mathematics

MDPI

Article

Enhancing Mirror and Glass Detection in Multimodal Images Based on Mathematical and Physical Methods

Jiyuan Qiu and Chen Jiang *

School of Aerospace Engineering, Tsinghua University, Beijing 100084, China; qiujy22@mails.tsinghua.edu.cn
* Correspondence: jc2017@mail.tsinghua.edu.cn

Abstract: The detection of mirrors and glass, which possess unique optical surface properties, has garnered significant attention in recent years. Due to their reflective and transparent nature, these surfaces are often difficult to distinguish from their surrounding environments, posing substantial challenges even for advanced deep learning models tasked with performing such detection. Current research primarily relies on complex network models that learn and fuse different modalities of images, such as RGB, depth, and thermal, to achieve mirror and glass detection. However, these approaches often overlook the inherent limitations in the raw data caused by sensor deficiencies when facing mirrors and glass surfaces. To address this issue, we applied mathematical and physical methods, such as three-point plane determination and steady-state heat conduction in two-dimensional planes, along with an RGB enhancement module, to reconstruct RGB, depth, and thermal data for mirrors and glass in two publicly available datasets: an RGB-D mirror detection dataset and an RGB-T glass detection dataset. Additionally, we synthesized four enhanced and ideal datasets. Furthermore, we propose a double weight Mamba fusion network (DWMFNet) that strengthens the model's global perception of image information by extracting low-level clue weights and high-level contextual weights from the input data using the prior fusion feature extraction module (PFFE) and the deep fusion feature guidance module (DFFG). This is complemented by the Mamba module, which efficiently captures long-range dependencies, facilitating information complementarity between multimodal features. Extensive experiments demonstrate that our data enhancement method significantly improves the model's capability in detecting mirrors and glass surfaces.

Keywords: mirror detection; glass detection; multimodal image; Mamba

MSC: 68T07; 68U10

Academic Editors: Xiaojiang Peng, Linlin Shen and Yang You

Received: 21 January 2025
Revised: 16 February 2025
Accepted: 24 February 2025
Published: 25 February 2025

Citation: Qiu, J.; Jiang, C. Enhancing Mirror and Glass Detection in Multimodal Images Based on Mathematical and Physical Methods. *Mathematics* **2025**, *13*, 747. https://doi.org/10.3390/math13050747

1. Introduction

Mirrors and glass are widely used in modern society and are nearly ubiquitous in our daily environments [1–3]. Although deep learning methods have been widely applied across various fields [4–7], their unique optical surface properties make their detection a significant challenge in the field of computer vision. The ability to effectively detect these objects has profound implications for areas such as visual navigation [8], robot perception [9], and autonomous driving [10]. In the visible light spectrum, mirrors are primarily characterized by their reflective properties, which cause them to project the surrounding environment onto their surfaces, making it difficult to distinguish real objects from reflections. The intensity and angle variations of the reflected light further complicate the detection of the mirror's surface. In contrast, the transparency of glass can cause

computer vision models to mistakenly detect objects passing through the glass surface, thus neglecting the presence of the glass itself. Despite the similar texture-absent nature of both mirror and glass surfaces, current data-driven deep learning methods for detecting these two materials tend to focus on different features [11–13]. Most models for mirror detection emphasize the specular reflection content and the geometric features of the boundary. High-quality specular reflections that distinguish from the surrounding environment help improve the model's ability to locate the mirror. For glass detection, due to the lack of reflective content, models often focus on the contrast and texture features of the glass surface. Enhancing the optical texture of the glass surface also aids in improving the model's ability to identify glass. Currently, methods for detecting mirrors and glass can be broadly classified into two categories: one based on single RGB images [14–16], and the other based on multimodal fusion using multiple different modalities [17–19]. The latter has become the mainstream approach, as it extends the observation of mirror and glass features beyond visible light. However, most existing multimodal methods for detecting mirrors and glass focus on complex model architectures and fusion techniques to enhance detection capabilities. They often overlook improving the quality of the raw data, which suffers from deficiencies due to sensor properties when collecting data from mirrors and glass surfaces. To address this issue, this paper aims to enhance the quality of raw multimodal data based on mathematical and physical principles to improve model performance in detecting mirrors and glass.

We first take two commonly used methods based on multimodal data fusion for mirror or glass surface detection as examples to illustrate the defects encountered when different sensors collect the corresponding modal data of mirrors or glass. As shown in Figure 1, a mirror surface detection method based on RGB and depth data fusion primarily collects the RGB modal data of mirrors through a monocular RGB camera and the depth modal data through a time-of-flight (ToF) depth camera. From the left side of Figure 1, we can see that the RGB data collection relies on the visible light reflected by the surrounding objects on the mirror. Therefore, the features reflected by the mirror in the RGB image entirely depend on the reflected objects, which leads to the difficulty of distinguishing the mirror surface from real objects in the RGB modality. For the depth modality, most of the data is obtained by the depth camera emitting laser beams to the object and measuring the time it takes for the laser to reflect back to the sensor, thus calculating the distance between the object and the camera using the speed of light. However, there are issues when measuring the depth data of the mirror surface in this way. Due to the reflective nature of the mirror's surface, the laser emitted by the depth camera is reflected into the surrounding environment, so the depth data obtained is not the actual depth data of the mirror surface. Therefore, the mirror information on the depth modality is not reliable. Additionally, the depth camera is susceptible to interference from external light sources, further preventing it from capturing accurate depth data from the mirror surface. Thus, the right side of Figure 1 shows the effects of RGB and depth modalities on the mirror. It can be observed that in the RGB modality, while the mirror surface can be observed by the human eye, it presents a significant challenge for neural networks. In the depth modality, the data reflected from the mirror surface is not the actual depth of the mirror; instead, it reflects the depth of surrounding objects, causing confusion for neural networks when recognizing the mirror in the depth modality.

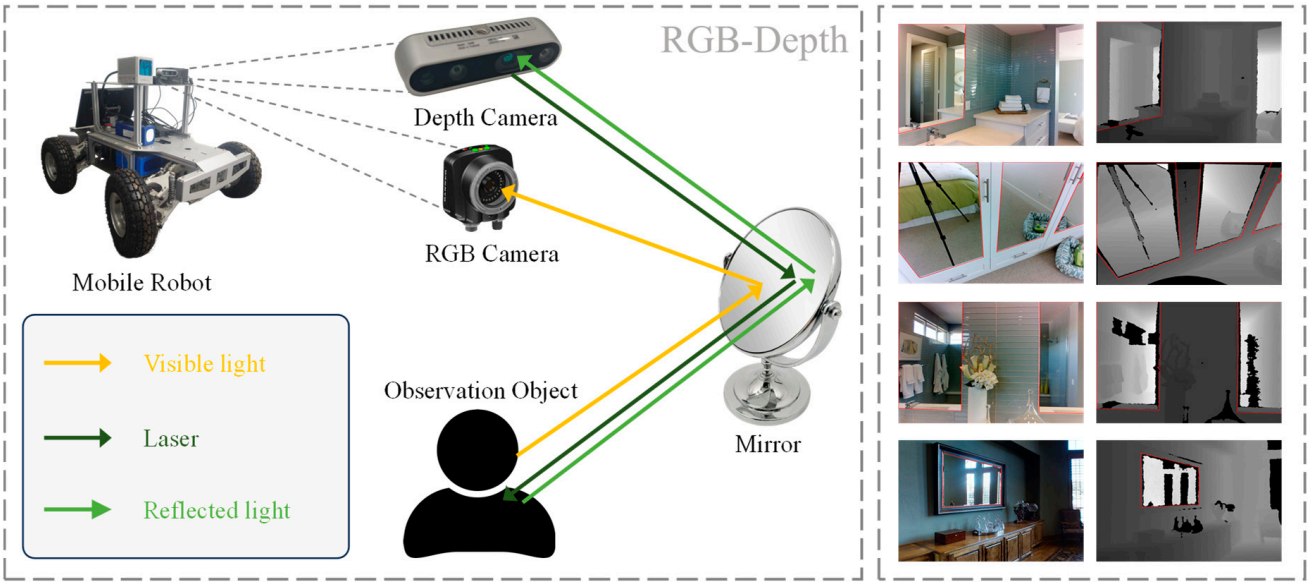


Figure 1. The acquisition of mirror surface data based on RGB-D modal. The left side of the figure illustrates the process of capturing RGB data using a monocular RGB camera and depth data using a depth camera. The right side of the figure shows how the mirror appears in both the RGB and depth images, with the mirror's outline highlighted by a red bounding box.

Figure 2 illustrates a glass detection method that relies on the fusion of RGB and thermal modalities. In this case, a monocular RGB camera collects visible light from the glass surface, but often what is collected is the visible light refracted from the objects behind the glass, making it difficult to distinguish the glass surface from the surrounding environment in the RGB modality. As for the thermal camera, since any object above absolute zero continuously emits infrared thermal radiation to its surrounding environment, it can provide more visual information compared to an RGB camera [20], relying on the principle that thermal energy does not pass through the glass surface, and collects infrared light reflected from the glass surface to capture the thermal data. While this approach can help distinguish the glass surface from the surrounding environment to some extent, it also reflects infrared light from other thermal sources near the glass surface, thereby introducing noise that affects the thermal data of the glass surface. The right side of Figure 2 shows the effects of the glass in both the RGB and thermal modalities. It can be seen that in the RGB modality, the glass often reflects or refracts the visible light of the surrounding objects, making it difficult to detect. In the thermal modality, it also reflects infrared light from the surrounding environment, and the data reflected does not accurately represent the thermal data of the glass surface.

