Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

MASTER THESIS

Reflection Detection in Inspection Images for Reactive Planning of Autonomous Inspections

Aditi Mhatre

MIN-Faculty

Department of Informatics

Course of Study: MSc Intelligent Adaptive Systems

Matriculation Number: 7593744

First Reviewer: Prof. Dr. Simone Frintrop

Second Reviewer: Dr. Marc Bestmann

Adviser: Dr. Ehsan Yaghoubi

Abstract

The adoption of automatic inspection systems is growing across various industries, such as manufacturing and energy, and is expected to expand significantly into other sectors, including aerospace. The potential for automation in this field is substantial, promising advancements in visual inspection to improve decision-making and performance. These systems often encounter challenges when inspecting highly reflective metallic surfaces, where varying light conditions can obscure critical surface details. Such limitations not only compromise inspection accuracy but also pose potential risks to safety. To address these issues, this thesis explores the implementation of various U-Net-based architectures for detecting specular light reflections in inspection images, facilitating reactive planning during autonomous inspections. A novel dataset comprising inspection images and corresponding masks of light reflections is introduced, serving as a foundation for training the U-Net models.

Key findings reveal that CNN-based U-Nets significantly outperform their Transformer-based counterparts, with U-Net++ featuring a ResNet-50 encoder yielding the highest Intersection over Union (IoU) and Dice Similarity Coefficient (DSC) scores. In contrast, the proposed UNETR-Attention Fusion (UNETR-AF) struggles to detect larger reflections but performs comparably for medium and smaller reflections. This research contributes valuable insights into industrial inspection applications focused on reflective surfaces. While the findings predominantly pertain to 2D RGB images, future work may explore the adaptation of these techniques to RGB-D images to capture additional depth information, potentially improving the efficacy of reactive planning in autonomous inspections. Furthermore, the application of generative AI could facilitate the creation of expansive datasets, while few-shot learning methods may be employed to mitigate data scarcity challenges.

Keywords: Specular Reflection Detection, Image Segmentation in Inspection, U-Net, Vision Transformer Architecture, Reflections in Inspection images, Inspection dataset, Autonomous Visual Inspections.

Contents

List of Figures

List of Tables

List of Abbreviations

Adam	Adaptive Moment Estimation.
AG	Attention Gate.
ASPP	Atrous Spatial Pooling Pyramid.
BCE	Binary Cross Entropy.
BER	Bit Error Tate.
CARNet	Channel Attention Refinement Network.
CBAM	Convolutional Block Attention Module.
CNN	Convolutional Neural Network.
DLR	Deutsches Zentrum für Luft- und Raumfahrt.
DRM	Dichromatic Reflection Model.
DSC	Dice Similarity Coefficient.
DSCFA	Dilated Spatial Contextual Feature Aggregation.
DVI	Detailed Visual Inspection.
ECA	Efficient Channel Attention.
ESCAB	Efficiency Spatial Channel Attention Block.
ET-HDR	Efficient Two-Stage Highlight and Removal Network.
FAA	Federal Aviation Administration.
FCN	Fully Convolutional Networks.
FFB	Feature Fusion Block.
FTL	Focal Tversky Loss.
GVI	General Visual Inspection.
HSV	Hue Saturation Value.
IoU	Intersection over Union.
JSHDR	Joint Specular Highlight Detection and Removal.
mIoU	Mean Intersection over Union.

MLP	Multi-layer Perceptron.
MRO	Maintenance, Repair, and Overhaul.
NDI	Non-Destructive Inspection.
PNG	Portable Network Graphics.
ReLU	Rectified Linear Unit.
ResNet	Residual Network.
RGB	Red Green Blue.
RGB-D	Red Green Blue Depth.
ROI	Region of Interest.
ROS 2	Robot Operating System 2.
SDVI	Special Detailed Visual Inspection.
SE	Squeeze-and-Excitation.
SGD	Stochastic Gradient Descent.
SHIQ	Specular Highlight Image Quadruples.
SIHRNet	Single Image Highlight Removal Network.
SNR	Signal-to-Noise Ratio.
SRG	Seeded Region Growing.
SSIM	Structural Similarity Index Metric.
t-SNE	t-Distributed Stochastic Neighbor Embedding.
TL	Transfer Learning.
UNETR	UNet Transformer.
UNETR-AF	UNet Transformer Attention Fusion.
ViT	Vision Transformer.
WAI	Walk-around Inspection.

Chapter 1

Introduction

Robots are increasingly being adopted by the industrial sector to perform automatic inspections. In the aviation industry, visual inspection constitutes more than 80% of inspection procedures [88], representing a substantial amount of labor hours. These inspections, which are still predominantly performed manually by human operators, require a significant workforce and contribute greatly to the overall operational costs. Moreover, the cost of maintenance during the operational phase is the highest and most unpredictable among all phases of an aircraft's life cycle [71]. This reliance on manual processes and the unpredictability of costs highlight the potential benefits of automated inspection solutions, which can reduce labor costs and improve efficiency.

Inspections form a large part of aircraft Maintenance, Repair, and Overhaul (MRO). MRO activities also include preventive, corrective and predictive maintenance, such as repairing or replacing system parts based on their quality condition [126]. Inspections are foundational to MRO operations, as they aid in identifying components that require replacement or repair to maintain optimal performance and safety standards. These inspections take place under varying lighting conditions, from the expansive external aircraft wings to narrow hydrogen fuel tanks. In such enclosed spaces, light sources produce reflections on metallic surfaces which can obscure critical details, compromising the reliability of the inspections. This poses significant safety risks, especially in the aviation industry where human lives are at stake. Visual inspections of aircraft involve multiple observation processes to identify irregularities and ensure the safety of the vehicle [126]. Therefore, developing methodologies that effectively identify these reflections is essential, allowing robots to revisit affected areas for more reliable and comprehensive inspections.

This thesis is completed in collaboration with the Deutsches Zentrum für Luft- und Raumfahrt (DLR) at the Institute for Maintenance, Repair and Overhaul, Hamburg as part of the ongoing research in autonomous inspections at the Robot-Assisted Inspection and Repair group.

1.1 Problem Statement

Machine vision is one of the key technologies in the field of visual inspection systems that enables automatic processing of images of an object through optical devices and sensors. Its efficiency and reliability has significantly improved with the integration of artificial intelligence. Due to this, it is possible that machine vision will partially replace the current manually performed visual inspections. The Federal Aviation Administration

(FAA) classifies visual inspections into four categories [24][126]: walk-around inspection, general visual inspection, detailed visual inspection and special detailed visual inspection.

- Level 1: Walk-around Inspection (WAI) is a general check conducted from ground level by either flight or maintenance personnel to identify discrepancies affecting aircraft performance; it is performed periodically and requires the aircraft to be clean and accessible. This includes checking for items that affect safety, legality, efficiency and comfort [24], such as checking general condition of the paint, observing the fuselage and aircraft wings and looking for major dents in the exterior of the aircraft.

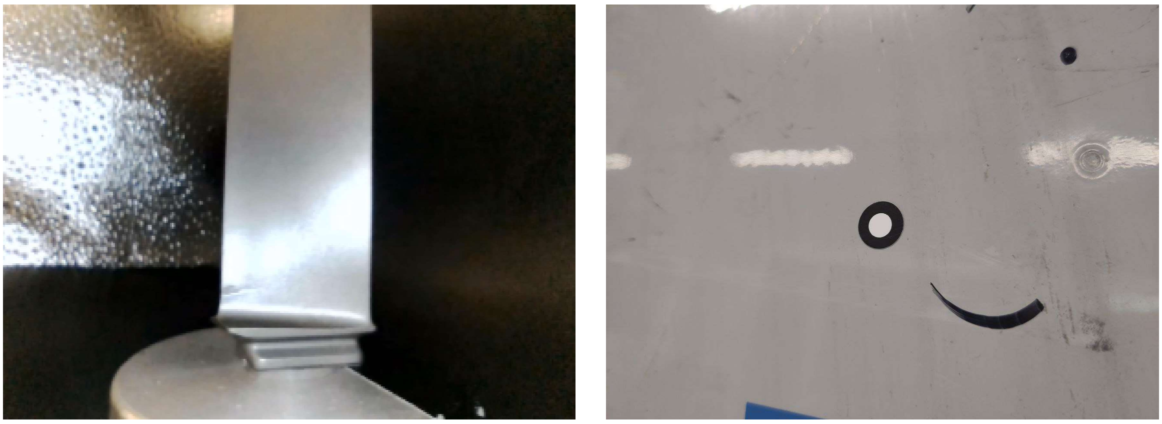
- Level 2: General Visual Inspection (GVI) involves a broader examination, often requiring tools for accessing panels and ensuring the aircraft's cleanliness, typically conducted when a specific problem is suspected. This level goes beyond observing and involves moving all parts possible, such as applying weight to load bearing components and viewing the object under different light conditions [24]. The findings of this level may lead to Level 3 or Level 4.

- Level 3: Detailed Visual Inspection (DVI) is an intensive evaluation of a specific area or system, utilizing a variety of specialized tools and requiring thorough preparation and documentation when further investigation is warranted. This requires reviewing the aircraft history and accident reports. This inspection occurs when a problem is detected, i.e corrosion or crack, and the surrounding area is inspected for failure, damage or irregularity [24]. New discoveries in this inspection may lead to Level 4.

- Level 4: Special Detailed Visual Inspection (SDVI) focuses on intricate components and often involves specialized techniques, complex disassembly, and advanced tools to ensure airworthiness, particularly for damage-tolerant aircraft. This tier is typically invoked based on prior inspections or specific directives [126]. The procedures carried out are centered on visual inspection but may incorporate Non-Destructive Inspection (NDI) techniques, such as dye penetrant testing or borescope imaging, to enhance the detection of irregularities in inaccessible areas. These inspections often target portions of the aircraft that require disassembly to access, such as lap joints and the interior surfaces of the aircraft wing skin [24].

Walk-around and general visual inspections are periodically performed by the aircraft maintenance and operating personnel to check for damages. Most of the damages caused to fuselages and the exterior of the aircraft are due to impacts with objects during flight or maintenance with objects, hails, lightning strikes and birds [126].

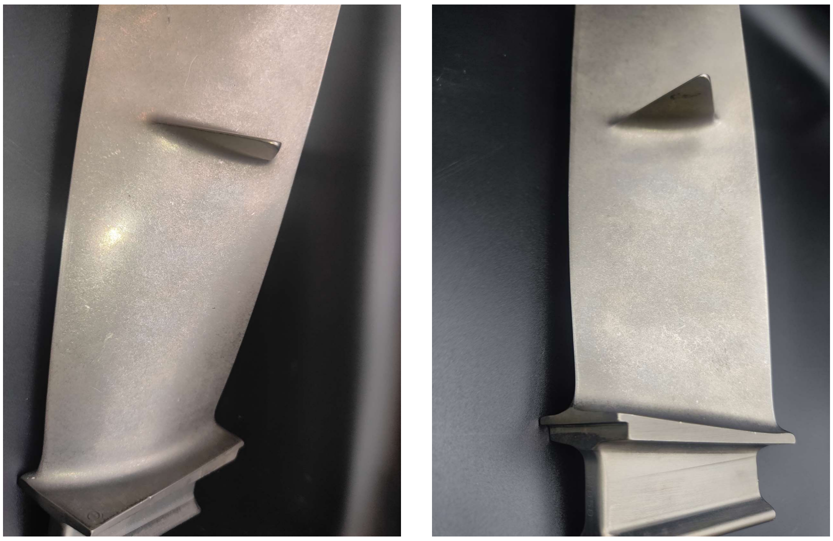
Aerospace and aviation industries widely use metal alloys or composites which have highly reflectively surfaces. These materials are often painted but the paint is also very reflective. These industries have high requirements for surface quality [92]. Autonomous inspection in this field are used to detect defects in metal and composite fuselage. An industrial visual inspections system consists of three components: optical illumination, image acquisition and image processing [92]. The hardware components are responsible for optical illumination and image acquisition tasks, which include light source, illumination modes, and image acquisition schemes used by the robot. While image processing is the software-based component which is responsible to find useful information from the captured images. These tasks include image preprocessing, classification, localization, and segmentation of defects. In most autonomous inspections robots, the bright field forward lighting illumination mode (see Section 2.1.2) is employed wherein the light source and the camera are located at the same side of the object. This illumination mode is preferred to capture the surface details however it produces specular reflections on reflective surfaces such as metal. Reflections are also created intentionally to identify some defects, especially to show small deviations in the surface. However, in certain cases, specular reflections can occlude the finer surface details which makes it challenging to reliably inspect the surfaces and meet the high inspection standards. A depiction of two use cases involving reflections - one where they are obstructive and another where they are intentional - is shown in Figure 1.1. Figure 1.1(a) illustrates an image of a turbine blade where a significant portion of the blade has reflections due to its proximity to the light source. The reflection hides the surfaced details of the object, making it challenging to inspect it. While Figure 1.1(b) illustrates a use case where reflections are intentional and helpful in revealing details of the surface.


(a) Reflections hide details

(b) Reflections reveal defects

Figure 1.1: Comparison of how reflections can either hide (left) or reveal (right) surface details in inspection images. In the left image, reflections obscure critical details, making it difficult to inspect the surface, while in the right image, reflections help highlight defects.

Various optical camera systems are widely used in industries to analyze strain and displacement fields in various materials and structures. These optical systems are affected by various external influences such as the test environment conditions and out-of-plane motion, due to which it is a considerable challenge to obtain high quality images [87]. The lighting conditions in the testing environment vary based on the area under examination. External surfaces like airplane wings are subjected to both environmental light and inspection lighting sources. In contrast, surfaces such as hydrogen fuel tanks are inspected indoors under a single light source, devoid of natural light. In industries with stringent surface quality standards, meticulous examination of the complete surface is essential. Areas where reflections obscure details must be reexamined to ensure thorough and reliable inspection. In autonomous inspection systems, it is vital to reposition the camera to find an optimal angle that allows for surface inspection without the interference of reflections. To achieve this, the system must first identify and localize reflections in the image, using this data to guide the camera's repositioning. This process enhances the system's reactive planning capabilities, enabling the robot to adjust its camera position dynamically for a more accurate and effective inspection.



(a) Image with Reflection (b) Image without Reflection

Figure 1.2: Comparison of an inspection image with reflection hides details (left) against the same image without reflection (right). In the left image, reflections make it difficult to understand whether that portion of the turbine blade has any defects or not, while in the right image captured from a different angle without reflection, it is easier to observe that there are no visible defects.

1.2 Research Question

Deep-learning and machine learning have been gaining popularity in the field of visual inspection. Recent research within computer vision include detection of cracks [125] [58] and aircraft dents [11]. Supervised deep-learning approaches are dependent on the quality of the data used. The presence of reflections in the images may deter the reliability of the results, which is detrimental for maintenance purposes.

The challenge of reflections in images has been studied across various domains, particularly in medical segmentation for procedures such as cervical cancer screenings [48] and endoscopic procedures like colonoscopy [72] where reflections pose a risk of misdiagnosis. The detection and removal of reflections in medical images has been widely researched, yielding numerous state-of-the-art methods. However, these methods are often focused on post-processing, where the primary goal is improving diagnostic accuracy after image acquisition. While effective in controlled environments, these methods do not account for real-time adjustments during image capture, which is crucial in other fields, such as industrial inspections.

In contrast, reflection detection in autonomous visual inspections is essential for real-time decision-making. Autonomous inspections necessitate the use of hybrid systems, which integrate various components: data acquisition systems for capturing and digitizing the inspected parts, sensors for detailed data collection, robots to automate sensor movement, and processing systems to analyze the data and detect irregularities or patterns [126]. Reflections not only obscure defects but also hinder the inspection process, requiring immediate intervention to ensure the inspection area is fully covered. The key challenge lies in handling these reflections dynamically during inspections to avoid compromising the safety and efficiency of the process. Reflection detection plays a critical role here—not just to enhance image clarity, but trigger reactive planning actions such as camera repositioning or adjusting inspection parameters. As environmental factors like lighting conditions, surface materials, and object orientation vary significantly in industrial settings [87], it can create challenges that standard post-processing techniques cannot address.

Therefore, the research question driving this work is:

How can semantic segmentation methods based on the U-Net architecture be adapted and optimized to effectively detect and classify reflections in inspection images under varying light conditions to support reactive planning in autonomous inspection systems?

1.3 Objective and Contribution of the Thesis

The primary objective of this thesis is to develop and evaluate a specialized reflection detection system using different semantic segmentation techniques. Given the real-time needs, semantic segmentation methods based on the U-Net architecture, offer a promising approach. U-Net's pixel-wise classification ability is well-suited for distinguishing reflections from key object features in images. However, the challenge is to adapt and optimize these models to manage the conditions found in industrial environments, where reflections can vary due to material properties and lighting changes. A key component of this research is the creation of a specialized dataset that mimics real-world inspection scenarios in aviation, focusing on reflection-heavy images taken under various conditions. This novel dataset comprises of RGB images along with their corresponding binary segmentation masks. Additionally, this thesis introduces a novel hybrid architecture that integrates a Vision Transformer with a Convolutional Neural Network, termed UNETR Attention Fusion (UNETR-AF).

To assess the effectiveness of different semantic segmentation models on this dataset, the models are trained and evaluated on the newly created inspection image dataset as well as on two existing datasets of specular reflection images of real-world objects captured in varying light conditions. Although these datasets are not focused on inspection images, it serves as a comparative baseline to determine the robustness and adaptability of the models to different types of reflections.

The thesis contributes by comparing the performance of models trained on all the datasets against a common test set of inspection images. This comparison reveals whether models trained on the inspection-specific dataset yield superior performance for the task at hand. Additionally, the thesis explores various semantic segmentation models to identify the one that performs best in detecting reflections within inspection images, which is critical for improving the accuracy and reliability of autonomous inspections.

1.4 Organization of the Thesis

This thesis is organized into several chapters that encompass a wide range of relevant topics in both breadth and depth. Chapter 1, titled Introduction, outlines the problem statement, research question, objectives, and contributions of the study, establishing a clear context for the research. Following this, Chapter 2, Background Information, delves into fundamental concepts related to light reflections and traditional image segmentation methods, as well as modern deep learning approaches. It includes discussions on various types of light reflections, illumination modes, neural network architectures, loss functions, optimizers, and evaluation metrics. Chapter 3, Related Works, reviews existing reflection detection methods, both traditional and deep-learning-based, alongside available specular reflection datasets relevant to the research.

In Chapter 4, Methodology, the thesis discusses the implementation details, covering the novel contributions such as the proposed UNETR-AF method and Inspection dataset, as well as other models employed, including U-Net, Attention U-Net, U-Net++, and UNETR. Chapter 5, Experiments and Results, presents the experimental setup, including datasets, model architectures, training processes, and the results obtained. The subsequent Chapter 6, Discussion, provides an in-depth analysis of the results, comparing them to state-of-the-art and traditional methods while discussing the strengths and limitations of each model. Finally, Chapter 7, Conclusion, summarizes the thesis's findings and proposes ideas for future work, reflecting on the overall contributions and implications of the research conducted.

Chapter 2

Background Information

Deep-learning involves a range of modules that assist in learning the task at hand, in addition to the model architecture. This chapter covers the background information needed to understand the components and functions implemented within the network architecture in deep-learning, and traditional image segmentation methods as well as fundamentals of the reflections in inspections.

2.1 Light Reflections

Inspections vary across different domains due to a number of factors like the shape and material of the object being inspected, the light conditions, and the type of data being captured. As the conditions for visual inspections differ depending on the application area, so does the types of reflections produced for any inspection task. This section covers the basic principles of reflections and the illumination modes used in machine inspections.

2.1.1 Types of Light Reflection

Light reflections occur when the light ray bounces off a surface instead of absorbing or transmitting through it. The nature of the light reflection depends on the surface's properties and the angle at which the light ray strikes the surface, known as the angle of incidence. This is explained by the law of reflections which states that the angle of incidence is equal to the angle of reflection [94]. The angles of incidence and reflection are measured with respect to a line perpendicular (normal) to the reflecting surface, as shown in Figure 2.1.

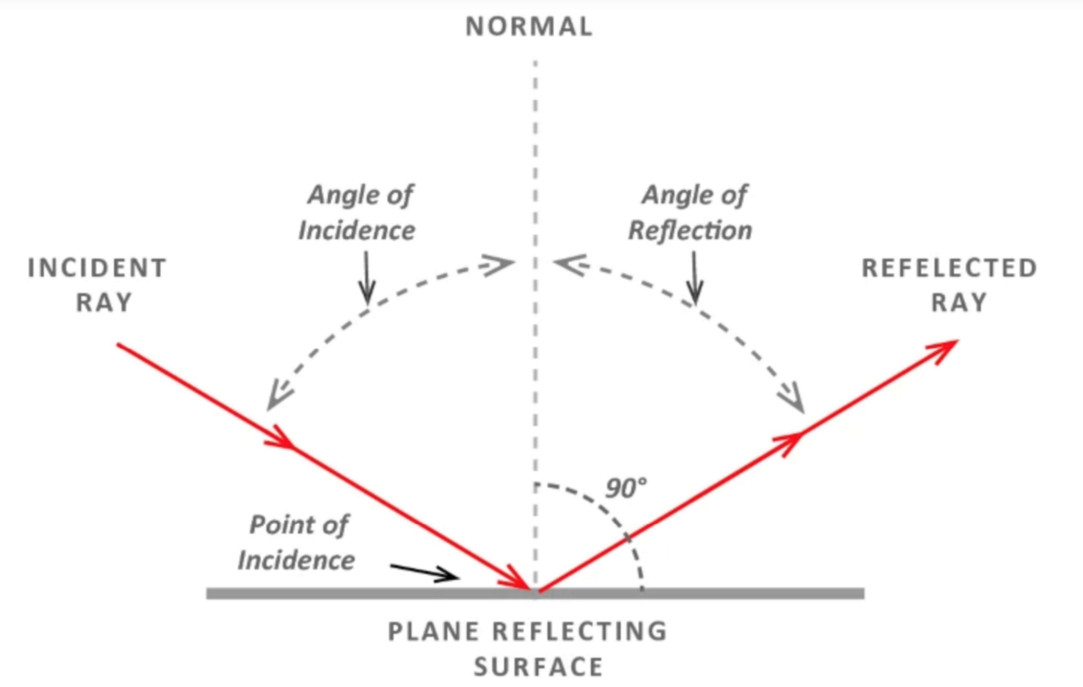


Figure 2.1: Law of Reflection: the angle of incidence is equal to the angle of reflection. The angles are measured from the normal or the perpendicular. The image was taken from [94].

There are two main types of light reflections: specular and diffuse. Specular reflection takes place when light reflects off a polished and shiny surface such as a mirror or a metallic surface in which the rays remain parallel [94] [108]. In this case, the angle of incidence (θ_i) equals the angle of reflection (θ_r), both measured relative to the surface normal ($\hat{n}$). This creates a sharp and clear reflection. This is expressed mathematically as:

$$\theta_i = \theta_r. \tag{2.1}$$

Diffuse reflection, on the other hand, occurs when light strikes a rough or uneven surface, causing the light rays to scatter and creating a blurred or soft reflection [94]. It follows Lambert's Cosine Law, [31], which states that the reflected light intensity (I) is proportional to the cosine of the angle between the incident light direction and the surface normal:

$$I = I_0 \cos(\theta), \tag{2.2}$$

where I_0 is the incident light intensity and θ is the angle between the incident light and the normal to the surface. This law is crucial for understanding how light interacts with non-specular surfaces, such as in industrial inspection, computer graphics, and remote sensing applications. According to the law, the illumination the surface is directly proportional to the cosine of the angle between the illuminating source and the normal. The concept is visualized in Figure 2.2.

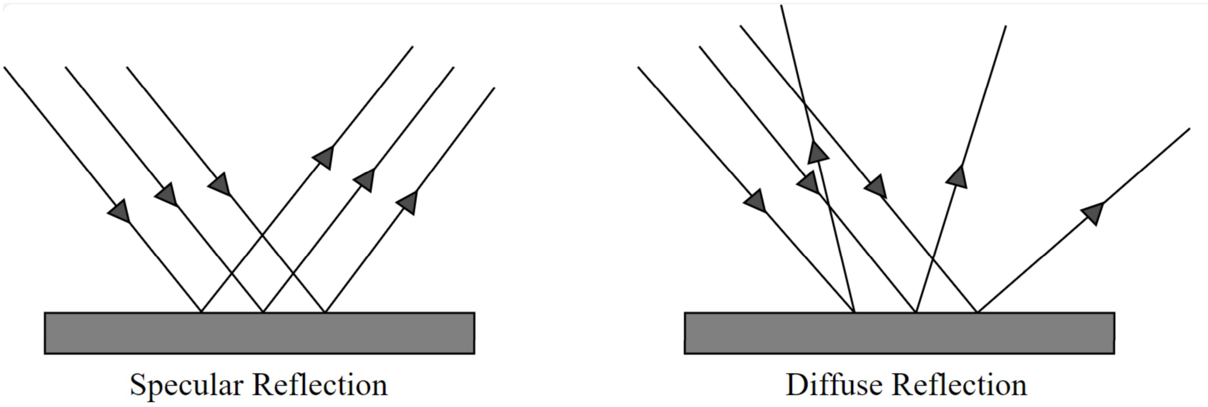


Figure 2.2: Light Reflection: Specular and Diffuse. The diagram is redrawn from [94].

Figure 2.3 illustrates the distinct appearances of specular and diffuse reflections in a real-world scenario. Specular reflections are highlighted with red bounding boxes, while diffuse reflections are indicated with blue bounding boxes. The object in the image comprises various materials. The specular reflections on the metallic screws make it challenging to discern its surface details. In contrast, surfaces with diffuse reflections reveal some level of surface texture, providing more visibility of the underlying details.

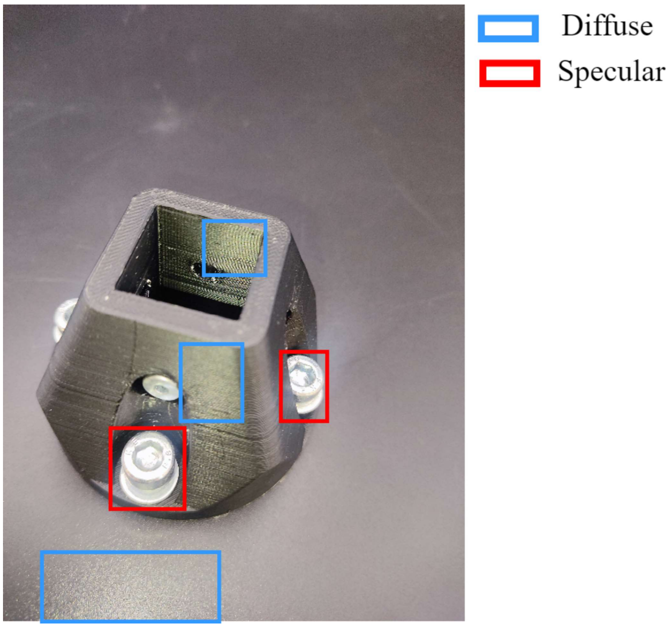


Figure 2.3: Light Reflection Example: Specular and Diffuse.

2.1.2 Fundamental Illumination Modes

There are numerous illumination options employed in industries in order to capture the best possible images of an object depending on the purpose. Some of the fundamental illumination modes are depicted in the Figure 2.4. The illumination modes reflect the positional relationship between the light source, camera and the object [92]. There are five key illumination modes: bright field forward lighting, dark field forward lighting, coaxial forward lighting, scattering forward lighting of dome structure, also known as diffuse lighting, and the back lighting, which are depicted in Figure 2.4.

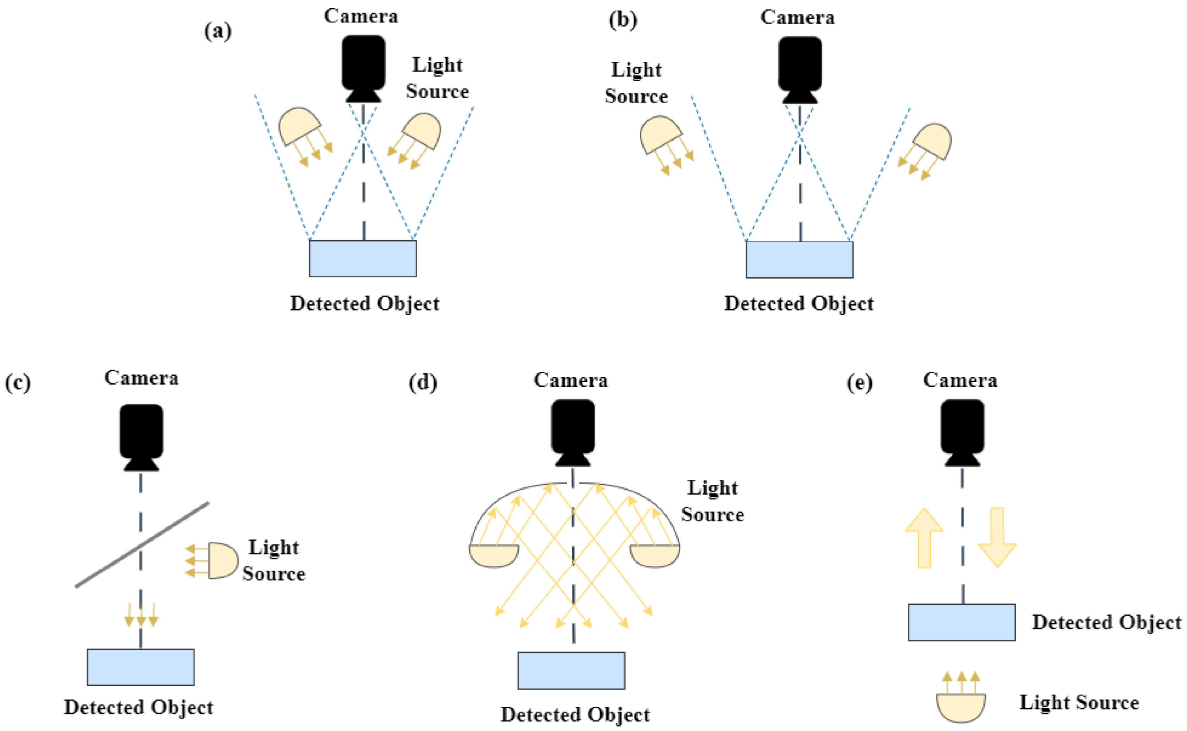


Figure 2.4: Types of Illumination modes: a) bright field forward lighting; b) dark field forward lighting; c) coaxial forward lighting; d) scattering forward lighting of dome structure or diffuse lighting; e) back lighting based on [92].

