as the thermal signatures provide a stable structural representation of the environment that is invariant to the presence or absence of visible illumination. However, because LWIR sensors capture emissive rather than reflective properties, the resulting images often lack the fine tex-

ture and high-frequency details found in RGB images, necessitating the supervision strategies discussed in section 2.2.

The sub-band LWIR corresponds to the peak thermal emission of objects at terrestrial temperatures (approximately 300 K or 27°C), as described by Planck’s Law of blackbody radiation [18], and is therefore suitable for practical scenarios.

Band	Abbreviation	Wavelength (μm)
Near-Infrared	NIR	$0.7\mu\text{m} - 1.4\mu\text{m}$
Short-Wave Infrared	SWIR	$1.4\mu\text{m} - 3\mu\text{m}$
Mid-Wave Infrared	MWIR	$3\mu\text{m} - 8\mu\text{m}$
Long-Wave Infrared	LWIR	$8\mu\text{m} - 15\mu\text{m}$
Far-Wave Infrared	FIR	$15\mu\text{m} - 1000\mu\text{m}$

Table 3.1: Sub-bands of the Infrared Spectrum [11]

3.2 Histogram Equalization

Histogram Equalization (HE) is a technique used to enhance image contrast by redistributing intensity levels to approximate a uniform distribution. This process "stretches" the dynamic range of an image by mapping a narrow cluster of frequently occurring input intensities across the entire available spectrum, commonly 0-255 for byte-valued pixels. This is achieved by using the normalized Cumulative Distribution Function (CDF) as the transformation function. As shown in Figure 3.1, the equalization process transforms the original non-linear CDF into a near-linear CDF, thus increasing the separation between intensity levels and revealing details in previously underexposed or overexposed regions [38].

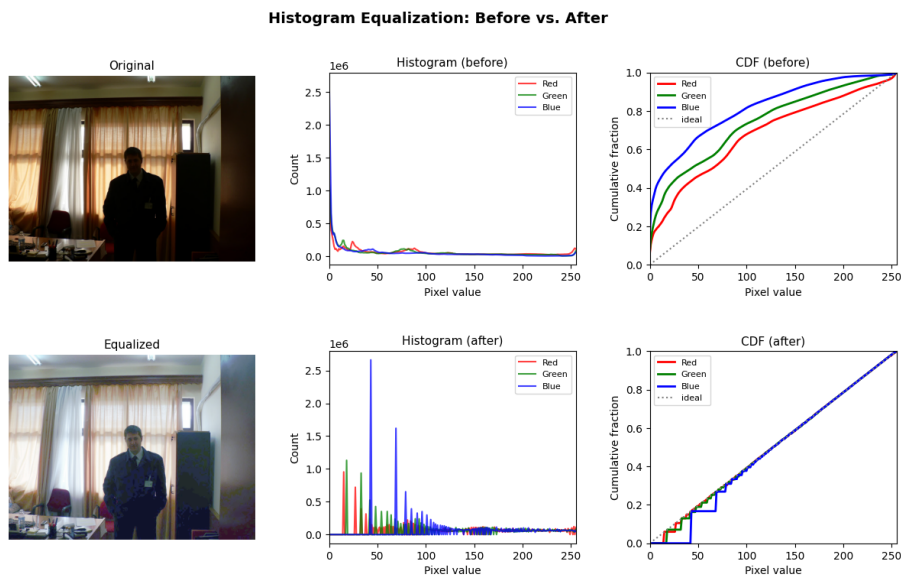


Figure 3.1: A low-light image alongside its corresponding intensity histogram and CDF before (top row) and after (bottom row) histogram equalization. Image from Zero-DCE [14]

3.3 Zero-DCE

Zero-Reference Deep Curve Estimation (Zero-DCE) is a method for low-light image enhancement that reframes the task as a pixel-wise curve estimation problem instead of the conventional image-to-image approach [14], as shown in Figure 3.2. Unlike conventional supervised deep learning approaches that require paired datasets of low-light and ground-truth normal-light images, Zero-DCE is trained using a set of differentiable loss functions that are not dependent on a well-lit reference image. This "zero-reference" approach allows the model to learn transformations directly from unpaired images, which avoids the biases and collection challenges inherent to synthetic or real-world paired datasets [14].

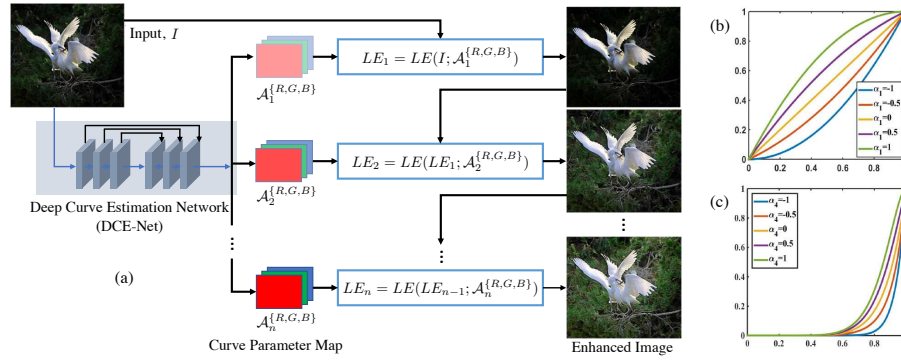


Figure 3.2: The Zero-DCE curve application process. By estimating pixel-wise LE-curves and iteratively applying them to the input image the enhanced image is produced. Figure from [14].

3.3.1 Light Enhancement Curves (LE-curves)

The design of LE-curves is inspired by the curve adjustment tools found in photo editing software. In such applications, users manipulate a curve to map the input pixel intensities to new output values, thereby adjusting the dynamic range and exposure of an image. Zero-DCE automates this process by training a network to estimate these curves automatically based solely on the characteristics of the input image [14].

Specifically, the network estimates a pixel-wise curve parameter map, $\mathcal{A}$, where each individual parameter $\alpha \in [-1, 1]$ acts as a gain factor [14]. This gain determines the magnitude of the curve's curvature and thus the degree of exposure adjustment for every pixel in the image. By predicting these parameters locally, the network can apply a high gain to brighten dark regions while maintaining or subtly adjusting well-lit areas.

The enhancement process is defined by a quadratic mapping that is applied iteratively to approximate higher-order curves, providing better dynamic range adjustment. The enhanced pixel value at iteration n , denoted as $LE_n(x)$, is calculated as Equation 3.1

$$LE_n(\mathbf{x}) = LE_{n-1}(\mathbf{x}) + \mathcal{A}_n(\mathbf{x})LE_{n-1}(\mathbf{x})(1 - LE_{n-1}(\mathbf{x})) \quad (3.1)$$

where $LE_0(x)$ is the original input pixel value and each A_n is the estimated parameter map for that specific iteration [14]. The final output of the model is the enhanced image obtained after n iterations (where n is typically set to 8).

3.3.2 DCE-Net

The *Deep Curve Estimation Network* (DCE-Net) is a plain CNN consisting of seven convolutional layers and is the model used in the original Zero-DCE [14]. Unlike other architectures like U-net which reduce spatial resolution to capture context, DCE-Net discards all downsampling (pooling) and batch normalization layers to preserve the relations between neighboring pixels. This ensures that the network maintains a full-resolution feature map throughout the entire process. A defining feature of DCE-Net is its use of symmetrical concatenation. In this configuration, feature maps from the first three layers are concatenated with the feature maps of the last three layers.

3.3.3 Losses

Since Zero-DCE is a zero-reference network, it does not rely on ground-truth images during training. Instead, the framework uses four specialized loss functions to evaluate the quality of the enhanced output.

Spatial Consistency Loss (L_{spa})

Encourages the network to maintain spatial coherence in the reconstructed image by preserving the relative differences between neighboring regions as they appeared in the original input. This discourages the network from "flattening" local contrast or introducing structural artifacts that deviate from the input image [14].

$$L_{spa} = \frac{1}{K} \sum_{i=1}^K \sum_{j \in \Omega(i)} \left(|(Y_i - Y_j)| - |(I_i - I_j)| \right)^2 \quad (3.2)$$

where K is the number of local regions, $\Omega(i)$ denotes the four neighboring regions (top, down, left, right) of region i , and Y and I represent the average intensity of the enhanced and input images, respectively.

Exposure Control Loss (L_{exp})

Regulates the dynamic range of the output by restraining regions that are either underexposed or overexposed. It functions as the "target" for the network, measuring the L1 distance between the average intensity of local regions and a predefined "well-exposedness" level, denoted as E [14].

$$L_{exp} = \frac{1}{M} \sum_{k=1}^M |Y_k - E| \quad (3.3)$$

M represents the number of non overlapping local regions of size 16×16 , Y_k represents the average exposure in region k , and E is set to 0.6 (the gray level in the RGB color space).

Color Constancy Loss (L_{col})

Corrects potential color deviations in the enhanced image. It is based on the Gray-World color constancy hypothesis, which assumes that the color in each sensor channel averages to gray when evaluated over the entire image. It prevents the model from disproportionately amplifying a specific color channel and therefore helps the model keep the inherent color balance of the input image [14].

$$L_{col} = \sum_{\forall (p,q) \in \varepsilon} (J^p - J^q)^2, \quad \varepsilon = \{(R, G), (R, B), (G, B)\} \quad (3.4)$$

where J^p and J^q denotes the average intensity value of channel p in the enhanced image, and (p, q) represents a pair of channels.

Illumination Smoothness Loss (L_{tv})

To prevent abrupt transitions in the enhancement curves and preserve monotonicity between neighboring pixels, a Total Variation (TV) loss is applied to the curve maps $\mathcal{A}$:

$$L_{tv\mathcal{A}} = \frac{1}{N} \sum_{n=1}^N \sum_{c \in \xi} (|\nabla_x \mathcal{A}_n^c| + |\nabla_y \mathcal{A}_n^c|)^2, \quad \xi = \{R, G, B\} \quad (3.5)$$

where N is the number of iteration steps (typically $N = 8$), and ∇_x, ∇_y denote the horizontal and vertical gradient operations [14].

Total Loss

$$L_{total} = L_{spa} + L_{exp} + W_{col}L_{col} + W_{tv\mathcal{A}}L_{tv\mathcal{A}} \quad (3.6)$$

The coefficients W_{col} and W_{tv} are used to balance the loss terms and are set to 0.5 and 20, respectively [14].

3.4 U-Net

U-Net is a convolutional neural network architecture commonly used for semantic segmentation, where the goal is to assign a class label to every pixel in an image [41]. Originally designed for biomedical image analysis, U-Net is characterized by its symmetric U-shaped structure consisting of a contracting *encoder* path and an expansive *decoder* path, as seen in Figure 3.3.

The architecture begins with a contracting path, which uses successive convolution and pooling layers to reduce spatial resolution while increasing feature depth. There is evidence that such a structure capture abstract high-level features for an image [41]. Unlike standard CNNs that only reduce resolution to produce a single classification, the expansive path of the U-Net performs upsampling (transposed convolutions) to recover the original dimensions of the input. The novelty of U-Net is the use of skip connections, which concatenate high-resolution feature maps from the contracting path directly onto the upsampled maps in the expansive path. This allows the network to integrate low-level texture with the high-level features from the encoder, thereby promoting the retention of spatial details and textures that might otherwise be lost in the bottleneck [41].

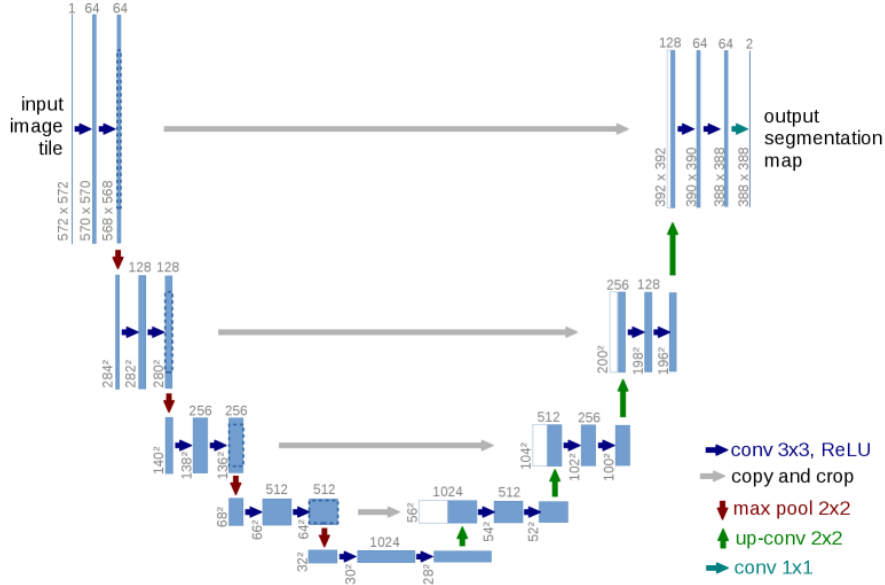


Figure 3.3: The U-Net architecture, consisting of an encoding (contracting) path and a decoding (expansive) path. Figure from [41].

The benefit of this architecture is its high spatial precision and its ability to perform well with limited training data. By processing the image through this symmetric network, it learns both the global context (what is in the image) and the local localization (where it is) simultaneously. While many segmentation models require massive datasets, U-Net’s heavy use of data augmentation and architectural efficiency makes it well-suited for tasks where labeled data is scarce [41].

Although U-net is commonly used for semantic segmentation, in this paper it is used to predict the pixel-wise LE-curves, similar to the approach used in Zero-DCE [14].