Therefore, to address the issue where the sensor's intrinsic properties prevent the effective collection of RGB, depth, and thermal modality data from mirror and glass surfaces, we employ mathematical and physical methods to reconstruct these three types of data on two publicly available datasets for RGB-D mirror detection [21] and RGB-T glass detection [22]. These two datasets, respectively, encompass mirrors and glass of varying shapes across diverse scenes, which is of significant importance for enhancing the diversity of data samples and improving the generalization capability of the model. This approach enhances the model's capability to detect mirrors and glass surfaces. Additionally, we incorporate the recently popular Mamba model [23], leveraging its ability to efficiently capture long-range dependencies, and propose a double weight Mamba fusion network (DWMFNet) to effectively accomplish the mirror and glass detection tasks.

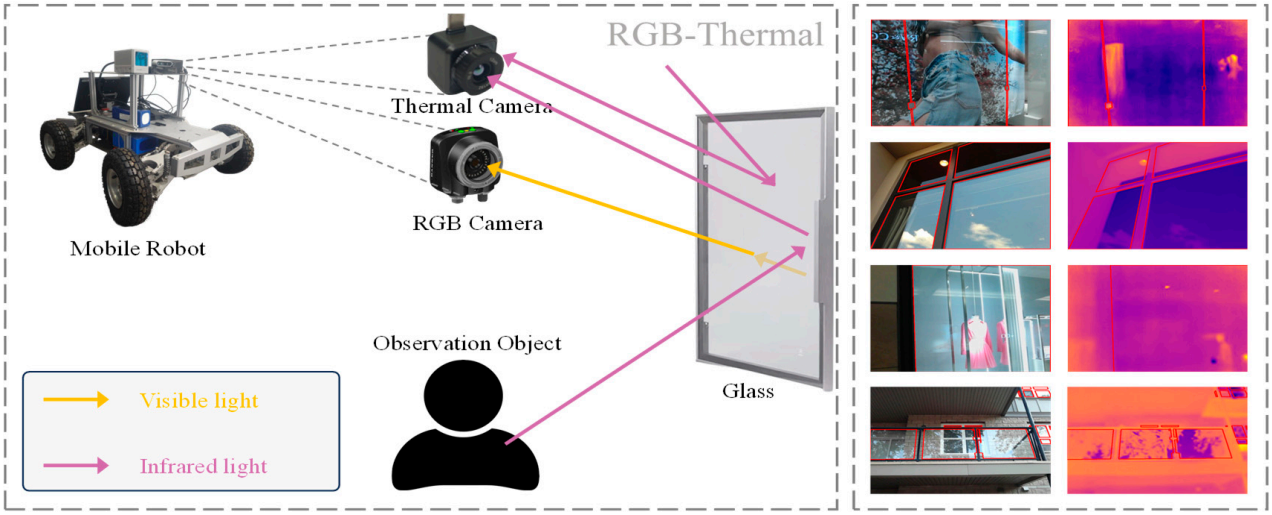


Figure 2. The acquisition of glass surface data based on RGB-T modal. The left side of the figure illustrates the process of capturing RGB data using a monocular RGB camera and thermal data using a thermal camera. The right side of the figure shows how the glass appears in both the RGB and thermal images, with the glass's outline highlighted by a red bounding box.

In summary, the contributions of this paper include the following:

- We completed the reconstruction of the RGB, depth, and thermal data for mirror and glass surfaces in the raw RGB-D mirror dataset and RGB-T glass dataset based on an RGB enhancement module, the three-point plane determination, and the steady-state heat transfer principle. Using this, we constructed four additional datasets, namely the enhanced RGB-D mirror dataset, ideal RGB-D mirror dataset, enhanced RGB-T glass dataset, and ideal RGB-T glass dataset.
- We proposed a double weight Mamba fusion network (DWMFNet) for multi-modal data fusion-based mirror and glass detection. By using the prior fusion feature extraction module (PFFE) and deep fusion feature guidance module (DFFG), we extracted the low-level clue weights and high-level context weights of the input data to enhance the model's global perception of image information, and combined them with the Mamba module to achieve information complementarity between multi-modal features.
- Extensive experimental results demonstrate that our model achieves satisfactory results on the six datasets mentioned above. Additionally, almost all models show varying degrees of performance improvement on the enhanced and ideal datasets.

The remainder of this paper is organized as follows: Section 2 reviews the current work on mirror detection, glass detection, and the joint application of multimodal data. Section 3 provides a detailed explanation of the process of building the enhanced and ideal datasets based on mathematical and physical principles, and describes the basic characteristics of all the datasets. Section 4 introduces the overall framework and component structure of the proposed DWMFNet. Section 5 presents our experimental setup and compares our method with other state-of-the-art methods. Finally, Section 6 concludes the paper.

2. Related Work

2.1. Mirror Detection

The detection of mirror surfaces is particularly challenging due to their reflective properties. As a result, there has been limited work on mirror surface detection using traditional methods. Chang et al. [24] suggested the use of an iPad application to manually label the mirror regions for indoor semantic segmentation based on RGB-D images.

Whelan et al. [1] used scanners and AprilTags [25] to generate image labels for mirrors and achieve mirror surface segmentation. As can be seen, traditional mirror detection methods often rely on manual annotation, which requires considerable time and resources, and the results are influenced by the subjective judgments of the annotators. With the development of artificial intelligence technologies, deep learning models, which rely on large-scale data for training and feature learning to achieve directional predictions, have provided an effective way to perform mirror detection. Yang et al. [26] first constructed the mirror segmentation dataset (MSD), a large dataset containing mirror detection in real-life scenes, and proposed the first deep learning model, MirrorNet, for mirror detection. This network takes a single RGB image as input and uses a CNN-based architecture to learn contextual information and multi-scale features from the input data to detect mirrors of varying sizes. Lin et al. [27] proposed a dataset that places more focus on the mirror surface's reflection phenomena, making the mirror surface less similar to its surrounding environment to improve the model's discriminative capability for mirrors. They also proposed an incremental mirror detection network that detects mirrors by learning multi-layer contextual contrasts between the mirror and its surroundings, effectively locating the mirror's boundary. Guan et al. [28] observed that mirrors are often placed in specific functional locations and proposed utilizing the semantic correlation between the mirror and surrounding objects for mirror surface localization. He et al. [29] proposed a multi-level heterogeneous network (HetNet) that focuses on high-precision mirror surface detection while also considering the real-time application of the model. Huang et al. [30] combined a transformer architecture to propose a dual-path symmetry-aware transformer-based mirror detection network (SATNet) for distinguishing between mirrors and real objects. Tan et al. [31] introduced VCNet, which uses dense visual chirality cues to detect mirror regions. Mei et al. [21] observed the discontinuity of depth information in both the mirror and non-mirror areas to detect mirrors, and proposed PDNet, which complements both RGB and depth modal information, significantly improving mirror detection performance. Additionally, some works [32,33] extended the mirror detection task to video sequences.

2.2. Glass Detection

For glass detection, similar to mirror detection, it faces various challenges due to the special transparent properties of the glass surface. Early glass detection often encountered difficulties. McHenry et al. [34] detected glass by identifying the contour of the glass in images. Wei et al. [35] used a method combining ultrasonic sensors and laser scanning data to detect the glass surface. Tibebu et al. [36] implemented glass localization using a laser scanner to predict occupancy grids. In recent years, most of the glass detection tasks rely on single-modal and multi-modal neural network models. Mei et al. [37] proposed the first deep learning-based glass detection method, GDNet, and constructed a large-scale RGB-based glass detection dataset (GDD). However, much of the data in this dataset consists of close-up shots, leading to a relatively homogeneous data distribution and simple data structure. To address this issue, Lin et al. [38] proposed the glass surface detection dataset (GSD), which includes close-ups, mid-range, and long-range images of the glass surface, and refined the glass boundary detection by leveraging reflection information from the glass surface. Zheng et al. [39] proposed a three-stream neural network, GlassNet, which decouples the original ground-truth (GT) map into internal diffusion and boundary diffusion maps. These newly generated maps are used to break the imbalance of object boundaries, thus improving glass detection quality. Fan et al. [40] proposed a reciprocal feature evolution network (RFENet), which reveals the mutual evolution requirements of semantic learning and boundary learning for mutual feature learning, thereby improving the model's ability to detect glass surfaces. Lin et al. [41] further expanded the detection

of glass in large-scale scenes by creating the glass surface detection-semantics (GSD-S) dataset, which extends the detection objects beyond the binary masks of the glass surface to multiple categories. Qi et al. [42] proposed a visual blur aggregation module that uses inherent visual blur cues from the glass surface to extract and aggregate multi-scale valuable features, guiding backbone features for accurate glass detection. In addition to the methods mentioned above, multi-modal fusion-based glass detection is also one of the commonly used approaches. Lin et al. [43] constructed a large-scale RGB-D glass surface detection dataset (RGB-D GSD). Huo et al. [22] combined RGB and thermal modalities to propose an RGB-T glass detection dataset, utilizing infrared light that does not penetrate the glass surface to supplement the visible light data. Mei et al. [44] combined RGB and polarization information to detect the glass surface. Yan et al. [45] used near-infrared images to suppress the reflection of the glass surface and proposed an NRGlassNet to merge the two modalities for glass surface detection.