Forward lighting is the most widely used method wherein the the camera and the light source are placed on the same side with the object across them. It is mainly used as it creates a good constrast and enhances the surface details. This is suitable for capturing the surface texture and detecting surfaces defects. Depending on the angle in which the light is reflected on the camera, this mode can be split into bright field forward and dark field forward lighting. Bright field light is when the light is directly placed in front of the object. This is the most commonly used approach for surface defect detection because it provides uniform illumination and ensures that surface irregularities are prominently highlighted, making defects easier to identify. Although it highlights the surface textural details, it produces specular reflections in reflective surfaces. In dark field lighting, the incident angle of the light is reduced, which helps to highlight the edges of the surface as well as surface concavity and convexity [92]. The difference between the two light settings on a peanut brittle bag are shown in Figure 2.5.

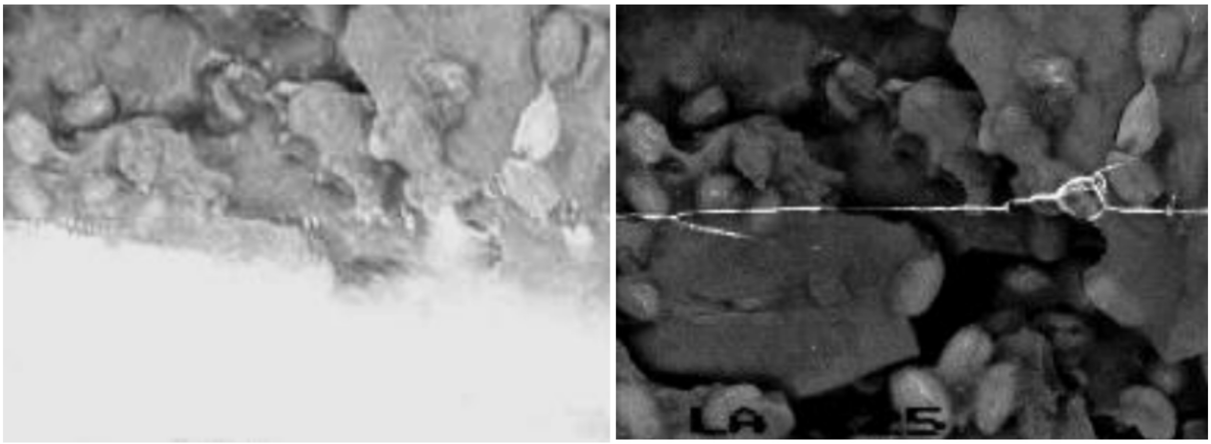


Figure 2.5: Peanut Brittle Bag from [68]. Left: under bright field forward lighting. Right: under dark field forward lighting - the seal is visible in this setting.

Coaxial light is another forward illumination mode that passes light through a half mirror to avoid strong reflections. This is suitable for detecting bumps, cracks and scratches on smooth surfaces. This is effective in illuminating textured features as instead of avoiding specular glare, it uses the glare to find details about a feature of interest. Figure 2.6 demonstrates a use case where diffuse light is used to inspect damage in the sealing surface of a bottle cap.

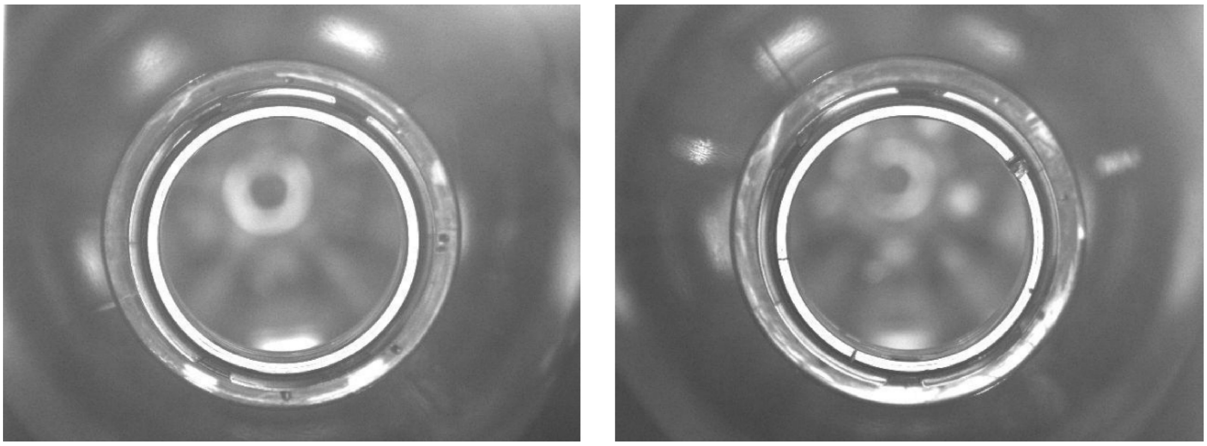


Figure 2.6: Sealing surface of two bottle caps under diffuse light from [68]. Left: Clean and undamaged surface. Right: Damaged surface. The damage is observed in the discontinuities within the white ring by the light.

Scattered forward lighting of dome structure or diffuse lighting is used to avoid direct light on surfaces by illuminating against a dome structure. In this mode the light passes through a diffuser which reflects or blocks part of the light, providing soft lighting to illuminate the scene however it requires multi directional lights.

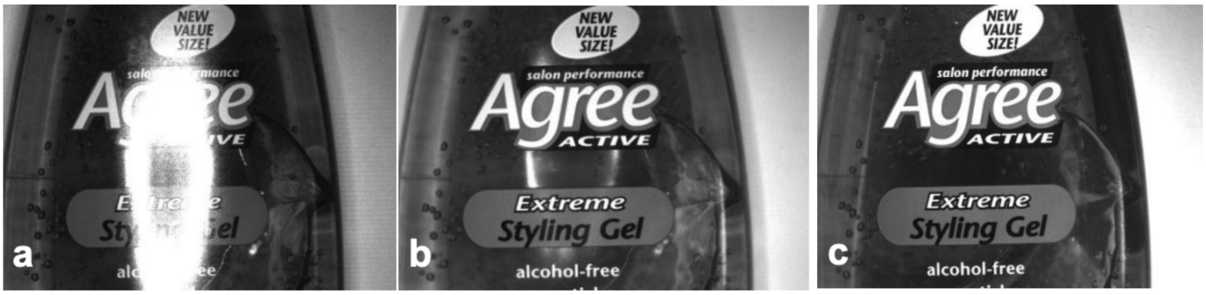


Figure 2.7: Specular vs Diffuse Lighting example from [68]: a. Ring Light without Polarizers, b. Ring Light with Polarizers c. Ring Light with Polarizers and in Diffuse mode. a. and b. are captured in bright field forward mode. a. produces strong reflection due to the illumination mode and the lack of polarizer. b. produces little glare even with polarizer, which filters out some of the reflections. c. produces no reflections due the combination of Diffuse Lighting mode and Polarizers.

Another significant mode to avoid reflections is back lighting wherein the light source is placed behind the object. This can highlight shadows of opaque objects and the interior of transparent objects. Due to these properties, it is useful for object shape detection and detecting the presence or absence of holes of gaps.

Mode	**Description**	**Advantages**	**Disadvantages**	**Application**
Bright Field	Light from the front.	High contrast for surface features.	Glare on reflective surfaces.	Surface inspection, label reading.
Dark Field	Shallow angle light for surface defects.	Reveals scratches, defects.	Ineffective on non-reflective surfaces.	Defect detection on shiny objects.
Coaxial	Light aligned with camera axis.	Reduces glare and shadows.	Limited to flat objects.	Inspection of reflective surfaces, glass.
Scattering (Dome)	Diffused light from all directions.	Eliminates shadows, reflections.	Low contrast for small defects.	Shiny, curved objects like bottles.
Back Lighting	Light placed behind the object.	High contrast for edges, holes.	No surface details.	Edge, hole detection, object profiling.

Table 2.1: Summary of Illumination Modes in Machine Vision.

2.2 Traditional Image Segmentation

Image segmentation is the process of dividing an image into regions based on high-level feature visual features like brightness or intensity. Traditional Image Segmentation methods consists of machine learning algorithms which are simple, efficient and computationally inexpensive. These include thresholding, edge detection, and region-based techniques.

2.2.1 Thresholding

Thresholding is a fundamental technique in image processing used to segment images into distinct regions by converting grayscale images into binary images. This process involves setting a specific intensity value, called the threshold, to differentiate between foreground and background pixels. Pixels with intensity values above the threshold are typically classified as foreground, while those below are classified as background. There are several thresholding methods, including global thresholding, where a single threshold value is applied across the entire image, and adaptive thresholding, which adjusts the threshold value based on local pixel neighborhoods.

Adaptive thresholding algorithms include Gaussian [18], mean [34], and Otsu [84]. In Gaussian thresholding, the threshold is set based on the local average which is a weighted average of the pixel values in the block, where the weights are a 2D Gaussian centered in the middle. Similarly in mean thresholding, the threshold for a region is set by averaging the pixel values. Otsu thresholding [84] selects the optimal threshold by maximizing the variance between the two classes of pixels that are separated by the threshold.

Figure 2.8 shows the results from each of the thresholding method on an image of coins, allowing for a visual comparison of how each technique segments the image.

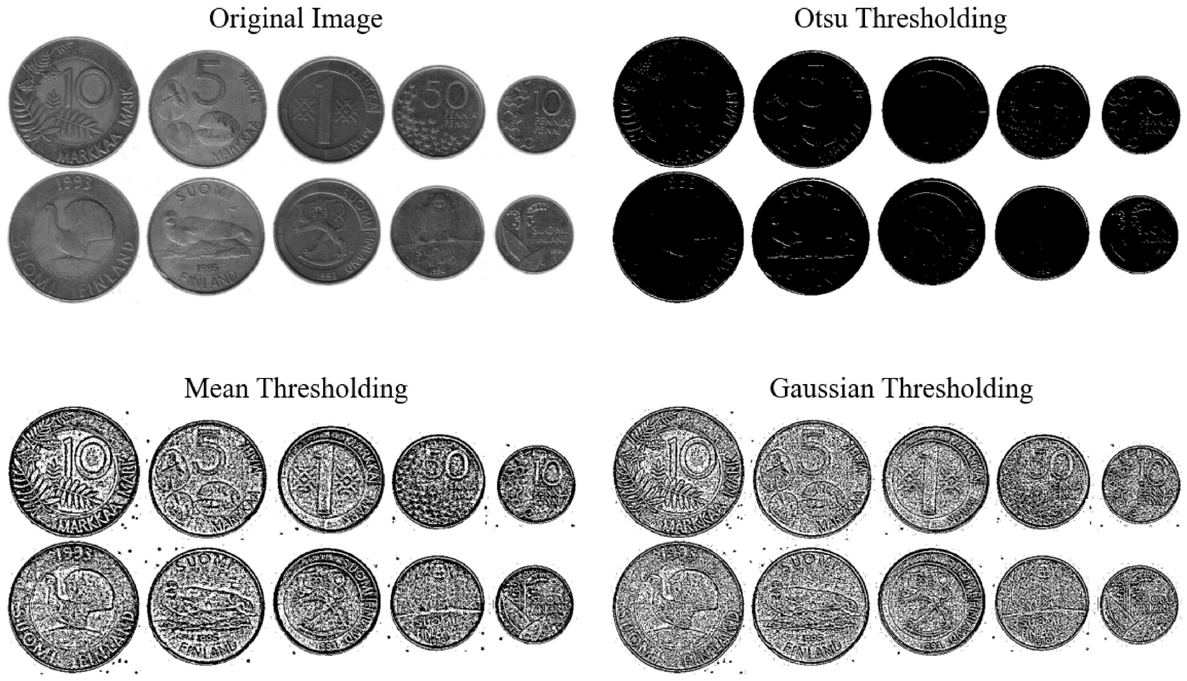


Figure 2.8: Comparison of Thresholding methods. The input coin image is taken from [56]. The segmentation were generated by me.

2.2.2 Edge Detection

Edge detection techniques identify the boundaries between regions by detecting the discontinuities in image intensities. The detected edges are linked together to form the contours that outline the boundaries of the object. Edge detection operators like Sobel [105], Canny [13] and Laplacian operators [67] are used to locate these discontinuities. The discontinuities between the regions are detected based on the types of intensity change. Sobel [105] is discrete differential operator used to compute the approximate gradient of an image. While canny [13] is an edge detection algorithm that combines edge detection, non-maximum suppression, and hysteresis thresholding to produce high-quality edges. Laplacian is a second-order differential operator used to detect zero crossings in an image, which correspond to edges [30].

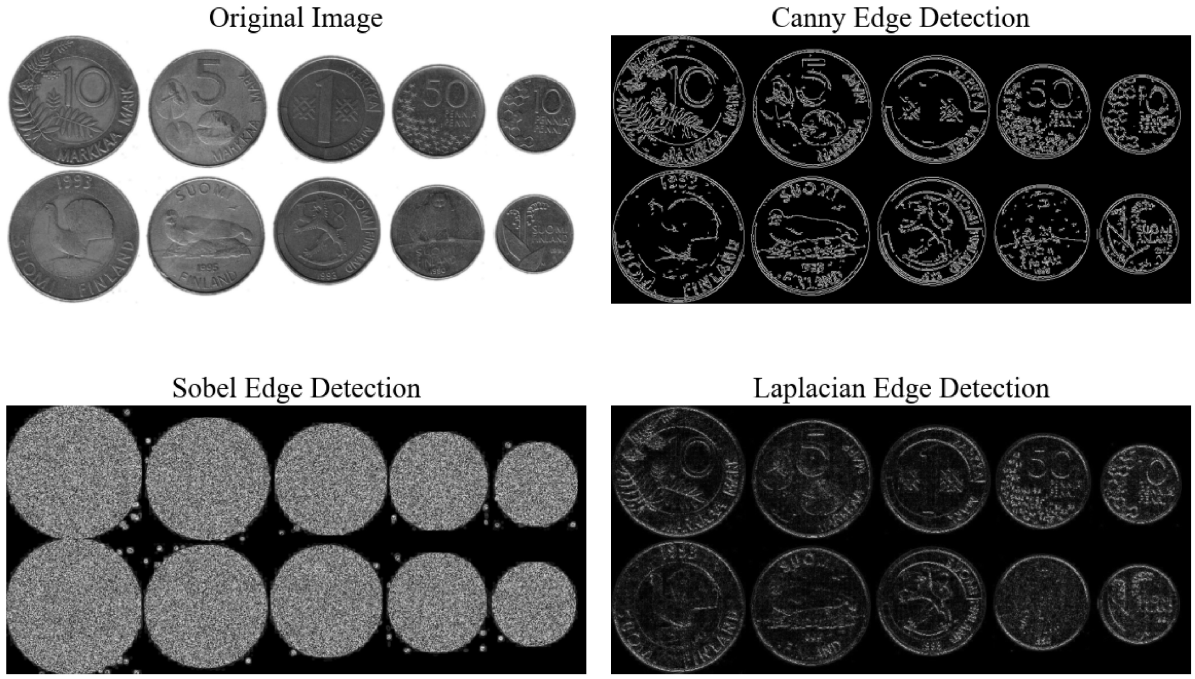


Figure 2.9: Comparison of Edge detection methods. The input coin image is taken from [56]. The edge detection images were generated by me.

Although these methods are suitable for simple segmentation, they are ineffective to segment regions with complex features. K-means clustering [65], which is primarily used for clustering data points, can also be applied to image segmentation by treating pixels as data points and clustering them based on their color or intensity values. This process effectively groups pixels into regions based on their similarity.

2.2.3 Region-based Segmentation

Region-based segmentation is a technique in image processing that involves grouping pixels based on shared characteristics, such as color, intensity, or texture. This method is particularly useful for distinguishing objects within an image where such properties can define clear boundaries. The two primary categories of region-based segmentation are region-growing and region-splitting.

In region-growing, the process starts with selecting seed pixels, which act as the initial points for segmentation. The algorithm then expands these regions by adding neighboring pixels that exhibit similar properties to the seed pixels, effectively "growing" the region. This approach is particularly effective in images with homogenous regions but may struggle with noisy or complex backgrounds that lack clear boundaries [30, 2]. A visualization of region-growing is shown in Figure 2.10.

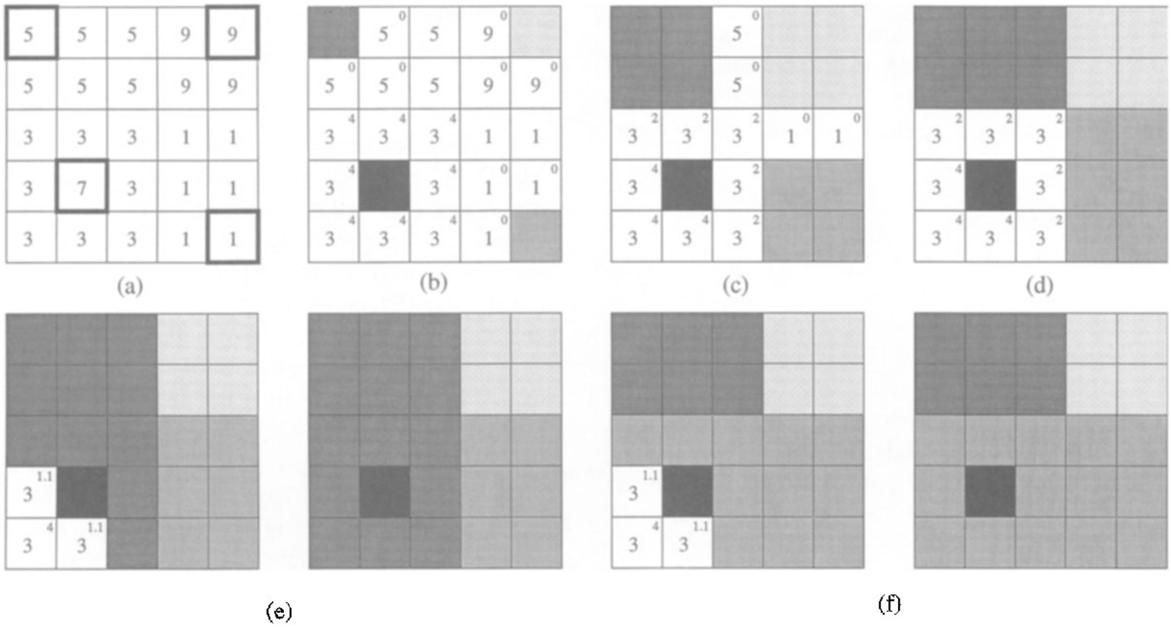


Figure 2.10: Illustration of the order dependency in the Seeded Region Growing (SRG) algorithm taken from [70]. (a) A gray-scale test image with four initial seed points marked. (b) Each pixel (x) is labeled with its corresponding value (3). (c) Result after 9 iterations of the algorithm. (d) Result after 13 iterations. (e) Final segmentation result based on one processing sequence. (f) Final segmentation result based on an alternate processing sequence, demonstrating how the order of processing affects the segmentation outcomes.

On the other hand, region-splitting treats the entire image as a single segment and recursively divides it into smaller regions based on predefined criteria. This method allows for a more structured segmentation process, particularly in images with varied textures or patterns. However, its effectiveness diminishes in cases where regions are not clearly defined [30].

A notable example of a region-growing algorithm is the Watershed segmentation [10] method. This technique models the image as a topographic surface, where the intensity values correspond to elevation. It simulates the way water would flow over this surface, accumulating in basins (regions) as it rains. This intuitive approach allows the algorithm to delineate regions based on their natural contours and gradients, making it particularly useful in applications like medical imaging, where anatomical structures may be difficult to segment using traditional methods [30].

2.3 Deep-learning Image Segmentation

Deep-learning-based image segmentation implement neural networks to learn complex features from the image data. This is widely used in computer vision tasks and different types of image segmentation namely semantic and instance segmentation. Semantic segmentation identifies and localizes an object in an image or a video. This process involves labeling each pixel in an image to a class based on its semantic features (i.e dog, person). It allows more precise locations and object boundaries, and adds more contextual information. Detection of reflection lies within the domain of semantic segmentation. While instance segmentation is specialized semantic segmentation that aims to distinguish between multiple instances of the same object class. There are numerous deep-learning architecture suitable for semantic segmentation, including Fully Convolutional Networks (FCN) [123] [95] [64] [25] and U-Net [109] architecture. In recent years, semantic segmentation has been a key component in diverse industries such as autonomous driving, remote sensing, medical image segmentation, and robotic vision.

2.3.1 Fully Convolutional Network

Fully convolutional networks (FCN) are a pivotal architecture introduced for image segmentation tasks, significantly improving the accuracy and efficiency of pixel-level predictions. Unlike convolutional neural networks (CNNs) [51], which are designed primarily for image classification, FCNs replace the fully connected layers with convolutional layers, allowing for input images of arbitrary size and producing spatially dense outputs. This transformation enables the network to make pixel-wise predictions, effectively segmenting images into distinct regions based on learned features [63]. This allows pixel-wise prediction when trained end-to-end on a dataset and has less parameters as it does not use dense layers. It consists of a downsampling path which extracts features and context from the image and the upsampling path which allows localization and give an output according to the size of the input image. Figure 2.11 shows the architecture proposed for segmentation tasks.

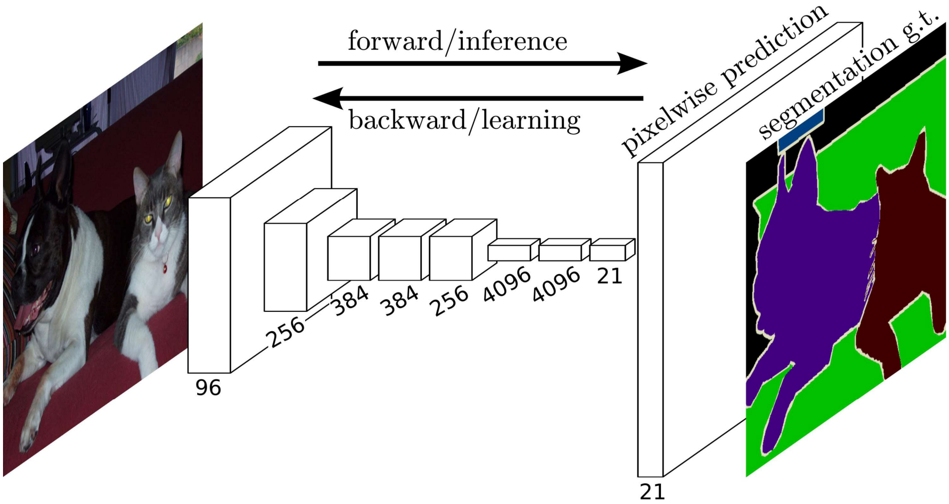


Figure 2.11: FCN Architecture proposed in 2015 [63].

FCNs have led to the development of more advanced architectures in segmentations tasks such as the U-Net.

2.3.2 U-Net

It is a fully convolutional network (FCN) architecture designed for image segmentation, particularly in medical image analysis. Its name is coined due to its shape which resembles a "U". It is one of the state-of-the-art models in medical segmentation due to its high accuracy in detecting objects with substantial shape variations, weak borders and inset or overlapping objects [93]. It improves upon the FCN by applying an encoder-decoder style of network that works in an end-to-end setting.

Encoder: it analyzes the image and derives the high-level features. The image size reduces as it passes through the encoder. This is also known as the downsampling path.

Decoder: it takes the compressed image from the encoder and expands it back to its original size. It incorporates information from the encoder through skip connections which preserve information that could be lost during the down sampling of the image. This is also referred to as the upsampling path.

U-Net has demonstrated remarkable performance across various applications, particularly in the medical imaging domain, where precise segmentation of structures like tumors or organs is essential [93]. Moreover, the U-Net architecture has inspired numerous variations and extensions, such as the U-Net++ [131] and Attention U-Net [82], which integrate advanced techniques to enhance performance and adaptability in different segmentation tasks. Recent works have explored its application beyond medical imaging, including satellite image segmentation [85] and agricultural applications [133], showcasing its versatility and robustness.

2.3.3 Vision Transformer

The Vision Transformer (ViT) [21] is a novel architecture that adapts the transformer model [114], originally designed for natural language processing, for computer vision tasks. Unlike traditional CNN, which process images using convolutional layers, ViT treats an image as a sequence of fixed-size patches. Each patch is flattened and linearly embedded into a token, allowing the transformer to learn spatial relationships through self-attention mechanisms.

ViT operates by stacking multiple transformer blocks, each consisting of multi-head self-attention and feed-forward layers. This design enables the model to capture long-range dependencies and global contextual information effectively. The positional encodings added to the patch embeddings help the model retain spatial information, which is crucial for understanding image content. For classification tasks, an additional learnable "classification token" is introduced to the sequence to help the model focus on the classification task. The architecture follows a standard Transformer approach as described by [114]. An illustration of the vision transformer is shown in Figure 2.12.

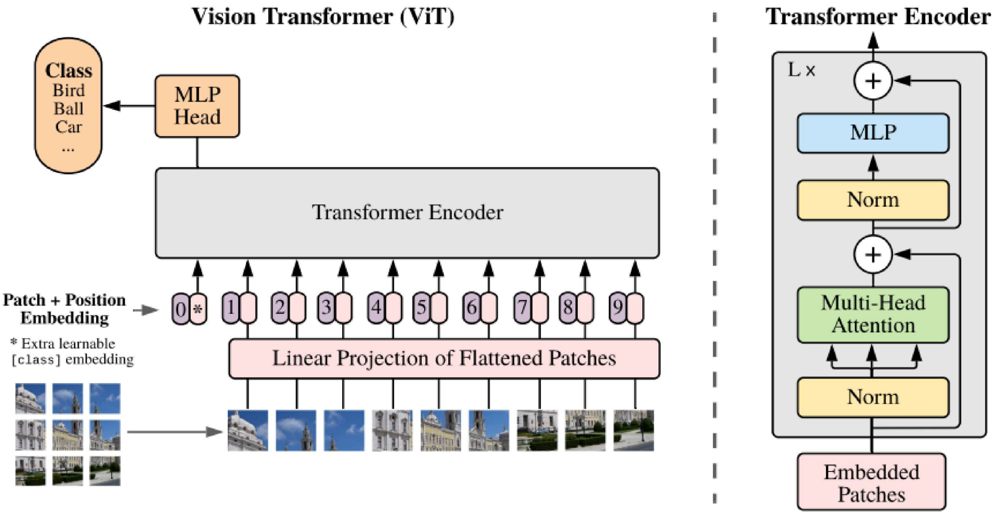


Figure 2.12: An image is split into fixed-sized patches into the Vision Transformer [21]. The Transformer encoder is inspired by the original Transformer in [114].

ViT has demonstrated state-of-the-art performance on various image classification and segmentation tasks, proving its effectiveness in leveraging the strengths of attention mechanisms for visual data.

2.3.4 ResNet-50

ResNet-50 [36] is a widely used convolutional neural network architecture known for its innovative use of residual connections, which address the vanishing gradient problem in deep networks. With 50 layers, ResNet-50 is designed to enable the training of very deep networks by introducing shortcut connections that skip one or more layers. These connections allow gradients to flow more easily during backpropagation, facilitating the training of deeper models without loss of performance.

The architecture consists of a series of convolutional layers grouped into residual blocks, each containing two or three convolutional layers. The output of these layers is combined with the input via identity mappings, creating a residual learning framework. This is demonstrated in Figure 2.13.

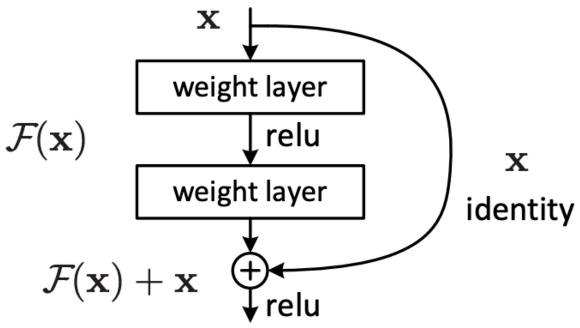

x
weight layer
$\mathcal{F}(\mathbf{x})$
relu
weight layer
x
identity
$\mathcal{F}(\mathbf{x}) + \mathbf{x}$
relu

Figure 2.13: Residual Learning Block [36].

ResNet-50 achieves high accuracy on image classification benchmarks and serves as a strong backbone for various vision tasks, including image segmentation. Its efficiency and effectiveness have made it a standard choice for many applications in computer vision.

2.4 Loss Function

Loss function are play a key role in designing the deep-learning methods as it defines the learning process of an algorithm. It measures the error margin between the actual value and the predicted value. The loss value returned reflects the accuracy of the model's performance. The range of this function is [0,1], with 0 denoting a perfect match and 1 indicating that nothing was learnt.

Binary Cross Entropy: Cross entropy is considered to be the difference between two probability distributions for a given random variable or set of events. Binary cross entropy [101] is one of the most commonly used loss functions in semantic segmentation tasks. It is derived from the Bernoulli distribution [44]. In equation 2.3, y is the true value and $\hat{y}$ is the predicted outcome.

$$L_{\text{BCE}}(y, \hat{y}) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})). \tag{2.3}$$

The range of BCE values is $[0, \infty]$, this values is passed through the sigmoid activation function which scales the loss value between 0 and 1. This prediction value corresponds to the likelihood of a data sample belonging to a class which is denoted by $\hat{y}$. $log(1-\hat{y})$ represents the probability of the negative class. The entire expression is negated because logarithms of probabilities produce negative values when the predicted probability is less than 1. Thus, by negating, we ensure that the loss is a positive value. BCE penalizes inaccurate results which results in higher loss values, motivating the model to minimize this loss during training. Although, it is widely used in segmentation tasks, its limitation lies its inability to consider class imbalances in the dataset. BCE treats all misclassifications equally regardless of their class frequencies.

Focal Tversky Loss: Focal Tversky loss [1] function is the focal loss function based on the Tversky index, which was introduced to provide a better trade-off between the precision and recall in semantic segmentation tasks. The Tversky index [111] is an asymmetrical similarity measure that compares the variant and its prototype on sets. It is a generalization of Dice score and addresses its limitations of class imbalance by allowing penalizing False Negatives and False Positives.

$$S(A,B) = \frac{|A \cap B|}{|A \cap B| + \alpha|A-B| + \beta|B-A|}. \tag{2.4}$$

In equation 2.4, the α controls the penalty for False Negatives while the β controls the penalty for the False Positives. This helps control class imbalance in applications like segmentation tasks. The Tversky index is then adapted to a loss function by minimizing $\sum_c (1 - TI_c)$ and parameterized with γ in the range [1,3]. The Focal Tversky loss function is depicted in equation 2.5.

$$FTL = \sum_c (1 - TI_c)^{\frac{1}{\gamma}}. \tag{2.5}$$

When $\gamma > 1$, the loss function focuses on the less accurate predictions which have been misclassified. In the Tversky index, when $\alpha = \beta = 0.5$, the FTL equates to the Dice coefficient, and when $\gamma = 1$, FTL is equal to the Tversky loss. It is observed that as α increases, it improves the model convergences by focusing on minimizing FN. According to [1], the experiments with $\gamma = \frac{4}{3}$ performed the best. FTL is found to perform better at learning hard examples with smaller regions of interest (ROI). In the study comparing loss function for segmentation tasks [44], the use of Focal Tversky loss resulted in one of the highest Dice score and Sensitivity results.