3.5 Attention Mechanisms

Attention mechanisms are integrated into convolutional neural networks to boost the model’s performance by enabling it to focusing on relevant features while suppressing less informative ones.

3.5.1 Squeeze-and-Excitation Attention

The Squeeze-and-Excitation (SE) block [20] is designed to weigh different channels. The SE mechanism consists of three stages:

1. **Squeeze:** Spatial information is squeezed into a channel descriptor using Global Average Pooling. This produces a vector $\mathbf{z} \in \mathbb{R}^{1 \times 1 \times C}$, where each element z_c represents the global average of the c -th channel.
2. **Excitation:** To capture non-linear dependencies between channels, the descriptor $\mathbf{z}$ is passed through a bottleneck structure consisting of two fully connected layers (FC).

A reduction ratio r is typically employed to limit model complexity. The excitation is calculated as:

$$\mathbf{s} = \sigma(W_2 \text{ReLU}(W_1 \mathbf{z})) \quad (3.7)$$

where $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in \mathbb{R}^{C \times \frac{C}{r}}$. The sigmoid activation σ ensures the resulting weights are in the range $[0, 1]$.

3. **Scale:** The final output is the product of the the original feature maps X and the learned weights $\mathbf{s}$ via channel-wise multiplication: $Y_c = s_c \cdot X_c$.

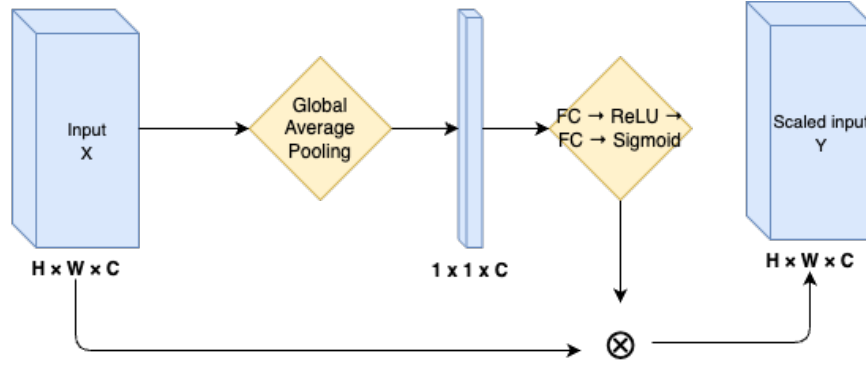


Figure 3.4: The structure of Squeeze-and-Excitation

3.5.2 Coordinate Attention

While SE blocks effectively capture channel-wise relationships, they do so by collapsing all spatial information, thereby ignoring local spatial distributions. Coordinate Attention (CA) [51] addresses this by embedding positional information into the attention maps. The CA mechanism follows a three-step process similar to Squeeze-and-Excitation:

1. **Squeeze (Coordinate Information Embedding):** Instead of using global pooling, CA decomposes the squeeze operation into two 1D average pooling operations. Features are averaged along the horizontal and vertical axes separately. This produces two direction-aware feature maps: $z^h \in \mathbb{R}^{C \times H \times 1}$ and $z^w \in \mathbb{R}^{C \times 1 \times W}$.

By pooling along one dimension, each vector captures long-range dependencies along that axis (e.g., horizontal context) while preserving precise positional information along the orthogonal axis (e.g., vertical coordinates).

2. **Excite:** The two feature maps are concatenated and passed through a 1×1 convolution that acts as a bottleneck to capture cross-channel interactions. The resulting intermediate tensor is then split back into two separate tensors. Two additional 1×1 convolutions and a sigmoid activation function are applied to transform these tensors into the attention weights $g^h \in \mathbb{R}^{C \times H \times 1}$ and $g^w \in \mathbb{R}^{C \times 1 \times W}$.
3. **Scale:** The final output Y is the product of the weights and the input X . For a channel c at coordinate (i, j) , the output is defined as:

$$Y_c(i, j) = X_c(i, j) \times g_c^h(i) \times g_c^w(j) \quad (3.8)$$

where $g_c^h(i)$ and $g_c^w(j)$ represent the weights for the height and width dimensions, respectively.

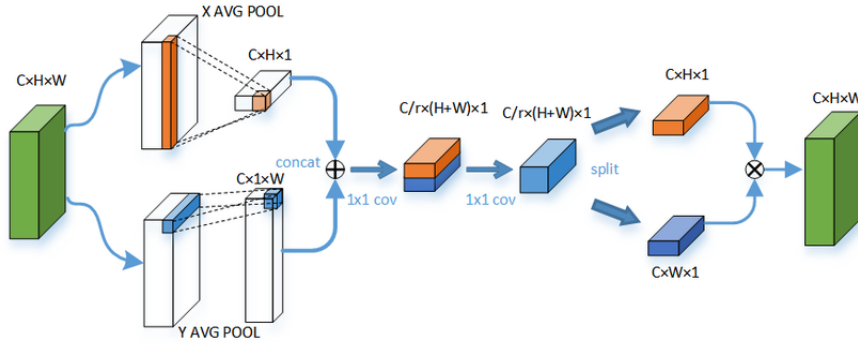


Figure 3.5: The structure of coordinate attention. Figure from [23].

3.6 DA-Zero-DCE

Denoising-Attention based Zero-Reference Deep Curve Estimation (DA-Zero-DCE) is an extension of the Zero-DCE method optimized for low-light enhancement in challenging environments such as transportation tunnels, which is a semi-closed environment affected by artificial light sources. The images are often captured by cameras from a long distance, making pedestrians appear blurry. This combination of poor lighting and blurry pedestrian targets makes it difficult for downstream image recognition algorithms to perform well.

While it maintains the core principle of pixel-wise curve estimation, DA-Zero-DCE introduces three primary architectural improvements to address the uneven illumination, local occlusion, and high noise levels characteristic of tunnel imagery. It swaps DCE-net to a U-net backbone, adds coordinate attention, and uses an additional denoiser module *NAF-Net* after the application of the LE-curves [51]. It also uses a denoising loss $Loss_{denoise}$ to train NAF-Net.

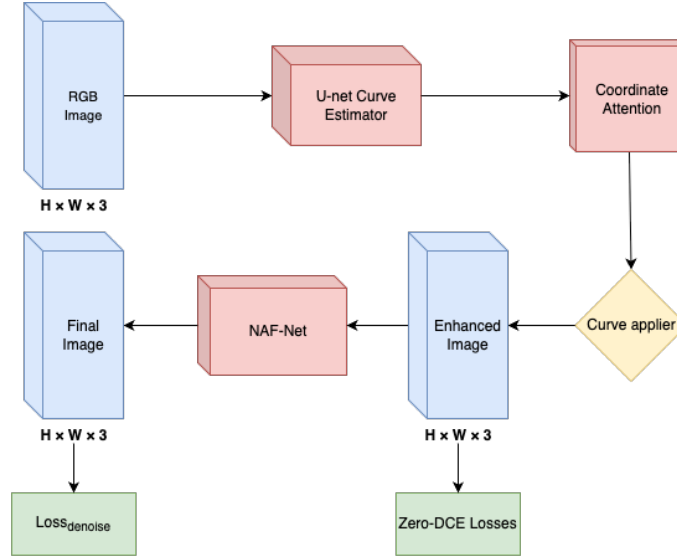


Figure 3.6: Overview of the DA-Zero-DCE architecture

3.7 Retinex Theory

Retinex theory is based on the human perceptual system and posits that an image is composed of two components: Luminance (L) and Reflectance (R). Reflectance describes the intrinsic properties of the objects within the image, such as their contour and color. Luminance represents the light falling upon those objects [26]. Given a source image S , Retinex theory states that it can be decomposed into its luminance and reflectance components as in Equation 3.9

$$S = R \cdot L \quad (3.9)$$

In the context of computer vision, luminance is generally understood to be spatially smooth, low-frequency gradients of the lighting in the image, whereas reflectance corresponds to sharp, high-frequency edges around structure and noise [25].

3.8 Inverse Problems

Understanding inverse problems is critical for understanding why low-light image enhancement is a difficult task in the first place. A forward problem is a well-defined, deterministic operation, such as darkening or inserting noise into a clean image. An inverse problem is the reverse process of trying to find the clean image which corresponds to the observed data. The problem being ill-posed means that, due to the degradation of information in the input, there is no stable or unique solution and there are multiple candidates for clean images which could have produced the exact same noisy output [12].

To frame such problems, deep models often employ Bayes theorem (Equation 3.10). Here $P(x|y)$ is called the *posterior* and represents the probability of x being the true image, given the noisy observation y . The aim is to find a model which maximizes the posterior. The term $P(y|x)$ is the *likelihood*, and represents how likely it is to observe the noisy output, if x is the true image. This term is therefore only based on the observed data. Lastly, $P(X)$ is called the

prior probability, or simply *prior*, and represents the likelihood of the clean image x existing in the real world, independent of the observations [13].

$$P(x|y) \propto P(y|x)P(x) \quad (3.10)$$

There are two primary ways in which models use the Bayesian equation to estimate the likelihood. The first is Maximum Likelihood Estimation (MLE), where the prior is assumed to be constant with respect to y , and dropped out of the equation entirely. The task then becomes to find parameters which maximize the likelihood based on the observed data. The other method is called Maximum A Posteriori (MAP) estimation, in which the parameters that maximize both the likelihood *and* the prior are found. The prior then acts as a regularization on the possible parameters which limits the outputs to what images “normally” look like [13].

3.9 Hallucination

While the term hallucination lacks a universally accepted definition within the field of computer vision, it commonly refers to the generation of plausible but incorrect structural or semantic information [3, 15]. In low-light image enhancement, the risk of hallucination is particularly high due to the lack of information in the input. Low-light conditions lead to low photon counts, sensor noise, and color distortion which degrades the input in exceptionally dark regions [6]. From a Bayesian perspective, Sampson et al. [42] argue that hallucinations occur when generation of the output is dominated by a data-driven prior, rather than being based mostly or entirely on the observation likelihood. This primarily occurs in the areas lacking strong input signal mentioned above, which causes the likelihood term to become weak.

Hallucinations can manifest themselves in multiple different ways. Structural and textural hallucinations can occur when heavy noise in low SNR images is misinterpreted as true high-frequency edges. Color distortions and chromatic noise often present in RAW images can lead to incorrect color mappings. In severe cases semantic hallucinations can occur, where entirely new features are invented due to them commonly appearing in the model’s training data. Addressing and mitigating these hallucinations remains a critical challenge in low-light image enhancement, particularly in domains where structural fidelity is strictly required, such as medical imaging and autonomous navigation in low-light environments.

3.10 RAW Images

RAW images contain the minimally processed data captured directly by the camera sensor. This is in contrast to RGB images which are the result of the camera’s internal image processing. RAW data is stored as a single-channel mosaic where each pixel records a single intensity value instead of full color information. Because of this, RAW images often have higher dynamic range and higher bit depth than processed alternatives [47].

RAW sensor data typically has a bit-depth of 12 or 16 bits, which offers a significantly higher dynamic range than the standard 8-bit RGB format. In extreme low-light scenarios,

8-bit quantization can introduce severe precision loss, and in very dark regions subtle signal variations are compressed or entirely clipped to pure black. High bit-depth RAW data preserves these subtle signals and provides a higher SNR for the model [4].

Standard RGB images are produced when a RAW file passes through an Image Signal Processor (ISP). While the ISP makes images visually appealing, it often removes the critical structural information needed for accurate reconstruction. By processing RAW data directly, models can generate structural and textural improvements based on the unadulterated sensor signal [4].

3.11 Sobel Filter

The Sobel filter is an operator used in computer vision to estimate the gradient of intensities in an image and therefore serves as a tool for edge detection [9]. The operation involves convolving the image I with two 3×3 kernels to calculate the gradient components in the horizontal and vertical directions, denoted as G_x and G_y respectively. The final edge map G is constructed by calculating the magnitude of the combined directional edge maps as defined in Equation 3.11.

$$G_x = \begin{bmatrix} 1.0 & 0.0 & -1.0 \\ 2.0 & 0.0 & -2.0 \\ 1.0 & 0.0 & -1.0 \end{bmatrix} * I \quad G_y = \begin{bmatrix} 1.0 & 2.0 & 1.0 \\ 0.0 & 0.0 & 0.0 \\ -1.0 & -2.0 & -1.0 \end{bmatrix} * I \quad G = \sqrt{G_x^2 + G_y^2} \quad (3.11)$$

Chapter 4

Method

Having established the theoretical basis of semi-supervised learning, and the utility of utilizing cross-modal information to improve the models enhancement capabilities, this thesis seeks to design such a network and apply it to images with challenging low-light conditions. This chapter details the design, implementation and training methodology of our proposed image enhancement pipeline. The architecture can conceptually be split into three main components: A curve parameter estimation backbone contributing to global illumination recovery, an *IR-based Enhancer* module which focuses on local detail reconstruction, and finally a custom cross-modal loss function which guides the model in recovering structure in low-light scenarios.

The remainder of this chapter is structured as follows. First, section 4.1 provides a high-level overview of the system architecture and design decisions. This is followed by section 4.2 which outlines the dataset used and preprocessing of the RAW images. Then, section 4.3 dives deeper into each component of the architecture, and section 4.4 describes our custom IR-supervised loss function, along with other losses used during training. The setup of the baselines which we compare our network against is specified in section 4.5. Lastly, in section 4.6, details concerning the specific training set-up and hyperparameters are presented.