2.3. The Joint Application of Multi-Modal Data

The joint application of multi-modal data has been widely used not only in mirror and glass detection but also in tasks such as scene parsing [46], salient object detection [47], and camouflaged object detection [48]. In the field of scene parsing, Hazirbas et al. [49] proposed FuseNet, an encoder–decoder network consisting of two branches, to fuse RGB and depth modalities for semantic segmentation of indoor scenes. Ha et al. [50] introduced MFNet, a real-time semantic segmentation network for autonomous vehicles designed for multi-spectral scenes, utilizing the thermal modality to facilitate road scene parsing under nighttime and adverse weather conditions. Sun et al. [51] also leveraged the thermal modality's resistance to external lighting conditions, proposing RTFNet to enhance autonomous vehicles' environmental perception by fusing RGB and thermal information. Chen et al. [52] developed a unified and efficient cross-modal guidance encoder that effectively calibrates RGB features while refining accurate depth information across multiple stages to improve the network's performance in RGB-D-based indoor and outdoor semantic segmentation. Zhou et al. [53] used RGB and depth fusion, proposing the three-stream self-attention network (TSNet) with two asymmetric input streams and a cross-modal distillation flow with a self-attention module for indoor scene modeling. Lan et al. [54] introduced MMNet, a multi-modal multi-stage network for RGB-T-based semantic segmentation, which divides the fusion of modal features into two stages to gradually complete the fusion details while avoiding cross-modal feature conflicts. Zhao et al. [55] proposed FDCNet, a novel RGB-T-based network, using a Siamese structure to extract high-level single-peak features from paired RGB and thermal images to bridge the gap between the low-dimensional features of the two modalities and ensure semantic consistency in the high-dimensional features. In the field of salient object detection, Chen et al. [56] introduced depth potential awareness, enabling their DPANet to leverage a gating controller for attention mechanisms, capturing long-range dependencies from a cross-modal perspective for RGB-D-based salient object detection. Tu et al. [57] proposed a novel collaborative graph learning algorithm to jointly learn the affinity and node saliency of both RGB and thermal modality images within a unified optimization framework. Song et al. [58] analyzed the pros and cons of RGB-D and RGB-T modality fusion methods, proposing a strategy that combines RGB, depth, and thermal modalities for salient object detection. Building on this, Qiu et al. [59] utilized a three-modal pre-training framework and introduced ETFormer, a lightweight network to perform fast and efficient RGB-depth-thermal-based salient object detection. In camouflaged object detection (COD), Wang et al. [60] introduced DaCOD, inspired by biology and evolutionary studies, integrating depth information as an additional clue to help break camouflage, separating spatial and texture information between the foreground and back-

ground of camouflaged objects. Yang et al. [61] extended RGB-D-based camouflaged object detection to the agricultural domain, using it to detect small and concealed crops, with a method employing multi-scale receptive fields to capture objects of different sizes and improving detection performance for dense objects by fusing depth features. Liu et al. [62] proposed a depth-weighted cross-attention fusion module to dynamically adjust the fusion weights of depth and RGB feature maps, adaptively combining these two modalities to strengthen the model's performance in camouflaged object detection tasks.

3. Dataset Construction

In this section, we first provide a basic description of the enhanced RGB-D mirror and enhanced RGB-T glass detection datasets. Then, we elaborate on the process of reconstructing data from different modalities within the datasets.

3.1. Dataset Description

The detection of mirrors with reflective properties is currently supported by several available datasets, including those using RGB modalities such as MSD [26] and PMD [27], and those utilizing both RGB and depth modalities, such as RGB-D mirror detection dataset [21]. In this study, we primarily focus on the RGB-D mirror dataset, which contains depth modality data and diverse real-world scene images. This dataset consists of 3049 pairs of RGB and depth images featuring mirror objects in everyday environments. The raw data is sourced from four commonly used RGB-D datasets (Matterport3D [24], SUNRGBD [63], ScanNet [64], and 2D3DS [65]), making it highly diverse in terms of mirror image samples. For a more detailed description of the RGB-D mirror dataset, please refer to [21]. To better differentiate the reflective mirror surface from the surrounding environment in the RGB images, we employ a small-scale network with a pre-segmentation approach, leveraging depth images to enhance the recognition of mirror surfaces. To address the issue of depth data loss or inaccuracy when capturing mirror surfaces with the depth camera, we first repair the raw depth data of the mirror surface using the three-point plane determination method. This is followed by an image restoration network to generate enhanced depth data. The specific principles behind these methods will be explained in the subsequent sections. Ultimately, we obtained a set of 3049 pairs of enhanced RGB and depth images with mirror objects, which we refer to as the enhanced RGB-D mirror dataset. The dataset obtained directly using the three-point plane determination method is referred to as the ideal RGB-D mirror dataset. A comparison of the raw, enhanced, and ideal depth data samples is shown in Figure 3, while a comparison of the raw, enhanced, and ideal RGB data samples is shown in Figure 4.

The detection of glass with transparent properties is currently supported by several available datasets, including those using the RGB modality, such as GDD [66], GSD [38], and Trans10k-v2 [2], and those utilizing both RGB and depth modalities, such as ClearGrasp [67] and GSD (2022) [43]. Additionally, datasets using RGB and thermal modalities, such as the RGB-T glass dataset [22], are also widely used. In this study, we primarily focus on the RGB-T glass detection dataset, which contains thermal modality data and includes numerous glass samples. The RGB-T glass detection dataset comprises 5551 pairs of RGB and thermal image data collected in real-world environments. Most of the images contain one or more glass samples, with 370 pairs of images that do not feature any glass. The raw data was collected using the FLIR ONE Pro camera. A more detailed description of the RGB-T glass dataset can be found in [22]. Similar to the processing of the RGB modality in the enhanced RGB-D mirror dataset, we used a small-scale network with a pre-segmentation approach to improve the recognition of the glass surface in the RGB images, combining the thermal images. To address the issues caused by artifacts from infrared light reflection in the thermal

images, we employed the principle of steady-state heat transfer in two-dimensional planes, combined with the Laplace equation. This method utilizes the temperature at the glass boundary to interpolate the temperature of the glass surface, resulting in a steady-state temperature distribution and eliminating artifacts on the glass surface. Afterward, we used an image restoration network to generate enhanced thermal data. The specific principles behind these methods will be explained in the following sections. Ultimately, we obtained a set of 5551 pairs of enhanced RGB and thermal images with glass samples, which we refer to as the enhanced RGB-T glass detection dataset. The dataset obtained using the principle of steady-state heat transfer in a two-dimensional plane is referred to as the ideal RGB-T glass detection dataset. The comparison of raw, enhanced, and ideal thermal data samples is shown in Figure 5, while the comparison of raw, enhanced, and ideal RGB data samples is shown in Figure 4.

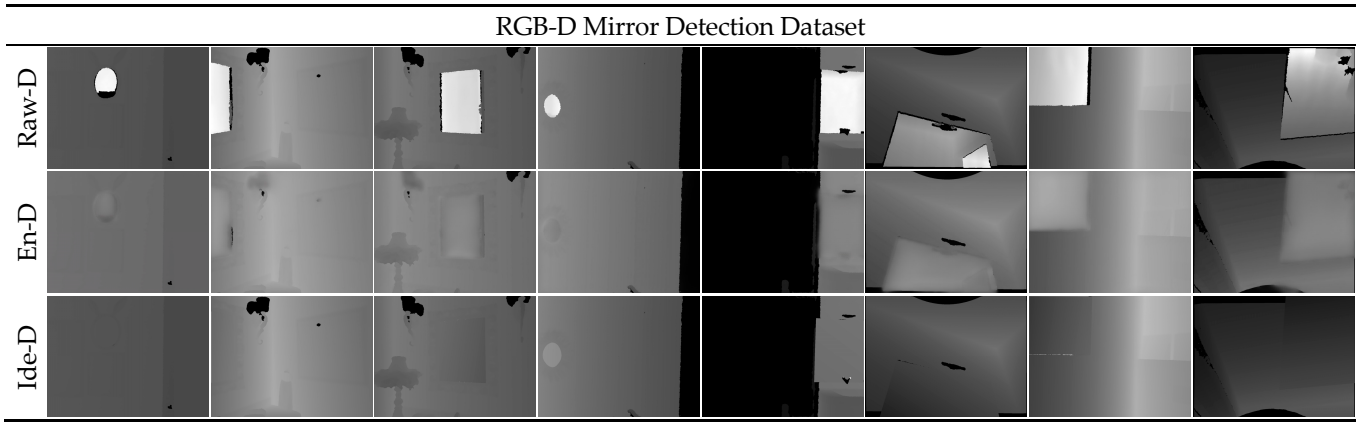


Figure 3. Examples of raw, enhanced, and ideal depth data samples used in the RGB-D mirror detection task.