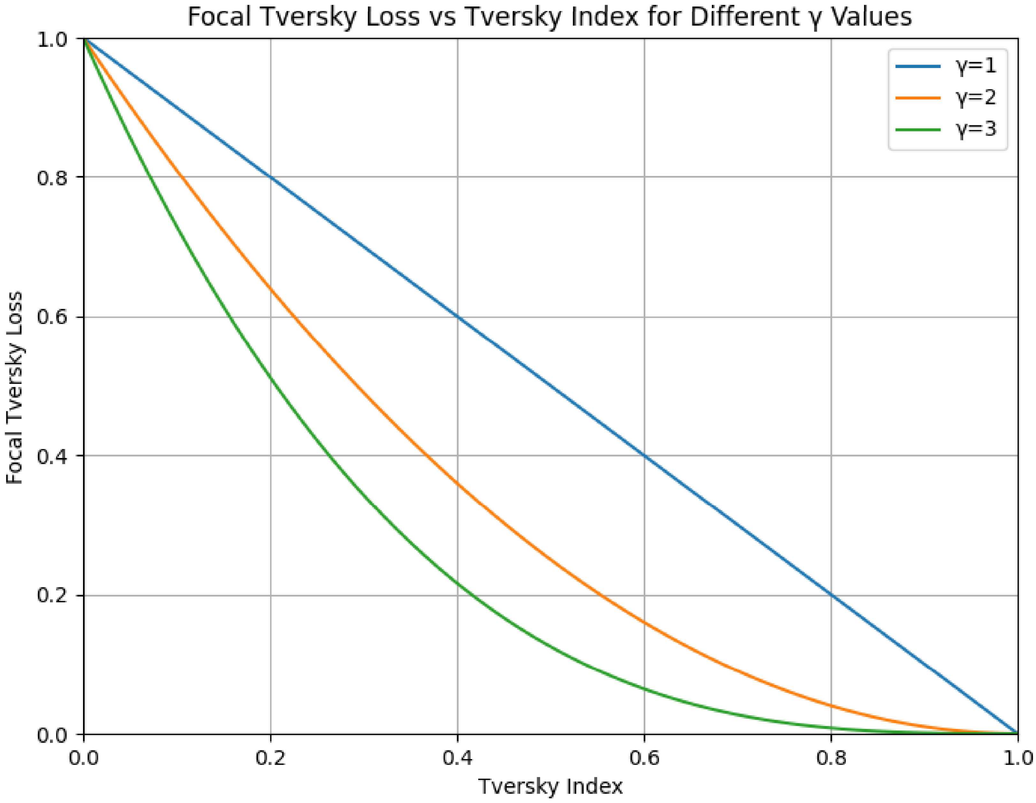


Figure 2.14: The plot illustrates how varying gamma influences the performance of both loss functions, highlighting the potential advantages of Focal Tversky Loss in handling class imbalance in image segmentation tasks.

The relationship between Focal Tversky Loss and the Tversky Index is illustrated in Figure 2.14. As shown, the gamma values significantly impact the behavior of both loss functions. Higher gamma values tend to place more emphasis on challenging samples, thereby improving the model's ability to handle class imbalance. This comparison highlights how tuning gamma can affect the efficacy of Focal Tversky Loss in optimizing segmentation performance.

2.5 Optimizer

Optimizers are using in the training phase, to adjust the parameters of the model to reduce the loss function. They determine how the weights of the neural network are updated based on the gradients calculated from the loss function on a given batch of data [74]. By controlling the learning rate, optimizers help the model converge efficiently towards the optimal solution, improving its performance in the given task. The optimizers covered in this thesis are all gradient-based algorithms which update the model weights in the direction that minimizes the loss. The choice of the optimizer can significantly affect the training speed and the model's performance.

Stochastic Gradient Descent [4] introduces randomness to update the model parameters based on the loss function, which helps converge the model faster. Along with the learning rate, SGD also uses Momentum to enhance its performance. Momentum accumulates the gradient of the past steps to determine the direction to go to. This helps accelerate convergence and the reduces oscillations, particularly in noisy gradients.

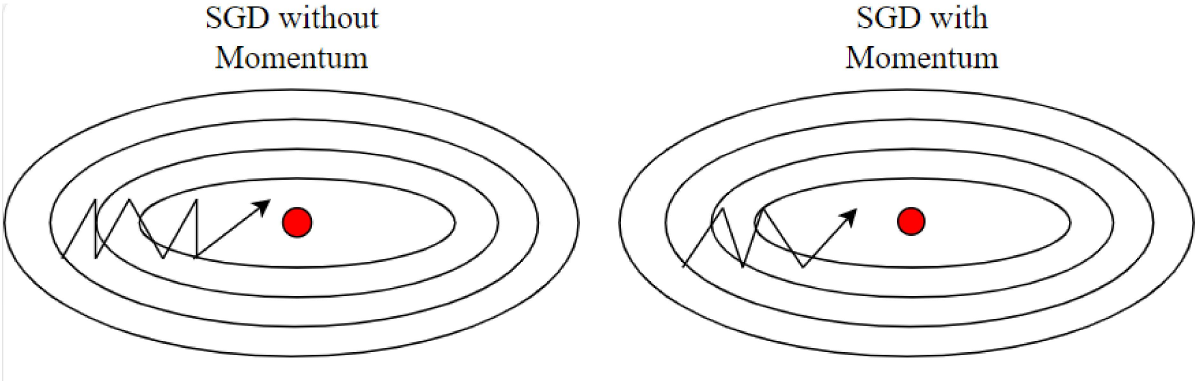


Figure 2.15: SGD with and without momentum. Redrawn from an illustration in [96].

Adam Adaptive Moment Estimation (Adam) [46] is a popular optimizer which combines the advantages of RMSProp [37] and AdaGrad [22]. It maintains running averages of both the gradients and their squares, allowing it to adaptively adjust the learning rates for each parameter. This feature makes Adam particularly effective for a wide range of problems and datasets. It typically requires less tuning than other optimizers and performs well across various tasks, making it one of the most popular choices in deep-learning. Adam is known for its efficiency in handling sparse gradients and is suitable for large-scale datasets.

2.6 Evaluation Metrics

Evaluation metrics are crucial to measure the performance of any deep-learning model. In semantic segmentation, the aim of these metrics is to calculate the correctly identified and segmented regions within an image. Each metric presents a different interpretation of the result. There are two key evaluation metrics in semantic segmentation.

2.6.1 Intersection over Union (IoU)

Intersection over Union (IoU) [43], also known as the Jaccard Index, measures the overlap between the predicted region and the ground truth. The IoU is calculated by dividing the area of intersection between the predicted segment and the ground truth segment by the area of their union. IoU can also be expressed in terms of the confusion matrix, considering true or false, positives or negatives.

$$\mathrm{IoU} = \frac{|A \cap B|}{|A \cup B|}, \qquad \mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{2.6}$$

The IoU value ranges from 0 to 1, where 1 indicates a perfect match, and 0 indicates no overlap. IoU is one of the most useful metrics for evaluating the performance of semantic segmentation models and is widely used in applications such as medical imaging, autonomous driving, and remote sensing imagery.

2.6.2 Dice Similarity Coefficient

Dice similarity coefficient (DSC) [20] measures the similarities between the data. It is particularly sensitive to small structures or regions, making it valuable in scenarios where detecting small objects or regions is important, such as in medical segmentation.

$$\mathrm{DSC} = \frac{2|A \cap B|}{|A| + |B|}, \qquad \mathrm{DSC} = \frac{2 \times \mathrm{TP}}{2 \times \mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{2.7}$$

Similar to IoU, the value of Dice coefficient ranges between 0 and 1, with 0 indicating no overlap and 1 indicating a perfect overlap. Unlike IoU, the denominator $|A|+|B|$ counts the area of A and B separately. This means if there is any overlap, it will be counted twice—once as part of A and once as part of B. This metric is particularly valuable in situations where the primary goal is to accurately identify the segmented regions, with less emphasis on the non-segmented areas.

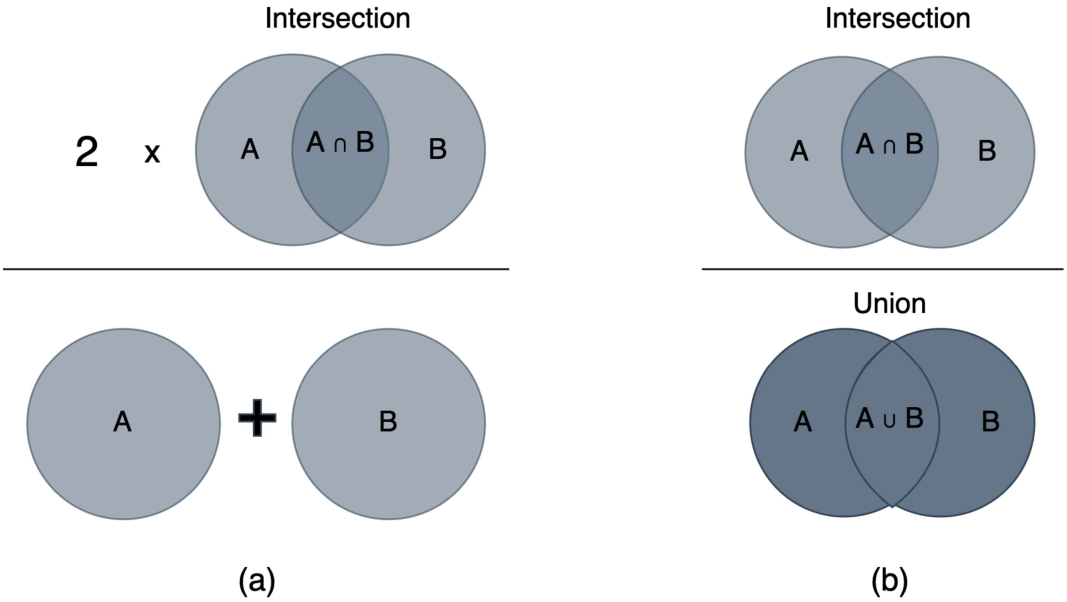


Figure 2.16: Calculation of Segmentation evaluation metrics: (a) Dice Similarity Coefficient and (b) Intersection over Union, based on [47].

2.7 Transfer Learning

Transfer Learning (TL) is a machine learning technique that which uses a pre-trained model from a machine learning task or dataset to improve the performance of a related task or dataset [121]. It essentially transfers the knowledge from one model to apply on a related domain. It is a widely used technique as it reduces the computational cost of training a model from scratch. By reusing the pre-trained models, the new model's training time and the training data to achieve a desirable result are reduced while also improving the performance. As highlighted previously, it is difficult to find publicly available large datasets and more so for specialized domains and is producing a dataset by manually labeling the data can be time-consuming and expensive. Transfer learning helps alleviate these difficulties. Moreover, TL also increases the model's generalizability since it involves retraining an existing model with a new dataset. This allows the model to retain knowledge from multiple datasets and domain. Due to these reasons, TL is one of the key domain adaptation methods applied in computer vision and natural language processing.

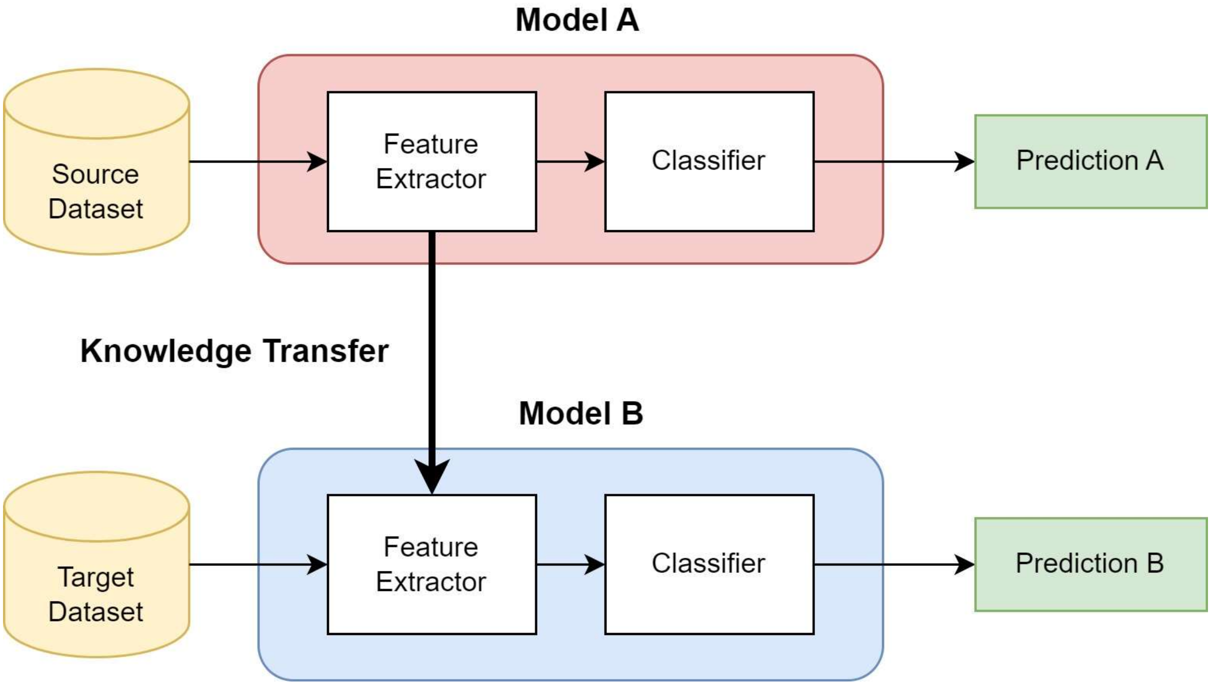


Figure 2.17: Schematic Representation of Transfer Learning. Source dataset is a larger dataset for a general task, while Target dataset is a smaller dataset for a specialized task. Knowledge transfer between the two domains allows the smaller model to learn from the larger model to solve a different but related problem.

2.8 Attention Mechanism in Deep Learning

Attention mechanisms have revolutionized deep learning by allowing models to focus on specific parts of the input data that are more relevant for the task at hand. By dynamically weighting the importance of different features, attention helps improve model performance across various applications, including natural language processing, computer vision, and medical image segmentation. The attention mechanism is primarily divided into hard attention and soft attention.

Soft Attention Soft attention computes a weighted sum of all input features, assigning probabilities that represent the importance of each part. This approach is differentiable, making it easy to optimize through backpropagation, and allows the model to softly attend to multiple relevant regions simultaneously [80]. For this work, soft attention is used.

Hard Attention Hard attention, on the other hand, selects a discrete subset of input regions, focusing only on specific parts. This approach is non-differentiable, making it more challenging to optimize but can lead to computational efficiency, as it ignores irrelevant areas. Reinforcement learning methods are often used to train models with hard attention [17].

Spatial Attention Spatial attention focuses on learning which regions of the input feature maps are most important for a given task. By generating spatial attention maps, the model can emphasize salient areas while suppressing less relevant regions. This is particularly useful in image segmentation, where the model needs to identify and differentiate various objects within an image. Spatial attention can be implemented through convolutional operations that aggregate information across channels to create a weighted representation of the spatial layout [122].

Channel Attention Channel attention operates by learning the importance of each channel in a feature map, emphasizing features that are more relevant for a task while downplaying less significant ones. It helps the network focus on informative channels (e.g., texture, color) and suppress irrelevant ones, improving feature representation across layers. Techniques like the Squeeze-and-Excitation (SE) block [39] implement channel attention by squeezing global information into each channel and exciting important ones [62] [106].

2.9 Attention Segmentation Modules

Attention segmentation modules are essential components in deep learning architectures aimed at precisely identifying and delineating objects in images. Various attention mechanisms can be integrated into segmentation modules to improve their performance.

2.9.1 Attention Gate

Attention gates (AG) guide the model's attention on the important regions while suppressing feature activation other redundant areas by introducing additive soft attention [82]. It is a light weight feature which does not significantly affect the model complexity as very few parameters are added. The attention gate takes two input vectors: g and x. The vector g comes from the next lowest level in the model containing better spatial feature representations, and has smaller dimensions. Vector x is the scaled skip connection. The two vectors are are summed element-wise, which results in aligned weights becoming larger and unaligned weights becoming smaller. The resultant vector is passed through the ReLU activation and a 1x1 convolution. This vector then goes through the sigmoid layer which scales the vectors between the range [0,1] to produce the attention coefficient. This coefficient represents how relevant the region is, with 0 denoting no relevance while 1 denoting high relevance. Trilinear interpolation is applied to upsample to the original dimension of the input feature map x. The coefficient is multiplied element-wise to the original x vector, which is then passed along with the skip connection [82].

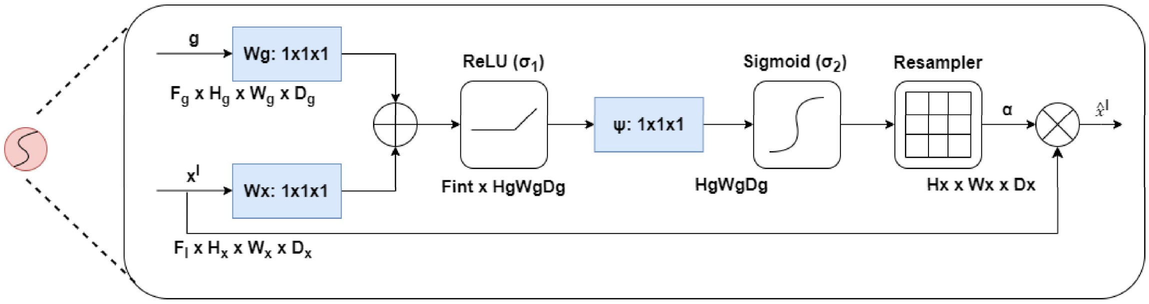


Figure 2.18: Attention Gate [82].

2.9.2 Convolutional Block Attention Module

The Convolutional Block Attention Module (CBAM) [122] is a lightweight module designed to enhance feature representation by applying both channel and spatial attention sequentially. First, it computes channel attention, which emphasizes informative features, followed by spatial attention to focus on significant regions in the feature maps. This dual attention mechanism allows CBAM to effectively refine feature maps, improving segmentation tasks by leveraging both spatial and channel information.

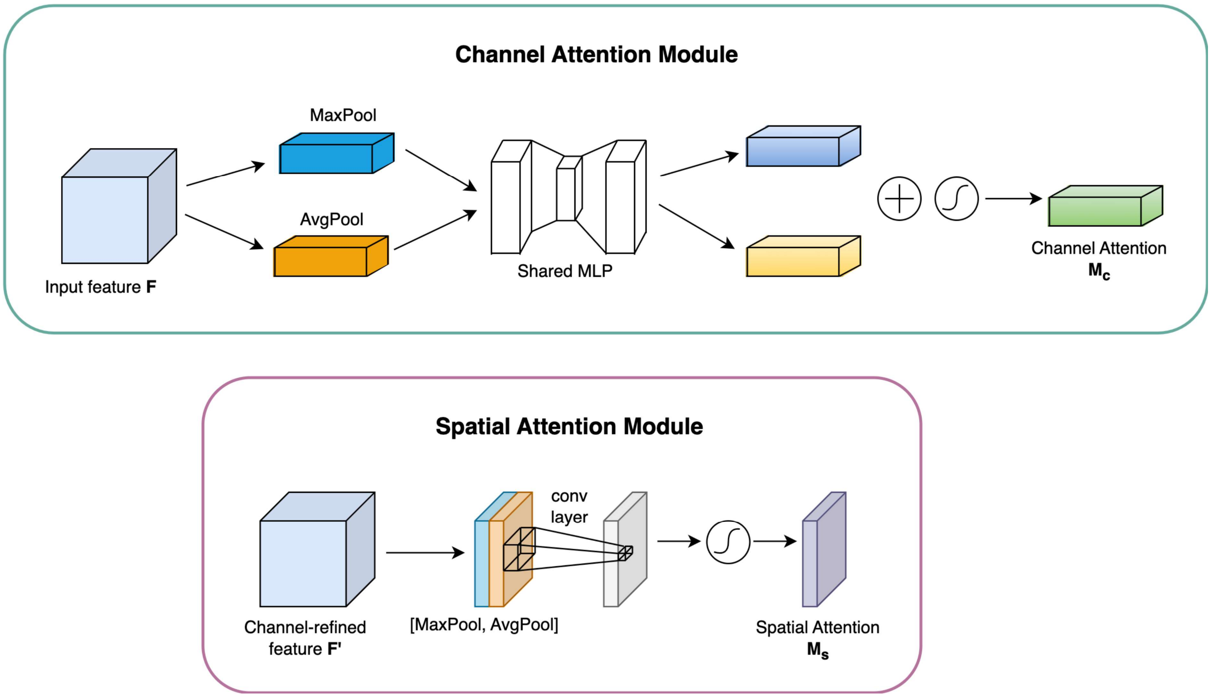


Figure 2.19: Convolutional Block Attention Module, redrawn from [122].

In practice, the integration of CBAM into existing architectures has demonstrated remarkable improvements in tasks such as image segmentation, object detection, and action recognition [80]. The module's lightweight nature ensures that it can be easily incorporated into a variety of networks without significantly increasing computational overhead, making it suitable for real-time applications in resource-constrained environments. Furthermore, CBAM's effectiveness has been validated across multiple datasets and benchmarks, solidifying its position as a valuable component in enhancing model performance in contemporary deep learning architectures.

2.9.3 Squeeze-and-Excitation Module

Squeeze-and-Excitation Networks (SE) [39] introduce an innovative mechanism for improving feature representation in convolutional neural networks by modeling channel-wise dependencies adaptively. The core concept of SE lies in the Squeeze-and-Excitation block, which operates in two main steps: squeezing and excitation. In the squeezing step, global average pooling is applied to the feature maps, generating a channel descriptor that summarizes the global spatial information of each channel. This process condenses the feature map into a compact representation, capturing the overall channel-wise information.

Next, in the excitation step, a fully connected layer is used to learn channel relationships by applying a set of weights to the channel descriptors. This generates an attention map that highlights the most informative channels while suppressing the less significant ones. By recalibrating the channel responses, SE enable the model to focus on the features that contribute most effectively to the task at hand, thus enhancing the overall representational power of the network.

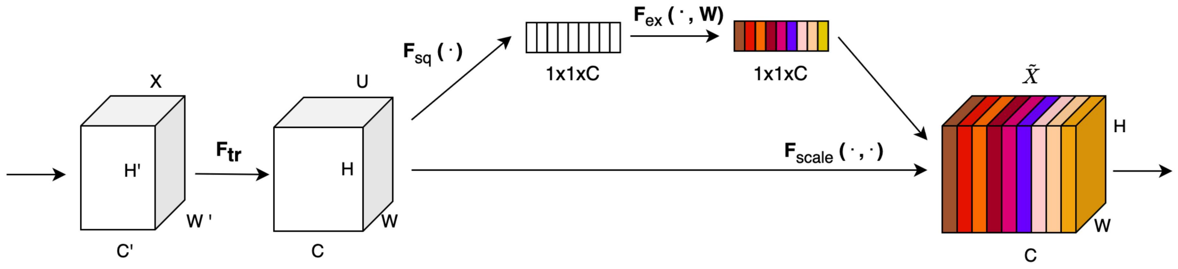


Figure 2.20: Squeeze-and-Excitation Block, redrawn from [39].

2.9.4 Efficient Channel Attention

Efficient Channel Attention (ECA) is an advanced attention mechanism designed to enhance the representational power of CNN while maintaining computational efficiency. Introduced in 2020 [118], ECA improves upon traditional channel attention mechanisms by simplifying the computation process. Instead of relying on complex operations like multilayer perceptrons (MLPs) to capture channel-wise dependencies, ECA uses a lightweight 1D convolutional layer, making it more efficient and faster. The core idea of ECA is to exploit local cross-channel interactions without the need for extensive computations. This is achieved by applying a kernel to aggregate channel information adaptively, allowing the model to weigh the importance of different channels based on their contribution to the feature maps. By leveraging this approach, ECA retains the ability to learn effective channel representations while reducing the overhead typically associated with attention mechanisms. The structure of ECA is depicted in Figure 2.21

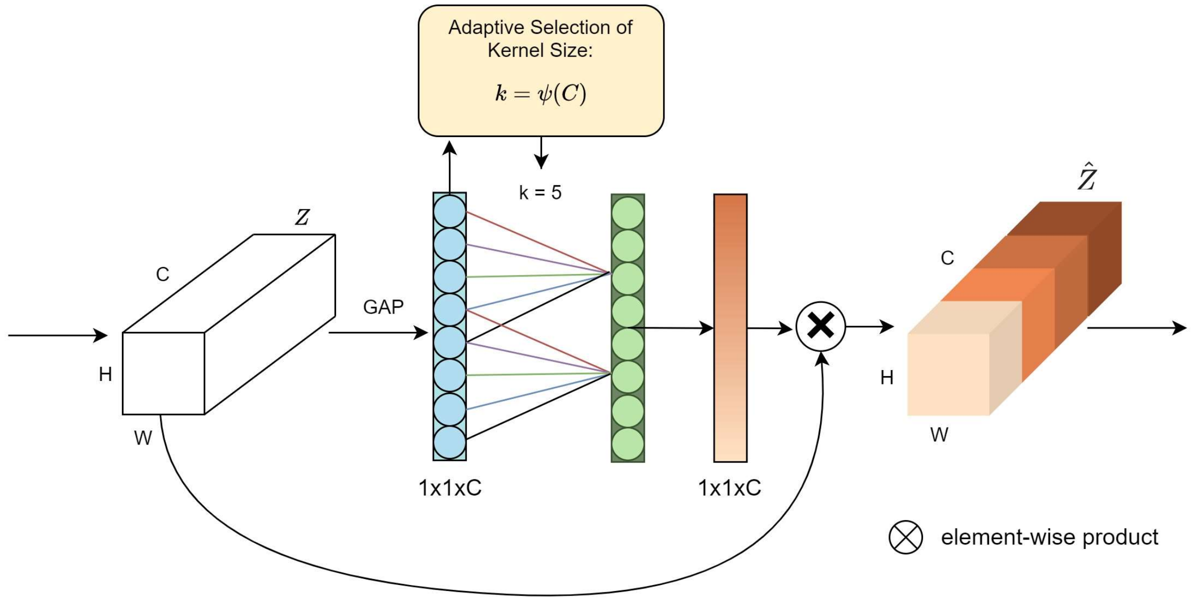


Figure 2.21: Efficient Channel Attention [118].

Studies have shown that ECA can lead to significant performance improvements in various tasks, such as image classification and object detection, while also being less resource-intensive than its predecessors. For instance, the authors demonstrated that ECA outperformed several existing attention models on benchmark datasets while requiring fewer computational resources, making it particularly suitable for real-time applications .

Segmentation Module	Spatial Attention	Channel Attention
Squeeze-and-Excitation (SE)		✓
Convolutional Block Attention Module (CBAM)	✓	✓
Attention Gate (AG)	✓	
Efficient Channel Attention (ECA)		✓

Table 2.2: Comparison of Segmentation Modules and Their Attention Types.

Chapter 3

Related Works

Reflections in images have been a critical issue in computer vision and image processing tasks like object recognition and localization [61]. The research on reflection detection been conducted for decades with this topic being relevant across numerous domains like medical segmentation [81] [6] [78], and inspections tasks [41] [130]. This chapter aims to explain the key concepts relevant for this topic, and the existing research in reflection detection across various domains.

3.1 Reflection Detection Methods

3.1.1 Traditional Methods

In exploring the historical evolution of techniques for detecting light reflections in images, traditional methods have played a foundational role. This section will describe these methods, tracing their development in the domain of image processing.

Thresholding Early works of specular reflection detection were conducted in medical image analysis which included traditional machine learning techniques like thresholding. Thresholding usually consisted of three processes: a pre-processing step to reduce the noise from varying lighting conditions, a thresholding algorithm to separate the reflected areas based on computed thresholds, and a post-processing step to reduce the number of false detections [73]. These techniques were applied on varying color spaces, like RGB and HSV. RGB (Red, Green, Blue) is an additive color model used to represent colors in digital formats by combining varying intensities of red, green, and blue light [104]. HSV (Hue, Saturation, Value) is a color model that describes colors in terms of their hue (color type), saturation (intensity), and value (brightness) [104], making it more intuitive for color selection and manipulation in applications like image processing. Past works have implemented both global thresholding [113] and adaptive thresholding [6] to detect specular reflections in RGB color. Although efficient, they detect regions which are more bright and not specular reflections, leading to high false positives especially in light colored or white surfaces.

Specularities are by definition regions of an image where pixel intensity is very high and where the color matches the illumination source. An early study on detection and correction of reflections in endoscope images implemented intensity and saturation-based thresholding [107]. To identify these areas, histogram decomposition is used by generating three histograms, each representing red, green and blue intensity to match high-intensity zones [98]. As specularities are more visible in the saturation component of HSV, a bi-dimensional histogram is built, where specularity is detected by the maximum value of intensity and saturation. The intensity is derived as the average of the red, green, and blue channels, while the saturation is computed by comparing the blue and red channels with the green channel. Pixels are then classified as part of a reflective region if their intensity is above 50% of the maximum intensity and their saturation is below 33% of the maximum saturation. Similarly, there are many instances of thresholding-based methodology for reflections detection such as subtracting minimum value of RGB plane from each pixel [102], binarization of image using single threshold [32]. In recent years, thresholding is used in combination with other techniques to produce highly accurate results. For instance, Nie et al. [78] introduced a unique approach using adaptive thresholding and brightness classification. The method categorized images based on average brightness and employs brightness component enhancement for low-brightness images, demonstrating superior performance and suitability for high-definition endoscopy images.