4.1 System Overview

The following architecture is designed to address the fundamental challenge of unsupervised low-light image enhancement without introducing hallucinations. In safety critical applications, generative models that synthesize rather than modulate existing pixel data may introduce an unacceptable risk of hallucinations. Therefore, the proposed architecture adopts Zero-DCE’s [14] approach of framing image enhancement as an illumination curve-mapping problem rather than a pixel generation task. By constraining the deep network’s output to simple exposure multipliers, along with enforcing losses that ensure structural adherence, the network is limited in it’s ability to hallucinate high-frequency details not present in the orig-

inal scene. To improve upon the original Zero-DCE, and enable the model to better focus its enhancements on especially low-light areas containing salient objects, this thesis draws architectural inspiration from DA-Zero-DCE [51], specifically adopting its U-Net-based curve estimator and Coordinate Attention mechanism. The U-Net backbone allows the network to process global illumination context while retaining local edge data. Simultaneously, Coordinate Attention provides directional and positional awareness to target specific spatial features. [51].

However, after the LE curve output, our architecture departs from the DA-Zero-DCE pipeline. The original DA-Zero-DCE architecture utilizes a deep NAF-net block to denoise the images, requiring training on paired, well-lit ground truth RGB images. In this thesis, one of our specific goals is to not rely on such supervision, and thus no such data is available. To resolve this, we replace the heavy NAF-Net with a constrained IR-based Enhancer. By limiting the depth to three convolutional layers and bounding the output via a restricted activation function, this architecture limits the ability to hallucinate non-existent features, forcing it to act as a localized edge sharpener. The entire system architecture, which we name IR-Zero-DCE, can be seen in Figure 4.1.

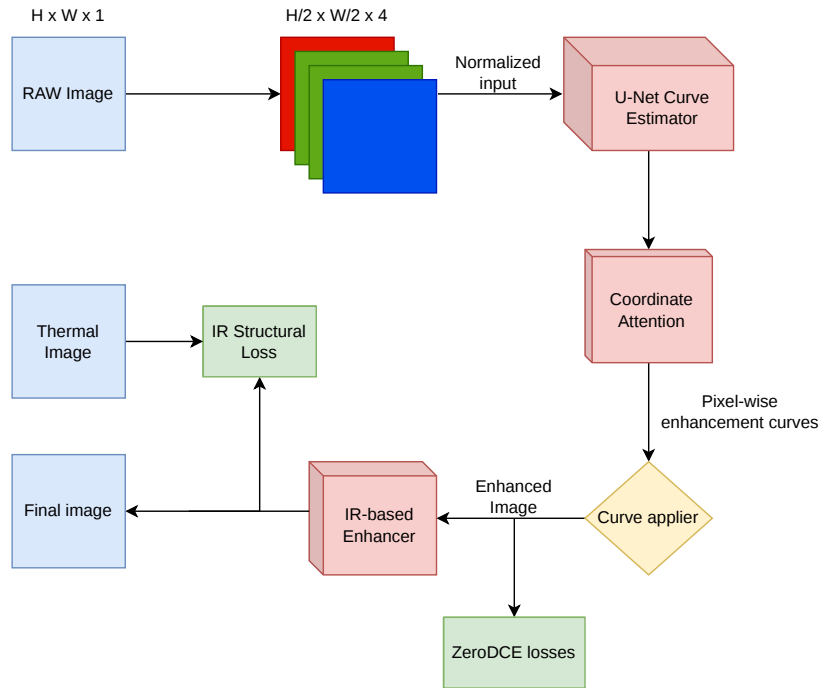


Figure 4.1: The system architecture of IR-Zero-DCE

4.2 Dataset and Preprocessing

Since our method has the goal of being applied in nighttime naval scenarios, we are interested in extremely low-light images. To the best of our knowledge, sufficiently dark images are not available in any public datasets. Therefore this thesis relies on data provided by Saab.

Our experiments are conducted on two distinct datasets. The first, $\text{Dataset}_{\text{Lab}}$, was collected in a controlled lab environment and consists of 74 RGB-IR pairs (example shown in Figure 4.2). This dataset is primarily used to illustrate the contributions of our losses and architecture as being able to control the environment enabled us to capture extreme low light scenarios. The second, $\text{Dataset}_{\text{Maritime}}$, is made up of images captured in a maritime environment after sunset (example shown in Figure 4.2). This dataset is substantially larger, consisting of 859 pairs, and is used to demonstrate the model’s performance in a more realistic setting. Both datasets consist of paired RAW RGB and IR images collected from cameras set a few centimeters apart. The RGB images are captured with 12-bit depth, while the IR images are captured with a 16-bit depth. RAW images are used for the benefits discussed in section 3.10.

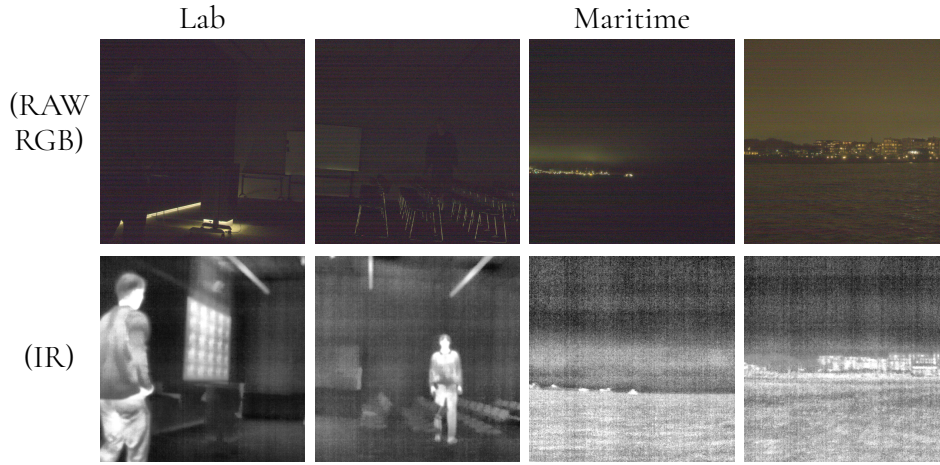


Figure 4.2: Sample image pairs from $\text{Dataset}_{\text{Lab}}$ and $\text{Dataset}_{\text{Maritime}}$. Raw inputs are gamma boosted for visibility. The IR images are normalized.

To prepare the data for model input the images were normalized to the $[0, 1]$ -range. To account for the different resolutions, aspect ratios, FOVs and physical placement of the visual spectrum and thermal cameras, the image pairs were cropped, scaled, resized and aligned. While automatic registration for the image pairs using Enhanced Correlation Coefficient (ECC) and feature-based matching was attempted, the resulting alignment proved to be inferior compared to obtaining the transformations for each camera pair through careful manual alignment, which was the method ultimately used.

4.3 Network Architecture

In this section we dive deeper into each component of the IR-Zero-DCE architecture. The network is divided into two distinct phases: First a curve estimation phase where the Zero-DCE LE-curves are produced and subsequently applied to each pixel of the input image. This phase is responsible for low-frequency global illumination enhancement. In the second phase our IR-based Enhancer takes the relit image and increases the contrast and sharpness of high-frequency cross-modal features while simultaneously dampening the noise.

4.3.1 U-Net Curve Estimator Backbone

Our curve estimation backbone is shown in Figure 4.3. As in the original DA-Zero-DCE architecture, we utilize a U-Net backbone to estimate the pixel-wise Light-Enhancement (LE) curves. By replacing regular Zero-DCE’s seven-convolutional layer DCE-Net with a deeper network which incorporates downsampling, we aim to increase the receptive field of the model and give it a better understanding of the global illumination context. Heavy downsampling does however degrade high-frequency edges in the input, and therefore our U-Net utilizes skip connections between layers to bypass this bottleneck and preserve such information. Rather than generating pixels, the U-Net estimates pixel-wise LE-curves. This approach constrains the model to a mapping-based transformation that preserves the original scene’s topology, which in combination with the total variation loss penalizes erratic spatial gradients [14].

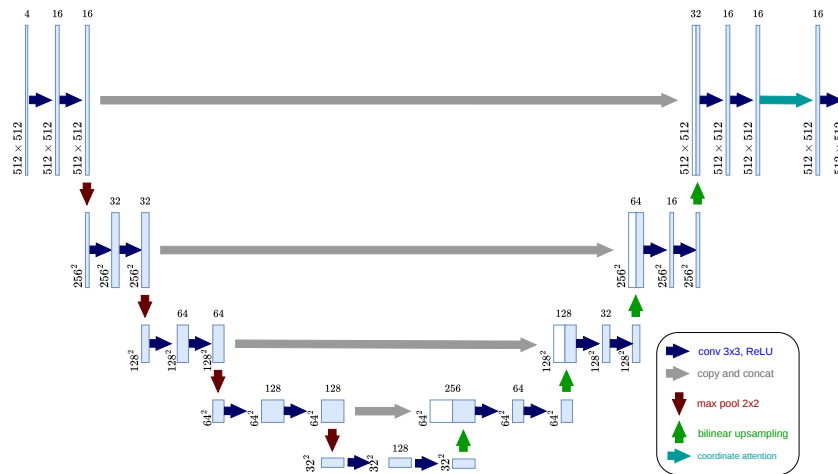


Figure 4.3: Structure of the U-Net curve estimation backbone. Coordinate attention is applied to the final feature map of the last decoding stage. Figure adapted from [41].

4.3.2 Coordinate Attention

Standard CNNs process images uniformly, applying the same filter across every pixel. However, this can lead to over-processing certain areas and under-correcting others. For this reason, using attention mechanisms that tell the model where to focus its adjustments, enabling more localized refinements, has been shown to improve the performance of deep networks [19].

For our application it is essential to retain not only which features are important, but also where they are spatially located within the image. The coordinate attention mechanism, shown in Figure 3.5, achieves this by separately applying vertical and horizontal pooling along the rows and columns of the feature map respectively. This instead produces two 1D vectors, retaining the exact coordinates of attended features. When these vectors are multiplied together in the last step of the CA-block their intersection acts as a precise crosshair of where the feature is located, producing the attention map. Figure 4.4 shows an example of how

the coordinate attention map can look during inference for a given input image. Note that the IR image is included only to highlight the attention map, and not used to inform the CA-module during training.

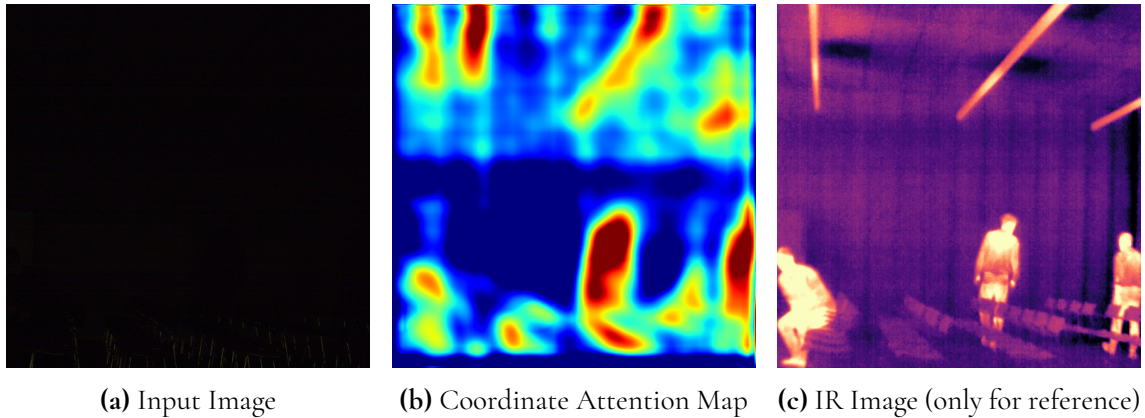


Figure 4.4: Visualization of the internal coordinate attention mechanism (b) during inference on an input image (a). The IR image (c) is only provided for reference to show the composition of the input scene, as it is too dark to see directly.

4.3.3 IR-based Enhancer

In lieu of the NAF-Net denoiser from the original DA-Zero-DCE architecture, as it requires paired well-lit images for training, IR-Zero-DCE introduces the *IR-based Enhancer* (IRE) to address the noise in the output from the curve estimator as well as increasing the sharpness and contrast of structural boundaries. The structure of the IRE can be seen in Figure 4.5. Prior to entering the IRE, the image is converted to grayscale and detached from the computational graph. The reason for this detachment is to prevent gradient updates stemming from the IR supervision from reaching the backbone. While the U-Net backbone outputs multiplicative pixel-wise LE-curves, the IRE instead outputs an additive residual map to the enhanced image. The final activation function used is the hyperbolic tangent, bounding the output to $[-1, 1]$ which is then further constrained by a multiplicative constant 0.5. The reason for this is to constrain the network's ability to hallucinate as the IRE module is capable of generating sharp, high-frequency adjustments which can disrupt the local spatial correlations of neighboring pixels in the input. The IRE performs two distinct tasks, guided by the IR-supervision, to improve the structural reconstruction capabilities of IR-Zero-DCE. The first task is edge amplification: Where there are geometric boundaries in the image, the enhancer outputs positive residuals to the brighter side of the edge and negative residuals to the darker side. This steepens the gradient creating a sharper, higher contrast boundary around the object of interest. The second task is noise suppression: Amplified sensor noise in the enhanced image creates thousands of high-frequency gradients. In regions where there are no features of interest in the corresponding IR image, e.g a dark background, the IR supervision punishes such edges. Here the IRE outputs residuals to dampen the noise spikes and smooth out the area. Crucially, the IRE is constrained from aggressively smoothing all non-IR features. Because thermal images inherently lack the fine textural and luminance details present

in RGB data (e.g. printed text or shadows), the noise dampening is calibrated to preserve these RGB-specific high-frequency textures, rather than forcing the output to strictly mimic the thermally flat regions of the IR ground truth.