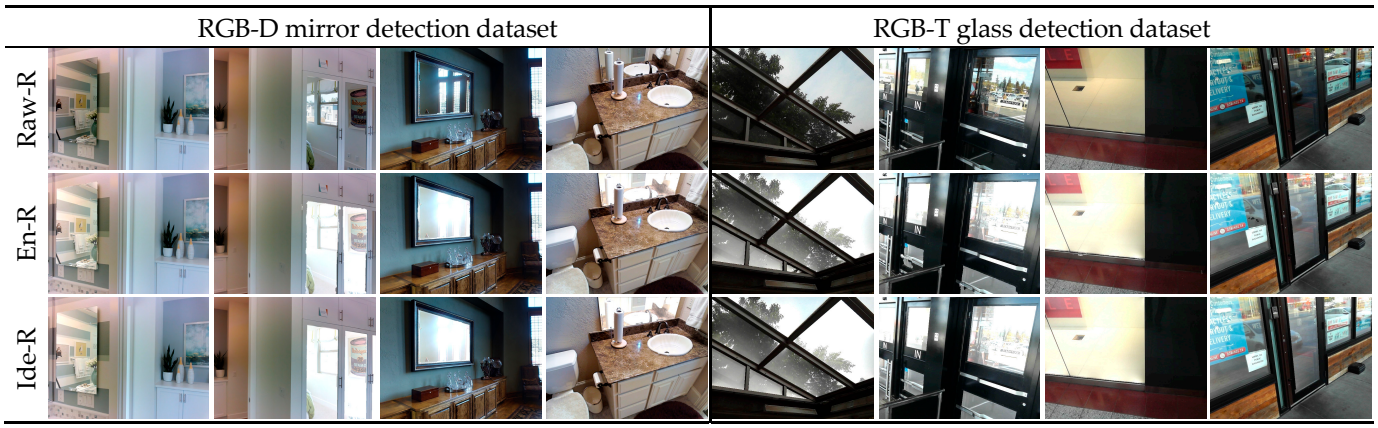


Figure 4. Examples of raw, enhanced, and ideal raw data samples used in the RGB-D mirror detection task and the RGB-T glass detection task, respectively.

To validate the effectiveness of our proposed method, we conducted experiments for mirror detection and glass detection using the RGB-D mirror and RGB-T glass detection datasets, respectively. Based on the extent of image restoration, the datasets were further divided into three different categories to demonstrate the significant improvement in model detection performance through our method. For the mirror detection experiment, we have three datasets: the raw RGB-D mirror dataset, the enhanced RGB-D mirror dataset, and the ideal RGB-D mirror dataset. The raw RGB-D mirror dataset refers to the dataset without any processing. The enhanced RGB-D mirror dataset refers to the dataset processed by an

intermediate network (pre-segmentation network and image restoration network) using manually processed data labels. The ideal RGB-D mirror dataset represents the dataset with manually processed labels after applying our proposed method, where all three modalities are in their optimal state. Similarly, for the glass detection experiment, we also used three datasets: the raw RGB-T glass dataset, the enhanced RGB-T glass dataset, and the ideal RGB-T glass dataset. Therefore, in this paper, we conducted experiments using the raw, enhanced, and ideal datasets for both tasks to validate the effectiveness of our method in improving the datasets.

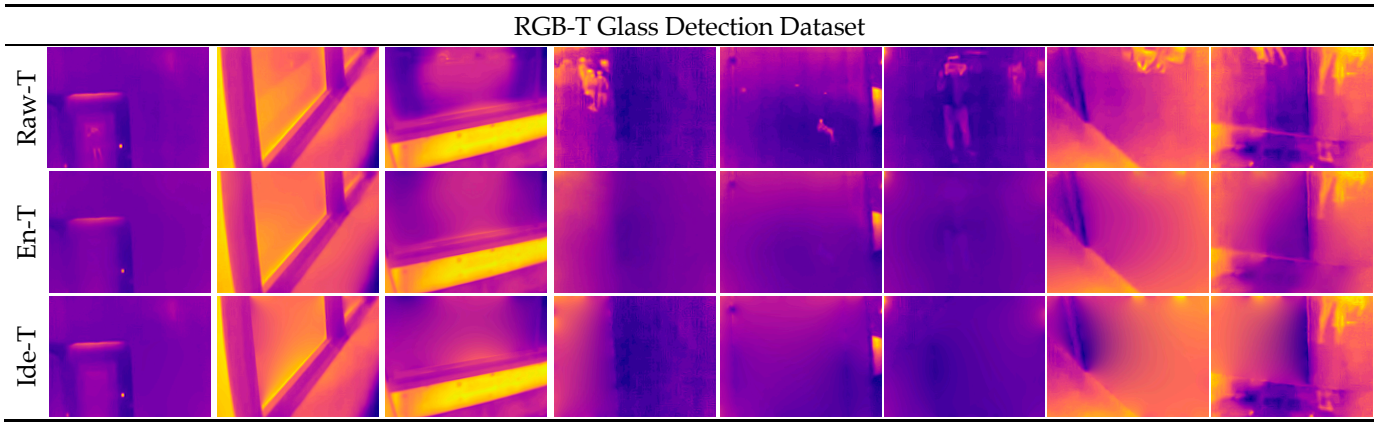


Figure 5. Examples of raw, enhanced, and ideal thermal data samples used in the RGB-T glass detection task.

3.2. Dataset Enhancement Principles

In this section, we will provide a detailed description of the data enhancement principles for the three modalities. For the RGB data enhancement, since both mirror detection and glass detection tasks involve this modality, we apply the same enhancement method for this modality in both tasks. We also use the same pre-segmentation network, image restoration network, and enhancement module for both detection tasks. On one hand, we chose the recently proposed U-Lite [68] for the pre-segmentation network, as it achieves excellent performance with a relatively simple structure. On the other hand, we used the latest DSRNet [69] as the image restoration network, which is a state-of-the-art (SOTA) model for image restoration tasks across multiple benchmark datasets.

3.2.1. Depth Data Enhancement

For the enhancement of depth data, our goal is to eliminate inaccuracies in the depth data of the mirror surface caused by laser reflection during measurement by the depth camera. Therefore, we first determine the ideal depth data based on the three-point plane determination method, and then use the image restoration network DSRNet to complete the raw depth data. The specific process is illustrated in Figure 6.

In three-dimensional space, a plane can typically be determined by three independent points $A(x_1,\ y_1, z_1)$, $B(x_2,\ y_2, z_2)$, and $C(x_3,\ y_3, z_3)$. First, by using the spatial coordinates of these three points, two vectors, **AB** and **AC**, can be defined. The normal vector of the plane, $\mathbf{n} = (n_x,\ n_y, n_z)$, can then be determined by the cross product of these two vectors, as shown in Equation (1):

$$\mathbf{n} = \mathbf{AB} \times \mathbf{AC} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \boldsymbol{k} \\ x_2 - x_1 & y_2 - y_1 & z_2 - z_1 \\ x_3 - x_1 & y_3 - y_1 & z_3 - z_1 \end{vmatrix} \tag{1}$$

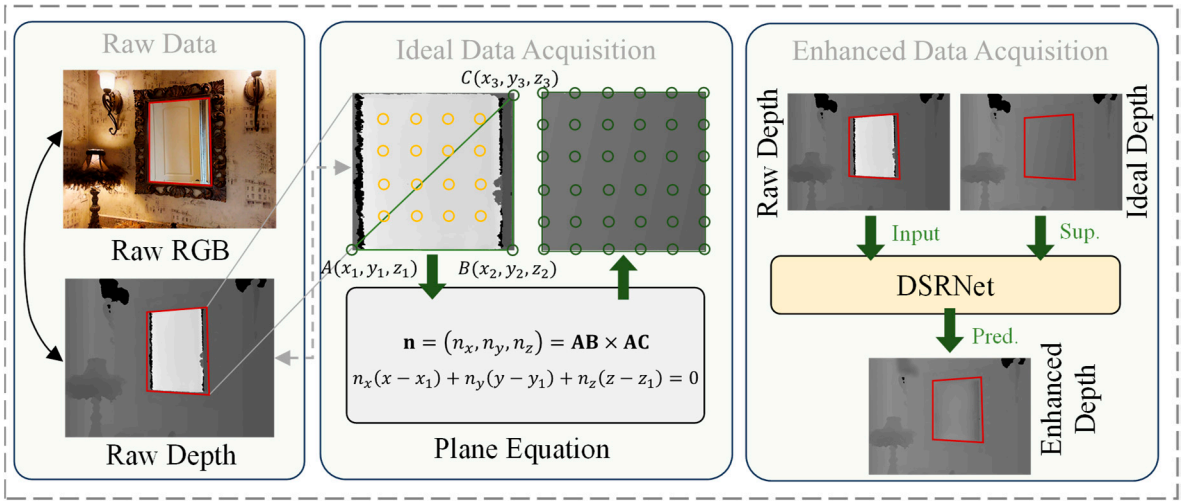


Figure 6. Depth data acquisition process for the mirror surface based on the three-point plane determination principle, including raw, enhanced, and ideal data.