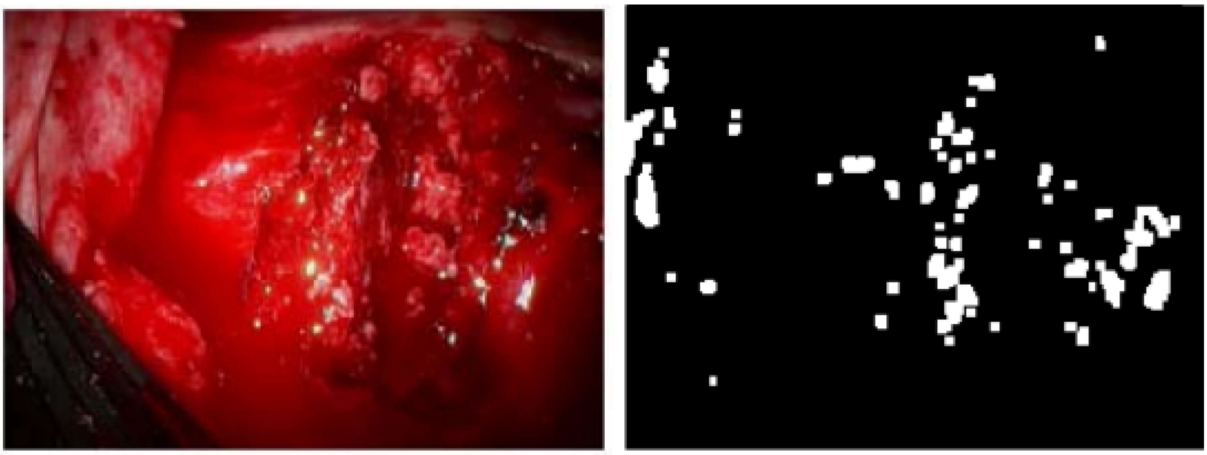


Figure 3.1: Results of real-time specular reflection detection in endoscopic images using intensity and saturation-based thresholding from [107].

Dichromatic Reflection Model Another approach is the principle of the Dichromatic Reflection Model (DRM). The DRM is based on the principle that radiation is composed of reflections from interface and surface body.

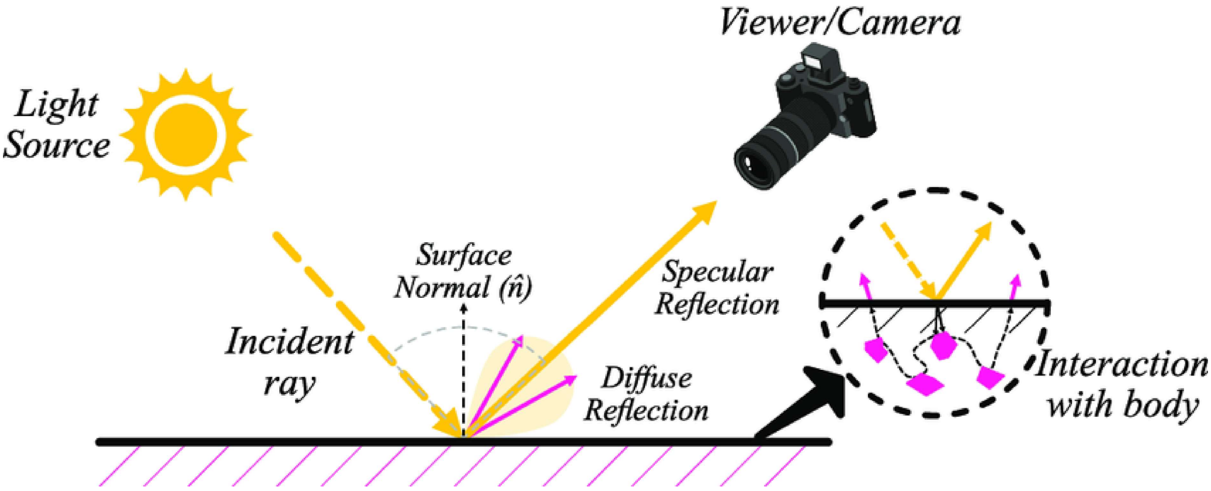


Figure 3.2: Dichromatic Reflection Model from [5].

The dichromatic model can be expressed with a linear sum of the spectra of the specular and diffuse reflection model as follows:

$$I = I_s + I_d, \tag{3.1}$$

where I $= (I_x, I_y, I_z)$, I_s denotes the luminance of the specular reflection component and I_d denotes the luminance of the diffuse reflection component [110].

The methods that apply this model aim to separate the combination of diffuse and specular reflections. It has the advantage of separating the specular reflection component from a single image in most instances however it cannot be separated from low-saturation colors like gray as well as the chromacity of the light source must be known [110]. In 2006 [127], a specular reflection detection method was proposed wherein the value of specularity invariant pixels and their ratio to separate diffuse components was calculated. Although this method was faster than other methods at the time, its application is limited to textured images and has reduced accuracy to approximation in normalization process.

Retinex Theory Retinex Theory is a computational model of color vision which proposes that the human vision system perceives color based on the relative reflectance of surfaces rather than absolute spectral properties [50]. According to this theory, unwanted illumination effects can be removed from an image by separating it into two components: reflectance and illumination [129]. The observed image can be expressed as follows [33]:

$$S(x, y) = R(x, y).I(x, y), \tag{3.2}$$

where (.) denotes element-wise multiplication, $S(x, y)$ os the captured image, $R(x, y)$ is the reflectance component and $I(x, y)$ is the illumination component. To find the $R(x, y)$, a Gaussian kernel function $G(x, y)$ is applied to estimate the illumination component. By taking the logarithm on both sides of Equation 3.2 and shifting the items, it leads to the following equation:

$$logR(x, y) = logS(x, y) - log[G(x, y) * S(x, y)], \tag{3.3}$$

where $*$ denotes the convolution operation.

In 2021, Asif et al. [7] presented a novel method based on intrinsic image layer separation (IILS), consisting of three steps, namely, pre-processing, reflected layer separation and specular reflection detection. The process begins by smoothening the image via a low-pass filter. The image is then normalized via extraction of high gradient area, which extracts the intrinsic layer containing all the reflected areas using the Retinex algorithm. This process is repeated on the input image to separate reflected areas by the RGB color model. Each layer is then subtracted from the first reflected layer to get all the reflected pixels. The detected reflected pixels are converted to a magenta color. This marked area is then reconstructed by image melding technique and a patch-based optimization function. The iterative algorithm in this last step searches for a patch and votes for a color at every scale. The experiment was conducted with 912 endoscopic images from CVC-EndSceneStill. This method effectively separated and detected reflected areas, demonstrating superior performance over other methods in detecting specular reflections in endoscopic images.

Image Segmentation or Classification Another approach is to identify pixels as reflections or non-reflections by image segmentation or classification. Levine and Bhattacharya [55], used Retinex algorithm in combination with the region-based segmentation to separate specular and shadow regions by initialization of seed pixel. They implemented the Support Vector Machine (SVM) [16] classifier to separate the reflected and non-reflected areas based on the shadow boundary found. One of the merits of the technique at the time was that there was no need for any camera specification. However, as the seeded region requires the seed as an additional input and the segmentation results are dependent on the choice of seeds and the noise in the image can cause the seeds to be placed poorly.

Method	**Key Features**	**Application**	**Data Type**	**Works**
Thresholding	Based on intensity-based thresholding. Includes global and adaptive thresholding. Prone to false positives in bright regions.	Medical image analysis, general reflection detection.	Endoscopic images, RGB, HSV color spaces.	[73], [113], [78], [107], [98]
Dichromatic Reflection Model (DRM)	Separates specular and diffuse reflections by decomposing them into surface body and interface components. Efficient in separation but struggles with low-saturation colors. Requires light source chromaticity knowledge.	Object surface analysis, specular reflection separation.	Textured images, RGB color space.	[110], [127]
Retinex Theory	Separates image based on reflectance and illumination components	Medical imaging, particularly in endoscopic/microscopic imaging.	Endoscopic images, RGB color model.	[50], [7]
Image Segmentation or Classification	Uses techniques like Seeded-growing region segmentation and support vector machine. The basic principle is to either divide the image into regions or grow and segment it.	Facial images	RGB and HSV color spaces	[55]

Table 3.1: Traditional Methods for Specular Reflection Detection.

3.1.2 Deep-learning-based Methods

The advancement of deep-learning techniques has catalyzed significant progress in specular reflection detection, leading to the development of various state-of-the-art methods tailored to address the challenges in this domain. A wide range of research has been conducted to detect specular light reflections in images across different surfaces with medical imaging being one of the most common applications alongside machine inspections. Current state-of-the-art methods incorporate various deep-learning architectures, including fully convolutional networks (FCN), U-Nets and attention mechanisms to detect these regions and even reconstruct the missing regions.

One notable approach is the SpecSeg network proposed by Anwar et al. [5], which is based on the U-Net architecture and focuses on specular highlight detection in real-world images from the Specular Highlight Image Quadruples (SHIQ) dataset. SpecSeg demonstrated robust performance across diverse materials and lighting conditions, generating masks closely resembling ground-truth images and surpassing classical methods in specular region detection. The encoder-decoder network performs downscaling and upscaling operations. SpecSeg consists of five encoder blocks and four decoder blocks following the classic U-net pattern. Each encoder block consists of two 2D convolutional layers with filters (k) =3 and stride (s) =3 and uses the ReLU activation in the output of each convolution layer. To improve the robustness of the learned features, an incremental dropout of 10%, 20% and 30% are introduced between the convolution layers. The decoder performs upscaling via 2D transpose convolution layers with filters k=2 and s=2. The network then outputs mask images of the input. This network uses a linear combination of Dice similarity coefficient (DSC), which measures the pixel-level similarity of two images, and Focal loss, which addresses the class imbalance, as its loss functions. SpecSeg was successfully able to detect specular regions across a wide range of materials and lighting conditions. It was found that it could generate masks closely resembling ground-truth images and its results were significantly better than other classical methods. The results are shown in Figure 3.3.

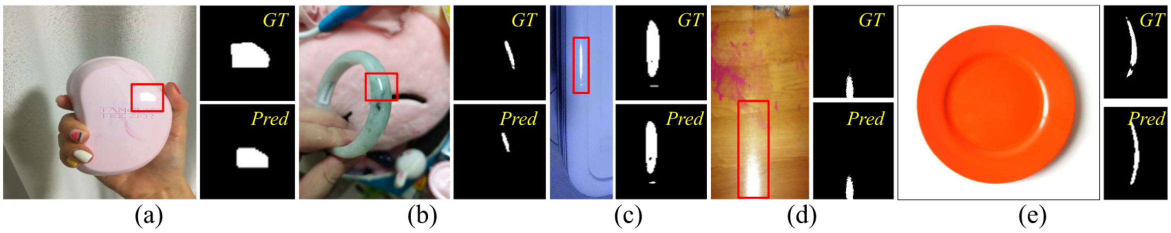


Figure 3.3: SpecSeg Results across different images [5].

Monkam et al. [72] introduced another prominent method is EasySpec, which employs a hybrid strategy utilizing Scaled-UNet for detection and GatedResUNet for suppression of specular reflections. It consists of two stages: detection and suppression. In the detection stage, the Scaled-UNet employs the benefits of weakly supervised and transfer learning concepts for specular detection. While the latter stage uses GatedResUNets which is based on the gated convolutions and deep impainting theory to restore specular reflection regions. The novel contributions of this paper include implementing gated convolutions to differentiate between specular and non-specular pixels, leverage U-Net architecture to mitigate the challenges of empirical parameter settings and to learn multi-level semantic representative features. Also, extensive empirical analyses have been conducted on a diverse dataset to validate the proposed framework's effectiveness. It was found that performance of EasySpec was superior to those of the state-of-the-art approaches in specular reflection suppression approaches.

Another U-Net based approach is the Efficient Two-Stage Highlight and Removal Network (ET-HDR) [61], featuring a Channel Attention Refinement Network (CARNet) and an efficiency spatial channel attention block (ESCAB) [128]. The first module is the U-Net which learns the context features of different scales in the highlight image. The second module is the Channel Attention Refinement Network (CARNet) which learns the spatial details of the image. Additionally, a feature fusion block (FFB) is developed to enrich the feature information. The Varifocal loss function is used to calculate the loss which improves model's ability to detect highlights leading to better training performance. A novelty of this method is that in the U-Net, its MDTA module in the transformer module was replaced with an efficiency spatial channel attention block (ESCAB) [128]. The ESCAB helps reduce the computational resources used and improve the inference speed of the model. The experiments were conducted on the Specular Highlight Detection and Removal Dataset (SHIQ) [27] and compared against methodologies which are trained on this dataset. The overall structure is depicted in Figure 3.4.

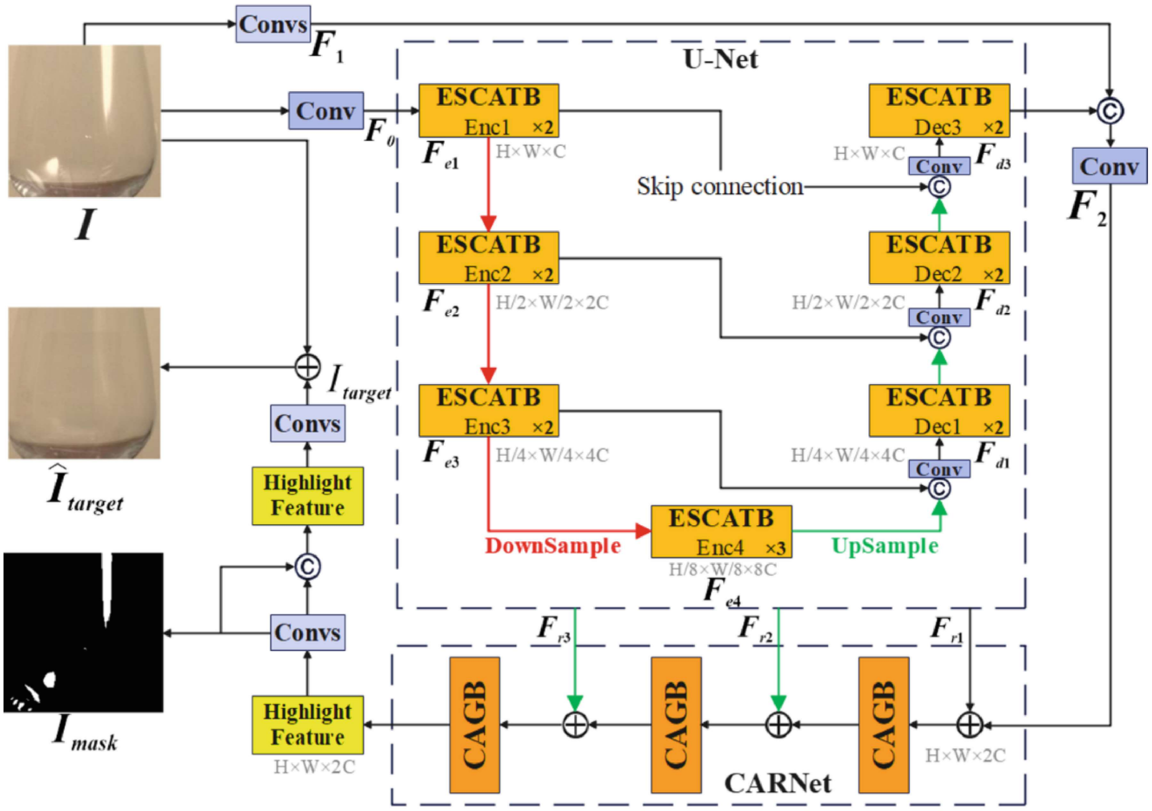


Figure 3.4: Overall structure of ET-HDR [61].

Zhou et al. [130] introduced a novel deep-learning-based framework, called DeepInspection, for automated defect detection on specular vehicle surface. The framework utilizes an attention-based fully convolutional neural network with Atrous Spatial Pyramid Polling (ASPP) [15]. The architecture consists of three parts: an encoder which learns the high-level features, an attention module which focuses on changing boundaries, and a decoder which reconstructs the spatial information. The encoder network is built on VGG-16 [103], composed of 13 convolutional layers, 5 down-sampling layers, and 3 fully connected layers. The attention module uses the attention gate (AG) to recognize salient regions in the image and retain the activations associated with the surface defects by pruning feature responses. The decoder uses the transposed convolutions from the encoder to reconstruct the spatial information. The feature maps derived from each component are concatenated, followed by ASPP to encode the rich semantic information. The experiment was conducted on the DeepInspection160 dataset with 160 manually labeled images. Various qualitative and quantitative assessments found that DeepInspection performed better than many state-of-the-art methods, by achieving an F1 score of 0.7513 (pixel level) and 0.8055 (component level).

In 2021, a multi-task network for Joint Specular Highlight Detection and Removal (JSHDR) consisting of the Dilated Spatial Contextual Feature Aggregation (DSCFA) module followed by the convolutional and ReLU layers to learn the attention map [27]. It is an encoder-decoder framework that uses skip connections to pass information between the encoder and decoder. It was found that the performance of JSHDR is comparable to state-of-the-art methods, especially in terms of handling spatially varying highlights via evaluation metrics like accuracy and bit error rate (BER). Furthermore, Esfahani and Wang [23] presented a deep network architecture for robust glare detection, which shares similarities with specular reflection detection. By considering both RGB and HSV representations and utilizing a modified U-Net architecture, this method demonstrated promising results for glare detection applications.

These state-of-the-art methods collectively showcase the advancements in specular reflection detection by leveraging deep-learning architectures. However, there is one major limitation in learning-based highlight detection. Typically, these methods are trained on synthetic data or very small datasets which are limited to a specific scenario; therefore, it is difficult to generalize the methods [27].

Method	Year	Dataset	Architecture	Metrics	Application	Key Features
SpecSeg [5]	2022	WHU [26], SHIQ [27]	U-Net based	Dice	Real-world	Encoder-decoder with 5 encoder blocks and 4 decoder blocks, incremental dropout, uses Focal loss. Effective across diverse materials and lighting conditions.
EasySpec [72]	2021	GLENDA [54]	Scaled-UNet, GatedResUNet	Dice, IoU, SNR, SSIM	Medical	Hybrid approach: detection (Scaled-UNet) and suppression (GatedResUNet). Employs gated convolutions, U-Net architecture, with extensive empirical validation. Outperforms state-of-the-art in suppression.
ET-HDR [61]	2023	SHIQ [27]	U-Net with CARNet, ESCAB	Accuracy, BER	Medical	Two-stage network with Channel Attention Refinement Network (CARNet), Efficiency Spatial Channel Attention Block (ESCAB), and Varifocal loss [89]. Improved highlight detection and computational efficiency.
DeepInspection [130]	2020	DeepInspection160	Attention-based FCN with ASPP	F1	Automobile Inspection	Utilizes VGG-16 based encoder, attention gate (AG), ASPP for defect detection on specular vehicle surfaces. Achieves high F1 scores.
JSHDR [27]	2021	WHU [26], SHIQ [27]	Encoder-decoder with DSCFA	Accuracy, BER	Real-world	Multi-task network for joint specular highlight detection and removal. Uses Dilated Spatial Contextual Feature Aggregation (DSCFA) module and skip connections for spatially varying highlights.
Esfahani and Wang [23]	2021	Glare Detection	Modified U-Net	Precision, Recall, F1, Accuracy	Defect Detection	Addresses glare detection using RGB and HSV representations with a modified U-Net architecture. Promising results demonstrated for glare detection applications.

Table 3.2: Summary of Deep-Learning Methods for Specular Reflection Detection.

3.2 Specular Reflection Datasets

Datasets play a fundamental role in the training, evaluation, and generalization of models. Deep-learning models learn patterns, features and relationships within the data. In supervised learning, datasets consist of labelled input-output pairs, where the model learns to map inputs to the correct outputs based on the provided examples. A diverse and representative dataset ensures that the model learns a wide range of features and generalizes to new data. Datasets are split into training, validation, and test sets to ensure effective model training, prevent overfitting, and evaluate performance. The training set is used to teach the model, the validation set helps tune hyperparameters and select the best model, and the test set provides an unbiased assessment of how the model performs on unseen data. Standard datasets allow for benchmarking and comparing different models and algorithms on the same data. This is essential for evaluating the relative performance of different approaches.

Datasets in computer vision and image processing primarily consists of images and videos for tasks such as object recognition, semantic segmentation and image generation. Some of the most popular datasets in this domain are MNIST [52] and ImageNet [19]. There are varying annotated methods to label the images and videos according to the application tasks. In object recognition, datasets are used to detect and localize multiple objects within an image. Each object is annotated with a bounding box and a class label. Semantic segmentation datasets are used to classify each pixel in an image into a predefined class. This task involves detailed pixel-level annotations, typically in the form of segmentation masks. However, it is difficult to find large datasets for specialized tasks such as medical segmentation or reflection detection. Research in specialized fields often requires generating custom datasets, such as for dent detection in aircraft using generative adversarial networks (GANs) [86]. Given the high cost and labor involved in manual data annotation, it is essential to train deep-learning models on existing, relevant datasets to address specific problem domains effectively.

Recent years have seen the development of numerous datasets for detecting reflections in images, spanning from medical imaging to real-world scenarios. Medical imaging contains various high-quality datasets for detecting reflections including 2D and 3D data. Some of the largest specular medical datasets include CVC-ClinicSpec [99] containing ground truth labels of 612 colonoscopy images and the DYY-Spec [79], which contains 1000 endoscopic specular images from various organs. One of the largest, the Specular Highlight Image Quadruples (SHIQ) [27], includes over 16,000 annotated images of specular highlights captured under various lighting conditions. Other notable datasets include the paired specular-diffuse dataset [124], the WHU-Specular dataset [26], and the multi-illumination images in the wild [76], which cover real-world scenes with metallic and shiny surfaces relevant for inspection tasks. These datasets will be used to evaluate proposed methods, with a summary provided in Table 3.3.

Dataset	Year	Category	Total Images	Size
CVC-ClinicSpec [99]	2017	Medical	612	-
Multi-Illumination [76]	2019	Real-world	25,000	6.7 GB
WHU-Specular [26]	2020	Real-world	4310	2 GB
Specular Highlight Image Quadruples (SHIQ) [27]	2021	Real-world	16,000	10.8 GB
Paired Specular-Diffuse Image [124]	2021	Real-world	13,380	7.1 GB
SIHRNet [120]	2022	Real-world	200	503 MB
DYY-Spec [79]	2023	Medical	1000	-

Table 3.3: Summary of Reflection Datasets.

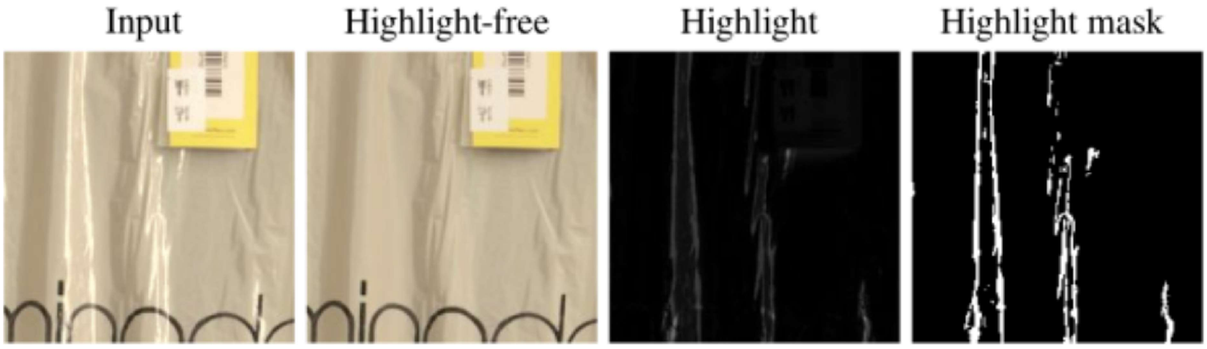


Figure 3.5: An illustration of the image quadruple in the SHIQ dataset containing highlight mask and highlight free images of the input [27].

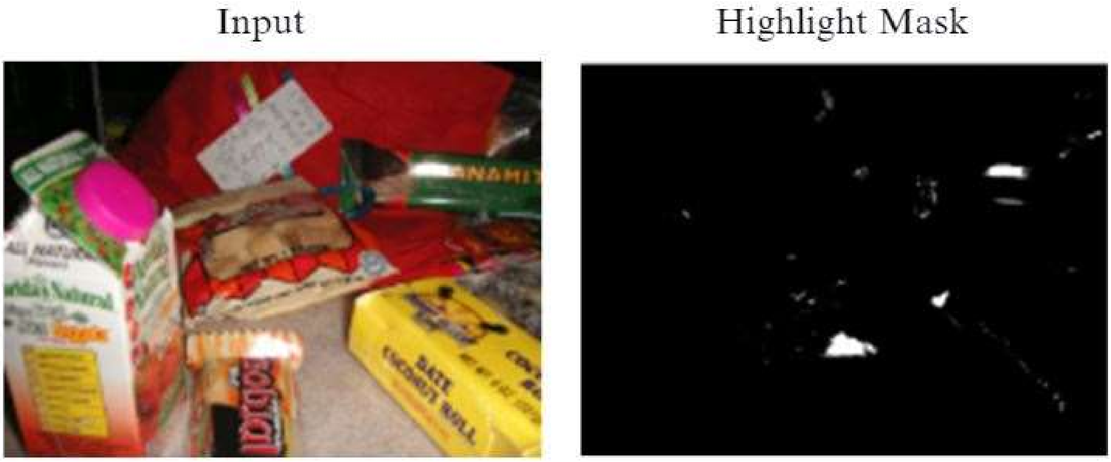


Figure 3.6: An illustration of the image pairs in the WHU-Specular dataset with the input image and its corresponding highlight mask [26].

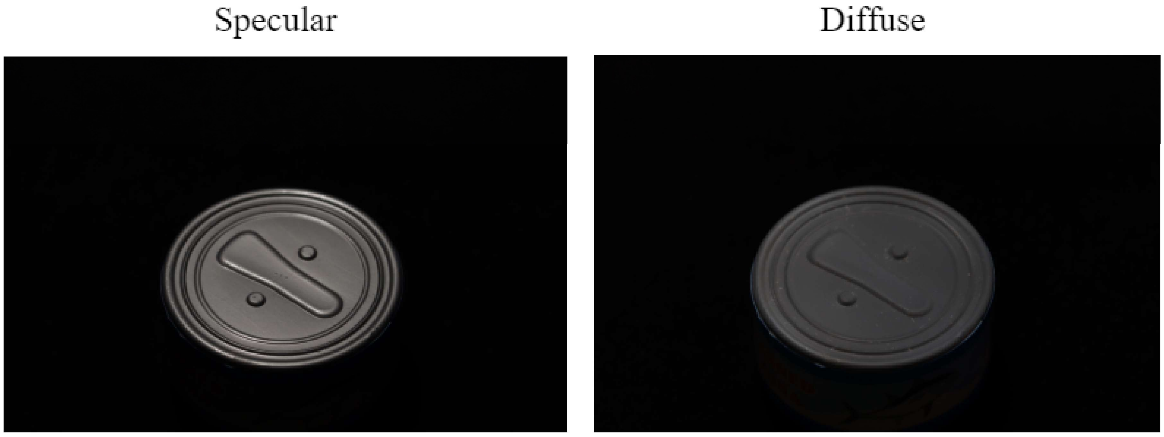


Figure 3.7: In PSD dataset [124], the images are captured in 2 polarization conditions, one with fixed polarization angles and the other with random polarization angles.

3.3 Reactive Planning for Autonomous Systems

Reactive planning is an adaptive approach used in robotics and artificial intelligence where systems respond dynamically to changes in their environment [29]. This is particularly valuable in unpredictable environments where conditions can rapidly evolve, such as in robot navigation around obstacles [9] or in unknown terrains [40]. In reactive planning, agents employ a combination of sensing, reasoning, and action execution. These sensors and data feed useful information to robots enabling them to plan their movements and react to their environment effectively.

Cameras, often combined with computer vision algorithms, allow robots to interpret visual information from their surroundings. They can detect objects, track movements, and recognize patterns, aiding in path planning and reactive behaviors. This is essential for applications in autonomous vehicles, where real-time object detection and classification of pedestrians and obstacles are crucial for safe navigation [3] [57].

A typical setup of a automatic visual inspection system consists of a robot manipulator, optical illumination, image acquisition subsystem, and image processing. By incorporating reactive planning, autonomous inspection systems can adapt their actions in response to real-time data, enhancing their flexibility and effectiveness. For example, when a robot encounters an area that is difficult to inspect due to environmental factors or unexpected conditions, reactive planning allows the system to modify its inspection path, adjust its sensors, or change its focus to ensure thorough coverage. An instance of this is demonstrated in a study about inspection of free-form specular surfaces [42]. The image acquisition subsystem captures the point cloud data and executes K-means algorithm to segment the free-form regions. These irregular shaped regions are passed as input to the path planning algorithm which outputs a scanning path based on the shortest path criteria and the acquisition model of line scan camera [42]. This adaptability is essential for maintaining high-quality inspections in complex and variable environments where pre-planned paths might not account for all possible issues. An example of a robot setup for inspection task in depicted in Figure 3.8.

The need for reactive planning arises from the unpredictability of real-world inspection scenarios and the deformities in the surfaces of the inspection objects. Without the

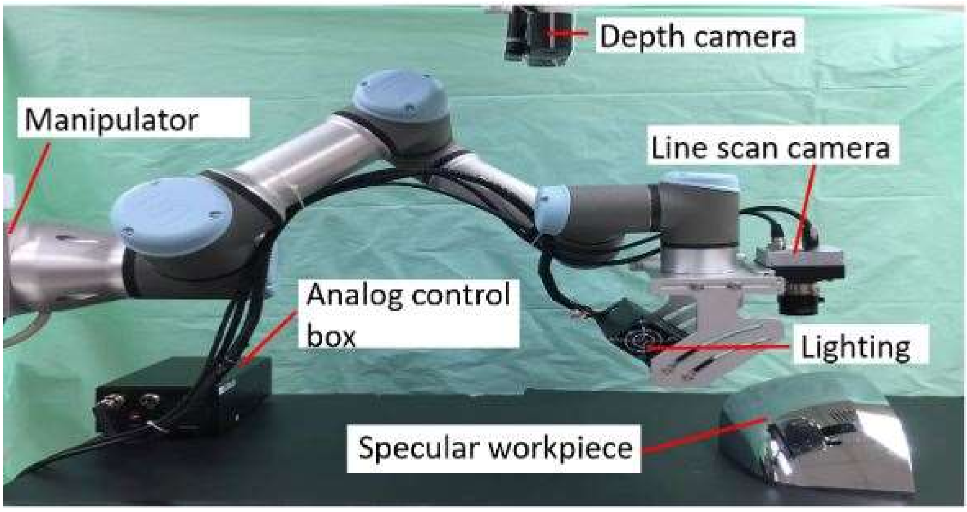


Figure 3.8: Experiment setup of a robot inspecting an object [42].

ability to react and adjust on the fly, autonomous systems risk missing critical defects or failing to adapt to unforeseen challenges, potentially compromising the inspection's accuracy and reliability. Reactive planning ensures that the system remains effective and responsive, improving overall inspection performance and reducing the likelihood of errors or omissions. Reflection detection aids reactive planning by identifying and localizing where reflections obscure features during an inspection. This allows the system to adjust the robot's camera angles to capture these hidden areas from different perspectives. Consequently, it ensures a more comprehensive and accurate inspection by revealing details that were previously obscured.