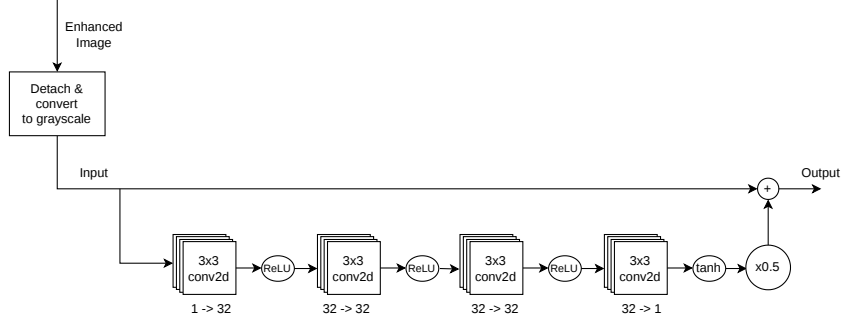


Figure 4.5: The IR-based Enhancer architecture

4.4 Loss Functions

Our model uses a composite loss function designed to balance global illumination enhancement with local structural recovery. While IR-Zero-DCE adopts the non-reference learning method of Zero-DCE [14], the spatial consistency loss (L_{spa}) proves insufficient when dealing with the extremely low signal to noise, and the low-contrast boundaries characteristic of RAW low-light data.

To address these limitations, we introduce a cross-modal structural loss that leverages infrared guidance to recover edges. This section details the mathematical formulation of these components and the rationale for their integration into the total loss function.

4.4.1 Edge-Enhancing Structural Loss

To guide the model towards enhancing cross-modal structural boundaries we present the following custom loss function. Given the input image I_{Input} , the predicted image I_{Pred} and the corresponding thermal image I_{IR} , the objective of the loss is twofold: First, produce an enhanced output I_{Pred} with sharp, high-contrast edges around structures which are found in I_{Input} or I_{IR} . Second, punish the network for producing sharp edges where none exist in either of the modalities.

By extracting gradients in both X and Y directions using Sobel filters as in Equation 3.11, we obtain the edge magnitude maps G_{Input} , G_{Pred} , and G_{IR} for I_{Input} , I_{Pred} , and I_{IR} respectively. To minimize the noise picked up by the Sobel filters, a Gaussian blur filter is applied to the images prior to extracting the edge maps.

To further prevent unwanted gradients caused by noise, as we are dealing with RAW images in extreme low-light conditions, we extract only the sharpest and most prominent edges of the input edge map and IR edge map by masking out all values in G_{Input} and G_{IR} under a certain threshold. This threshold is dynamically determined for each image using PyTorch’s `torch.quantile` function, and computed as $P(G_{Input} \leq T_{Input}) = q$ and $P(G_{IR} \leq T_{IR}) = q$

to find a threshold T for the q^{th} quantile. For our datasets a good value for q was experimentally determined to be 0.97. Utilizing this threshold a binary mask $M_{edges} = M_{input} \cup M_{IR}$ is created, where M_{input} and M_{IR} are calculated according to Equation 4.1.

$$M(G, T) = \begin{cases} 1, & \text{if } G > T \\ 0, & \text{otherwise} \end{cases} \quad (4.1)$$

To account for misalignment between input and thermal images, M_{IR} is expanded using max pooling. The resulting thresholded binary masks are shown in Figure 4.6.

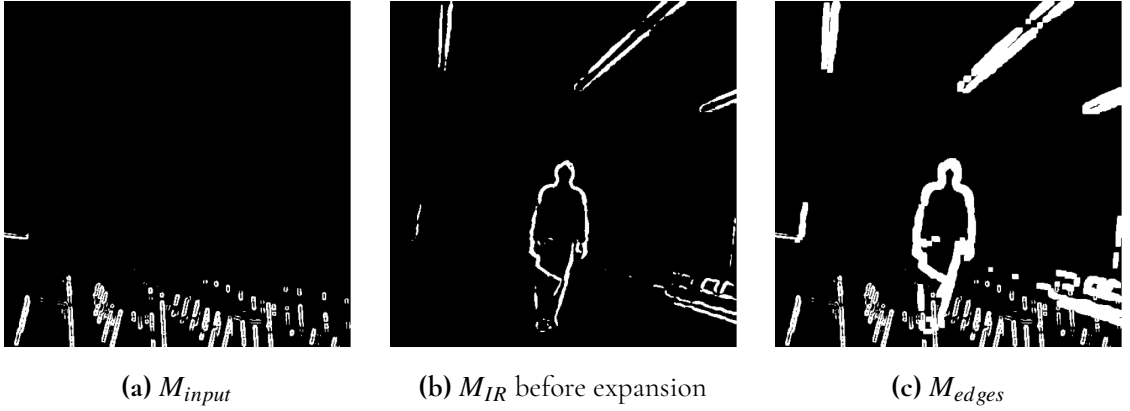


Figure 4.6: Visualization of the edge mask generation process. (a) Binary mask M_{input} from the input image, (b) binary mask M_{IR} derived from the thermal image before expansion, and (c) the final combined and expanded edge mask M_{edges} ($q = 0.97$).

Using the binary mask, the first objective of the loss, to reward the model for enhancing object contours, is formulated as Equation 4.2 where i denotes the pixel index. This corresponds to the L1 distance between the predicted gradients and a target gradient magnitude of 0.5, normalized by the total edge area. The reason for targeting a softer 0.5 gradient magnitude at the edges rather than for example 1.0, is that such sharp edges almost never appear in ordinary images, and therefore too steep targets result in unnatural looking outputs.

$$L_{edges} = \frac{\sum_i M_{edges}^{(i)} \cdot |G_{pred}^{(i)} - 0.5|}{\sum_i M_{edges}^{(i)}} \quad (4.2)$$

The second objective is to punish the model for producing edges where none exist in the thermal image. To perform this noise suppression, while still allowing the preservation of textures and features *only* present in the input’s visual spectrum (but not the thermal spectrum of the supervising IR image), the following method is employed:

1. To determine the signal amplification of the network, given an input image I_{Input} , we calculate a scalar boost factor $\beta = \frac{\mu_{Pred}}{\mu_{Input} + \epsilon}$, where μ is the respective image’s mean pixel value and ϵ a small constant for stability.
2. The gradients G_{input} of the input image are extracted, and scaled with the boost factor:
 $\tilde{G}_{Input} = \beta \cdot G_{Input}$.

3. The network’s noise amplification N_{amp} is calculated (shown in Figure 4.7a), which determines how much the predicted image’s background gradients have scaled *beyond* the input’s scaled gradients: $N_{amp} = \max(0, G_{Pred} - \tilde{G}_{Input})$.

Finally, by inverting the binary mask $M_{bg} = \neg M_{edges}$, and only punishing the noise amplification, the background component of the loss is formulated as Equation 4.3. Similarly to L_{edges} , this corresponds to the L1 distance between the noise amplification and zero, normalized by the total background area.

$$L_{bg} = \frac{\sum_i M_{bg}^{(i)} \cdot N_{amp}^{(i)}}{\sum_i M_{bg}^{(i)}} \quad (4.3)$$

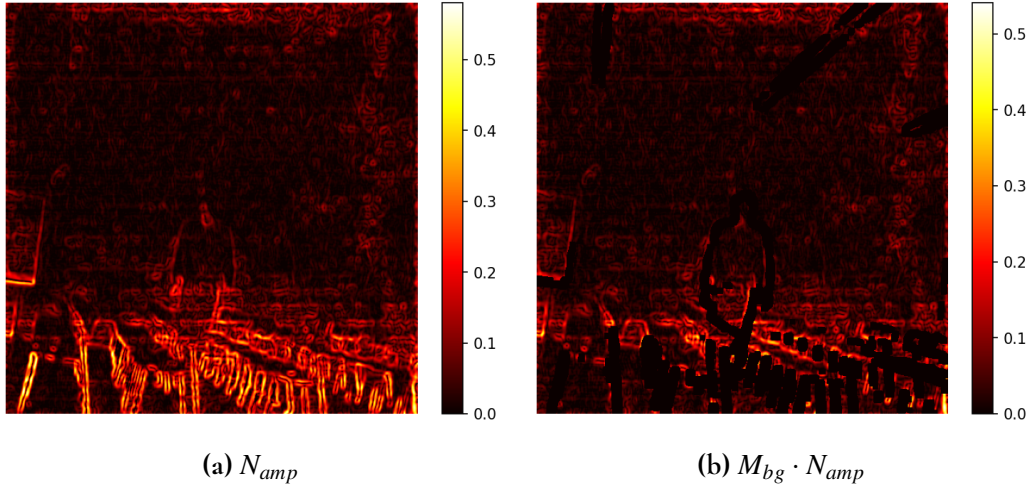


Figure 4.7: (a) shows the gradients which have been amplified beyond the naturally scaled input. In (b) the IR and RGB edges have been masked out, as we do not want to suppress the edges that correspond to true structure. What remains in (b) is the noise that the background loss punishes.

The total loss is then formulated as:

$$L_{struct} = L_{edges} + L_{bg} \quad (4.4)$$

4.4.2 Vertical Total Variation Loss

To address the extreme noise in the input images, often presenting as horizontal “banding” or readout noise, we introduce a simple Vertical Total Variation Loss. The purpose is to reduce the banding that is otherwise amplified in the enhanced image, as can be seen in Figure 4.8.



Figure 4.8: The input images contain prominent horizontal banding noise. Note that the image in the figure is gamma boosted to highlight this.

This loss is functionally similar to the Illumination Smoothness Loss Equation 3.5, but operates on the final output image from the IRE rather than on the LE-curve maps. The loss calculates the absolute difference between vertically adjacent pixels in each channel, and then takes the overall average over all those differences, as described by Equation 4.5.

$$L_{vert} = \frac{1}{N} \sum_c \sum_y \sum_x |I_{c,y,x} - I_{c,y+1,x}| \quad (4.5)$$

Note that this loss is not part of the default architecture for our network, but is still included as an optional add-on to produce smoother and arguably more visually pleasing images.

4.4.3 Total Loss

The total loss is described as:

$$L_{total} = W_{exp}L_{exp} + W_{col}L_{col} + W_{tv,\mathcal{A}}L_{tv,\mathcal{A}} + W_{struct}L_{struct} \quad (4.6)$$

Where L_{struct} is our custom Edge-Enhancing Structural Loss which replaces the Spatial Consistency Loss (Equation 3.2) from regular Zero-DCE.

4.5 Baselines

To contextualize our results, we evaluate IR-Zero-DCE against three baseline methods: Zero-DCE, DA-Zero-DCE, and Histogram Equalization. To ensure a fair comparison while pre-

serving the core characteristics of the original architectures, we made only the minimal adjustments necessary to accommodate our RAW data format and training pipeline. These modifications are detailed below for each baseline.

Zero-DCE

To adapt the Zero-DCE curve estimation network (DCE-Net) to our four-channel RGGB RAW input, we modified the initial convolutional layer to accept four input channels rather than the standard three. Consequently, because the network estimates pixel-wise curve parameters for each color channel across 8 iterations, the output layer was adjusted to produce 32 parameter maps ($4 \times 8 = 32$) instead of the original 24 ($3 \times 8 = 24$). Finally, to make comparisons with our model’s grayscale output possible, we implemented a straightforward RGGB-to-grayscale conversion at the end of the Zero-DCE pipeline. We first consolidate the two green channels of the RAW format by averaging them, and then apply standard luminance weighting to the resulting R, G, and B components. The final grayscale intensity Y is computed as:

$$Y = 0.299R + 0.587\left(\frac{G_1 + G_2}{2}\right) + 0.114B \quad (4.7)$$

where G_1 and G_2 represent the two green channels in the RGGB Bayer pattern.

Otherwise the original DCE-Net architecture is kept intact, and the parameters of the network are initialized with zero mean and a 0.02 standard deviation Gaussian function, as in the original paper [14].

DA-Zero-DCE

Because there is no official, publicly available codebase for DA-Zero-DCE, we implemented the network based on our best interpretation of the original paper, leveraging standard implementations for its two core components: the U-Net and Coordinate Attention modules. A critical deviation from the original architecture involves the NAF-Net denoising module. In the original work, this module is trained using paired, well-lit RGB images. Since relying on paired data strictly contradicts our objective of training without such supervision, we omitted the NAF-Net module entirely.

The U-Net backbone of DA-Zero-DCE mirrors the one used in our proposed model, illustrated in Figure 4.3. One notable departure from the original DA-Zero-DCE implementation is that we reduced the channel capacity of each layer by a factor of four. This reduction was introduced to accommodate computational constraints, mitigate the risk of overfitting on our relatively small dataset, and because the original U-Net capacity is arguably excessive for the comparatively simple task of estimating Light Enhancement-curves. Furthermore, initial experiments using transposed convolutions for upsampling, as in the original DA-Zero-DCE, produced severe artifacts and heavy vignetting in the output images. To resolve this, and to further reduce the risk of overfitting by reducing the total parameter count, we replaced the `ConvTranspose2d` layers in the U-Net backbone with bilinear upsampling followed by standard convolutions, as recommended by Odena et al.[33]. Similarly to the regular Zero-DCE baseline, as a final step we convert the image to grayscale (Equation 4.7).

Histogram Equalization

Prior to equalization the images are passed through the same preprocessing pipeline used for the models' inputs. Then HE is applied to each channel independently, and the equalized channels are stacked to produce the final image. For comparison with the other methods the image is then converted to grayscale (Equation 4.7).