Thus, by using the normal vector **n** and any point such as $A(x_1,\ y_1, z_1)$ on the plane, the plane equation can be obtained using the general form of the plane equation as shown in Equation (2). Consequently, by knowing the x and y coordinates of any point on the plane, the z coordinate, which is also the depth data, can be obtained using the plane equation in Equation (2):

$$n_x(x - x_1) + n_y(y - y_1) + n_z(z - z_1) = 0 \tag{2}$$

For the depth data enhancement of the mirror surface in our study, the goal is to fill in the depth data gaps caused by the reflection of the mirror that result in missing measurements from the depth camera. By analyzing the distribution of data in the raw RGB-D mirror dataset, we observed that the missing depth data is concentrated around the mirror surface and at the center of the mirror. Therefore, based on the principle of the three-point plane, we manually selected the border or wall surrounding the mirror, which is coplanar with the mirror, to determine the equation of the mirror surface in three-dimensional space. This allowed us to approximate the depth data for all points on the mirror surface, thereby replacing the erroneous data with ideal depth data.

To ensure the coherence and integrity of our method, and to eliminate the need for manual data processing in practical applications, we used the raw depth data and ideal depth data. By inputting the raw depth data into an image restoration network (DSRNet) and using the ideal depth data as labels to supervise network training, we developed a model that can predict the missing depth data on the mirror surface during actual predictions. The resulting data is referred to as the enhanced depth data. Consequently, we obtained three types of depth data: the raw depth data, enhanced depth data, and ideal depth data. Samples of these three types of depth data are shown in Figure 3.

3.2.2. Thermal Data Enhancement

For the enhancement of thermal data, we utilize the boundary thermal data from the raw thermal data, and based on the principle of steady-state heat transfer in a two-dimensional plane and the Laplace equation, we determine the ideal thermal data. We then employ the image restoration network, DSRNet, to eliminate artifacts on the glass surface in the raw thermal data. The specific process is shown in Figure 7.

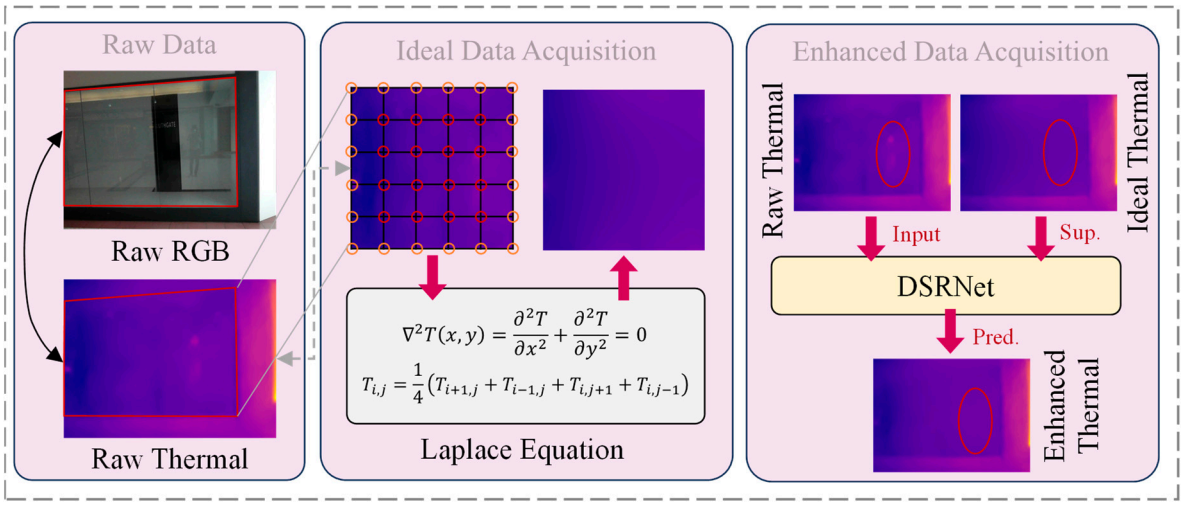


Figure 7. Thermal data acquisition process for the glass surface based on the principle of steady-state heat transfer in a two-dimensional plane, including raw, enhanced, and ideal data.

In a two-dimensional planar medium, steady-state heat conduction assumes that the temperature does not change with time, satisfying the steady-state condition. The thermal conductivity of the medium is uniform, and there are no significant heat sources. Therefore, according to Fourier's law of heat conduction and the principle of energy conservation, the temperature distribution $T(x, y)$ satisfies the Laplace equation. For the glass surface, the specific case in the raw RGB-T glass dataset suggests that we can assume the glass in the dataset satisfies the steady-state heat conduction conditions. That is, in the absence of significant heat sources, the temperature distribution on the glass surface remains in steady-state. Hence, the temperature at each point on the glass surface can be understood as the balance between the heat flowing into the point and the heat flowing out. Therefore, the temperature distribution on the glass surface, $T(x, y)$, can be described by the partial differential equation in Equation (3) as follows:

$$\nabla^2 T(x,y) = \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} = 0 \tag{3}$$

In Equation (3), ∇^2 represents the Laplace operator, $\frac{\partial^2 T}{\partial x^2}$ is the second-order temperature derivative in the x direction, and $\frac{\partial^2 T}{\partial y^2}$ is the second-order temperature derivative in the y direction.

For solving this second-order partial differential equation, we can discretize it with the boundary conditions specified at the boundary of the glass. Discretization methods can be categorized into the five-point difference scheme and the nine-point difference scheme. The five-point difference scheme approximates the second derivative using central differences, while the nine-point difference scheme improves discretization accuracy by incorporating neighboring points in the diagonal directions. The core idea behind the nine-point difference scheme is to construct higher-order difference approximations through Taylor series expansion. Since the primary objective of our study is to eliminate artifacts on the glass surface in the original thermal data, the requirement for numerical precision in the thermal data is relatively low. Additionally, considering computational costs, we have chosen to discretize the second-order partial differential equation using the five-point difference scheme in this context. This process is illustrated in Figure 7. Additionally,

Equation (3) can be expressed in the form of an algebraic system of equations, as shown in Equation (4):

$$\frac{T_{i+1,j} - 2T_{i,j} + T_{i-1,j}}{(\Delta x)^2} + \frac{T_{i,j+1} - 2T_{i,j} + T_{i,j-1}}{(\Delta y)^2} = 0 \tag{4}$$

Here, $T_{i,j}$ represents the temperature at the node with coordinates (i, j) on the plane, and $T_{i+1,j}$, $T_{i-1,j}$, $T_{i,j+1}$, $T_{i,j-1}$ refers to the temperature at the neighboring four nodes around this node. The above equation can be further simplified to Equation (5) as follows:

$$T_{i,j} = \frac{1}{4}(T_{i+1,j} + T_{i-1,j} + T_{i,j+1} + T_{i,j-1}) \tag{5}$$

Therefore, for the temperature $T_{i,j}$ at any point on the glass surface, it can be obtained from the discretized temperatures of the surrounding four nodes. By applying the boundary conditions and initializing the temperature at the center point, we can iteratively compute the temperature distribution across the entire glass surface. This approach allows us to eliminate artifacts caused by infrared light reflection in the thermal image of the glass surface. The data obtained through this method are referred to as the ideal thermal data.

Similarly, by combining the raw thermal data and the ideal thermal data, along with DSRNet training and prediction, we obtain the enhanced thermal data. Thus, we have three types of thermal data: the raw thermal data, enhanced thermal data, and ideal thermal data. Samples of these three types of thermal data are shown in Figure 5.

3.2.3. RGB Data Enhancement

For RGB data enhancement, we propose an RGB enhancement module, as illustrated in Figure 8. The module consists of two parts: one is the pre-segmentation part based on U-Lite, and the other is the directional feature enhancement module.

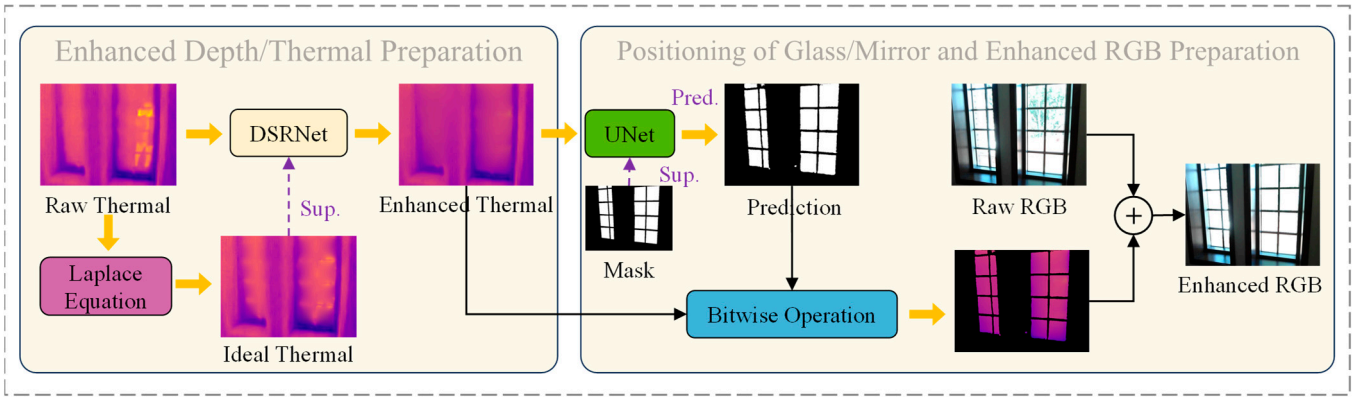


Figure 8. RGB data acquisition process for the mirror surface and glass surface based on the RGB enhancement module, including raw, enhanced, and ideal data.