Chapter 4

Methodology

This chapter explains the novel contributions of this thesis including the proposed transformer-based segmentation model UNETR Attention Fusion (UNETR-AF) and the Inspection dataset that consists of inspection images and its corresponding reflection highlight. This chapter also delves into the other U-Net based segmentation models implemented in this thesis. The source code and dataset for this thesis is available on GitHub [1].

4.1 UNETR-Attention Fusion (UNETR-AF)

The hypothesis for this study centers on the idea that semantic segmentation models based on the U-Net architecture can be adapted and optimized to effectively detect and classify reflections in inspection images. Most of the current state-of-the-art models for reflection detection implement U-Net architecture as discussed in Chapter 3. Fully convolutional networtks, particularly U-Net, have been highly effective in image segmentation tasks across many fields which deal with segmentation tasks of varying sizes from biomedical applications [8] [93] to inspection tasks [66]. Its design efficiently captures spatial information, making it a logical choice for detecting reflections in inspection images, which often feature subtle boundaries and irregular shapes. U-Net's down-sampling encoder and up-sampling decoder pipeline are instrumental in extracting features while preserving spatial details, essential for precise reflection detection.

Detecting reflections under various lighting conditions requires distinguishing between subtle changes in intensity and texture. Standard U-Net may struggle with this due to its reliance on local receptive fields, which can limit its ability to capture global contextual information in complex reflective environments. This limitation led to the inclusion of variants that incorporate transformers, such as UNETR [35], TransUNet [14] and attention mechanisms [82] to improve the model's ability to detect nuanced features across different scales and improve segmentation performance under varying illumination.

The Vision Transformer (ViT) [21] (see Section 2.3.3) has shown promise in capturing long-range dependencies due to its attention-based architecture. By incorporating ViT as the encoder in the U-Net, the model gains the ability to interpret relationships across entire images, which is particularly beneficial in handling variations in lighting and reflection patterns. This is observed in, UNETR [35], which uses ViT as the encoder and the TransUNet [14] which implements a CNN-ViT hybrid encoder. ViT's global receptive field enables it to understand contextual relationships that convolutional layers might

[1] Thesis GitHub Link: `https://github.com/Aditi-Mhatre/reflection-detection`

miss, which is crucial for detecting reflections that may vary based on angle and intensity. As a pure transformer cannot directly accept an image input (H x W), the image is divided into patches ($\frac{H}{P}$ x $\frac{W}{P}$) and which are converted to patch embeddings. These patch embeddings are a key feature of ViT, which allows a pure transformer to process images by applying a simple linear transformation to the flattened pixel values of the patch [21]. However as the patches ($\frac{H}{P}$ x $\frac{W}{P}$) are smaller than the original image resolution (H x W), it results in a loss of low-level details such as the boundaries and shape of the reflections, even with the skip connections.

To address the identified challenges, the UNETR Attention Fusion (UNETR-AF) model is proposed as an extension of the UNETR architecture [35]. This model leverages the Vision Transformer (ViT) as its encoder, utilizing its 12 attention heads to achieve a global context and capture dependencies across the entire image. To efficiently integrate information from the encoder to the decoder, skip connections are added after every three attention heads, linking to four convolutional decoder blocks (CNNs). These skip connections allow the model to fuse detailed spatial information progressively from the transformer encoder into the CNN-based decoder, which helps enhance the segmentation of reflections. However, ViT-based architectures can struggle with local feature learning due to the loss of image resolution in patches, especially in smaller datasets with reflections. This is where the need for additional attention mechanisms, such as spatial and channel-wise attention, becomes critical. Attention mechanisms help refine and balance both global and local features, ensuring that fine details are not lost during encoding and decoding. By focusing on the most relevant channels and spatial regions, attention blocks can compensate for ViT's limitations in spatial localization during tasks that involve dense prediction, like segmentation. In models that incorporate ViT within U-Net structures, there is a trend toward embedding more attention mechanisms to improve performance in segmentation tasks such as CFATransUnet [115], which introduced channel-wise cross fusion attention and TSCA-Net [28], that used spatial-channel attention blocks. A new version of UNETR was also proposed recently called UNETR++ [100] which introduced efficient paired attention (EPA) block to for improved spatial and channel attention in volumetric medical data.

Therefore two attention modules are integrated, namely the Squeeze-and-Excitation (SE) [39] module and the Convolutional Block Attention Module (CBAM) [122]—into both the skip connections and decoder stages. These attention modules aim to balance the global and local feature learning of the model.

Skip Connections with Squeeze-and-Excitation (SE): UNETR-AF retains the powerful self-attention mechanism of ViT in the encoder, which excels at capturing global dependencies in images. However, a known limitation of ViT is its difficulty in learning local spatial features, particularly on smaller datasets. To address this, SE modules [39] are integrated into the skip connections. These modules emphasize important channel-wise features, allowing the model to selectively focus on the most relevant information at each resolution level during the encoding process. This channel-wise refinement compensates for ViT's limited ability to capture local context. This allows the model to emphasize important channels while suppressing less relevant ones, enhancing the feature maps transferred from the encoder to the decoder. SE modules adaptively refine the features extracted at different layers, helping the model focus on the most critical information as it progresses through the network.

Decoder with Convolutional Block Attention Module (CBAM): CBAM [122] is used to enhance both spatial and channel attention in the decoder. By applying attention hierarchically, CBAM refines spatial feature extraction during the upsampling stages, where fine-grained details must be reconstructed. This spatial refinement is critical because ViT architectures tend to lose spatial resolution in upsampling, whereas CNN-based decoders typically excel in reconstructing spatial details.

The UNETR Attention Fusion (UNETR-AF) is depicted in Figure 4.1 where the SE blocks and CBAM module are included in the skip connection and the decoder.

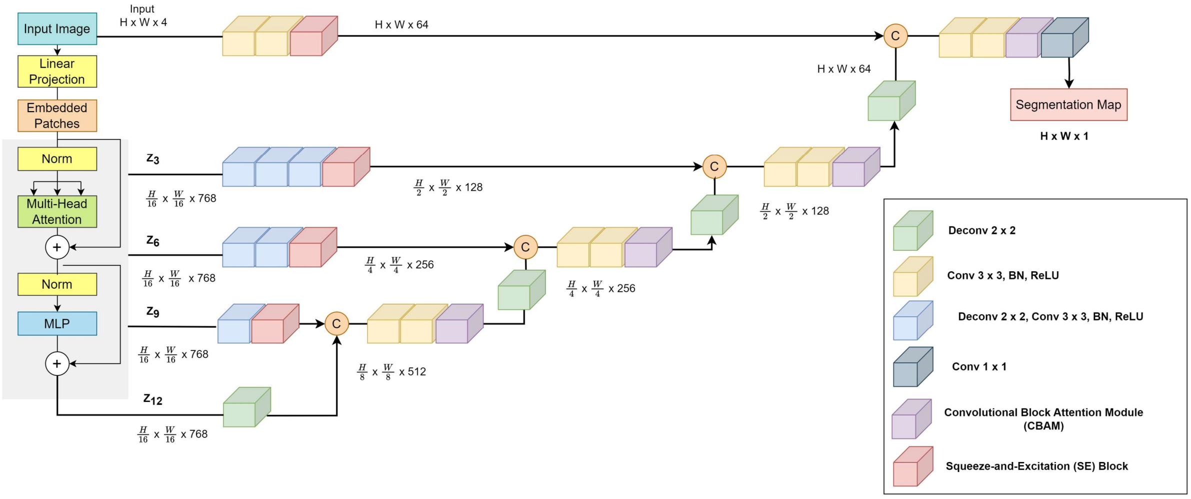

Input Image
Linear Projection
Embedded Patches
Norm
Multi-Head Attention
+
Norm
MLP
+
Input
H x W x 4
H x W x 64
C
H x W x 64
Segmentation Map
H x W x 1
z_3
$\frac{H}{16}$ x $\frac{W}{16}$ x 768
$\frac{H}{2}$ x $\frac{W}{2}$ x 128
C
$\frac{H}{2}$ x $\frac{W}{2}$ x 128
z_6
$\frac{H}{16}$ x $\frac{W}{16}$ x 768
$\frac{H}{4}$ x $\frac{W}{4}$ x 256
C
$\frac{H}{4}$ x $\frac{W}{4}$ x 256
z_9
$\frac{H}{16}$ x $\frac{W}{16}$ x 768
C
$\frac{H}{8}$ x $\frac{W}{8}$ x 512
z_{12}
$\frac{H}{16}$ x $\frac{W}{16}$ x 768
Deconv 2 x 2
Conv 3 x 3, BN, ReLU
Deconv 2 x 2, Conv 3 x 3, BN, ReLU
Conv 1 x 1
Convolutional Block Attention Module (CBAM)
Squeeze-and-Excitation (SE) Block

Figure 4.1: Proposed UNETR Attention Fusion (UNETR-AF).

By combining SE for channel refinement in the skip connections and CBAM for spatial and channel attention in the decoder, UNETR-AF achieves a balance between global attention from the ViT encoder and local detail preservation in the CNN decoder. This addresses a common challenge in hybrid ViT-CNN architectures, ensuring that both local and global features are effectively represented, leading to improved segmentation results, particularly in complex or noisy regions. This proposed architecture, UNETR-AF, is tailored to optimize both global context and local detail in reflection segmentation, providing a valuable tool for robust and reactive planning in autonomous inspection systems under diverse environmental conditions. Moreover, CBAM and SE are lightweight modules therefore it does significantly increase the model parameters and complexity compared to UNETR.

4.2 Inspection Dataset

Datasets are required to train segmentation models as they provide ground truth to understand the regions that need to be segmented. Reflections in images is a specialized field, which is further narrowed down in inspection images. As there are no publicly available datasets for reflections in inspection images, it is important to prepare an Inspection dataset for this purpose. The preparation of this dataset is one of the novel contributions of this thesis. The objective is to prepare an Inspection dataset containing inspection images of different objects and materials captured under varying light conditions. This section describes the preparation of the dataset from data collection, annotation to the characteristics of the dataset.

4.2.1 Design and Requirements

There are various existing datasets for reflections based on medical images or real-world objects which are mentioned in Section 3.2. Most of these datasets, especially SHIQ and WHU datasets, were referred to while designing this inspection dataset. The first step was to conceptualize the illumination scenes, which included the object and the light conditions under which it would be captured. The primary requirements were to select objects composed of materials commonly found in aircraft inspections which were: aircraft wing, fuel tank, turbine blade, other laboratory tools and equipment such as screws, hexagon wrench, screw plate, swivel caster wheels, T-slot aluminium. These objects are made of metals like aluminium and stainless steel, which are highly reflective as well as non-reflective materials like polyurethane that produce softer reflections.

The next step was to define the light conditions and the illumination modes. From figure 2.4, two illumination modes were incorporated: bright field forward and diffuse lighting. Each of the objects mentioned were captured in different light conditions and angles to produce a diverse dataset. For diffuse lighting, the objects were photographed in a 60x60x60 light box, that had white LEDs fixed at the top of the box over a translucent cloth which helped diffuse the light. The inner surface of the box was lined with aluminium foil which was covered with a black Polyvinyl Chloride (PVC) background film for certain scenarios. These RGB images were captured by the RealSense D435i embedded in the Eeloscope inspection robot and a mobile phone camera. The RealSense allowed to photograph images with two settings: high exposure which also included an external light source and the auto-exposure where no additonal light source was used. The Eeloscope is equipped with a LED light within the robot that supports bright field forward lighting. In addition, there were two more light sources, the flashlight of the mobile phone and a LED video light panel. The light panel has a wide range of color temperature from 3200k to 5600k as well as the light intensity, which ranges between 1 and 6. While collecting the images, the light intensity was kept between 2 and 6, as it was the ideal range to produce varying reflections. The light conditions in the scenarios included a mix of direct light from the external light sources, natural light and low light.

Table 4.1 summarizes the scenarios for the dataset.

Object	**Material**	**Light Conditions**	**Camera**
Aircraft Wing	Metal	Direct LED, Natural, Low light	Mobile, RealSense
Fuel Tank	Metal	Direct LED, Low Light	Mobile, RealSense
Turbine Blade	Metal	Direct LED, Natural, Diffuse Lighting	Mobile, RealSense
Hexagon Wrench	Metal	Direct LED, Natural, Diffuse Lighting	Mobile
Swivel Castor	Metal, Rubber	Direct LED, Natural, Diffuse Lighting	Mobile
T-slot Aluminium	Metal	Direct LED, Natural, Diffuse Lighting	Mobile
Screws Plate	Metal, Composite	Direct LED, Natural, Diffuse Lighting	Mobile

Table 4.1: Dataset Preparation: Objects and Light Conditions.

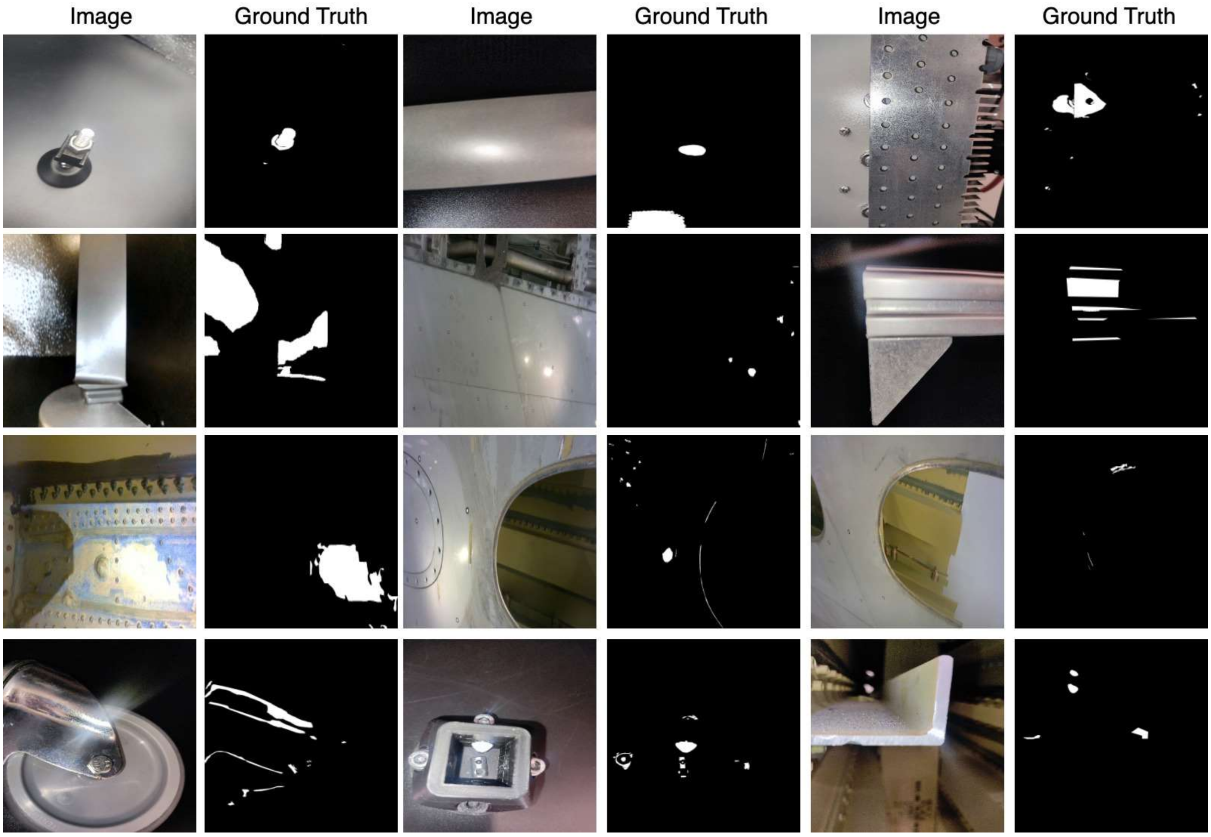


Figure 4.2: A few sample image-mask pairs from the Inspection dataset. The images captured here are of different inspection objects with varying levels of light reflections.

4.2.2 Data Collection and Annotation

Based on the scenarios in 4.1, a total of 1025 images were captured. These RGB images depicted both specular and diffuse reflections as well as no reflections. Once this data was collected, the next step was to process the images by renaming them according to the identification numbers and recording the metadata such as light conditions, camera type. The images are then annotated. The data was annotated using Labelbox [49], an online data labelling platform. An annotation pipeline was created on the platform to annotate the reflection and non-reflection segments for image segmentation tasks. The labelling pipeline was divided into the following stages: to label and to review, which enabled to verify the labels and redo the rejected tasks. Every image is segmented into "reflections" and "no reflections" using the draw tool. The labelling guidelines were defined as follows:

1. A region is marked as "reflection" for specular reflections and for diffuse reflections, when it is difficult to see the surface details. The rest of the region is marked as "no reflection".

2. The regions labelled as "reflection" are colored in white (#FFF) and the regions labelled as "no reflections" are colored in black (#000).

Moreover, metadata was added to each image to indicate its light conditions, color temperature , in cases where the panel light was used, and the camera type. The labelling data from the Labelbox was saved into JSON format for further post-processing. As a smartphone was also used to capture the images, the orientation from Exchangeable Image

File (EXIF) data was analysed and processed to ensure that orientation of the images and the generated masks are consistent. The data collection and the labelling process of the dataset is shown in Figure 4.3.

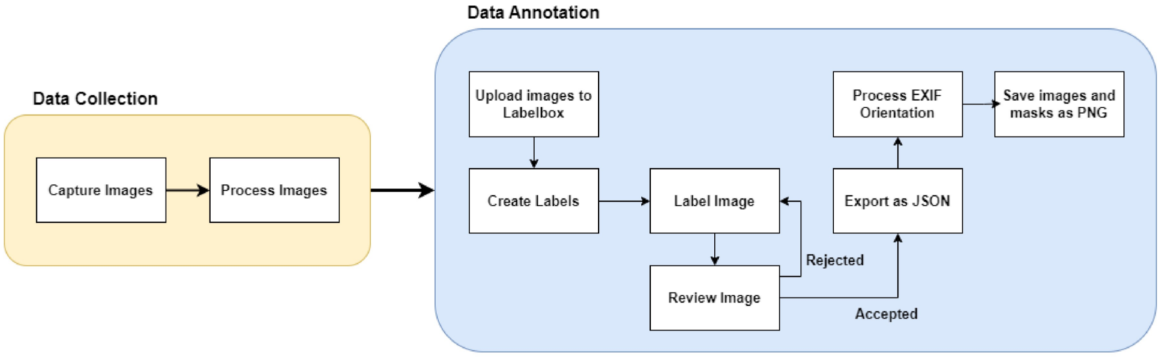


Figure 4.3: Inspection Dataset Pipeline consists of two parts: Data Collection and Data Annotation.

The binary segmentation masks generated are similar to the highlight masks of SHIQ [27] and WHU datasets [26]. All the images and masks are saved in the portable network graphics (PNG) format. The Inspection Dataset consists of 1025 image-mask pairs. An illustration of an image-mask pair from the Inspection dataset is depicted in Figure 4.4.

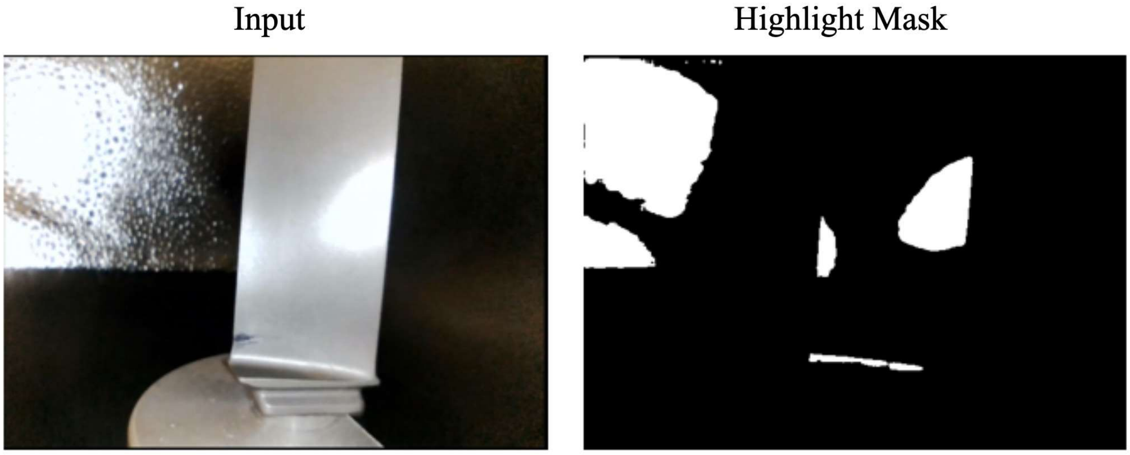


Figure 4.4: An illustration of the image pair in the Inspection Dataset with the input image and its corresponding ground truth.

The dataset is further split into train and test subsets in a ratio of 80:20. The train subset contains 820 images and is used for training purposes due to the diversity of reflections captured. The test subset consists of 205 images, most of which are of airplane wing, the interior of hydrogen fuel tank and the turbine blades. The split was decided based on the t-SNE (t-distributed stochastic neighbor embedding) [112] representation of the dataset which is used for feature visualization of images.

4.3 Models and Architecture

Various state-of-the-art models implement different U-Net architecture variants for segmentation. In this section, the U-Net based models and architectures applied for this thesis are explained. These architectures allow precise detection of reflections due to their structure which effectively learns the features from the datasets discussed in the previous section. The architectures mentioned are U-Net, Attention U-Net, U-Net++ and UNETR, which will be employed to evaluate the segmentation performance against UNETR-AF.

4.3.1 U-Net

U-Net was originally proposed for localization tasks in biomedical image processing and is based on a fully convolutional network (FCN), designed to achieve precise segmentation with fewer training images and through data augmentation [93]. The network architecture, illustrated in Figure 4.5, consists of a contracting path (left) and an expansive path (right). The contracting path is a typical CNN architecture which includes repeated application of two 3x3 convolutions, each followed by a rectified linear unit (ReLU), and a 2x2 max-pooling operation with stride 2 for downsampling. The number of feature channels doubles at each downsampling step, progressively capturing more complex features. In the expansive path, each step starts with upsampling the feature map, followed by a 2x2 convolution that halves the number of feature channels. This is then concatenated with the corresponding cropped feature map from the contracting path. Two 3x3 convolutions, each followed by a ReLU, are applied. At the final layer, a 1x1 convolution maps each 64-channel feature vector to the desired number of classes [93]. As the contracting and expansive paths are symmetrical, it results in the characteristic U-shaped architecture of the network. One of its novelty is that it passes the contextual information throughout the network, allowing it to segment objects using context from a broader surrounding area.

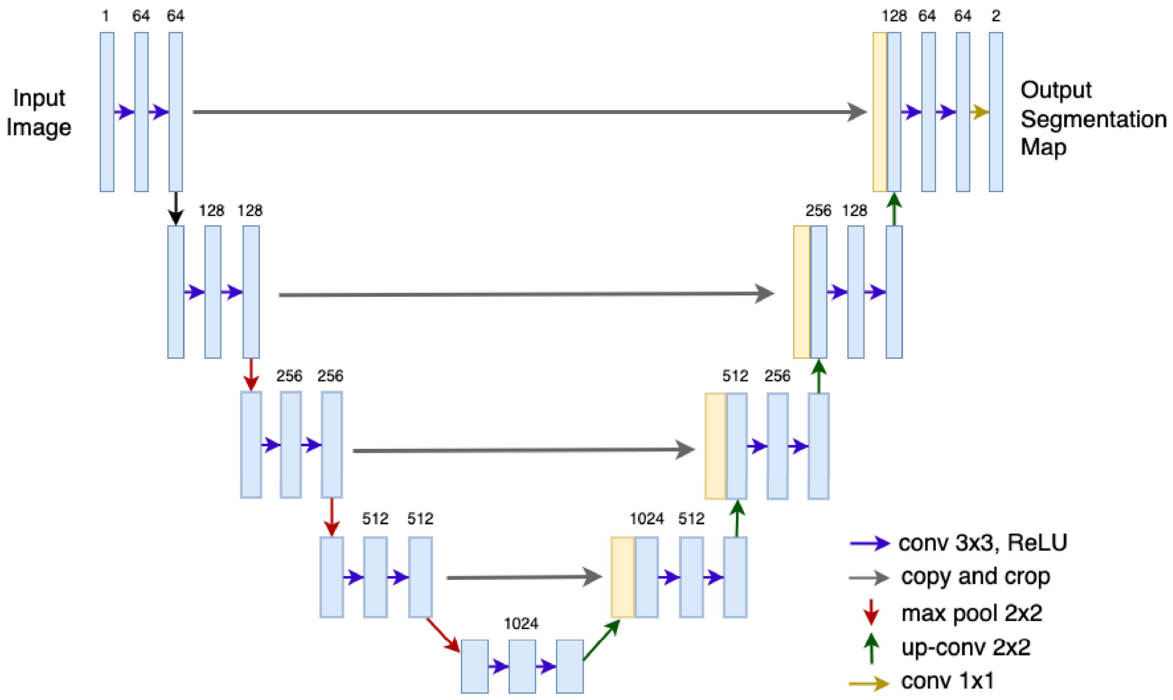


Figure 4.5: U-Net Architecture, redrawn from [93].

The energy function is computed using a pixel-wise softmax over the final feature map, combined with the cross-entropy loss. The softmax function is defined as:

$$p_k(x) = \frac{\exp(a_k(x))}{\sum_{k'=1}^{K} \exp(a_{k'}(x))}, \tag{4.1}$$

where $a_k(x)$ is the activation at pixel $x \in \Omega$ for class k, and K is the number of classes. The softmax approximates a maximum function, with $p_k(x) \approx 1$ for the class with the highest activation and $p_k(x) \approx 0$ for all others.

The cross-entropy loss penalizes deviations of $p_{\tau(x)}(x)$ from 1, where $\tau : \Omega \to \{1, \ldots, K\}$ represents the true label of each pixel. The energy function is given by:

$$E = \sum_{x \in \Omega} w(x) \log(p_{\tau(x)}(x)), \tag{4.2}$$

where $w(x)$ is a weight map that assigns different importance to each pixel during training.

U-Net has been applied as the base model to various areas of segmentation tasks for including transparent object detection [59] [97], industrial defect detection systems [86] [117] [69] [77], and medical images segmentation [48] [8].

4.3.2 Attention U-Net

In image segmentation, attention is used to highlight important regions in an image more than other regions. This helps in increasing efficient use of computational resources by not wasting it on irrelevant areas. Attention can be categorized into hard and soft. Hard attention process only one region at a time, which makes it non-differentiable and difficult to train. This means that for any given image, it can either pay attention or no attention at all to a region. Soft attention is probabilistic and differentiable. It weighs different parts of the image by assigning weights to a region depending on its relevance. During training, there is more focus given to regions with higher weights. As it is differentiable, it can be trained with standard back propagation, which allows the model to be better at deciding which regions to pay more attention as the training process continues. In the U-Net, skip connections are used to preserve spatial information from the downsampling path to the upsampling path, however this also transfers most of the redundant low-level feature extractions that do not contribute to better segmentation [82]. To overcome this issue, attention gates are introduced to add soft attention to the skip connections.

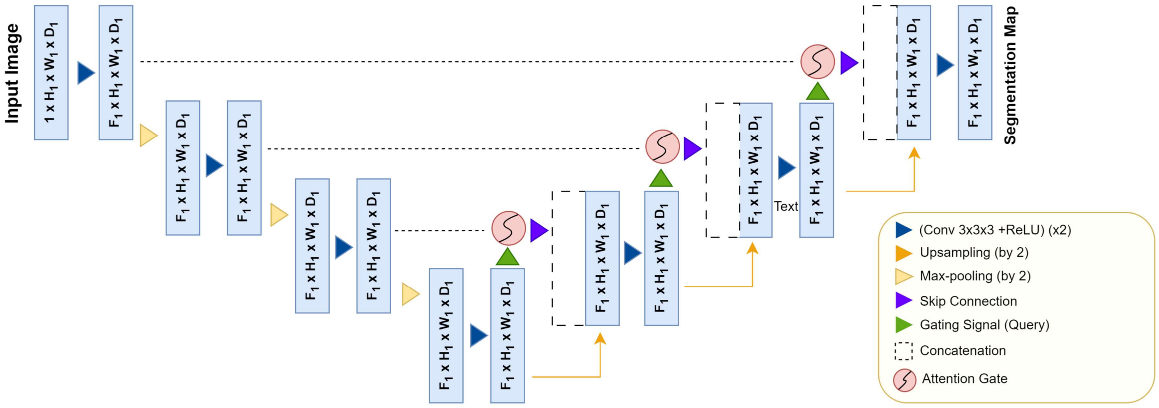


Figure 4.6: Attention U-Net Architecture [82].

The authors in the original paper [82] visualized the attention mechanism in a grid, which depicts the effect of soft attention as the training process continues.

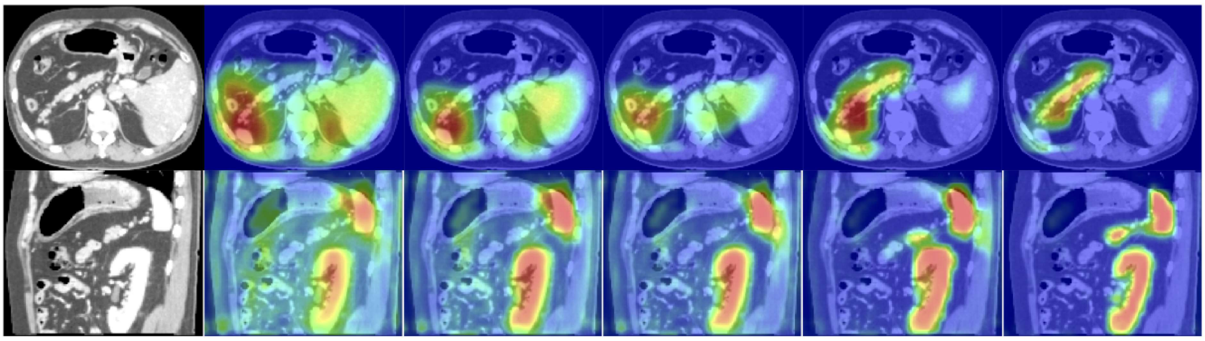


Figure 4.7: Attention Coefficient during Training - the areas marked by red show soft attention [82].