4.6 Experiments

The experiments and tests were performed on a system with an RTX 4070 GPU. The training procedure included training for 50 epochs with a batch size of 8, a learning rate of 0.0001 and an AdamW optimizer with weight decay of 0.0001 to reduce the risk of overfitting on our relatively small datasets. An unconventionally small batch size was used due to computational constraints. The weights for the losses are presented in Table 4.1, where the biggest difference is our implementation replacing the Spatial Consistency Loss from the baselines with our custom Edge-Enhancing Structural Loss.

Weight Name	Loss Weights	
	Our Model	Zero-DCE & DA-Zero-DCE
W_{exp}	10.0	10.0
W_{str}	N/A	1.0
W_{col}	5.0	5.0
$W_{tv_{\mathcal{A}}}L_{tv_{\mathcal{A}}}$	200.0	200.0
W_{struct}	5.0	N/A

Table 4.1: Weighting of loss components for our model and the baselines

Chapter 5

Results & Discussion

5.1 Evaluation Metrics

Common evaluation metrics used in low-light image enhancement include full-reference metrics such as PSNR and SSIM. These metrics require a well-lit ground truth image as reference, which is not available for the images within our datasets. Furthermore, as discussed earlier, existing public benchmark datasets do not contain images in the same domain, format and extremely dark lighting conditions as our model was trained on. One approach would be to try and evaluate the structural similarity between the model output and the corresponding IR, but standard SSIM fails in this regard due to the completely different domains of the images. Another approach is to extract some common features from both the prediction and thermal image, then compare that set of features. This is exactly what our Edge-Enhancing Structural Loss does, acting as a quantitative metric of hallucination, but simultaneously making it a bad measure for comparison. As Goodhart’s law states: When a measure becomes a target, it ceases to be a good measure [46]. Due to this our thesis prioritizes qualitative visual analysis and quality assessment.

5.2 Ablation Study

We present three versions of IR-Zero-DCE to highlight the impact of the different design decisions concerning the loss functions. The *Default* version uses the exact architecture and losses described in chapter 4, excluding the Vertical Total Variation (VTV) Loss. The second version, *Denoising*, is identical to Default, except that it utilizes the VTV Loss, weighted at $W_{VTV} = 20.0$. The last version, *Edge-Only*, is identical to the default setup, but discards the background noise suppression component of the Edge-Enhancing Structural Loss, thereby making this loss equal to L_{edges} (section 4.4). Figure 5.1 shows how the different models perform on data from *Dataset_{Lab}*.

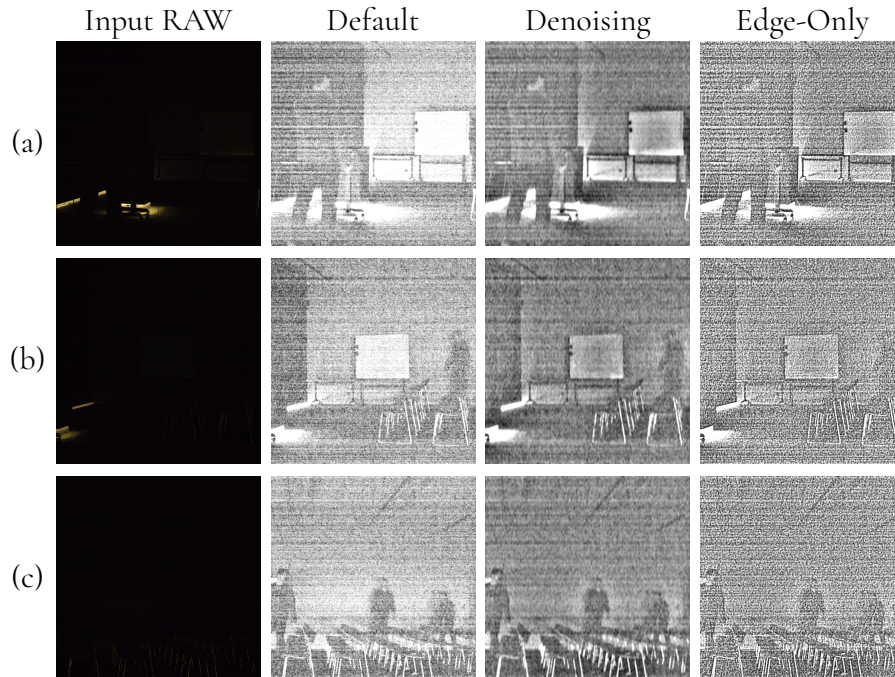


Figure 5.1: Visual ablation study of three different loss setups

While managing to suppress some noise, the denoising version fails to adequately address the complex noise patterns in the input. Without any reference ground truth RGB images, the model does not learn to distinguish noise from true signal in the image, and instead slightly blurs the entire image. The third version, which only rewards enhancing edges, is free to over-amplify background noise in the output. Without the background noise suppression component, the model produces a pattern of artifacts which might best satisfy the loss, but clearly degrades the background from a qualitative perspective. Through this analysis we find that our default setup performs the best, and is thereby the model we use for our comparison with the baselines

5.3 Comparison with Current Methods

To evaluate the efficacy of our proposed solution, we compare its performance to the traditional method Histogram Equalization (HE), as well as the deep learning frameworks our work is based on, Zero-DCE and DA-Zero-DCE. The results demonstrate that our method provides superior structural definition when working with grayscale images. Notably, HE performs surprisingly well which is further discussed in section 5.6.

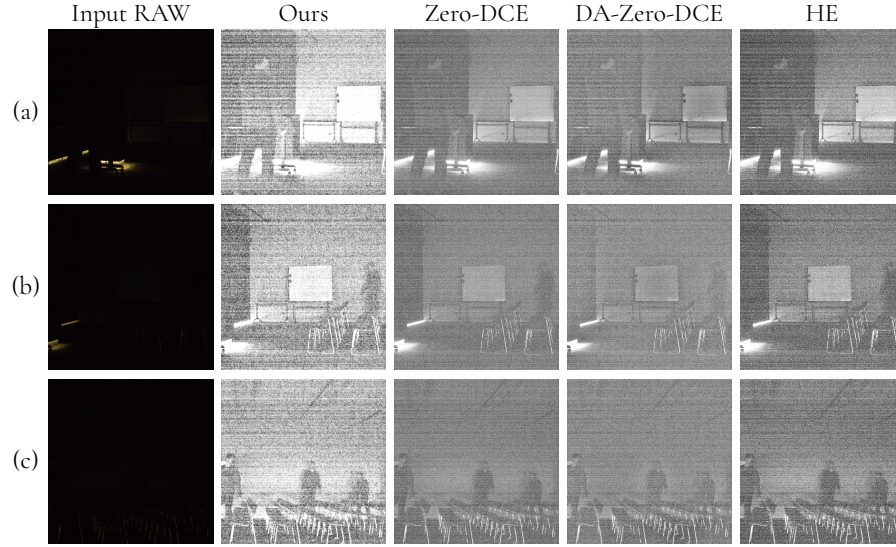


Figure 5.2: Visual comparison of baseline methods on images from *Dataset_{Lab}*. Compared to baseline methods, our proposed model increases contrast and clarity around structures and objects, such as people who emit thermal radiation.

In Figure 5.2(a) and (b), the proposed model demonstrates sharper local contrast around the whiteboard compared to the baselines. Similarly, in image (c) our method achieves a more distinct separation between the persons and the background, with improved contrast.

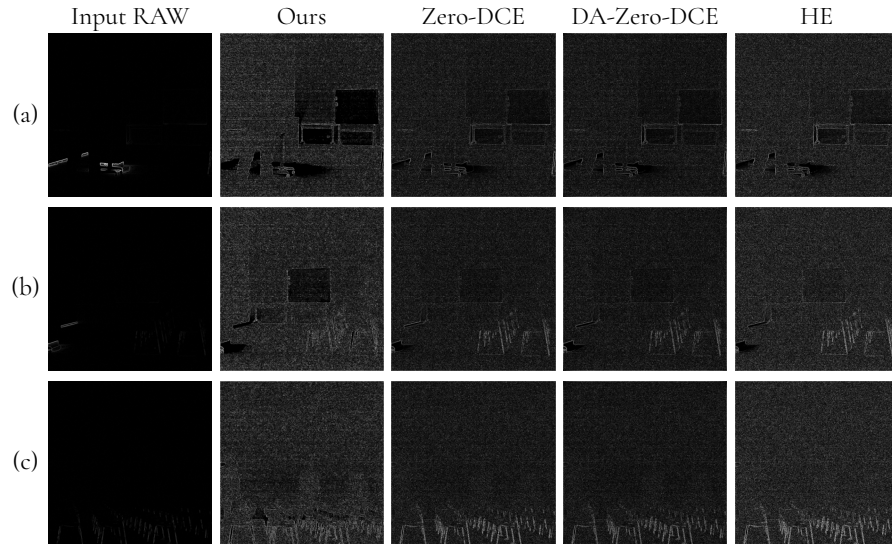


Figure 5.3: Edge maps of the outputs from Figure 5.2, computed with a 3×3 Sobel operator and visualized as the gradient magnitude $\sqrt{G_x^2 + G_y^2}$. All maps in this figure are normalized by a single shared maximum so that magnitudes are directly comparable across methods.

The Sobel gradient maps in Figure 5.3 demonstrate that IR-Zero-DCE achieves supe-

rior structural recovery. By leveraging spatial information from the IR images, the proposed method enhances the edges of the subjects. While this leads to a higher noise floor in the background compared to the baselines, the final output shows greater boundary clarity. This suggests that the model prioritizes the preservation of object geometries even at the cost of some background noise amplification. To highlight the improved boundary of the person, a cropped and enlarged view of image (c) from Figure 5.3 is presented in Figure 5.4.

Although the Sobel operator is an internal component of the L_{struct} loss, these results confirm that the network effectively translates cross-modal structural data into the enhanced output, reconstructing contours that are otherwise lost in the baseline methods.

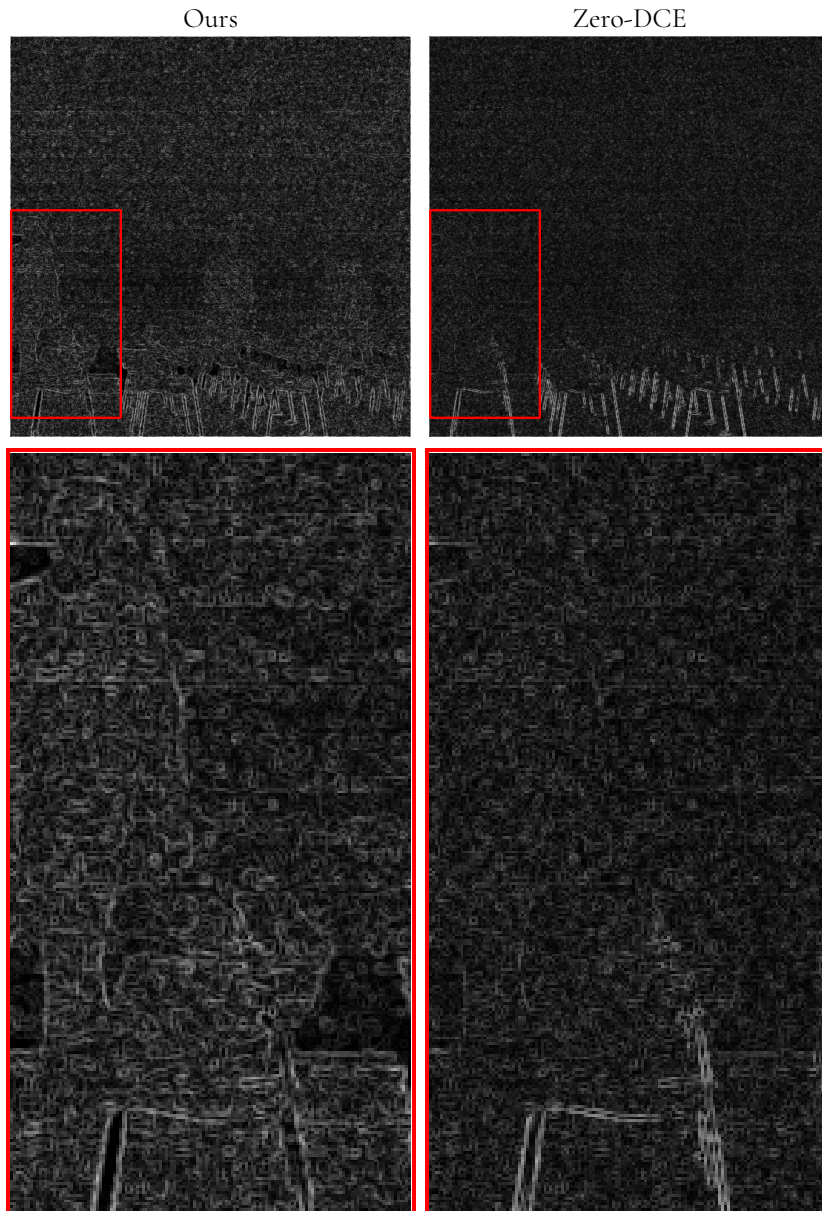


Figure 5.4: Cropped and zoomed in version of image (c) from Figure 5.3. The edges of the person is more prevalent in our version, although at the cost of amplified background noise.

5.4 LE-curves & IRE Residual

To showcase the separation of responsibilities between the curve-estimation backbone and the IR-based Enhancer, Figure 5.5 illustrates the effect the two different modules have on the image. Figure 5.5c shows the curve maps applied to the input image, while Figure 5.5d shows the residual which is added by the IR-based Enhancer.