To improve the distinction between the mirror or glass surface and the surrounding environment in RGB images, we fuse the RGB image with another modality of image data based on the principles of image fusion tasks. This fusion allows us to extract complementary information between the two modalities, thereby enhancing the quality of the features in the image.

Based on the aforementioned feature principles, the flow of our method is shown in Figure 8. First, we input the enhanced depth/thermal data processed by DSRNet into the U-Lite network to obtain an initial mirror/glass mask from a single modality. This mask can then be used to preliminarily determine the location and extent of the mirror/glass in the image. Using this mask, we extract the mirror/glass portion from the normalized

depth/thermal image and combine it with the RGB image, allowing us to enhance only the features of the mirror/glass surface in the RGB image. Thus, we obtain the enhanced RGB data.

To validate the effectiveness of our method, we replace the mask predicted by the U-Lite network with a manually labeled mask to obtain the most accurate feature-enhanced data for the mirror/glass surface. The data obtained through this method is referred to as the ideal RGB data. Therefore, we obtain three types of RGB data: the raw RGB data, enhanced RGB data, and ideal RGB data. Samples of these three types of RGB data for mirror detection and glass detection are shown in Figure 4.

4. Methodology

Since our research focuses on RGB-D mirror detection and RGB-T glass detection tasks, our approach primarily revolves around the use of bimodal input data, specifically RGB-depth or RGB-thermal modalities. In this section, we first introduce the overall framework of the model, followed by a detailed explanation of the model, with a focus on its modules and loss functions.

4.1. Architecture Overview

Most existing methods for recognizing objects with special optical surface properties, such as mirrors or glass, typically explore the combination of different modalities. They extract semantic information about mirrors or glass from different modalities and fuse the information from different dimensions and modalities through strategies like shallow fusion, intermediate fusion, and deep fusion. These fused cues are then compared with the semantic information of the surrounding environment to detect these objects. However, these methods mainly focus on improving the model's ability to extract and fuse information from different modalities, without addressing the inherent deficiencies of sensors in detecting objects like mirrors or glass with special surface properties. In other words, they do not consider how to improve model performance by enhancing and cleaning the dataset. Therefore, our work first focuses on using mathematical- and physical-based methods to enhance the surfaces of mirrors and glass in RGB, depth, and thermal modality data. We then use two parallel backbone networks to extract features from the enhanced RGB and depth/thermal modalities. We chose Res2Net [70] as our backbone network. During the encoding process, we use the prior fusion feature extraction (PFFE) module and the deep fusion feature guidance (DFFG) module to extract and fuse shallow feature weights and deep feature weights from both modalities. Unlike most models that employ various complex fusion methods during the encoding process to fuse two modality datasets, our approach uses the PEEF module in the early stages of the network to obtain prior deep fusion feature weights for both modality datasets. These weights are then interacted with the RGB features extracted at each stage of the encoder and simply added to the depth/thermal branch to improve its performance. The DFFG module then aggregates the final layer deep features extracted by both encoders and acquires high-level semantic information weights to better complement the fusion of modal representations during the decoding process. For the decoding process, we integrate the highly-discussed Mamba module and propose the bidirectional state space fusion (BSSF) block, which leverages the Mamba module's strong global modeling capability and combines it with the deep weights extracted by the DFFG module. This enables the stepwise restoration of fused encoding features in reverse order, generating the predicted output map. Throughout this process, we mainly extract and combine low-level clue weights and high-level context weights for the two modalities using the PEEF and DFFG modules. This allows us to model the global relationship between the two modalities. Additionally, we use the BSSF block based on the Mamba module

to complete the decoding process. Thus, we name our model the double weight Mamba fusion network (DWMFNet), designed to detect mirrors and glass. The overall architecture of the model is shown in Figure 9, and we use the RGB-T glass detection task as an example.

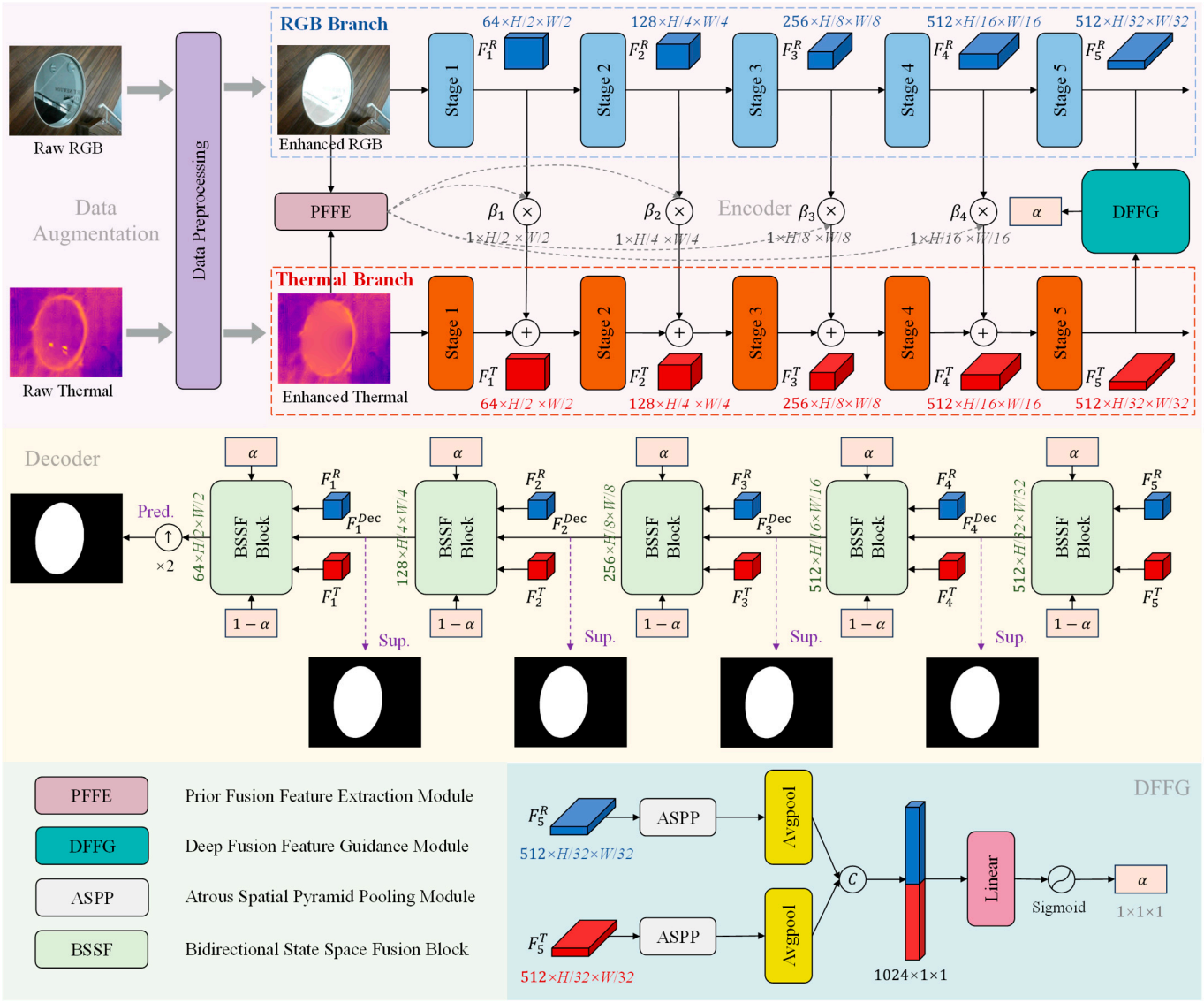


Figure 9. Overall architecture and DFFG module of the proposed DWMFNet.