4.3.3 U-Net++

U-Net++ is an advanced variant built upon the U-Net architecture which uses nested and dense skip-connections to connect the encoder and decoder sub-network [131]. It introduces a nested U structure, re-designed skip pathways and deep supervision. The aim is to reduce the semantic gap in the feature maps between the encoder and decoder at different scales which would lead to learning the task more easily. The architecture is illustrated in Figure 4.8.

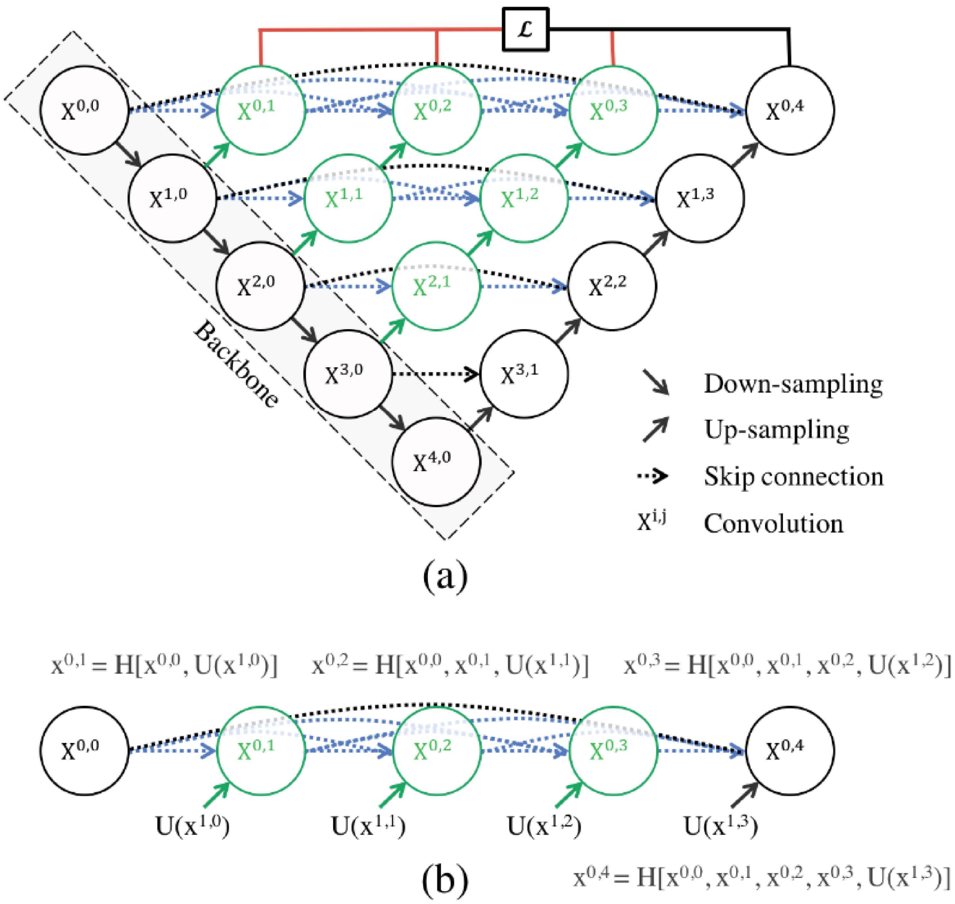


Figure 4.8: UNet++ Architecture [131]. (a) The encoder and decoder are connected through a series of nested dense convolutional blocks to bridge the semantic gap in the feature maps with three convolution layers. The convolutional layers are depicted in green with the dense skip connections on skip pathways shown in blue and deep supervision indicated by red. The convolutional blocks in black are from the original U-Net structure. (b) The first skip pathaway of UNet++ is visualized.

The architecture also incorporates two key features: re-designed pathways and deep supervision.

Re-designed Pathways: Instead of directly passing the feature maps from the encoder to the decoder as in the original U-Net, U-Net++ applies multiple intermediate convolutions at each level of the skip connection before being passed to the decoder. This way, the feature maps passed through the skip connections carry more refined, semantically meaningful information. It is hypothesized that this would make the optimization problem easier for the optimizer when the encoder feature map and the corresponding decoder feature map are semantically similar [131]. This skip connection is depicted in Figure 4.8b.

Deep Supervision: Deep Supervision is a regularilization technique introduced in [53]. The core idea behind it is to apply intermediate loss functions at multiple stages within the network, in addition to the final loss at the output. This encourages the model to make accurate predictions at different levels of its architecture rather than only at the final layer. In U-Net++, deep supervision allows the model to function in two modes: (1) accurate mode, where the outputs from all segmentation branches are averaged for higher precision, and (2) fast mode, where the final segmentation map is obtained from a single segmentation branch, with the branch selection controlling the degree of model pruning and speed improvement [131].

4.3.4 UNETR

UNEt TRansformer (UNETR) [35] is a vision transformer-based deep learning architecture originally proposed for 3D medical segmentation. The idea is to utilize the vision transformer (ViT) [21] as the encoder and combine it with the original U-Net-based CNN decoders. It leverages the powerful self-attention mechanism of transformers to capture global dependencies in images, addressing some limitations of traditional CNN-based encoders used in U-Net. The encoder has 12 multi-attention heads with a skip connection to the decode after every three attention heads. In the original paper, the skip connections were for multi-head layers 3,6,9 and 12. The structure is represented in Figure 4.9.

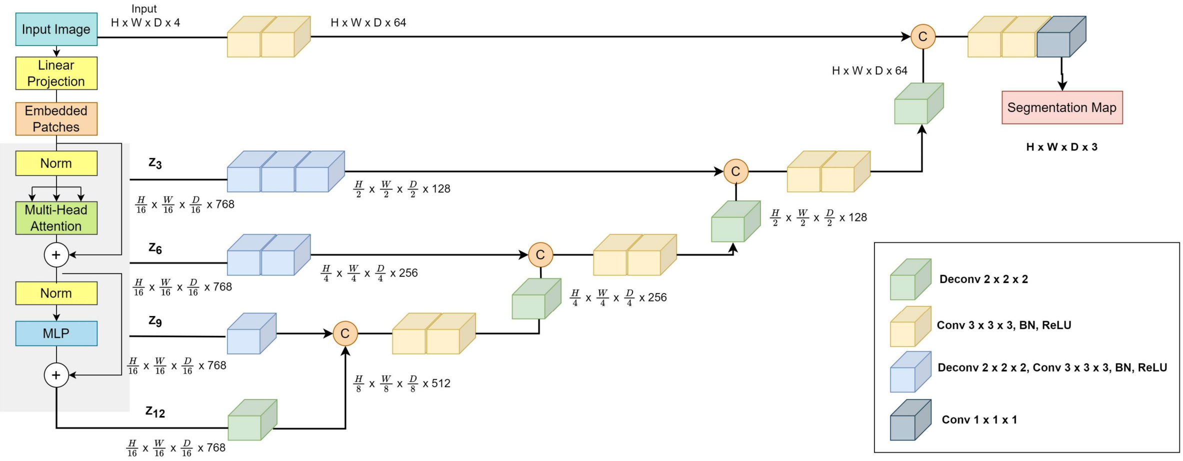

Input Image
Input
H x W x D x 4
H x W x D x 64
Linear Projection
Embedded Patches
Norm
Multi-Head Attention
+
Norm
MLP
+
z_3
$\frac{H}{16}$ x $\frac{W}{16}$ x $\frac{D}{16}$ x 768
$\frac{H}{2}$ x $\frac{W}{2}$ x $\frac{D}{2}$ x 128
z_6
$\frac{H}{16}$ x $\frac{W}{16}$ x $\frac{D}{16}$ x 768
$\frac{H}{4}$ x $\frac{W}{4}$ x $\frac{D}{4}$ x 256
z_9
$\frac{H}{16}$ x $\frac{W}{16}$ x $\frac{D}{16}$ x 768
$\frac{H}{8}$ x $\frac{W}{8}$ x $\frac{D}{8}$ x 512
z_{12}
$\frac{H}{16}$ x $\frac{W}{16}$ x $\frac{D}{16}$ x 768
C
C
C
C
$\frac{H}{4}$ x $\frac{W}{4}$ x $\frac{D}{4}$ x 256
$\frac{H}{2}$ x $\frac{W}{2}$ x $\frac{D}{2}$ x 128
H x W x D x 64
Segmentation Map
H x W x D x 3
Deconv 2 x 2 x 2
Conv 3 x 3 x 3, BN, ReLU
Deconv 2 x 2 x 2, Conv 3 x 3 x 3, BN, ReLU
Conv 1 x 1 x 1

Figure 4.9: UNETR Architecture based on [35].

Since this thesis focuses on 2D images, the UNETR architecture has been adapted to process 2D inputs and produce corresponding segmentation maps. The adapted UNETR follows the same operations as the 3D UNETR but removes the depth dimension. The ViT splits the input image into non-overlapping patches, using self-attention mechanisms to capture long-range dependencies across the entire image. The encoder's skip connections at different layers feed into the decoder, combining high-level global features with fine-grained spatial details. In addition to its novel architecture, UNETR also benefits from multi-scale feature learning, where the ability to combine global attention from the transformer with the hierarchical decoding structure improves the segmentation of both large and small structures in medical images. This ability to handle multi-scale features makes UNETR highly effective for volumetric medical images. Furthermore, experiments in the original paper demonstrated that UNETR significantly outperforms conventional U-Net models on brain tumor and spleen segmentation tasks, showcasing its effectiveness in real-world clinical settings.

Chapter 5

Experiments and Results

5.1 Experimental Setup

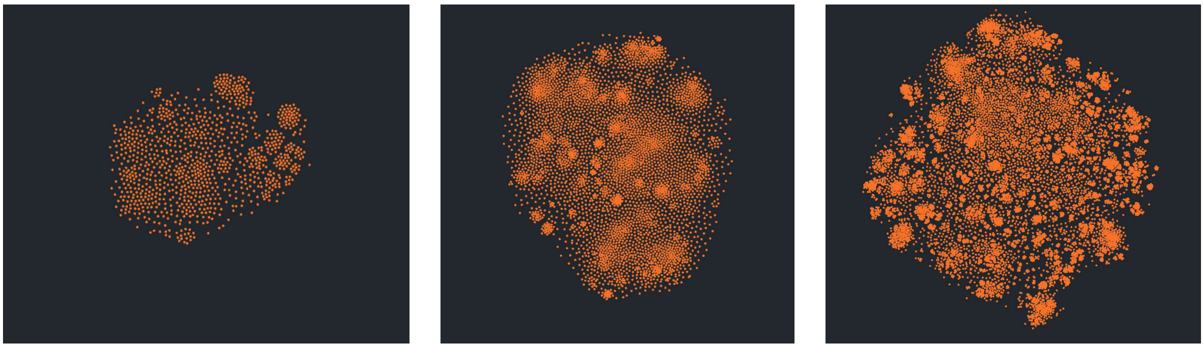
This chapter outlines the procedures and configurations used to conduct the experiments in this work. It details the datasets, model architectures, training processes, and evaluation methods applied to achieve the results. The experimental setup is designed to test the performance of various models on different datasets, with a particular focus on the segmentation tasks. Additionally, this section explains the hardware and software environments used to facilitate the experiments, including pre-processing and post-processing steps to ensure the reliability and robustness of the results. Through a structured methodology, the aim is to provide a clear understanding of the workflow that led to the final outcomes.

5.1.1 Datasets

The outcomes of this study heavily rely on the characteristics of the datasets used for experimentation. The experiments are run on three datasets, namely: Specular Highlight Image Quadruples (SHIQ), WHU, and the Inspection datasets. The purpose of the thesis is to detect reflections in inspection images, therefore the specialized inspection dataset is introduced, which consists of objects and scenes found during industrial visual inspections. The dataset contains 1025 images, which are partitioned into a training set of 820 images and a test set of 205 images. The train set is further randomly split into 80:20 ratio, for the training and model evaluation. As mentioned in Table 3.3, SHIQ dataset consists of quadruple which also depicts highlight-free and highlight intensity images to aid in the removal of specular highlight. However, as the aim of this study is to detect specular reflections, only the input image and the highlight mask are used for the experiments. The SHIQ dataset is divided into separate train and test subsets, with the test subset used for validation in this study. Similarly, WHU is a dataset that captures reflections in real world scenarios. The dataset consists of 5000 images, which were randomly split into train and validation sets in 80:20 ratio for model training and evaluation purposes.

SHIQ and WHU were included for training because they are among the largest datasets available for real-world reflection images. The goal was to determine whether models trained on these generalized datasets could accurately detect reflections in industrial inspection scenarios, which involve varied lighting conditions and a limited range of materials. The models trained on each of these datasets is tested on the test dataset of inspection images to compare its performance. The aim is to observe how models trained on larger and generalized datasets like SHIQ and WHU perform on inspection images as compared to the inspection dataset, and evaluate whether the domain shift between the three datasets has any effect on the performance of the models.

t-distributed stochastic neighbor embedding (t-SNE) [112] is a nonlinear dimensionality reduction method to visualize high-dimensional dataset on a lower dimensional space of two or three-dimensional. The t-SNE visualization of the Inspection dataset is shown in 5.1 (a). Based on this distribution, the dataset is divided into train and test. The three clusters on the extreme right in the figure as well as the slightly larger cluster on top right sampled as the test dataset, while the remaining are placed in the train dataset. t-SNE is implemented to visualize the differences in the three datasets, and derive the domain shift amongst from distributions. The differences are visualized in the Figure 5.1 below.



(a) Inspection Dataset (b) WHU Dataset (c) SHIQ Dataset

Figure 5.1: t-SNE Visualizations of Inspection, WHU, and SHIQ Datasets. The Inspection dataset shows a more compact distribution, while the WHU and SHIQ datasets have wider distributions.

5.1.2 Model Architecture

U-Net (see Section 4.3.1)is one of the models implemented in this study due to it being the base architecture for the current state-of-the-art reflection detection methods. There are five types of models implemented which can be categorized as CNN-based and ViT-based. U-Net, Attention U-Net and U-Net++ are categorized as CNN-based models while UNETR and UNETR-AF are ViT-based models. The U-Net implemented contains 4 encoder and decoder blocks along with a bottleneck and 4 skip connections between the encoder and decoder based on the original paper in [93]. The Attention U-Net [82] has the same structure as the U-Net but with an addition of an attention gate at each skip connection to pass on additional information derived from the attention mechanism. The U-Net++ [131] is built upon the U-Net by introducing nested dense skip connections which further adds to the model complexity. The next architecture is UNETR [35] which retains the original U-Net architecture but replaces the CNN encoders with Vision Transformer (ViT) [21]. This forms the basic architecture of the proposed method UNETR-AF which includes attention modules SE and CBAM to add channel and spatial attention between the skip connections and the decoders.

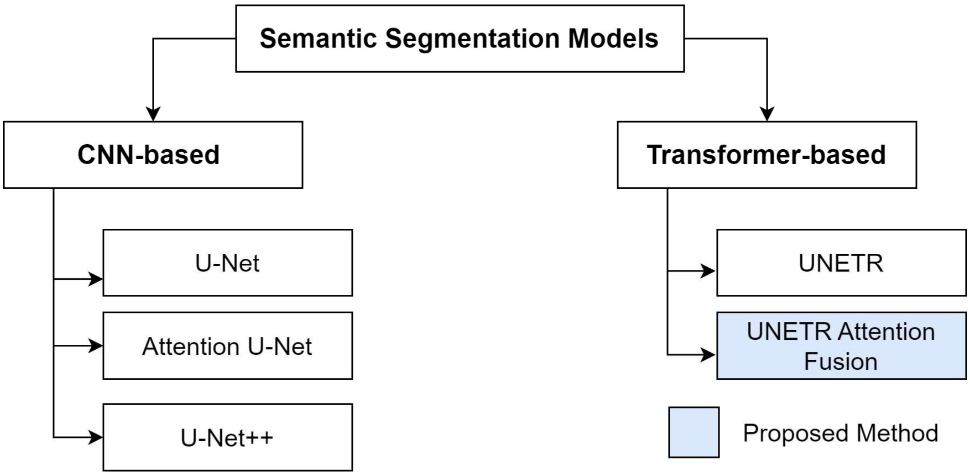


Figure 5.2: Model Architectures.

5.1.3 Training

The training process was conducted to optimize the performance of all the models on all the datasets previously mentioned. The model architecture implemented were U-Net and Attention U-Net, which are known to be effective for segmentation tasks. The inspection dataset was split into training (60%), validation (20%), and test (20%), ensuring a balanced distribution of images under varying lighting conditions and object types. The test set was reserved for inference and model evaluation.

Pre-Processing: The images from the three datasets varied in both width and height. To ensure compatibility with the model architecture, all images were resized to a uniform dimension of 224x224 during the data loading process. This resizing step standardizes the input sizes, enabling the model to process the images efficiently without sacrificing the underlying structure required for segmentation. Moreover, the images and its corresponding masks are normalized which helps standardize the pixel values of images across the dataset

Data Augmentation: To increase the diversity of training samples and prevent overfitting, various data augmentation techniques were applied, including rotations, and flips. These augmentations simulate real-world variations and help the model generalize better to unseen scenarios.

Training Configuration: The aim was to attain the best possible performance on each model and each dataset by implementing different combinations of hyperparameters. The models were trained using a hyperparameter sweep conducted via Weights and Biases. Random sweeps were performed with the number of epochs set to 20, 30, 50, and 100. The loss functions used were binary cross entropy and Focal Tversky, along with the optimizers Adam and SGD. SGD was implemented with a momentum 0.9. For experiments conducted with Focal Tversky loss, the alpha was 0.7, beta 0.3 and the gamma was 2. The batch size was 16 for all CNN-based experiments and 8 for ViT-based models. Learning rates were initialized from a range of values: 3e-2, 1e-2, 3e-3, 1e-3, and 1e-4. To prevent overfitting, early stopping was employed by monitoring the validation loss, halting training if no improvement was seen over a predefined patience period. Moreover, the models were trained from scratch across the datasets for U-Net and Attention U-Net. While the UNet++ was trained with a ResNet-50 backbone and UNETR implemented a pretrained ViT model. Both the models were pretrained on the ImageNet-21k [19] dataset, which consists of 14 million images and 21000 classes.

Transfer Learning Experiments: In addition to training models from scratch, transfer learning was applied to further explore domain adaptation. In the first experiment, a pre-trained U-Net model, initially trained on the SHIQ dataset, was fine-tuned on the Inspection dataset by freezing the decoder layers, allowing only the encoder to adapt to the new domain. A lower learning rate of 1e-4 was used to train for 200 epochs. This method aimed to transfer the knowledge from a larger dataset within the real-world reflection domain (SHIQ) to the Inspection dataset, which is a smaller dataset aimed at reflections in inspection images.

In the second setup, the decoder layers were frozen, while the bottleneck and encoder layers were fine-tuned. This allowed the model to leverage the SHIQ pre-trained decoder to reconstruct outputs while adapting the encoder and bottleneck to better learn the feature representations of the Inspection dataset. The intuition is that the deeper levels in the U-Net capture more feature information which are relevant for the task, as the bottleneck is the last level, retraining allows for further domain-specific fine-tuning which can be derived from the SHIQ dataset to handle the complexities of reflection detection in inspection images. The two transfer learning strategies are demonstrated with schematic diagrams in Figure 5.3.

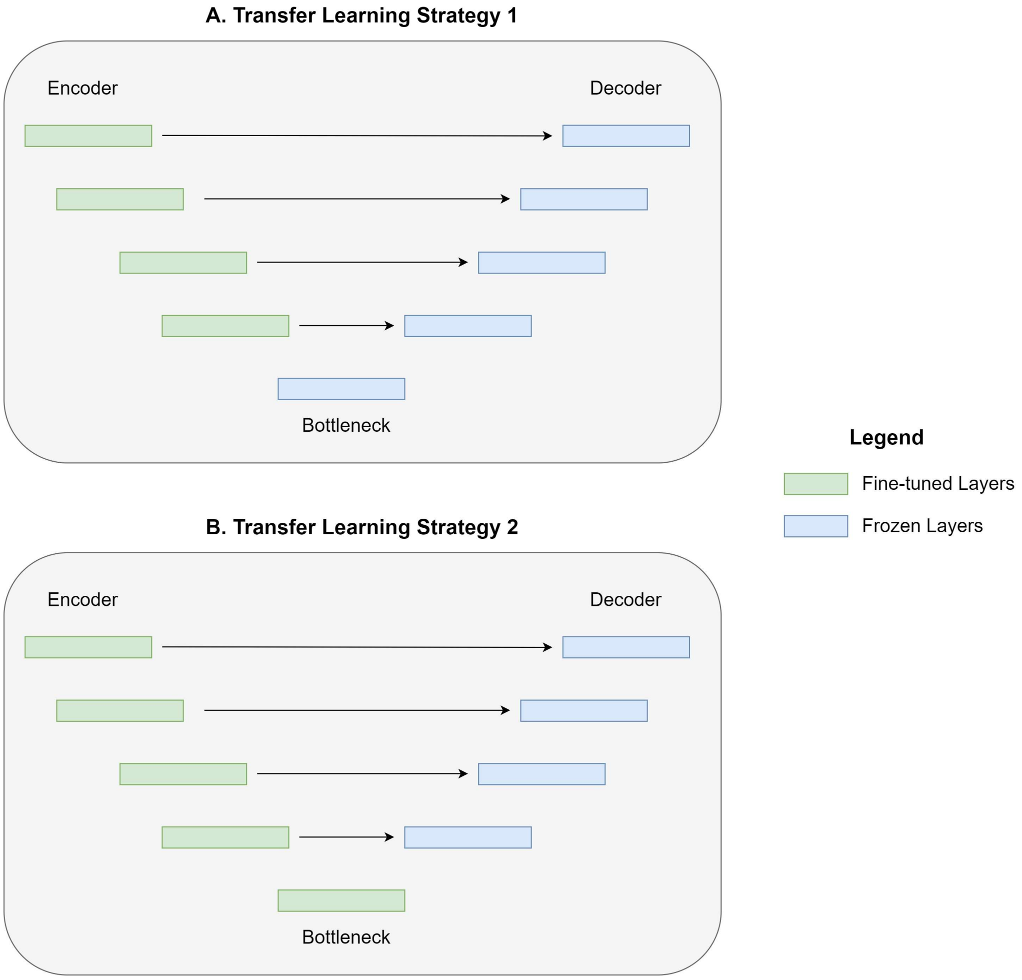


Figure 5.3: Schematic Diagram of Transfer Learning Strategies: A. Only the encoder layers are fine-tuned, the rest are frozen. B. The encoder and the bottleneck are fine-tuned while the decoder layers are frozen. The strategies are applied on the U-Net architecture.

Hardware and Environment: Training was performed using an NVIDIA GeForce RTX 4070 SUPER GPU, utilizing the PyTorch framework. The duration of each experiment varied depending on the dataset, with the SHIQ dataset taking approximately 18 hours, the WHU dataset requiring up to 24 hours, and the inspection dataset taking up to 11 hours.

Performance Metrics: Throughout training, key performance metrics, including training loss, validation loss, accuracy, Dice coefficient and IoU score were tracked using the Weights and Biases library. The model's convergence was monitored using validation loss, and the final model was selected based on the highest mean IoU score on the validation dataset.

5.1.4 Inference and Post-Processing

During the inference phase, the trained models were used to predict segmentation masks on unseen data from the test set. To ensure consistency with the training process, the test images were pre-processed in the same way, including resizing to 224x224. The models generated predictions in the form of binary masks, which were then post-processed for evaluation. For each dataset, the predictions were compared to the ground truth masks using metrics such as IoU and Dice coefficient. These metrics provided a quantitative assessment of the model's ability to accurately segment objects under various lighting conditions and image complexities. The inference speed and memory usage were also tracked, providing insight into the practical feasibility of deploying the models in real-world autonomous inspection systems.

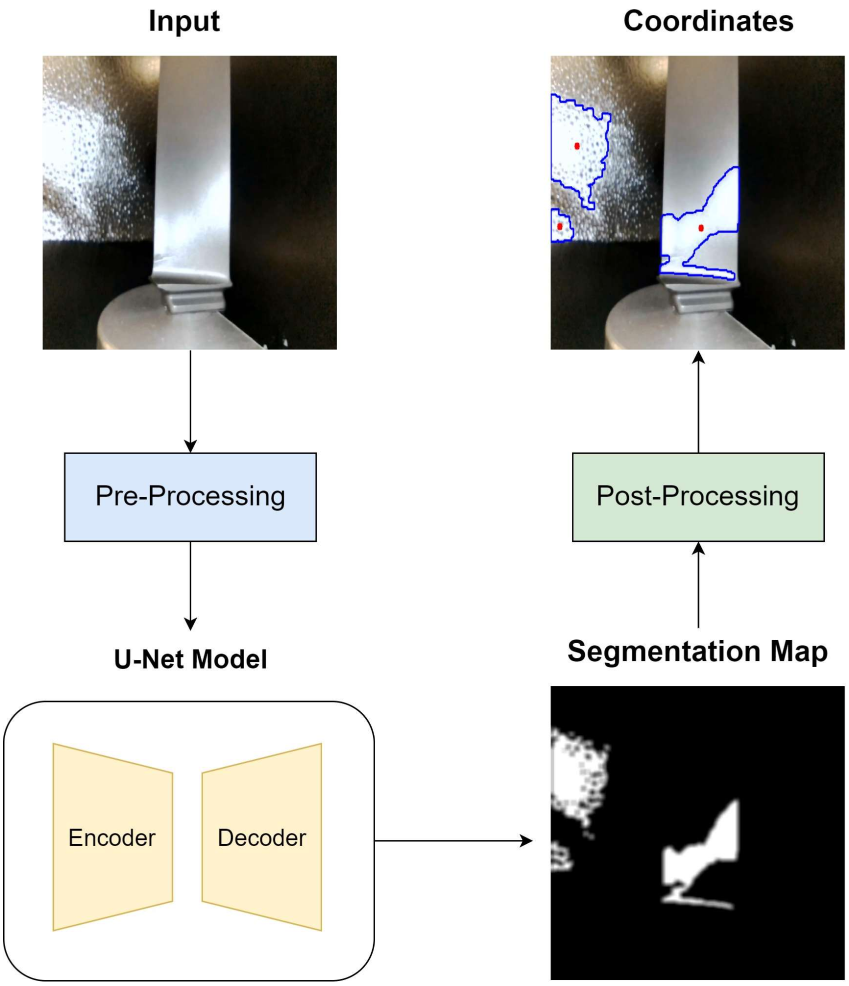


Figure 5.4: Inference Process Flowchart.

After inference, post-processing techniques were applied to extract valuable information from the predicted segmentation masks. A key aspect of this process is identifying the centroid coordinates of localized reflection segments, which is crucial for the reactive planning of autonomous inspections. To achieve this, contours of the reflective areas are drawn from the segmentation masks, enabling the calculation of centroids. To filter out insignificant reflections that may not impact inspection decisions, a threshold is applied: reflection areas smaller than 20 pixels are disregarded for centroid calculation, as they are deemed too minor to warrant further attention. The resulting centroid coordinates are initially derived from the 224x224 binary segmentation masks and are subsequently rescaled to align with the original image dimensions. These coordinates serve as reference points for navigating the robot, allowing it to re-inspect areas obscured by reflections effectively. The post-processing steps are visualized in Figure 5.5.

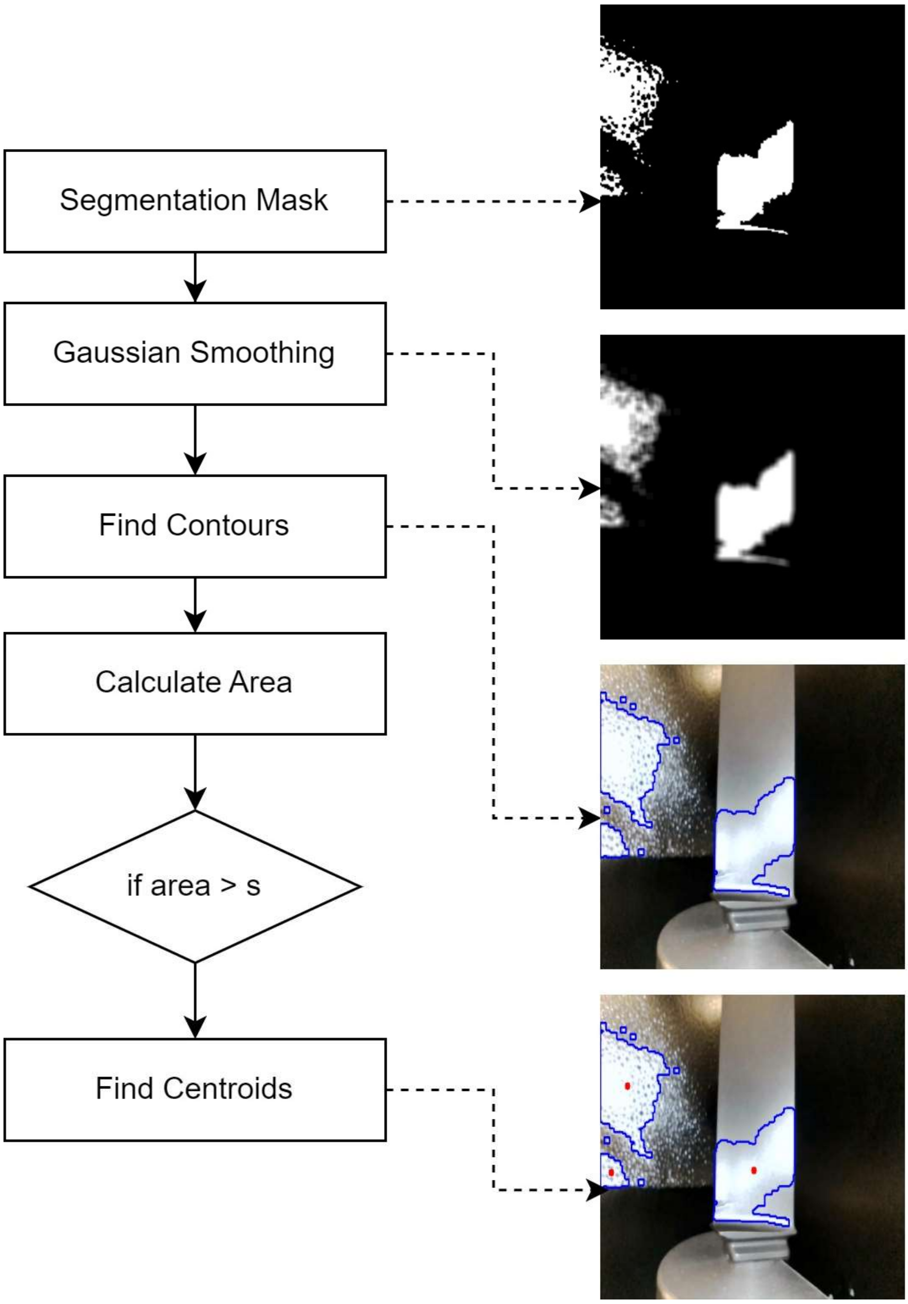


Figure 5.5: Post-Processing steps where s is the threshold set. Each step maps to its corresponding image.