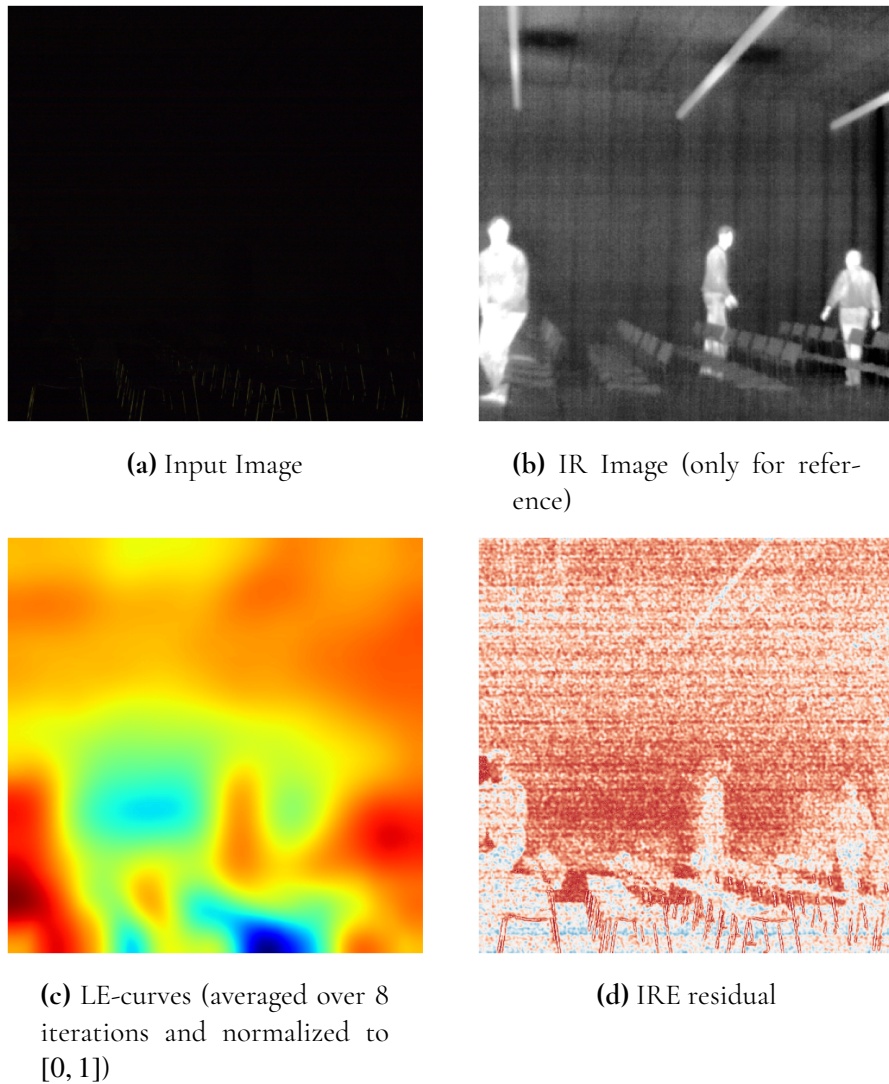


Figure 5.5: The different components of the enhancement pipeline visualized, when applied to an input image (a). The thermal image (b) is not used as input to the network, and is only provided for reference to show how the network learns to focus its enhancements. (c) shows an average over the curve parameter maps, $\mathcal{A}$, generated by the backbone. (d) shows the additive residual generated from the IR-based Enhancer.

This example shows a clear separation of responsibilities within our network. The backbone outputs smooth curve parameter maps responsible for the global illumination recovery

of the scene. The IR-based Enhancer on the other hand is responsible for localized, high-frequency adjustments and the sharpening of edges around objects.

This is similar to the two components which Retinex theory states an image is composed of, spatially smooth illumination (L), and high-frequency structural reflectance (R) [25]. Our architecture is constrained to separate these components, making it easy to reason about where different types of enhancements stem from. The application of LE-curves is strictly multiplicative, and acts as a scaling factor of the images illumination. Additionally, the heavily weighted TV loss encourages the backbone to output low-frequency parameter maps. The visualization in Figure 5.5c shows that the curve maps are entirely smooth, confirming that the luminance mapping which this part of the architecture performs resembles (L).

The IRE on the other hand is limited by a 9×9 receptive field, making it blind to any global context. Through guidance from the two components of the Edge-Enhancing Structural Loss, the additive residual which it outputs (Figure 5.5d) instead contains only high-frequency gradients which enhance edges around structures and people, as well as counteracting noise in the background, corresponding to (R). By separating responsibilities between the two stages of the pipeline and training them independently with specialized loss functions, we demonstrate how the model approximates Retinex decoupling in its image enhancement.

5.5 Hallucination Resistance

In lieu of quantitative measures of hallucination, here we establish an argument to why our model is hallucination resistant and highlight how the different components of our architecture contribute to our goals of providing both image-wide exposure correction and localized edge refinement around structures.

Image Enhancement is an Ill-Posed Inverse Problem

First, we reiterate that image enhancement is an ill-posed inverse problem, meaning the model is tasked with reconstructing the “true” image from degraded, noisy, and incomplete data in the input. In such problems, even small variation in noise can cause instability and drastically change the output, as there is no unique, stable solution due to the insufficient information in the input. To provide stability, and limit the output space to only the plausible solutions, image restoration models therefore often rely on either explicit prior knowledge or priors learned from distributions in the training data [12].

Why The Curve Estimation Backbone is Hallucination Resistant

With this background in mind, we can see that how the model manipulates the input, and what information it relies on to produce the output, is of utmost importance for reasoning about the model’s ability to hallucinate. To show how the generation of curve parameter maps in the backbone of our network differs from traditional image-to-image pixel generation, we compare this architecture to other methods such as Generative Adversarial Network’s (GAN)

and Diffusion models. These models often rely on encoder-decoder architectures, which compress an input image into a latent space, losing high frequency detail and noise information but capturing broader semantic meaning, and then recreate the image to produce a sharper, brighter, or noise-free output. The reconstruction performed is then based on both the input, but also the models’ semantic understanding of what an image is supposed to look like, its prior. In areas where the input lacks information, such as extremely dark areas, the prior dominates the generation of the output resulting in higher risk of hallucination [42]. Furthermore, due to the complexity, high-parameter counts and non-linear activation functions of many generative models, even a small difference in the noise pattern of the input can cause drastic variations in the output, something which is especially relevant in low-light RAW images where random shot noise is prevalent.

The generation of pixel-wise LE-curves, as performed in our architecture, is drastically different than the pixel-to-pixel approaches. Firstly, the model does not predict pixels directly, but instead estimates curves that are constrained by a predictable and limited mathematical formula (Equation 3.1). The manipulation of the pixels themselves only occurs in the curve application phase, where each output pixel’s value is directly based on the corresponding input pixel’s value, applied to the curve formula. Furthermore, the output values are strictly limited to the range $[0, 1]$ and due to the monotonicity of the curves, the relative contrast between neighboring pixels will be preserved [14]. The predictability and rigidity of how the actual pixel values are manipulated means small variations in noise does not have a considerable impact on the output of the model. It does however also mean that the model is limited in its ability to denoise, as it can not understand the semantics of the image and predict sharp high-frequency curves to negate the noise.

Decoupled Architecture

A natural critique of the IR-Zero-DCE architecture then arises regarding the IR-based Enhancer. While the curve estimator outputs smooth, monotonic curves, the IRE is by design capable of sharp, high-frequency adjustments. Furthermore, as it is trained with IR-supervision through the Edge-Enhancing Structural Loss, it must be examined whether this module learns a data-driven prior, which we have established is a root cause of hallucination. While these concerns are valid, our architecture mitigates the risk of hallucination through three specific constraints: mathematical bounding, limited receptive field, and gradient isolation during learning.

Firstly, similarly to the curve estimator, the IRE is limited in its output capacity. Instead of performing direct image-to-image generation, the IRE outputs a residual map which is then added to the enhanced image. This residual is further restricted by a scalar multiplier acting as a restraint on the magnitude of the residual and keeping the output grounded to the input.

Secondly, the receptive field of the IRE is restricted to a 9×9 window, limiting it from learning any global semantic prior. By restricting it to local adjustments, it lacks the spatial awareness to learn any semantics of the image. Instead what the network learns is highly localized gradient enhancement, a likelihood $P(x|IR)$ which teaches it to differentiate weak structural edges from pure noise. As argued by Sampson et al. the likelihood $P(x|IR)$ will not induce any hallucinations [42].

Finally, the training pipeline achieves strict decoupling between the IRE and the curve es-

timator backbone through gradient detachment. Gradients stemming from the Edge-Enhancing Structural Loss are explicitly prevented from flowing back from the IRE to the backbone. This ensures that the curve estimator U-Net, which has a large receptive field and is capable of semantic understanding, remains completely isolated from the IR-supervision. The backbone is therefore prevented from learning any structural priors, maintaining its role as an unsupervised hallucination-resistant global illumination enhancer.

Loss Function Tug of War

The robustness of the proposed network stems from a deliberate tension between the curve estimator backbone and the IR-based Enhancer. By separating the objectives of the network and training the two modules in isolation, we achieve both improved spatially aware global illumination and high contrast amplification without cross-contamination.

On the one side the backbone which is constrained by its zero-reference loss functions, specifically the TV loss, focuses on uniform, spatially aware illumination correction. This allows the network to deal with complex lighting scenarios containing both under and over-exposed regions. It possesses a global understanding of the scene, but is bound in its adjustment capabilities through the curve application mathematics and detachment from the IR supervision.

The IRE on the other hand lacks semantic understanding and is restricted to local adjustments, but exploits the IR-based guidance to enhance object contours in extremely low-light regions beyond what approaches like vanilla Zero-DCE is capable of.

This tension allows the network to accomplish two inherently contradictory goals: Maintaining smooth illumination gradients and sharply reconstructing structural details in low-light regions, while simultaneously ensuring both components involved in achieving these goals are hallucination resistant.

Perception-Distortion Tradeoff in Hallucination-Resistant Networks

While the proposed architecture is designed to remain resistant to hallucination, the strict adherence to the input signal introduces an inherent trade-off in visual quality. Because the network is constrained from inventing information where the input signal is weak, heavy noise present in the RAW input remains in the output.

It is important to note that the method through which generative image-to-image approaches such as GAN and Diffusion models achieve visually pleasing results is precisely through their reliance on a prior. Hallucinations in these models are an intended feature which allows the network to produce sharp, high-quality output without being entirely tethered to the input [42]. Through their training data they learn “dictionaries” which map noisy, low-quality semantic content in the input to their well-lit, sharp counterpart.

The Perception-Distortion Tradeoff [27] explicitly outlines the phenomenon that distortion and perceptual quality are at odds with each other in image generation tasks. Models cannot achieve both low distortion and high perceptual quality. The rise of perceptual loss functions [24], and architectures such as GAN [28] stems in part from the realization that traditional distortion metrics such as PSNR and SSIM do not align with aesthetic quality.

Our network therefore produces images that are noisier and less visually pleasing than what would be achieved with other state-of-the-art approaches such as Diffusion or GAN. However, in safety critical domains where hallucination of false structure is unacceptable, this sacrifice may prove necessary.

5.6 Analyzing Histogram Equalization Performance

As shown in Figure 5.3, HE performs surprisingly well compared to the deep learning methods in restoring structural visibility. A comparative analysis of HE and Zero-DCE helps clarify why this is the case, particularly by highlighting how the performance gap narrows when color information is disregarded.



Figure 5.6: Visual comparison of Histogram Equalization (HE) and Zero-DCE

In extremely dark image enhancement, the primary challenge is to map a highly compressed input distribution, where signal values are mostly near zero, into the full dynamic range of the output. This objective is fundamentally aligned with the mathematical operation of HE, which flattens the image histogram to maximize global contrast. Because `DatasetLab` is extremely dark and has a very narrow dynamic range, this contrast stretching is highly effective at revealing objects in the scene.

Furthermore, the transition to grayscale simplifies the noise profile by removing the chromatic artifacts that HE would otherwise amplify. While the deep learning methods like Zero-DCE do not struggle with this due to their learned exposure and smoothness constraints, this advantage is diminished when the images are evaluated in grayscale.

5.7 Limitations

While the qualitative and theoretical assessment of our method shows promise in both its enhancement capabilities and hallucination resistance, there are several constraints which

limit the performance. The ill-posed nature of low-light image enhancement and the tradeoff between noise suppression and structural fidelity affects the perceptual quality of our images and potentially the performance on downstream tasks such as object detection. In this section we critically evaluate the limitations in our approach, specifically in regards to performance on maritime domains and alignment issues between the RGB and IR images.

Performance on Maritime Dataset

To showcase how our model performs in a varied and realistic domain, Figure 5.7 compares our model to the baselines on images from *Dataset_{Maritime}*. It can be seen that the scenes are more complex, with a majority of the fine structural detail located in a small portion of the image along the horizon. Due to this, the misalignment of RGB-IR image pairs in our data becomes a greater issue, and our mitigation techniques of expanding the edge-maps to account for that misalignment instead results in the model blurring out details in order to match the broad edges (Figure 5.8). We can however see promise in our model in this setting as well. While producing slightly blurred result, it performs on par with HE in terms of increasing contrast around buildings and masts, making them stand out more from the background compared to Zero-DCE and DA-Zero-DCE. Where HE struggles with varied lighting conditions, our model retains natural lighting without over-exposing bright areas and under-exposing dark ones.