4.2. Prior Fusion Feature Extraction Module

Whether for the RGB-D mirror detection task or the RGB-T glass detection task, both are essentially pixel-level classification tasks for two modalities, which require pixel-wise traversal of the input data to determine the specific category of each pixel. Since mirrors and glass are difficult to distinguish from the surrounding environment in each individual modality, it is necessary to effectively complement and correspond the pixel-level information from both modalities. Therefore, we propose the prior fusion feature extraction (PFFE) module, which combines these two types of modal data in the early stages of the model to more comprehensively acquire environmental information and enhance the complementarity of feature representations. The structure of the PFFE module is shown in Figure 10. Inspired by [71,72], we employ an early fusion strategy from [71] to directly add, multiply, and concatenate the features from different modalities. This operation enables the model to understand the relationship between the modalities at a lower level. This fusion method, while maintaining a low computational complexity, can

significantly improve the model's performance. Especially for tasks like mirror or glass detection, where there is a strong correlation between the two modalities, it ensures that the model captures the relationship between RGB and depth/thermal information from the outset. Subsequently, we input the concatenated features into a series of convolutional and average-pooling down sampling modules. Unlike [72], to reduce the model's computational load, we only use the weight β, obtained through the sigmoid function, to guide the fusion of RGB branch information into depth/thermal information. As shown in Figure 9, the extracted RGB features are expressed as $F_i^R\left(C_i^R \times H_i^R \times W_i^R\right)$, while thermal features are $F_i^T\left(C_i^T \times H_i^T \times W_i^T\right)$, where $i \in \{1,2,3,4,5\}$, and C, H, and W represent the channels, height, and width of the features at each stage, respectively, with the specific values shown in Figure 9. Meanwhile, the weight β obtained from the PFFE module has a final size of $(1 \times H/4 \times W/4)$. We apply weighted fusion only to the RGB flow features at stages $i \in \{1,2,3,4\}$. Prior to weighting the RGB flow features, we use an upsampling operation to adjust β to match the size of the RGB features at each stage. Therefore, the above process can be described by Equation (6) as follows:

$$\begin{aligned} \beta_i &= \mathrm{S}(Up(\mathrm{PFFE}(R_{in},\ T_{in}))) \\ F_{i+1}^D &= F_i^D \oplus \left(\beta_i \otimes F_i^R\right) \end{aligned} \tag{6}$$

where $\mathrm{PFFE}(\cdot)$ represents the PFFE module, $Up(\cdot)$ denotes the upsampling operation, and $\mathrm{S}(\cdot)$ indicates the Sigmoid function. The $\oplus$ and $\otimes$ represent element-wise summation and multiplication, respectively.

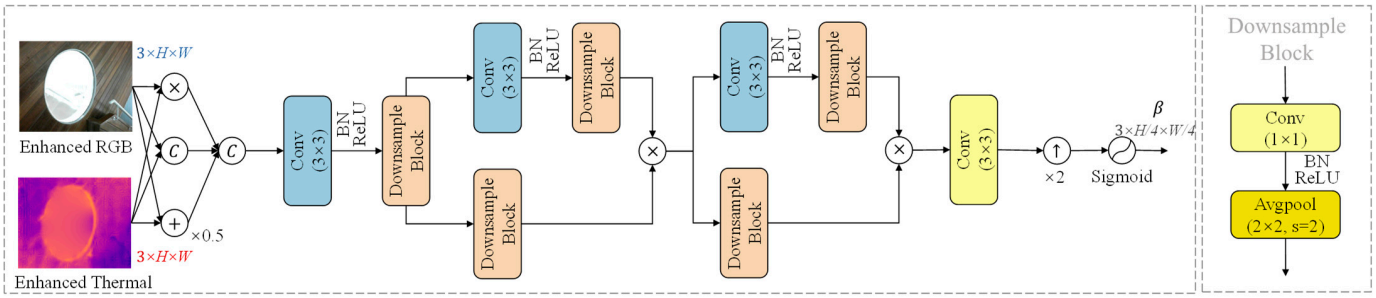


Figure 10. Illustration of prior fusion feature extraction module (PFFE).

4.3. Deep Fusion Feature Guidance Module

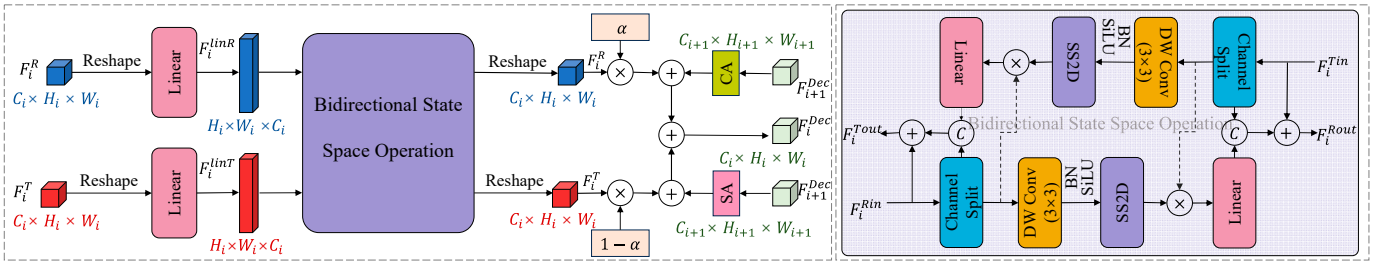
For the deep features F_5^R and F_5^T extracted from the last layer of the two encoder branches, we refer to the structure in [72] and use the deep fusion feature guidance (DFFG) module to aggregate the deep information from the two modalities. The structure of the DFFG module is shown in Figure 9. First, we apply the atrous spatial pyramid pooling (ASPP) module with multi-scale convolution operations to capture both the global and local details of the deep features from the two modalities. Then, the features from both modalities are downsampled to a size of 1×1, allowing the features from each channel to aggregate into a global statistical representation. This not only captures the global representative information of the feature map but also enhances the model's spatial invariance. Finally, through a series of fully connected layers and the Sigmoid function, we obtain a weight α of size $1 \times 1 \times 1$ to complement the fusion of the modality representations during the decoding process. The above operations can be represented by Equation (7) as follows:

$$\alpha = \mathrm{S}\left(FC\left(Concat\left(GAP\left(\mathrm{ASPP}\left(F_5^R\right)\right), GAP\left(\mathrm{ASPP}\left(F_5^T\right)\right)\right)\right)\right) \tag{7}$$

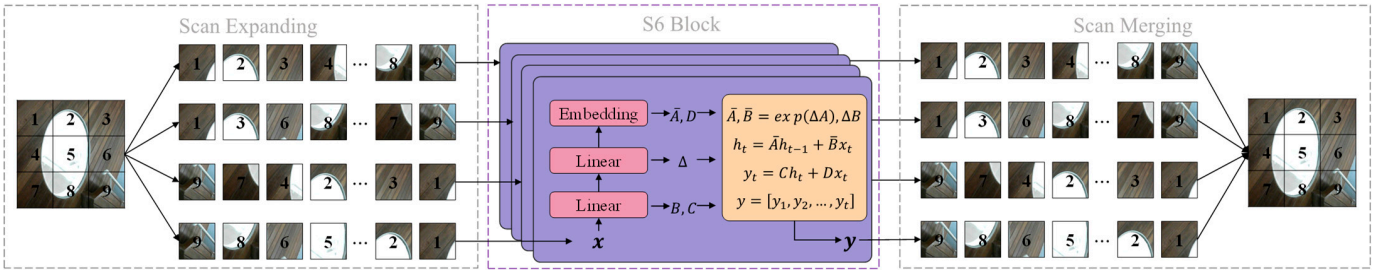
where ASPP$(\cdot)$ represents the ASPP module, $GAP(\cdot)$ denotes the global average pooling operation, $Concat(\cdot)$ refers to the concatenation operation, $FC(\cdot)$ indicates the fully connected layer, and S$(\cdot)$ represents the Sigmoid function.

4.4. Bidirectional State Space Fusion Block

Unlike most existing works that focus on modality fusion during the encoding phase, our DWMFNet emphasizes the fusion of features from two modalities during the generation of the final output, thereby improving the accuracy, detail retention, and global context understanding of the output. In other words, the fusion of the two modalities occurs during the decoding phase, aiming to integrate the features extracted by the network during the encoding phase. We incorporate the Mamba model, which efficiently captures long-range dependencies, and embed it into the BSSF block of our decoder. With the bidirectional feature fusion structure, we progressively restore and refine the features extracted during the encoding phase, enabling the model to generate more accurate outputs. The schematic of the BSSF block is shown in Figure 11a.



(**a**) Bidirectional state space fusion block



(**b**) SS2D operation

Figure 11. Illustration of bidirectional state space fusion block (BSSF) and SS2D operation.

The Mamba model is fundamentally a state space model (SSM). The SSM model originates from the Kalman filter [73] and is often used to represent linear time-invariant systems, where the internal state of the system is typically modeled as a set of linear equations [74]. For a one-dimensional input sequence $x(t) \in \mathbb{R}$, the model can obtain the output $y(t) \in \mathbb{R}$ through an intermediate state $h(t) \in \mathbb{R}^N$, where N is the dimensionality of the hidden layer. Specifically, the SSM is usually expressed in the form of Equation (8) as a linear ordinary differential equation (ODE) as follows:

$$\begin{aligned} h'(t) &= Ah(t) + Bx(t) \\ y(t) &= Ch(t) + Dx(t) \end{aligned} \tag{8}$$

where $A \in \mathbb{R}^{N\times N}$ represents the state transition matrix, $B \in \mathbb{R}^{N\times 1}$, and $C \in \mathbb{R}^{1\times N}$ represents the projection parameters, and $D \in \mathbb{C}^1$ represents the skip connection.