5.1.5 Real-Time Inference Using ROS 2 Integration

Once the models were trained and evaluated on images (including post-processing for contours and centroid extraction), it was essential to test their performance in a real-time scenario, mimicking actual inspection conditions. For this purpose, the RealSense camera was used to capture image frames in real time, and a flashlight was employed to simulate the reflective surface inspection environment. The inspection scenario involved both the exterior and interior of an aircraft wing fuel tank.

To facilitate this, ROS 2 (Robot Operating System 2) [83] was used for integration. The system utilized a publisher-subscriber framework, where the camera acted as the publisher, streaming image frames at 30 frames per second to a dedicated topic. The trained neural network models subscribed to this topic to process the incoming frames, detect reflections, and output the contours and centroids of the reflective regions in real time. This setup provided valuable insights into the practical viability of the models in terms of speed and accuracy under real-world conditions. Figure 5.6 illustrates the real-time inspection of the interior of a fuel tank using the U-Net++ model. The model accurately detects specular reflections, highlighting their contours and centroids, which are critical for localizing reflective surfaces. This demonstrates the model's capability to operate effectively in practical inspection scenarios, providing reliable outputs for potential robotic repositioning and adaptive planning.

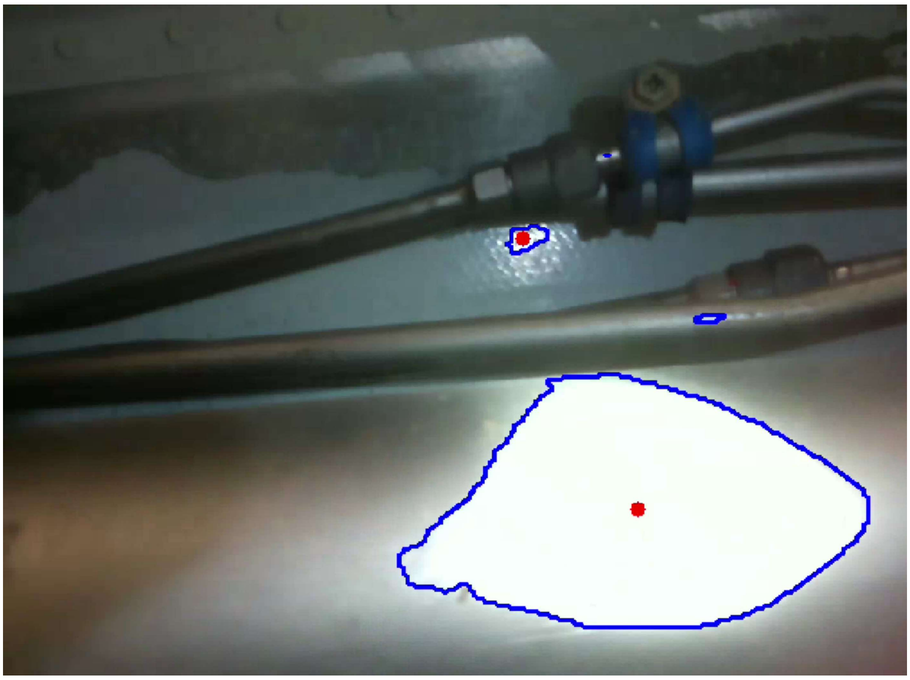


Figure 5.6: Reflection Detection in a Real-Time Inspection Scenario.

5.2 Results

The results derived from the experiments conducted in the previous section are reported in this section. The results are categorized as quantitative and qualitative. The quantitative results are calculated with the evaluation metrics discussed in the previous chapters. The qualitative results are the segmentation maps generated by the model when inferred on a given image.

5.2.1 Baseline Result

Before evaluating the results of the implemented models, the algorithm from a traditional segmentation method [107] was applied to understand whether saturation and intensity-based thresholding is sufficient to reach the desired output. The sample inspection images and the results are shown in Figure 5.7. Most of these techniques work only for its specific use case and do not generalize well across different domains. For instance, this method was originally applied for real-time detection of specularity in endoscope images, however it failed to correctly identify the reflections in the sample inspection images as shown in Figure 5.7.

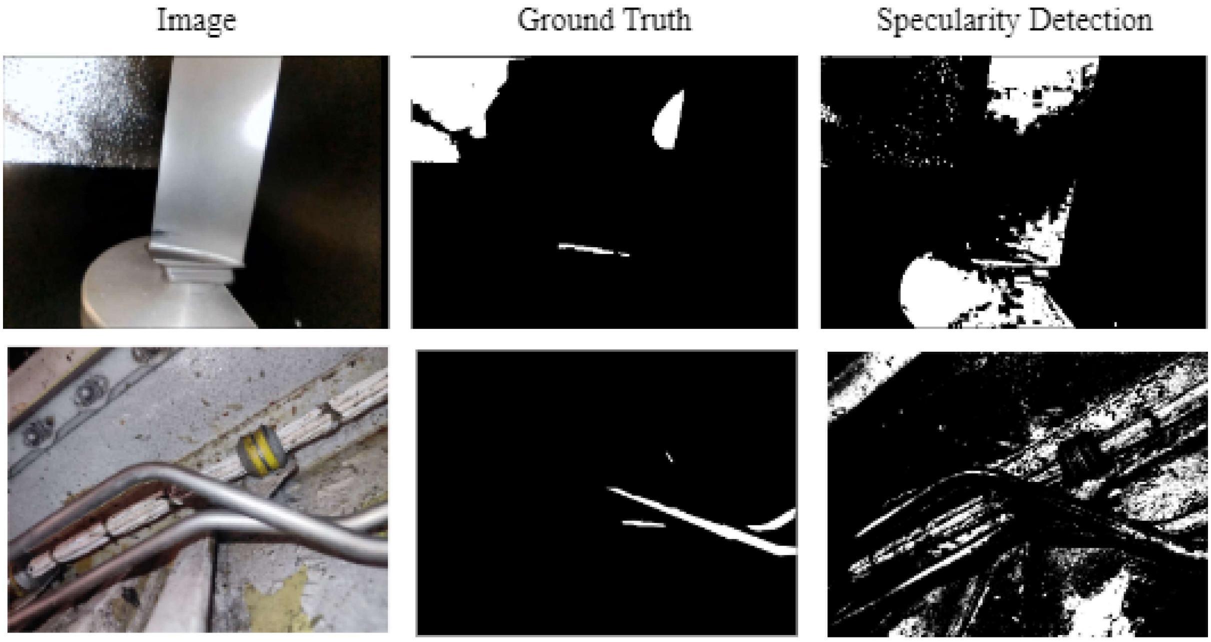


Figure 5.7: Results of the real-time intensity and saturation-based threshold technique from [107] on two inspection images.

5.2.2 Quantitative Results

The quantitative evaluation of the models was conducted using key performance metrics such as the Intersection over Union (IoU) and Dice coefficient. As the average of the IOU score is calculated on the test dataset, the mean IoU (mIoU) is reported. These metrics were computed for the validation sets across all the datasets and the test set from the Inspection dataset, providing a comprehensive measure of the models' segmentation accuracy. The results of all the models are demonstrated in Table 5.1.

Model	**Dataset**	Validation			Test		
		Acc	**mIoU**	**DSC**	**Acc**	**mIoU**	**DSC**
U-Net [93]	SHIQ	98.5	57.5	72.4	89.46	29.65	40.18
	WHU	99.3	47.2	63.8	89.55	32.14	41.48
	Inspection	93.5	41.0	58.0	95.30	43.8	53.76
Attention U-Net [82]	SHIQ	98.5	55.6	70.8	93.32	34.68	45.95
	WHU	99.2	36.9	53.2	90.01	35.00	43.70
	Inspection	90.8	41.9	56.3	96.96	44.63	53.24
U-Net++[131]	SHIQ	98.5	55.43	70.7	88.60	25.3	34.27
	Inspection	94.28	44.32	61.18	**97.72**	**48.65**	**58.13**
UNETR [35]	Inspection	93.38	34.95	51.14	89.23	12.90	17.38
	SHIQ	98.34	49.85	65.12	89.97	26.06	34.86
UNETR-AF (Ours)	Inspection	93.38	34.95	51.14	89.33	17.94	24.66
	SHIQ	98.34	37.89	53.39	89.20	19.02	26.63
	All	98.34	37.89	53.39	89.06	22.06	30.44

Table 5.1: Optimized Results of all the U-Net variants across different datasets.

Table 5.2 provides the configuration parameters used for all the models including U-

Net, Attention U-Net, U-Net++, UNETR, and UNETR-AF, across different datasets: SHIQ, WHU, and the Inspection dataset. The configuration includes the number of training epochs, the optimizer algorithm, the learning rate, and the batch size used in each experiment.

Model	Dataset	Epochs	Loss	Optimizer	Learning Rate	Batch Size
U-Net [93]	SHIQ	100	BCE	Adam	3e-2	16
	WHU	30	BCE	SGD	1e-2	16
	Inspection	100	BCE	Adam	1e-2	16
Attention U-Net [82]	SHIQ	100	BCE	Adam	1e-2	16
	WHU	100	BCE	SGD	1e-2	16
	Inspection	50	BCE	SGD	3e-3	16
U-Net++ [131]	SHIQ	50	BCE	Adam	1e-3	16
	Inspection	50	BCE	Adam	1e-3	16
UNETR [35]	SHIQ	100	FTL	Adam	1e-3	8
	Inspection	100	FTL	Adam	1e-3	8
UNETR-AF (Ours)	SHIQ	100	FTL	Adam	1e-3	8
	Inspection	100	FTL	Adam	1e-3	8

Table 5.2: Configuration Parameters for Optimized Models on Each Dataset.

There were additional transformer configurations used for the transformer-based models: UNETR and UNETR-AF. The transformer models were configured with an input image size of 224x224 pixels to ensure uniformity and compatibility with pre-trained weights. Each image was divided into 16x16 pixel patches to strike a balance between detail and computational efficiency. The embedding dimension was set to 768, providing sufficient space for representing each patch, while 12 attention heads were used to capture multiple features from different image regions simultaneously. A Multi-Layer Perceptron (MLP) dimension of 3072 was selected to capture complex, non-linear relationships, and the model depth was set to 12 transformer layers to maintain a balance between representational power and generalization. The configuration is listed out in Table 5.3.

Parameter	Configuration
Input Image Size	224x224 pixels
Patch Size	16x16 pixels
Embedding Dimension	768
Attention Heads	12
MLP Dimension	3072
Transformer Depth (Layers)	12

Table 5.3: Transformer Model Configuration.

For the U-Net and Attention U-Net models, additional experiments were conducted using transfer learning. While the model trained on the SHIQ dataset performed better than that on the Inspection dataset during validation, it did not yield similar results on the test dataset, likely due to a domain shift between the two datasets. Therefore, it was decided to utilize the SHIQ-trained U-Net and retrain the model with the Inspection dataset to assess the impact on performance. Given that the SHIQ dataset is considerably larger and had previously performed well during validation, it was selected as the base model for training. The transfer learning results are demonstrated in Table 5.4.

Model	**Configuration**	**Dataset**	**Validation**			**Test**		
			Acc	**mIoU**	**DSC**	**Acc**	**mIoU**	**DSC**
U-Net [93]	Decoder frozen, Encoder + Bottleneck fine-tuned	SHIQ and Inspection	94.3	43.2	60.0	93.89	37.11	48.24
	Decoder and Bottleneck frozen, Encoder fine-tuned	SHIQ and Inspection	94.3	43.2	60.0	96.00	47.36	57.23
Attention U-Net [82]	Decoder frozen, Encoder + Bottleneck fine-tuned	SHIQ and Inspection	93.73	34.49	51.10	97.01	42.76	52.26

Table 5.4: Transfer Learning Results of U-Net on SHIQ and Inspection Datasets.

Three different configurations were tested with transfer learning:

- **U-Net:**
  - Decoder frozen, Encoder + Bottleneck fine-tuned
  - Decoder and Bottleneck frozen, Encoder fine-tuned
- **Attention U-Net:**
  - Decoder frozen, Encoder + Bottleneck fine-tuned

To mitigate overfitting, a lower learning rate of 1×10^{-4} was employed, and the model was trained for extended periods of 100 and 200 epochs. The batch size, optimizer, and loss function remained consistent throughout the experiments. It was observed that the configuration with only the encoder fine-tuned yielded the best performance, enhancing the mIoU score of the Inspection-trained U-Net from 43.8 to 47.36 and improving the DSC from 53.76 to 57.23.

Model	**Config**	**Dataset**	**Epochs**	**Optimizer**	**Learning Rate**	**Batch Size**
U-Net [93]	Only Encoder Fine-tuned	SHIQ and Inspection	200	Adam	1e-4	16
	Encoder and Bottleneck Fine-tuned	SHIQ and Inspection	100	Adam	1e-4	16
Attention U-Net [82]	Encoder and Bottleneck Fine-tuned	SHIQ and Inspection	100	Adam	1e-4	16

Table 5.5: Configuration Parameters for Transfer Learning Experiments.

The real-time inference times of all trained models were evaluated to assess their computational efficiency in practical scenarios. Table 5.6 summarizes the measured inference times (in milliseconds) for each model.

Model	Inference Time (ms)
U-Net [93]	1.37
Attention U-Net [82]	1.76
U-Net++ [131]	3.40
UNETR [35]	3.51
UNETR-AF (Ours)	4.61

Table 5.6: Inference times of the models in a real-time inspection scenario.

Among the tested models, the U-Net exhibited the fastest inference time of 1.37 ms, owing to its relatively lightweight architecture. The Attention U-Net followed closely at 1.76 ms, reflecting the minor additional computational overhead introduced by the attention mechanisms. The U-Net++ model demonstrated a significant increase in inference time, taking 3.40 ms, primarily due to its nested dense skip connections, which increase computational requirements.

The transformer-based models, UNETR and UNETR-AF, required the longest inference times, with 3.51 ms and 4.61 ms, respectively. While their inference times are comparable to U-Net++ in practical terms, the additional delay in UNETR-AF can be attributed to the inclusion of attention mechanisms such as CBAM and Squeeze-and-Excitation modules.

5.2.3 Qualitative results

In addition to quantitative metrics, a qualitative analysis was performed to visually assess the segmentation outputs. Sample predictions from each model were compared to the ground truth masks, highlighting the models' ability to detect intricate reflections. The visual results illustrate the strengths and limitations of the models, particularly in scenarios with complex lighting and occlusions.

The segmentation output undergoes post-processing where the centroids of the segmented areas are calculated and returned. For demonstration purposes, the final results are shown using a sample prediction from the SHIQ model, with contours drawn around the reflection areas and centroids marked. Since the models produce a segmentation mask at 224x224 resolution, the centroid coordinates are scaled to match the original image dimensions.

Figure 5.8 shows the size of the image at each stage, as well as how the image is processed in the post-processing stage with contours and centroids.

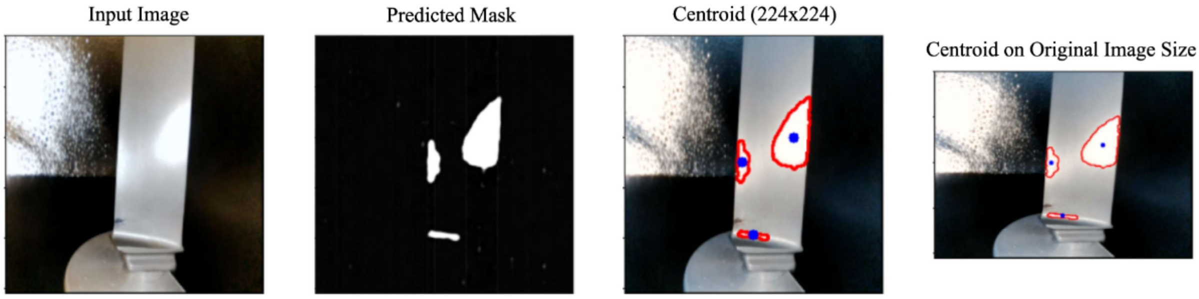


Figure 5.8: Inferring U-Net on a Test Image.

The segmentation map generated from the models are shown in Figure 5.9.

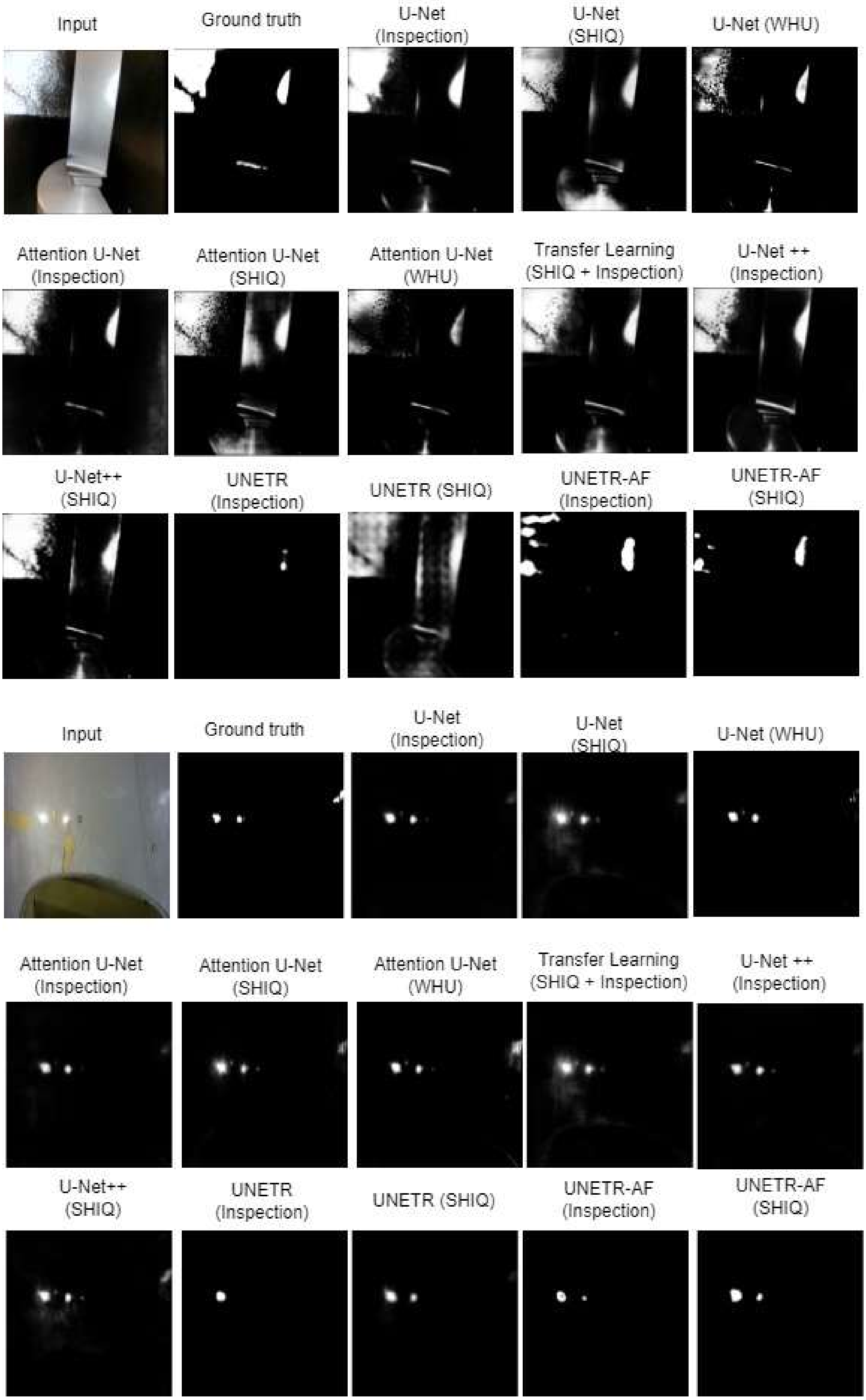


Figure 5.9: Qualitative Results of all the models.

5.2.4 Ablation Study

An ablation study examines the performance impact of individual components within a deep-learning model by selectively removing or modifying them, thereby identifying each component's contribution to the system. In this study, only the Inspection dataset was used to assess the role of each component in the model architecture. First, for the U-Net and Attention U-Net models, the loss function was changed from BCE to Focal Tversky Loss to observe any potential improvement in performance. Results showed a decrease in performance by approximately 30% to 33%, with the mIoU score for U-Net dropping from 43.8 to 31.44 and that of Attention U-Net falling from 44.63 to 29.49.

Subsequently, skip connections between the encoder and decoder were assessed. All U-Net-based models in the primary experiments use four skip connections, so an ablation test was conducted by removing two of these connections. For the UNETR-AF model, two additional experiments examined the effects of removing one of the attention modules. In the first experiment, CBAM was removed from the decoder, leading to a significant drop of 50% in performance, with DSC decreasing from 23 to 10.33 and mIoU from 16 to 7.31. In the second experiment, the SE blocks were removed from the skip connections, resulting in improved performance. These findings are detailed in Table 5.7.

Model	**Dataset**	**Change**	**Acc**	**mIoU**	**DSC**
U-Net [93]	Inspection	Focal Tversky Loss	93.30	31.44	41.37
U-Net [93]	Inspection	Remove 2 skip connections	88.95	9.80	14.59
Attention U-Net [82]	Inspection	Focal Tversky Loss	91.91	29.49	38.97
UNETR-AF (Ours)	Inspection	Remove CBAM Block	89.94	7.31	10.33
UNETR-AF (Ours)	Inspection	Remove SE Block	90.13	22.17	30.98

Table 5.7: Ablation Study Results.

Chapter 6

Discussion

In this chapter, the models and results reported in the previous chapter are discussed to understand their implications. According to the results for U-Net and Attention U-Net in Table 5.1, both models trained with SHIQ [27] performed best on the validation dataset. However, the models trained on the Inspection dataset achieved the best results on the test data. Among all the architectures, those trained on the SHIQ dataset performed best in the validation datasets, while those trained on the Inspection dataset excelled in the test dataset. This difference in performance can be attributed to the variation in size and distribution of the datasets. As previously noted, the SHIQ dataset contains over 16,000 images of specular reflections in real-world environments and objects, while the Inspection dataset is smaller, with only 1,025 images, further split into an 80:20 ratio for training and testing. Due to its specific use case, it may be challenging for models trained on the Inspection dataset to generalize well to other datasets. Moreover, the domain shift between datasets is important to consider. Since the test dataset comes from the Inspection dataset, it belongs to a different domain compared to SHIQ and WHU. Therefore, despite the validation metrics (such as accuracy, mIoU, and DSC) being higher for the SHIQ dataset, the performance fell short of meeting or surpassing that of the Inspection-trained models on the test dataset.

To address the issue of a smaller dataset, the U-Net model trained on the larger SHIQ dataset was used as a base model for transfer learning, retraining U-Net and Attention U-Net with the Inspection dataset. Two different configurations were employed: in the first, the decoder was frozen while the encoder and bottleneck were fine-tuned, and in the second, the decoder and bottleneck were frozen, allowing only the encoder to be fine-tuned. It was observed that performance metrics for mIoU and DSC improved when only the encoder was fine-tuned, whereas performance decreased when both the encoder and the bottleneck were fine-tuned. The performance of Attention U-Net remained relatively unchanged, but it improved for U-Net. Compared to U-Net trained only on the Inspection dataset, transfer learning from SHIQ to Inspection improved the mIoU score by 3.5 and the DSC by 4.

The next set of experiments involved the U-Net++ [131] architecture with a ResNet-50 [36] backbone. ResNet-50 was pretrained on the ImageNet [19] dataset. The performance metrics for the validation dataset were comparable to those of the previous models, but the test dataset metrics were higher when trained with the Inspection dataset. It achieved the highest evaluation metrics among all experiments, including those with transfer learning, with DSC and mIoU scores of 58.13 and 48.65, respectively. However, there was a significant drop in DSC and mIoU scores of 6 and 4 points, respectively, for U-Net++

trained on the SHIQ dataset. This drop indicates that domain shift has a considerable impact on U-Net++ performance.

The next experiments were conducted with the transformer-based UNETR [35]. The Vision Transformer (ViT) encoder was pretrained on the ImageNet dataset [19]. Compared to the other models, it had a much lower mIoU and DSC scores with 11 and 22 respectively on the validation datasets of both SHIQ and Inspection datasets. This could be due to ViT's limitations in effectively segmenting images when trained on smaller datasets. ViT generally perform well on larger datasets, which may explain why it underperformed compared to other models. Additionally, in the original paper [35], UNETR was applied to 3D medical segmentation, whereas in this study, it was adapted for 2D image segmentation. This change in data type may also have affected its performance, as less information about the data is processed. Similarily, UNETR-AF struggled to match the performance of the CNN-based models. However, when trained on the SHIQ dataset, the attention-focused model performed slightly better than the UNETR as the DSC score improved from 22 to 26.62.

In terms of qualitative results, models trained with the Inspection dataset outperformed those trained on SHIQ, particularly in detecting light reflections. As shown in Figure 5.9, the U-Net++ model and the transfer-learned U-Net (SHIQ + Inspection) demonstrated more precise detection of smaller reflections. The ground truth highlights four areas of reflection, and all model predictions were able to detect the smaller reflection at the base of the turbine blade. However, many models inaccurately identified larger areas at the base as reflections, whereas U-Net++ and the transfer-learned U-Net provided a more precise identification of this subtle reflection.

In general, the predicted segmentation maps across all models demonstrated high precision in detecting reflections. Interestingly, while the ground truth annotations focused on larger reflection areas, omitting finer details such as the fading reflection in the background of the turbine blade, the model predictions captured these nuanced effects more effectively. For instance, in the first input image in Figure 5.9, reflections with varying intensities are visible, some stronger and others gradually fading out. While the ground truth did not label these subtler areas— because only the coordinates of larger reflections are necessary for reactive planning—most models, especially U-Net++ and fine-tuned U-Net, captured these effects accurately. Since these subtle differences between the ground truth and the model predictions are not reflected in the evaluation metrics (mIoU and DSC), it suggests that the actual performance of these models might be higher than what the current metrics indicate.

It was also noted that the CNN-based models were able to detect the larger reflections precisely with higher mIoU and DSC, compared to ViT-based models, particularly UNETR. This can be observed in more scenarios captured in Figure 6.1.

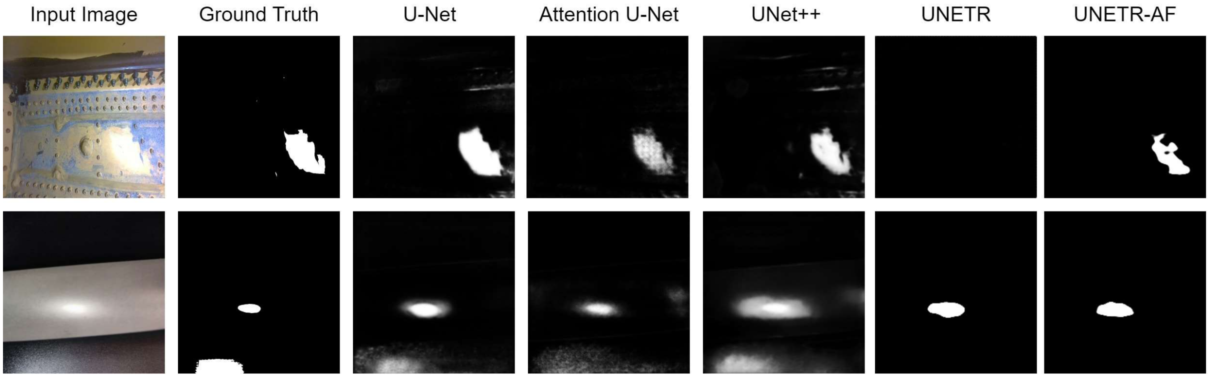


Figure 6.1: Qualitative results with Inspection dataset: the segmentation maps depict results of all the models trained on the Inspection dataset in two different scenarios.

A notable advantage of transformer-based models is their versatility in segmenting 3D data, as demonstrated by the original UNETR, which was designed for 3D medical image segmentation. This capability could prove highly beneficial when using RGB-D images, allowing the capture of more precise distance information and enhancing reactive planning in autonomous inspections.

Model Name	**Strengths**	**Limitations**	**Observations**
U-Net [93]	Strong performance on various datasets	Limited detail in very small reflections	Good baseline for comparison
Attention U-Net [82]	Effective in capturing fine details	Higher computational cost	Better at identifying nuanced features than standard U-Net
U-Net++ [131]	Excellent for capturing fine details	Higher computational cost	Achieved the best overall performance
UNETR [35]	Good at detecting smaller reflections	Struggles with larger reflections	ViT-based model needs larger datasets for effective training
UNETR-AF (Ours)	Improved performance in detecting small reflections	Lower overall performance compared to CNN models	More effective than UNETR in detecting minute details

Table 6.1: Comparison of Model Characteristics and Insights.

In terms of model complexity, there is a significant difference in computational demands and parameter sizes across the U-Net variations. Due to the difference in its structure and modules, such as addition of attention modules and nested dense skip connections, the models have varying model complexity which ultimately affects their performance. Model complexity is a key to understand the capabilities and limitations of a model. The factors that influence this include the number of parameters and filters in a model, as well as the data complexity [116]. The number of parameters are the sum of the weights and biases in the neural networks. Parameters "learn" the training data in the network and is referred to as a measure of how well the model performs. The higher the number of parameters, the higher is the model complexity. A higher number of parameters generally leads to better performance but increases the risk of overfitting on smaller datasets [116].

The number of Multiply-Accumulate operations (MACs) are computed to understand the computational complexity of the model. MACs estimate the number of arithmetic calculations that involve multiple two numbers and adding the results. This operation is often used in linear algebra for computations such as convolutions, dot products, and matrix multiplication, which are fundamental to various deep-learning models. This provides useful insights into the energy consumption and memory resources required by the model to process the input and produce meaningful results. The number of MACs and parameters not only provides a measure of model complexity, but also offers insights into the expected computational load, often serving as an indicator of how long it may take for a neural network to train on a given dataset. Generally, models with higher MACs and parameters require more computational resources and longer training times, particularly when working with large or complex datasets [45]. To compare the model complexity, the number of parameters and MACs are computed and summarized in 6.2. The MACs are calculate based on a RGB image of size 224x224.