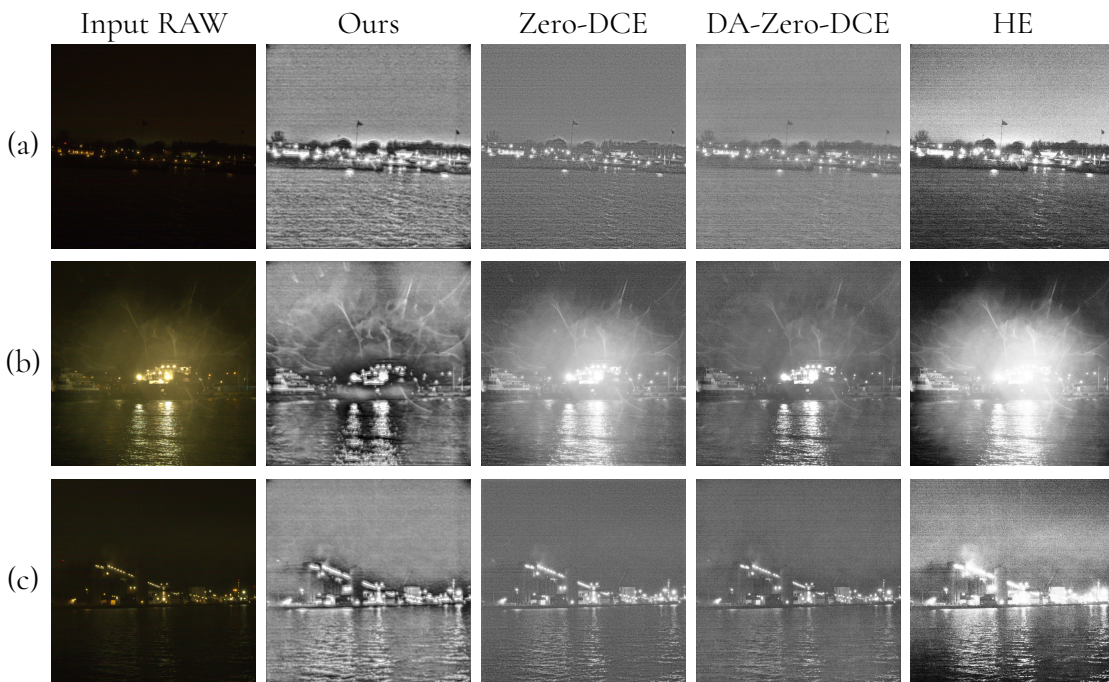


Figure 5.7: Visual comparison of baseline methods on images from *Dataset_{Maritime}*

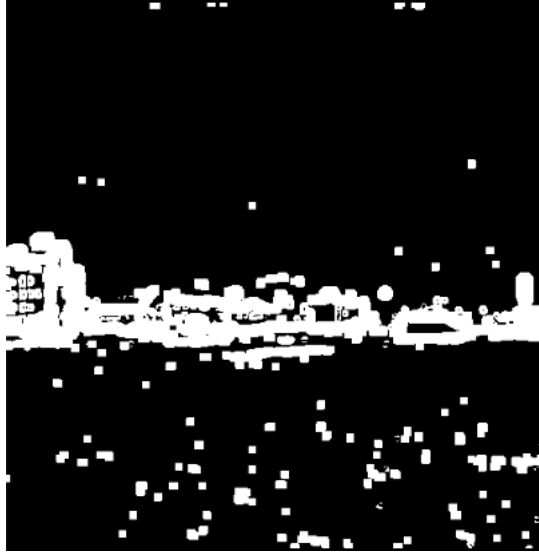


Figure 5.8: Due to the dilation of the edge map, when a lot of fine structural detail is concentrated in a small portion of the picture, the model is incentivized to blur the image in order to match the target edges.

In Figure 5.7 we see an example of a failure case for our model. Looking at the building to the left in image (c) we see that while attempting to increase the contrast of its borders, the facade of the building is a flat surface (which also doesn't contain any texture in the RGB) and therefore the background loss flattens out the area making the building "disappear" (Figure 5.9).

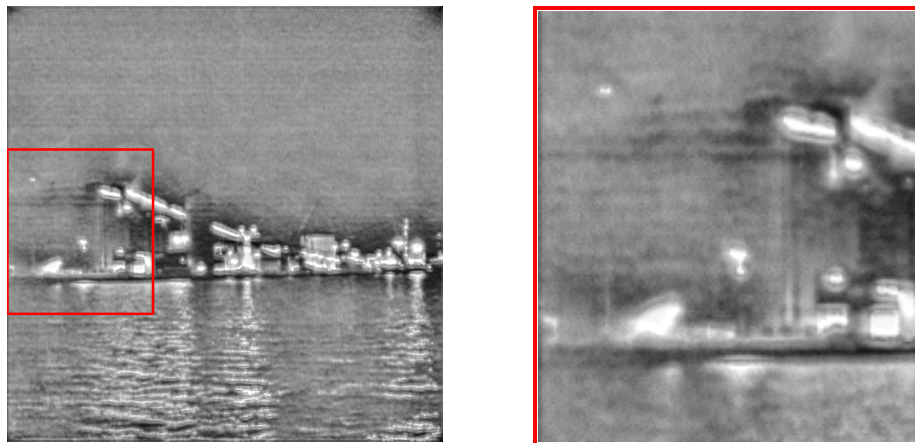


Figure 5.9: The model manages to bring out the edges of the building against the background, but is not guided towards also increasing the contrast of the facade.

In image (b) we can see a promising scenario for our model. It is the only model which attempts to suppress the difficult head-on lighting and extract the ship's structure behind it. While the result is not necessarily satisfactory in this particular case, a model that has the awareness to perform such an enhancement and keep the ship visible is clearly useful for many downstream applications such as detection, navigation and distance measuring.

Alignment Issues

One bottleneck for our solution is potential misalignment between the RGB and IR image pairs. As our Edge-Enhancing Structural Loss function compares gradients in the two images, even slight misalignment can cause a large spike in the loss, pushing the model to find alternative ways of satisfying the loss which do not correspond with our objective of increased contrast and sharp edges around structures. One illustrative example of this can be seen if the expansion of the target edge map M_{edges} is decreased. This expansion is performed to provide tolerance for some misalignment of geometric elements, and by decreasing it from 3 pixels in either direction (Figure 5.10a) to 1 pixel (Figure 5.10b), the model produces a blurry output as it is strictly punished for sharpening edges in the input which are not aligned with edges in the target map.

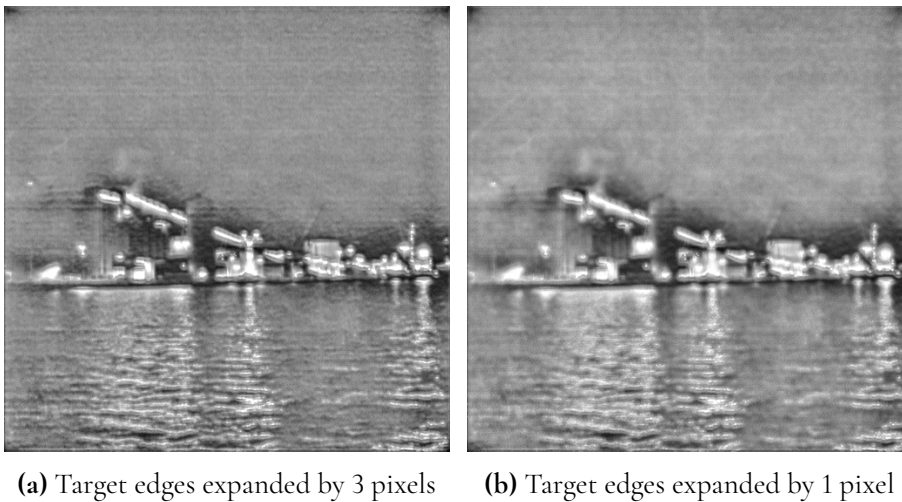


Figure 5.10: When the target edges are not wide enough to account for the misalignment between the RGB inputs and IR targets, the model blurs the output in an attempt to align the edges (b). By expanding the target edges, the model is free to create sharper edges in the output, as they now align with the expanded target (a).

Chapter 6

Future Work

This thesis has explored the complex challenge of extreme low-light image enhancement for safety critical applications where hallucinations pose a major risk. While we present a theoretically sound architecture to tackle such problems, several challenges remain for future work to examine.

6.1 Improved Alignment

A primary complication throughout our work has been the registration of RGB-IR image pairs. Due to the differences in domain, resolution, aspect ratio and FOV between the different cameras, both manual and automatic alignment has proved difficult. To account for the misalignment, several measures such as blurring and edge-map expansion were undertaken. While these were successful in improving the performance of the model on our dataset, we expect significant performance improvements if the images were perfectly aligned and the measures above would not need to be taken in the first place. This is especially important when, as is the case for our Edge-Enhancing Structural Loss, the loss relies on precise conformity between edges in both images. Measures such as mask dilation to account for misalignment will inherently make such a loss less precise and produce results which do not align with the true goal of sharp and clear structural boundaries. We believe the most robust way to achieve cross-modal pixel perfect registration for our application is to perform geometric camera calibration for the camera pair to calculate both the intrinsic and extrinsic parameters of the cameras. This would allow alignment that accounts for both differences in FOV, aspect ratio and distortion as well as physical offset and rotation.

An alternative approach involves moving away from an alignment sensitive loss function altogether. This could be achieved by utilizing a pre-trained feature extractor such as DINO [5] to distill the IR and RGB images into a shared feature-space. A loss function can then be formulated comparing the two sets of high-level features. By utilizing a dataset of paired IR and well-lit RGB images and training a small layer to project the IR feature space to RGB

feature space, such a loss could also lead our model to produce outputs which better capture the illumination, spatial relations and semantics of bright images. This is similar to what perceptual losses, such as those based on VGG, aim to capture [24].

6.2 Quantitative Evaluation

Downstream Performance and Image Quality

One of the main applications of our model is to improve performance on downstream tasks such as object detection, tracking, and classification. While we hypothesize that increased structural contrast and clarity would be beneficial, a key expansion on our work is to apply our model to such downstream pipelines and evaluate the difference in performance.

We could validate this performance using an established benchmark like the KAIST Multispectral Pedestrian Detection dataset [21]. Although these images are not in RAW format, are brighter than our use case, and capture urban rather than maritime environments, the dataset is relevant as it features paired RGB and IR images along with object detection annotations. Evaluating our model on the KAIST dataset would allow us to measure how the structural improvements impact the precision of downstream object detection networks. Testing object detection performance on this dataset would provide a proxy for semantic image quality and allow for a quantitative comparison against existing models that have already been benchmarked on the same data. It should be noted however, that performance on this benchmark does not necessarily generalize to extremely dark maritime environments.

Another evaluation framework for model comparison could utilize well-lit, paired RGB-IR datasets by applying synthetic degradation. By artificially darkening the well-lit images, we could utilize standard full-reference metrics such as PSNR or SSIM to measure the similarity between our model's enhanced output and the original, well-lit ground truth. While synthetically darkened images do not fully capture the complex noise characteristics of a true night scene, this approach could still provide a valuable comparison between models for image restoration.

Hallucinations

In this thesis we present a theoretical justification for the hallucination resistance of the IR-Zero-DCE architecture, especially against prior-induced hallucinations. By developing a cross-modal objective via a custom loss function, the model is grounded by IR-supervision to enhance structure strictly where it exists in the physical scene.

However, the qualitative nature of this evaluation presents an opportunity for continued research. Future work could be dedicated to developing a quantitative framework for measuring the degree to which low-light image enhancement networks hallucinate. We see two primary challenges in doing this for our application: First, we operate on datasets without reference ground truth images. This makes it difficult to distinguish fabricated details from the true structure of the scene. Second, if utilizing the paired IR images as a proxy for a structural ground truth during evaluation, there is a risk of optimizing the network for the measure itself. Because our proposed network is already trained to align predicted spatial gradients with thermal images via its loss function, using thermal alignment as an independent

test metric conflates the measure with the target. One potential solution to these challenges is to once again use the performance on an object detection benchmark as a proxy metric for hallucinations. It can then be argued that a model which performs well on such a benchmark likely does not inject false semantic or structural elements.

6.3 Train on Larger Datasets

This thesis was naturally limited in scope in terms of data acquisition and processing due to both time and resource constraints. While the performance of our model on our datasets shows a promising direction for the architecture, training on larger and more diverse data would likely lead to even better performance and highlight areas of the training process where the pipeline would need to be refined.

6.4 Reduce Noise in Edge Maps

The current implementation of the Edge-Enhancing Structural Loss uses a static percentile-based thresholding strategy which preserves the top 3% of the strongest gradients in each image. This introduces a vulnerability to inconsistent SNR in the dataset as it may inadvertently preserve noise induced artifacts and not just legitimate physical boundaries. To mitigate this, future research could explore different denoising strategies to be used before edge extraction, such as bilateral filtering.

6.5 Alternative architectures

In our approach we utilized structural features extracted from the IR images as privileged information directly applied to the output of the network through the Edge-Enhancing Structural Loss. This ensures the network learns structural constraints during training while only requiring RGB images at inference. One alternative approach to incorporate the thermal images during training is to employ a fusion architecture. These architectures utilize a “teacher” network trained on both RGB and IR input to extract some set of common latent features between the modalities, which are then used to guide the training of a “student” network. A comparison between this alternative approach of learning under privileged information through feature-based knowledge distillation, and the one presented in this thesis, could provide valuable insight into how the incorporation of IR features in different stages of the pipeline affects model performance. Particularly how our loss-guided architecture which enforces constraints directly on the output compares to the distillation-based fusion networks, which learn a latent representation, in terms of hallucinations.

Chapter 7

Conclusion

Low-light conditions remain a critical challenge for many tasks within computer vision, including autonomous systems relying on camera sensors for tasks such as navigation and object detection. While deep learning methods have vastly improved performance in image enhancement tasks, their reliance on difficult to acquire paired data remains a challenge. Furthermore, generative approaches which are not constrained to ground their output in the input signal may introduce unacceptable risk of hallucination in safety critical tasks. In this thesis we propose a decoupled architecture *IR-Zero-DCE* and a semi-supervised training approach utilizing a novel cross-modal Edge-Enhancing Structural Loss which incorporates guidance from thermal images. The proposed method is designed to achieve superior recovery of structure and contrast, while remaining resistant to hallucination. The primary objectives of this thesis were formalized in three research questions. Based on the evaluation of the proposed framework, these questions are answered as follows:

RQ1: *To what extent can IR images be used as privileged information during training to guide high-frequency enhancement of structure and contrast?*

Our results suggest that IR imagery can be effectively leveraged as privileged information to guide the enhancement of structural details. By incorporating IR-supervision through the Edge-Enhancing Structural Loss, we demonstrate that structural information from the thermal images can be successfully distilled into the network during training. Qualitative evaluations indicate that this approach increases the visibility and contrast of structures against the background, and outperforms the selected baselines in challenging low-light scenarios

RQ2: *How can an architecture be designed to provide both global illumination enhancement and local structural refinement in extreme low-light scenarios?*

We propose a decoupled architecture which separates the responsibility of low-frequency luminance estimation and high-frequency structural enhancement, similar to Retinex theory. By adopting the Zero-DCE curve estimation formulation for the backbone,

and residual learning strategy for the IR-based Enhancer, the model independently performs the two distinct types of enhancements.

RQ3: *Can loss functions and training constraints be formulated to punish hallucinated structure, while still encouraging strong contrast and detail recovery?*

By mathematically tethering the the output capacity of the model to the input signal, the network is constrained in its alteration capabilities. Furthermore, by restricting the receptive field of the IRE and blocking gradient updates stemming from the IR-supervision from reaching the backbone, the model is resistant to semantic or structural prior-induced hallucinations. By exploiting the superior structural information found in thermal images in low-light scenarios, the Edge-Enhancing Structural Loss aims to prevent the network from hallucinating structure in areas only containing pure noise.

The qualitative analysis in this thesis shows promising results for the IR-Zero-DCE framework on two difficult datasets with extremely low-light images. However, both IR-guidance and hallucination suppression are relatively unexplored in the literature, and more work is likely needed before methods such as the one presented here can reliably be applied for downstream tasks.

References

- [1] Eashan Adhikarla, Yixin Liu, and Brian D Davison. Diffusion models for low-light image enhancement: A multi-perspective taxonomy and performance analysis. *arXiv preprint arXiv:2510.05976*, 2025.
- [2] Mahdi Alehdaghi, Arthur Josi, Rafael MO Cruz, and Eric Granger. Visible-infrared person re-identification using privileged intermediate information. In *European conference on computer vision*, pages 720–737. Springer, 2022.
- [3] Sayantan Bhadra, Varun A Kelkar, Frank J Brooks, and Mark A Anastasio. On hallucinations in tomographic image reconstruction. *IEEE transactions on medical imaging*, 40(11):3249–3260, 2021.
- [4] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T. Barron. Unprocessing images for learned raw denoising. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11028–11037, 2019.
- [5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *CoRR*, abs/2104.14294, 2021.
- [6] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. *CoRR*, abs/1805.01934, 2018.
- [7] Joshua Cho, Sara Aghajanzadeh, Zhen Zhu, and D. A. Forsyth. Zero-shot low light image enhancement with diffusion prior, 2025.
- [8] Regev Cohen, Idan Kligvasser, Ehud Rivlin, and Daniel Freedman. Looks too good to be true: An information-theoretic analysis of hallucinations in generative restoration models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.
- [9] Jules R. Dim and Tamio Takamura. Alternative approach for satellite cloud classification: Edge gradient application. *Advances in Meteorology*, 2013(1):584816, 2013.

- [10] Xueyang Fu, Delu Zeng, Yue Huang, Yinghao Liao, Xinghao Ding, and John Paisley. A fusion-based enhancing method for weakly illuminated images. *Signal Processing*, 129:82–96, 2016.
- [11] Rikke Gade and Thomas B. Moeslund. Thermal cameras and applications: a survey. *Machine Vision and Applications*, 25(1):245–262, 2014.
- [12] Angela F Gao, Oscar Leong, He Sun, and Katherine L Bouman. Image reconstruction without explicit priors. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.
- [13] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [14] 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*, pages 1780–1789, 2020.
- [15] Hang Guo, Tao Dai, Zhihao Ouyang, Taolin Zhang, Yaohua Zha, Bin Chen, and Shu-tao Xia. Refir: Grounding large restoration models with retrieval augmentation. *Advances in Neural Information Processing Systems*, 37:46593–46621, 2024.
- [16] Majed El Helou and Sabine Süsstrunk. Bigprior: Towards decoupling learned prior hallucination and data fidelity in image restoration. *CoRR*, abs/2011.01406, 2020.
- [17] Judy Hoffman, Saurabh Gupta, and Trevor Darrell. Learning with side information through modality hallucination. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 826–834, 2016.
- [18] Gerald C. Holst. *Common Sense Approach to Thermal Imaging*. JCD Publishing and SPIE Optical Engineering Press, 2000.
- [19] Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. *CoRR*, abs/2103.02907, 2021.
- [20] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.
- [21] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multi-spectral pedestrian detection: Benchmark dataset and baselines. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [22] 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. *CoRR*, abs/1906.06972, 2019.
- [23] Yinan Jiang, Chaoliang Zhong, and Botao Zhang. Agd-linknet: A road semantic segmentation model for high resolution remote sensing images integrating attention mechanism, gated decoding block and dilated convolution. *IEEE Access*, PP:1–1, 01 2023.

-
- [24] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *CoRR*, abs/1603.08155, 2016.
 - [25] Ron Kimmel, Michael Elad, Doron Shaked, Renato Keshet, and Irwin Sobel. A variational framework for retinex. *International Journal of Computer Vision*, 52:7–23, 10 2003.
 - [26] Edwin H. Land. The retinex theory of color vision. *Scientific American*, 237(6):108–129, 1977.
 - [27] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016.
 - [28] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802, 2016.
 - [29] Yu Liu, Xun Chen, Juan Cheng, Hu Peng, and Zengfu Wang. Infrared and visible image fusion with convolutional neural networks. *International Journal of Wavelets, Multiresolution and Information Processing*, 16(03):1850018, 2018.
 - [30] Kin Gwn Lore, Adedotun Akintayo, and Soumik Sarkar. Llnet: A deep autoencoder approach to natural low-light image enhancement. *CoRR*, abs/1511.03995, 2015.
 - [31] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: A survey. *Information Fusion*, 45:153–178, 2019.
 - [32] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019.
 - [33] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016.
 - [34] Niharika Pachori. Low-light image super-resolution using GANs: A comprehensive comparative review. *Journal of Information Systems Engineering and Management*, 10:1023–1031, 05 2025.
 - [35] Vrushabh Patil, Savita Bagewadi, and M. Uma. RetinexGAN: A hybrid approach for low-light image enhancement. *International Journal of Communication, Science, and Technology*, 1:16–23, 01 2025.
 - [36] Zhen Pei, Jinbo Lu, Yongliang Qian, Lihua Fan, Hongyan Wang, and Jinling Chen. A new method for fusing infrared and visible images in low-light environments based on visual perception and attention mechanism. *Optics and Lasers in Engineering*, 186:108800, 2025.
-

- [37] Stephen M. Pizer, E. Philip Amburn, John D. Austin, Robert Cromartie, Ari Geselowitz, Trey Greer, Bart ter Haar Romeny, John B. Zimmerman, and Karel Zuiderveld. Adaptive histogram equalization and its variations. *Computer Vision, Graphics, and Image Processing*, 39(3):355–368, 1987.
- [38] R Priyah. and S. Kamalakkannan. Contrast limited adaptive histogram equalization approach contrast enhancement in liver tumor. In *2025 9th International Conference on Inventive Systems and Control (ICISC)*, pages 1–7, 2025.
- [39] Z. Rahman, D.J. Jobson, and G.A. Woodell. Multi-scale retinex for color image enhancement. In *Proceedings of 3rd IEEE International Conference on Image Processing*, volume 3, pages 1003–1006 vol.3, 1996.
- [40] Manuel J. C. S. Reis. Low-light image enhancement using deep learning: A lightweight network with synthetic and benchmark dataset evaluation. *Applied Sciences*, 15(11), 2025.
- [41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.
- [42] Matt L Sampson and Peter Melchior. Spotting hallucinations in inverse problems with data-driven priors. *arXiv preprint arXiv:2306.13272*, 2023.
- [43] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, 2009. Advances in Neural Networks Research: IJCNN2009.
- [44] Jörg Wagner, Volker Fischer, Michael Herman, Sven Behnke, et al. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *ESANN*, volume 587, pages 509–514, 2016.
- [45] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *CoRR*, abs/1808.04560, 2018.
- [46] Wikipedia contributors. Goodhart’s law — Wikipedia, the free encyclopedia, 2026. [Online; accessed 17-April-2026].
- [47] Yazhou Xing, Zian Qian, and Qifeng Chen. Invertible image signal processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6287–6296, June 2021.
- [48] Maoxun Yuan, Xiaorong Shi, Nan Wang, Yinyan Wang, and Xingxing Wei. Improving rgb-infrared object detection with cascade alignment-guided transformer. *Information Fusion*, 105:102246, 2024.
- [49] Syed Waqas Zamir, Aditya Arora, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. *CoRR*, abs/2111.09881, 2021.
- [50] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017.

- [51] Tong Zhou, Dongchun Li, Yudan Tian, and Yuxuan Li. Da-zero- dce. In *2023 China Automation Congress (CAC)*, pages 4249–4255, 2023.

Appendices

EXAMENSARBETE Exploiting Infrared Images for Hallucination Resistant Low-Light Image Enhancement**STUDENTER** Samuel Naumann, Karl Hammar**HANDLEDARE** Simon Kristofferson Lind (LTH), Mårten Lager (Saab)**EXAMINATOR** Eren Aksoy (LTH)

AI som ser i mörkret utan att hallucinera

POPULÄRVETENSKAPLIG SAMMANFATTNING Samuel Naumann, Karl Hammar

Moderna självstyrande system är ofta beroende av kameror, men deras förmåga att tolka omgivningen försämras drastiskt i mörkret. Genom att låta vår AI-modell dra nytta av infrarött ljus kan vi återskapa ljusa bilder ur extremt mörker. Detta ger framtidens autonoma teknik det mörkerseende som krävs för att ta säkrare beslut.

Många har upplevt hur AI-baserade chatbottar kan hitta på information som inte överensstämmer med verkligheten. På samma sätt kan AI-modeller som förbättrar bilder uppfinna saker som inte finns i originalbilden. Detta är ett stort problem för exempelvis självstyrande fordon som förlitar sig på kameror för att analysera sin omgivning. När en vanlig kamera placeras i ljusfattiga förhållanden så saknas informationen som den behöver för att fånga en bra bild. Genom att utnyttja infrarött ljus som fångar värmesignaturer i mörkret kan AI:n lära sig var objekt faktiskt befinner sig och därmed förstärka det riktiga motivet.

I vårt arbete har vi tränat ett neuronät på både mörka bilder och deras infraröda motsvarighet samtidigt. Istället för att gissa sig fram till var en byggnad eller person finns i omgivningen, så kan IR-data användas som ett facit. Den resulterande modellen kan förbättra synligheten av intressanta objekt i extremt mörka bilder, utan att hallucinera.

Mörkerseende är avgörande för att autonoma fordon och robotar ska kunna navigera och undvika hinder. Det handlar inte bara om vad som är precis framför fordonet, den måste också se objekt på långt håll för att kunna ta effektiva beslut. Om AI-modellen hallucinerar och skapar falska objekt i bilden, riskerar det att leda till felaktiga och di-

rekt farliga beslut.

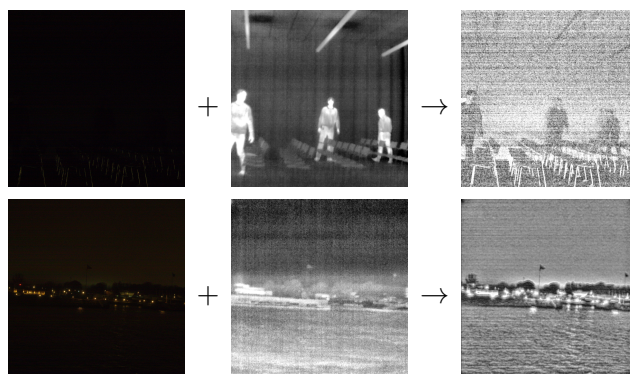


Bild 1: Genom att kombinera den vanliga kamerabilden (vänster) med infraröd data (mitten) kan AI-modellen återskapa en tydlig bild (höger) trots extremt mörker.

Vi lever i en värld där AI kan generera uppfinningsrika bilder med ett enda knapptryck. Men när ett fartyg navigerar okänt vatten mitt i natten så kan precis samma förmåga att fabricera vara livsfarlig. Genom att utveckla AI som håller sig till sanningen bygger vi inte bara system som kan se i mörkret, vi bygger system man kan lita på när det verkligen gäller.