Model	**MAC (G)**	**Parameters (M)**	**Inference Time (ms)**
U-Net [93]	10.95	31.04	1.37
Attention U-Net [82]	11.60	31.74	1.76
U-Net++ [131]	35.28	48.99	3.40
UNETR [35]	41.61	105.26	3.51
UNETR-AF (Ours)	41.62	105.34	4.61

Table 6.2: Computational complexity, model size and inference time of the U-Net Versions.

As illustrated in Table 6.2, the standard U-Net is the most computationally efficient, requiring only 10.95 billion MACs and about 31.04 million parameters. In contrast, UNETR, which incorporates the ViT [21], requires a substantial increase in computational resources with 41.61 (G) MACs and 105.26 million parameters. This jump in complexity highlights the heavy computational burden of transformer-based architectures, which, while effective at capturing global dependencies, demand significantly more resources.

Attention U-Net, despite its integration of attention mechanisms, is relatively efficient with 11.60 (G) MACs and 31.74 million parameters. While it presents a notable increase in complexity compared to the standard U-Net, it remains more manageable than UNETR. The figures for U-Net++ lies between U-Net and UNETR, given its dense skip connections which notably adds more MACs compared to attention modules. The differences in complexity among the models are crucial when considering their practical applications. While UNETR and UNETR-AF can yield accurate results by capturing detailed spatial information, their increased computational cost may limit their usability in real-time or resource-constrained environments. Therefore, it is essential to balance performance improvements with computational efficiency when selecting a model for deployment. Additionally, vision transformers require substantial datasets for optimal learning; given the smaller sizes of the datasets used in this study, their performance —even with pre-trained weights — was not as effective as anticipated. Furthermore, the training epochs for transformer models were comparable to those for CNN-based models, suggesting that transformer architectures might need extended training periods to achieve results comparable to CNN-based U-Nets.

When comparing inference times in real-time inspection environment, which are listed in Table 6.2, the standard U-Net achieves the fastest time of 1.37 ms, reflecting its relatively lightweight architecture. Attention U-Net computes results in 1.76 ms, which is slightly longer than U-Net due to the addition of attention gates. U-Net++, on the other hand, has nearly double the inference time of Attention U-Net at 3.40 ms. Interestingly, despite having a larger parameter count, the UNETR achieves an inference time of 3.51 ms, which is comparable to U-Net++, with only a difference of 0.11 ms. Although the ResNet-50 encoder in the U-Net++ has fewer parameters than the ViT in UNETR, the nested skip connections in U-Net++ significantly increase the computations, leading to a higher MAC count, while most of the MACs of UNETR are due to the computations performed by the ViT and not its skip connections. UNETR-AF has the highest inference time of 4.61 ms. Despite UNETR and UNETR-AF having a similar number of MAC operations and parameters, there is a difference of 1.1 ms between their inference times. This difference is due to the addition of the attention modules: CBAM and Squeeze-and-Excitation.

Testing in real-time highlighted that all models processed the frames efficiently, with only minor differences in inference time, demonstrating their suitability for rapid decision-making in inspection scenarios. The detected contours and centroids can be pivotal for an inspection robot to achieve reactive planning. Once the reflection's centroid is identified in the camera's frame, it can serve as a reference point for positioning adjustments. This data can be integrated into a feedback loop for the robot to reposition itself dynamically, ensuring optimal inspection angles or better lighting conditions for further analysis. The integration also underscores the potential for broader use cases, such as optimizing robotic movements during inspection tasks or improving inspection accuracy in environments with limited visibility. By incorporating this real-time feedback loop, inspection systems can achieve greater adaptability and efficiency.

Chapter 7

Conclusion

This thesis conducted experiments with different architectures and methods, categorized as CNN-based and Transformer-based. These included U-Net, Attention U-Net, U-Net++, UNETR, and UNETR-AF, which were evaluated on the test dataset of the Inspection dataset. The U-Net and Attention U-Net were trained across three datasets: SHIQ, WHU, and Inspection dataset, while the transformer-based UNETR and UNETR-AF were trained with SHIQ and Inspection. The SHIQ [27] and WHU [27] are large specular reflection datasets that capture reflections on different materials and objects found in the real-world environment. On the hand, an Inspection dataset is introduced which captures reflections in different illumination modes of the inspection objects. Due to the different domain and distribution of the three datasets, the best performing U-Net trained on SHIQ was used for transfer learning with Inspection dataset. The thesis adapts the UNETR architecture [35], which was originally implemented for 3D medical segmentation to a 2D image segmentation. The thesis also proposed an extension of the UNETR architecture, named UNETR Attention Fusion (UNETR-AF), which incorporates squeeze-and-excitation (SE) blocks and the convolution block attention module (CBAM) in the skip connections and the decoder for improved channel attention.

7.1 Findings

This thesis performed U-Net based semantic segmentation to detect specular light reflections in inspection images. A comparison of the quantitative results revealed that the U-Net trained on the SHIQ dataset produced the highest mIoU and DSC scores when evaluated on the test dataset. However, when considering models adapted from the U-Net architecture, it was found that the U-Net++ model outperformed all others, achieving superior metrics in both the validation and test datasets. The experiments highlighted the limitations of the UNETR-AF model, which, despite its innovative architecture incorporating Squeeze-and-Excitation (SE) blocks and Convolutional Block Attention Modules (CBAM), did not yield satisfactory results. This performance may stem from its adaptation from the UNETR architecture, initially designed for 3D medical segmentation, and the use of a pretrained Vision Transformer (ViT) encoder. The ViT's performance on smaller 2D datasets appeared to be a significant factor in this underperformance. Despite this, the proposed UNETR-AF outperformed UNETR in detecting medium-sized reflections, which is attributed to the additional spatial and channel attention modules in the skip connections and the decoders. Moreover, all the models were also tested in a real-time inspection scenario, wherein all the models demonstrated efficient performance, detecting reflections in a timely manner. The differences in inference times between the models were minimal, only a few milliseconds, indicating that all models are viable for real-time deployment in inspection tasks. In summary, while the U-Net model demonstrated solid performance, particularly with the SHIQ dataset, the U-Net++ model proved to be the most effective for this application. The findings suggest that while transformer-based models like UNETR and UNETR-AF hold potential, their integration into specific tasks requires careful consideration of architectural suitability and dataset characteristics.

7.2 Future Work

This section takes a look at future work that could further improve upon the challenge addressed. Firstly, there is a need for larger, diverse and high-quality datasets consisting of inspection images to achieve state-of-the-art results. These diverse datasets can be generated with the help of variational autoencoders (VAEs) [75] and generative adversarial networks (GANs) [60] that can create realistic images and allow data augmentation. Similarly, synthetic images with specular reflections can be produced using raytracing simulators [38][132], which facilitate the automatic creation of large datasets by identifying specular reflection rays. Moreover, the model can be trained on limited data with the help of few-shot or zero-shot learning [12] [91]. Few-shot learning can also be applied to detect reflections in the images which learns to segment in a query image based on a few pixel-wise annotated support images [119] [125].

Future work could focus on processing RGB-D data in deep-learning networks, as the inclusion of depth information would enable better extraction of spatial and geometric details, improving the system's ability to accurately localize reflections and support robust reactive planning. By combining RGB-D inputs with reflection contours and centroids, methods such as stereo triangulation [90] could be employed to calculate optimal repositioning strategies, allowing robots to dynamically adjust their position for improved inspection angles and thorough evaluations. Integrating the proposed models with real-time inspection systems and robotics platforms offers promising applications, as reflection localization data could guide inspection robots to adapt to changing conditions, such as varying light sources or obstructed views, while enhancing both efficiency and coverage through reactive planning. Expanding the model's capabilities to detect additional use cases, such as surface defects or environmental anomalies, could significantly broaden its industrial applications, optimizing maintenance and repair operations in aviation.

Bibliography

[1] Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 683–687, 2019.

[2] Rolf Adams and Leanne Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641–647, 1994.

[3] Gamze Akyol, Alperen Kantarcı, Ali Eren Çelik, and Abdullah Cihan Ak. Deep learning based, real-time object detection for autonomous driving. In *2020 28th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4, 2020.

[4] Shun-ichi Amari. Backpropagation and stochastic gradient descent method. *Neurocomputing*, 5(4-5):185–196, 1993.

[5] Atif Anwer, Samia Ainouz, Mohamad Naufal Mohamad Saad, Syed Saad Azhar Ali, and Fabrice Meriaudeau. SpecSeg Network for Specular Highlight Detection and Segmentation in Real-World Images. *Sensors*, 22(17):6552, August 2022.

[6] Mirko Arnold, Anarta Ghosh, Stefan Ameling, and Gerard Lacey. Automatic Segmentation and Inpainting of Specular Highlights for Endoscopic Imaging. *EURASIP Journal on Image and Video Processing*, 2010:1–12, 2010.

[7] Muhammad Asif, Hong Song, Lei Chen, Jian Yang, and Alejandro F. Frangi. Intrinsic layer based automatic specular reflection detection in endoscopic images. *Computers in Biology and Medicine*, 128:104106, 01 2021.

[8] Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10076–10095, 2024.

[9] David Ball, Ben Upcroft, Gordon Wyeth, Peter Corke, Andrew English, Patrick Ross, Tim Patten, Robert Fitch, Salah Sukkarieh, and Andrew Bate. Vision-based obstacle detection and navigation for an agricultural robot. *Journal of Field Robotics*, 33(8):1107–1130, 2016.

[10] Andre Bleau and L. Joshua Leon. Watershed-based segmentation and region merging. *Computer Vision and Image Understanding*, 77(3):317–370, 2000.

[11] Soufiane Bouarfa, Anıl Doğru, Ridwan Arizar, Reyhan Aydogan, and Joselito Serafico. Towards automated aircraft maintenance inspection. a use case of detecting aircraft dents using mask r-cnn. 01 2020.

[12] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32:468–479.

[13] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986.

[14] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew P. Lungren, Shaoting Zhang, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024.

[15] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin P. Murphy, and Alan Loddon Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:834–848, 2016.

[16] Corinna Cortes and Vladimir Naumovich Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[17] Alana de Santana Correia and Esther Luna Colombini. Attention, please! a survey of neural attention models in deep learning. *Artificial Intelligence Review*, 55(8):6037–6124, 2022.

[18] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (methodological)*, 39(1):1–22, 1977.

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[20] Lee Raymond Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 10 2020.

[22] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[23] Mahdi Abolfazli Esfahani and Han Wang. Robust glare detection: Review, analysis, and dataset release. *ArXiv*, October 2021. 10.48550/arXiv.2110.06006.

[24] Federal Aviation Administration. *Visual Inspection for Aircraft*, 1997. Advisory Circular ACNO. 43-204.

[25] Gang Fu, Changjun Liu, Rong Zhou, Tao Sun, and Qijian Zhang. Classification for high resolution remote sensing imagery using a fully convolutional network. *Remote Sensing*, 9(5):498, 2017.

[26] Gang Fu, Qing Zhang, Qifeng Lin, Lei Zhu, and Chunxia Xiao. Learning to detect specular highlights from real-world images. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1873–1881, New York, NY, USA, 2020. Association for Computing Machinery.

[27] Gang Fu, Qing Zhang, Lei Zhu, Ping Li, and Chunxia Xiao. A Multi-Task Network for Joint Specular Highlight Detection and Removal. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7748–7757, June 2021. ISSN: 2575-7075.

[28] Yinghua Fu, Junfeng Liu, and Jun Shi. Tsca-net: Transformer based spatial-channel attention segmentation network for medical images. *Computers in Biology and Medicine*, 170:107938, March 2024.

[29] Michael P Georgeff and Amy Lansky. Reactive reasoning and planning. In *AAAI'87 Proceedings of the Sixth National Conference on Artificial Intelligence*, volume 2 of *6*, pages 677–682. AAAI Press, July 1987.

[30] Rafael Gonzalez and Richard Woods. *Digital Image Processing Global Edition*. Pearson Deutschland, 2017.

[31] Jeremy John Gray. Johann heinrich lambert, mathematician and scientist, 1728-1777. *Historia mathematica*, 5(1):13–41, 1978.

[32] Jian-Jhih Guo, Day-Fann Shen, Guo-Shiang Lin, Jen-Chun Huang, Kai-Che Liu, and Wen-Nung Lie. A specular reflection suppression method for endoscopic images. In *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)*, pages 125–128, 2016.

[33] Xiaojie Guo, Yang Yang, Chaoyue Wang, and Jiayi Ma. Image dehazing via enhancement, restoration, and fusion: A survey. *Information Fusion*, 86(C):146–170, October 2022.

[34] Lalit Gupta and Thotsapon Sortrakul. A gaussian-mixture-based image segmentation algorithm. *Pattern recognition*, 31(3):315–325, 1998.

[35] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 574–584, 2022.

[36] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[37] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. page 2, 2012.

[38] Antti Hirvonen, Atte Seppälä, Maksim Aizenshtein, and Niklas Smal. Accurate real-time specular reflections with radiance caching. In Eric Haines and Tomas Akenine-Möller, editors, *Ray Tracing Gems: High-Quality and Real-Time Rendering with DXR and Other APIs*, pages 571–607. Apress.

[39] 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.

[40] Jiunn-Kai Huang and Jessy W. Grizzle. Efficient anytime clf reactive planning system for a bipedal robot on undulating terrain. *IEEE Transactions on Robotics*, 39(3):2093–2110, 2023.

[41] Shengzeng Huo, David Navarro-Alarcon, and David TW Chik. A robotic defect inspection system for free-form specular surfaces. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11364–11370, 2021.

[42] Shengzeng Huo, David Navarro-Alarcon, and David TW Chik. A robotic defect inspection system for free-form specular surfaces. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11364–11370, 2021.

[43] Paul Jaccard. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50, 1912.

[44] Shruti Jadon. A survey of loss functions for semantic segmentation. In *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7.

[45] Daniel Justus, John Brennan, Stephen Bonner, and Andrew Stephen McGough. Predicting the computational cost of deep learning models. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 3873–3882, 2018.

[46] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[47] Ihor Konovalenko, Pavlo Maruschak, Janette Brezinová, Olegas Prentkovskis, and Jakub Brezina. Research of u-net-based cnn architectures for metal surface defect detection. *Machines*, 10(5), 2022.

[48] Vidya Kudva, Keerthana Prasad, and Shyamal Guruvare. Detection of specular reflection and segmentation of cervix region in uterine cervix images for cervical cancer screening. *Irbm*, 38(5):281–291, 2017.

[49] Labelbox. Labelbox. `https://labelbox.com`, 2024. Online.

[50] Edwin Herbert Land and John J. McCann. Lightness and retinex theory. *Journal of the Optical Society of America*, 61 1:1–11, 1971.

[51] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

[52] Yann LeCun and Corinna Cortes. The mnist database of handwritten digits. 2005.

[53] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial Intelligence and Statistics*, pages 562–570, 2015.

[54] Andreas Leibetseder, Sabrina Kletz, Klaus Schoeffmann, Simon Keckstein, and Jörg Keckstein. Glenda: gynecologic laparoscopy endometriosis dataset. In *International Conference on Multimedia Modeling*, pages 439–450. Springer, 2019.

[55] Martin D Levine and Jisnu Bhattacharyya. Detecting and removing specularities in facial images. *Computer vision and image understanding*, 100(3):330–356, 2005.

[56] Lexas.biz. Die Finnische Mark – Währungseinheit Finnlands. `https://www.lexas.biz/waehrungen/mark/finnische_mark.aspx`, 2024.

[57] Baojun Li, Shun Liu, Weichao Xu, and Wei Qiu. Real-time object detection and semantic segmentation for autonomous driving. In *MIPPR 2017: Automatic Target Recognition and Navigation*, volume 10608, pages 167–174. SPIE, 2018.

[58] Yadan Li, Zhenqi Han, Haoyu Xu, Lizhuang Liu, Xiaoqiang Li, and Keke Zhang. Yolov3-lite: A lightweight crack detection network for aircraft structure based on depthwise separable convolutions. *Applied Sciences*, 9(18), 2019.

[59] Zhonglin Li, Yang Luo, Yunxiang Jiang, Bi Zhang, Guoli Song, Xingang Zhao, and Yiwen Zhao. A Novel Transparent Object Detection Approach Based on Segmentation-Depth Reconstruction. In *2023 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pages 853–858, July 2023.

[60] Ziqiang Li, Beihao Xia, Jing Zhang, Chaoyue Wang, and Bin Li. A comprehensive survey on data-efficient gans in image generation. *arXiv preprint arXiv:2204.08329*, 2022.

[61] Yuyang Lin, Yan Yang, Yongquan Jiang, Xiaobo Zhang, and Pengyun Song. ET-HDR: An Efficient Two-Stage Network for Specular Highlight Detection and Removal. In Huimin Lu, Michael Blumenstein, Sung-Bae Cho, Cheng-Lin Liu, Yasushi Yagi, and Tohru Kamiya, editors, *Pattern Recognition*, pages 273–287, Cham, 2023. Springer Nature Switzerland.

[62] Huaiyu Liu, Yueyuan Zhang, and Yiyang Chen. A symmetric efficient spatial and channel attention (esca) module based on convolutional neural networks. *Symmetry*, 16(8):952, 2024.

[63] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[64] Philipp Lottes, Jens Behley, Andres Milioto, and Cyrill Stachniss. Fully convolutional networks with sequential information for robust crop and weed detection in precision farming. *IEEE Robotics and Automation Letters*, 3(4):2870–2877, 2018.

[65] J MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*, 1967.

[66] Daniel Maestro-Watson, Julen Balzategui, Luka Eciolaza, and Nestor Arana-Arexolaleiba. Deflectometric data segmentation for surface inspection: a fully convolutional neural network approach. 29(4):041007. Publisher: SPIE.

[67] David Marr and Ellen Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167):187–217, 1980.

[68] Daryl Martin. A practical guide to machine vision lighting. *Midwest Sales and Support Manager, Adv Illum2007*, pages 1–3, 2007.

[69] Maziar Jamshidi, Mamdouh El-Badry, and Navid Nourian. Improving Concrete Crack Segmentation Networks through CutMix Data Synthesis and Temporal Data Fusion. *Italian National Conference on Sensors*, 23(1):504–504, January 2023. MAG ID: 4313518939 S2ID: 50017d443f93e8b90c84cc1bed0f12d39d788833.

[70] Andrew Mehnert and Paul T. Jackway. An improved seeded region growing algorithm. *Pattern Recognit. Lett.*, 18:1065–107, 1997.

[71] Tseko Mofokeng, Paul T. Mativenga, and Annlizé Marnewick. Analysis of aircraft maintenance processes and cost. *Procedia CIRP*, 90:467–472, 2020. 27th CIRP Life Cycle Engineering Conference (LCE2020) Advancing Life Cycle Engineering : from technological eco-efficiency to technology that supports a world that meets the development goals and the absolute sustainability.

[72] Patrice Monkam, Jing Wu, Wenkai Lu, Wenjun Shan, Hao Chen, and Yuhao Zhai. EasySpec: Automatic Specular Reflection Detection and Suppression From Endoscopic Images. *IEEE Transactions on Computational Imaging*, 7:1031–1043, 2021.

[73] Alexandre Morgand and Mohamed Tamaazousti. Generic and real-time detection of specular reflections in images. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 1, pages 274–282, January 2014.

[74] Aliasghar Mortazi, Vedat Cicek, Elif Keles, and Ulas Bagci. Selecting the best optimizers for deep learning–based medical image segmentation. *Frontiers in Radiology*, 3:1175473, 2023.

[75] Fatemeh Mostofi, Onur Behzat Tokdemir, and Vedat Toğan. Generating synthetic data with variational autoencoder to address class imbalance of graph attention network prediction model for construction management. *Advanced Engineering Informatics*, 62:102606, 2024.

[76] Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. A multi-illumination dataset of indoor object appearance. In *2019 IEEE International Conference on Computer Vision (ICCV)*, Oct 2019.

[77] Nastaran Enshaei, Safwan Ahmad, and Farnoosh Naderkhani. Automated detection of textured-surface defects using UNet-based semantic segmentation network. *International Conference on Prognostics and Health Management*, pages 1–5, 2020. MAG ID: 3084325977 S2ID: ba860a2149913c6de28d6f203ae19f478a869906.

[78] Chao Nie, Chao Xu, Zhengping Li, Lingling Chu, and Yunxue Hu. Specular Reflections Detection and Removal for Endoscopic Images Based on Brightness Classification. *Sensors (Basel, Switzerland)*, 23(2):974, January 2023.

[79] Chao Nie, Chao Xu, Zhengping Li, Lingling Chu, and Yunxue Hu. Specular reflections detection and removal for endoscopic images based on brightness classification. *Sensors*, 23(2):974, 2023.

[80] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

[81] Pratik Oak and Brijesh Iyer. Specular reflection detection and substitution: A key for accurate medical image analysis. In Amit Kumar and Stefan Mozar, editors, *ICCCE 2019*, pages 223–241, Singapore, 2020. Springer Singapore.

[82] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, M. J. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention u-net: Learning where to look for the pancreas. *ArXiv*, abs/1804.03999, 2018.

[83] Open Source Robotics Foundation (OSRF). Ros2. `https://index.ros.org/`, 2024. Online.

[84] Nobuyuki Otsu et al. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296):23–27, 1975.

[85] Zhuokun Pan, Jiashu Xu, Yubin Guo, Yueming Hu, and Guangxing Wang. Deep learning segmentation and classification for urban village using a worldview satellite image based on u-net. *Remote Sensing*, 12(10):1574, 2020.

[86] Pasquale Lafiosca, Pasquale Lafiosca, Ip-Shing Fan, Ip-Shing Fan, Nicolas P. Avdelidis, and Nicolas P. Avdelidis. Automatic Segmentation of Aircraft Dents in Point Clouds (SAE Paper 2022-01-0022). *SAE technical paper series*, March 2022. ARXIV_ID: 2205.01614 MAG ID: 4220929863 S2ID: 60f5050b7dd168441bb67009458468a55077d5bd.

[87] Florian Paysan, Eric Dietrich, and Eric Breitbarth. A robot-assisted microscopy system for digital image correlation in fatigue crack growth testing. *Experimental Mechanics*, 63(6):975–986, 2023.

[88] Angelos Plastropoulos, Kostas Bardis, George Yazigi, Nicolas P. Avdelidis, and Mark Droznika. Aircraft skin machine learning-based defect detection and size estimation in visual inspections. *Technologies*, 12(9), 2024.

[89] Yulei Qin, Juan Wen, Hao Zheng, Xiaolin Huang, Jie Yang, Ning Song, Yue-Min Zhu, Lingqian Wu, and Guang-Zhong Yang. Varifocal-net: A chromosome classification approach using deep convolutional networks. *IEEE transactions on medical imaging*, 38(11):2569–2581, 2019.

[90] Zengyi Qin, Jinglu Wang, and Yan Lu. Triangulation learning network: from monocular to stereo 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7615–7623, 2019.

[91] Wenqi Ren, Yang Tang, Qiyu Sun, Chaoqiang Zhao, and Qing-Long Han. Visual semantic segmentation based on few/zero-shot learning: An overview. *IEEE/CAA Journal of Automatica Sinica*, 11(5):1106–1126, 2024.

[92] Zhonghe Ren, Fengzhou Fang, Ning Yan, and You Wu. State of the art in defect detection based on machine vision. *International Journal of Precision Engineering and Manufacturing-Green Technology*, 9:661–691, 2021.

[93] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing.

[94] Thomas D. Rossing and Christopher J. Chiaverina. *Ray Optics: Reflection, Mirrors, and Kaleidoscopes*, pages 51–88. Springer International Publishing, Cham, 2019.

[95] Holger R Roth, Hirohisa Oda, Xiangrong Zhou, Natsuki Shimizu, Ying Yang, Yuichiro Hayashi, Masahiro Oda, Michitaka Fujiwara, Kazunari Misawa, and Kensaku Mori. An application of cascaded 3d fully convolutional networks for medical image segmentation. *Computerized Medical Imaging and Graphics*, 66:90–99, 2018.

[96] Sebastian Ruder. An overview of gradient descent optimization algorithms. *ArXiv*, abs/1609.04747, 2016.

[97] Philipp Ruppel, Michael Görner, Norman Hendrich, and Jianwei Zhang. Detection and Reconstruction of Transparent Objects with Infrared Projection-Based RGB-D Cameras. In Fuchun Sun, Huaping Liu, and Bin Fang, editors, *Cognitive Systems and Signal Processing*, pages 558–569, Singapore, 2021. Springer.

[98] Charles-Auguste Saint-Pierre, Jonathan Boisvert, Guy Grimard, and Farida Cheriet. Detection and correction of specular reflections for automatic surgical tool segmentation in thoracoscopic images. *Machine Vision and Applications*, 22(1):171–180, January 2011.

[99] F Javier Sánchez, Jorge Bernal, Cristina Sánchez-Montes, Cristina Rodríguez de Miguel, and Gloria Fernández-Esparrach. Bright spot regions segmentation and classification for specular highlights detection in colonoscopy videos. *Machine Vision and Applications*, 28(8):917–936, 2017.

[100] Abdelrahman M Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Unetr++: delving into efficient and accurate 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024.

[101] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[102] Hui-Liang Shen and Qing-Yuan Cai. Simple and efficient method for specularity removal in an image. *Applied optics*, 48:2711–2719, 6 2009.

[103] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[104] Alvy Ray Smith. Color gamut transform pairs. *ACM Siggraph Computer Graphics*, 12(3):12–19, 1978.

[105] Irwin Sobel and Gary Feldman. A 3×3 isotropic gradient operator for image processing. *Pattern Classification and Scene Analysis*, pages 271–272, 01 1973.

[106] Chull Hwan Song, Hye Joo Han, and Yannis Avrithis. All the attention you need: Global-local, spatial-channel attention for image retrieval. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2754–2763, 2022.

[107] Stephane Tchoulack, J.M. Pierre Langlois, and Farida Cheriet. A video stream processor for real-time detection and correction of specular reflections in endoscopic images. In *2008 Joint 6th International IEEE Northeast Workshop on Circuits and Systems and TAISA Conference*, pages 49–52, 2008.

[108] Attila Temun, Lars Mattsson, and Irma Heikkilä. Localizing micro-defects on rough metal surfaces. In *4M 2006-Second International Conference on Multi-Material Micro Manufacture*, pages 169–172. Elsevier, 2006.

[109] Le-Anh Tran and My-Ha Le. Robust u-net-based road lane markings detection for autonomous driving. In *2019 International Conference on System Science and Engineering (ICSSE)*, pages 62–66. IEEE, 2019.

[110] Toshiaki Tsuji. Specular reflection removal on high-speed camera for robot vision. In *2010 IEEE International Conference on Robotics and Automation*, pages 1542–1547, 2010.

[111] Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

[112] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[113] Joost van der Putten, Jeroen de Groof, Fons van der Sommen, Maarten Struyvenberg, Svitlana Zinger, Wouter Curvers, Erik Schoon, Jacques Bergman, and Peter H.N. de With. Informative Frame Classification of Endoscopic Videos Using Convolutional Neural Networks and Hidden Markov Models. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 380–384, September 2019. ISSN: 2381-8549.

[114] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[115] Cheng Wang, Le Wang, Nuoqi Wang, Xiaoling Wei, Ting Feng, Minfeng Wu, Qi Yao, and Rongjun Zhang. Cfatransunet: Channel-wise cross fusion attention and transformer for 2d medical image segmentation. *Computers in Biology and Medicine*, 168:107803, 2024.

[116] Chuan-Chi Wang, Ying-Chiao Liao, Ming-Chang Kao, Wen-Yew Liang, and Shih-Hao Hung. Toward accurate platform-aware performance modeling for deep neural networks. *ACM SIGAPP Applied Computing Review*, 21:50–61, 2020.

[117] Donghuan Wang, Hong Xiao, and Shengqin Huang. Automatic Defect Recognition and Localization for Aeroengine Turbine Blades Based on Deep Learning. *Aerospace*, 10(2):178, February 2023.

[118] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020.

[119] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1130–1140, 2023.

[120] Xucheng Wang, Chenning Tao, Xiao Tao, and Zhenrong Zheng. SIHRNet: a fully convolutional network for single image highlight removal with a real-world dataset. *Journal of Electronic Imaging*, 31(3):033013, 2022.

[121] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3:1–40, 2016.

[122] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

[123] Bichen Wu, Forrest Iandola, Peter H Jin, and Kurt Keutzer. Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 129–137, 2017.

[124] Zhongqi Wu, Chuanqing Zhuang, Jian Shi, Jianwei Guo, Jun Xiao, Xiaopeng Zhang, and Dong-Ming Yan. Single-Image Specular Highlight Removal via Real-World Dataset Construction. *IEEE Transactions on Multimedia*, 24:3782–3793, 2022.

[125] Zhen Yao, Jiawei Xu, Shuhang Hou, and Mooi Choo Chuah. Cracknex: a few-shot low-light crack segmentation model based on retinex theory for uav inspections. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11155–11162, 2024.

[126] Yuri D.V. Yasuda, Fabio A.M. Cappabianco, Luiz Eduardo G. Martins, and Jorge A.B. Gripp. Aircraft visual inspection: A systematic literature review. *Computers in Industry*, 141:103695, 2022.

[127] Kuk-jin Yoon, Yoojin Choi, and In So Kweon. Fast separation of reflection components using a specularity-invariant image representation. In *2006 International Conference on Image Processing*, pages 973–976, 2006.

[128] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[129] Huaxia Zhao, Chuangbai Xiao, and Hongyu Zhao. Chapter 16 - night color image enhancement via statistical law and retinex. In Leonidas Deligiannidis and Hamid R. Arabnia, editors, *Emerging Trends in Image Processing, Computer Vision and Pattern Recognition*, pages 249–261. Morgan Kaufmann, Boston, 2015.

[130] Qinbang Zhou, Renwen Chen, Bin Huang, Wang Xu, and Jie Yu. Deepinspection: Deep learning based hierarchical network for specular surface inspection. *Measurement*, 160:107834, 2020.

[131] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pages 3–11. Springer, 2018.

[132] Jingsen Zhu, Fujun Luan, Yuchi Huo, Zihao Lin, Zhihua Zhong, Dianbing Xi, Rui Wang, Hujun Bao, Jiaxiang Zheng, and Rui Tang. Learning-based inverse rendering of complex indoor scenes with differentiable monte carlo raytracing. In *SIGGRAPH Asia 2022 Conference Papers*, SA '22, New York, NY, USA, 2022. Association for Computing Machinery.

[133] Kunlin Zou, Xin Chen, Yonglin Wang, Chunlong Zhang, and Fan Zhang. A modified u-net with a specific data argumentation method for semantic segmentation of weed images in the field. *Computers and Electronics in Agriculture*, 187:106242, 2021.