To apply the SSM to more complex deep learning models, it is necessary to convert the continuous state space model into a discrete form. Specifically, this involves introducing a time-scale parameter Δ, which converts the continuous A and B into discrete $\overline{A}$ and $\overline{B}$.

During this process, the zero-order hold (ZOH) principle is often widely used [75], and the above process can be expressed as Equation (9) as follows:

$$\begin{gathered} h_t = \overline{A}h_{t-1} + \overline{B}x_t \\ y(t) = \overline{C}h_t + Dx_t \\ \overline{A} = e^{\Delta A} \\ \overline{B} = (\Delta A)^{-1}\left(e^{\Delta A} - I\right)\Delta B \\ \overline{C} = C \end{gathered} \tag{9}$$

Finally, the output of the SSM can be represented as a convolution operation as Equation (10) as follows:

$$\begin{gathered} y = x * \overline{K} \\ \overline{K} = \left(C\overline{B}, C\overline{A}\overline{B}, \ldots, C\overline{A}^{L-1}\overline{B}\right) \end{gathered} \tag{10}$$

Based on the above principles, Mamba [8] divides sequence data into multiple sub-sequences, and the length of each sub-sequence can be adjusted according to the specific requirements, thereby achieving flexible processing of sequence data.

To extend the mechanism of Mamba from 1D sequence data to 2D image data, VMamba [76] proposed the 2D selective scanning (SS2D) mechanism, which expands image patches in four different directions to generate four independent sequences. Figure 11b provides a basic overview of SS2D. SS2D first divides the input 2D image into several blocks, then scans along four directions to generate four distinct sequences. Afterward, the selective scanning spatial state sequence model (S6) [23] is used to process each sequence. Finally, the processed sequences are re-stitched to form a new 2D image. SS2D is an important component of the visual state space model (VSSM), which is primarily capable of capturing both the global and detailed information of the image by perceiving the overall dimensionality of the input data [77]. Further descriptions of SS2D can be found in [76]. We utilize the SS2D mechanism to perceive the overall dimensions of the input spatial features, thereby capturing both global and local information of the image. Additionally, during the feature extraction phase, we extract features from the input modality data at both shallow and deep layers of the model. After fusion, using learnable weights α and β, this approach helps mitigate the high sensitivity of the SS2D operation within the Mamba module to spatial structures in the BSSF block.

As for our BSSF block, its schematic is shown in Figure 11a. First, for the input features F_i^R and F_i^T from the encoder, we first adjust their dimensions and pass the data flow through a fully connected layer. Then, the obtained features from the two modalities are input into the bidirectional state space operation, i.e., F_i^{Rin} and F_i^{Tin}. The above operation can be expressed as Equation (11):

$$\begin{aligned} F_i^{Rin} &= \text{reshape}\left(FC\left(F_i^R\right)\right) \\ F_i^{Tin} &= \text{reshape}\left(FC\left(F_i^T\right)\right) \end{aligned} \tag{11}$$

where $FC(\cdot)$ represents the fully connected layer, and $\text{reshape}(\cdot)$ indicates the dimension adjustment operation. In this case, we adjust the dimensions of F_i^R from its original size of $\left(C_i^T \times H_i^T \times W_i^T\right)$ to $\left(H_i^T \times W_i^T \times C_i^T\right)$.

Next, in the BSSO, although the input data to our model consists of images, which, unlike text and time-series data, lack explicit order, the inherent spatial proximity within images implies the potential for bidirectional interactions. Therefore, our BSSO does not assume a fixed order, but rather processes RGB and depth/thermal data through symmetric parallel paths. The two opposing branches respectively learn different aggregation

directions of local features, and ultimately, dynamic weight fusion adaptively balances the bidirectional information. This design enables the model to autonomously discover latent interaction patterns within the data, rather than relying on explicit sequences. We perform equal channel splitting on F_i^{Rin} and F_i^{Tin}, dividing them into $F1_i^{Rin}$, $F2_i^{Rin}$, $F1_i^{Tin}$, and $F2_i^{Tin}$ to allow for better interaction between the two modules along the channel dimension. Then, $F1_i^{Rin}$ and $F1_i^{Tin}$ are respectively passed through successive depthwise separable convolutions [78], the SS2D, and linear layers. Prior to entering the linear layers, the two features interact through a simple element-wise multiplication to allow for cross-modal interaction. Subsequently, the features are concatenated with the previously split $F2_i^{Tin}$ and $F2_i^{Rin}$, and added to the original F_i^{Tin} and F_i^{Rin} to complete the entire BSSO operation. The final outputs, F_i^{Rout} and F_i^{Tout}, maintain the same dimensions and size as the original inputs F_i^{Rin} and F_i^{Tin}. The above process can be represented as Equation (12):

$$\begin{aligned} F1_i^{Rin},\ F2_i^{Rin} &= \text{split}\left(F_i^{Rin}\right) \\ F1_i^{Tin},\ F2_i^{Tin} &= \text{split}\left(F_i^{Tin}\right) \\ F_i^{Rout} &= F_i^{Tin} \oplus Concat\left(F2_i^{Tin} \oplus FC\left(F1_i^{Tin} \otimes \text{SS2D}\left(DW\left(F1_i^{Rin}\right)\right)\right)\right) \\ F_i^{Tout} &= F_i^{Rin} \oplus Concat\left(F2_i^{Rin} \oplus FC\left(F1_i^{Rin} \otimes \text{SS2D}\left(DW\left(F1_i^{Tin}\right)\right)\right)\right) \end{aligned} \tag{12}$$

where $\text{split}(\cdot)$ denotes the operation of splitting the features along the channel dimension, $DW(\cdot)$ refers to the depthwise separable convolution operation, and $\text{SS2D}(\cdot)$ represents the 2D selective scanning operation.

After passing through the BSSO, the features of the two modalities undergo semantic-level fusion. We then reshape them to match the original dimensions of F_i^R and F_i^T, and further augment these features using the deep-level weight information α obtained from the DFFG module. This results in the enhanced features F_i^{EnR} and F_i^{EnT}. To further improve the propagation and interactivity between decoder layers, we also employ channel attention (CA) [79] and spatial attention (SA) [80] mechanisms to strengthen the features from the previous decoder layer F_{i+1}^{Dec}. These are then added to the enhanced features F_i^R and F_i^T to produce the final output of the BSSF block, denoted as F_i^{Dec}. The above process can be described as Equation (13):

$$\begin{aligned} F_i^{EnR} &= \alpha \otimes F_i^R \\ F_i^{EnT} &= (1-\alpha) \otimes F_i^R \\ F_i^{Dec} &= \text{CA}\left(F_{i+1}^{Dec}\right) \oplus \text{SA}\left(F_{i+1}^{Dec}\right) \oplus F_i^{EnR} \oplus F_i^{EnT} \end{aligned} \tag{13}$$

where $\text{CA}(\cdot)$ and $\text{SA}(\cdot)$ represent the channel attention and spatial attention operations, respectively.

4.5. Loss Function

During the training process, we use a hybrid loss with multi-level supervision [81] to train our network. For the given side predictions $P_i, i \in \{1,2,3,4,5\}$, and the ground truth G, we combine the binary cross entropy (BCE) loss l_{bce} [82] and the intersection over union (IoU) loss l_{iou} [83] to form the total loss function L_{total}. Specifically, l_{bce} is used to compute the difference between the probability distribution P predicted by the model and the ground truth G, while l_{iou} calculates the overlap between the predicted probability distribution P and the ground truth G. The expression for the total loss function L_{total} is given by Equation (14) as follows:

$$L_{\text{total}} = \sum_{i=1}^{5} l_{\text{bce}}(G, P_i) + \sum_{i=1}^{5} l_{\text{iou}}(G, P_i) \tag{14}$$

As for the expressions of l_{bce} and l_{iou}, they are given by Equations (15) and (16), respectively:

$$l_{\text{bce}} = -\sum_{j=1}^{N}\left[G * \log(P_j) + (1-G) * \log(1-P_j)\right] \tag{15}$$

$$l_{\text{iou}} = 1 - \frac{\sum_{j=1}^{N}\left[G * P_j\right]}{\sum_{j=1}^{N}\left[G + P_j - G * P_j\right]} \tag{16}$$

where N is the total number of pixels in P.

5. Experiments

5.1. Experimental Setup

5.1.1. Datasets

The main objective of this work is to investigate the effectiveness of the mathematical and physical methods we employ for the reconstruction of RGB, depth, and thermal images of mirror/glass objects in enhancing the performance of the model. For the RGB-D mirror detection task, we established three datasets: the raw RGB-D mirror dataset, the enhanced RGB-D mirror dataset, and the ideal RGB-D mirror dataset. We then observed whether the model's performance improves across these three datasets. If the same model shows improvement on these datasets, it validates the effectiveness of our approach. For all the aforementioned RGB-D mirror datasets, the authors of the raw RGB-D mirror dataset implemented the registration and alignment of the two-modal images, and performed normalization on the depth images. We followed the setup from [21], where 2000 image pairs were used for training and 1049 image pairs for testing. Similarly, for the RGB-T glass detection task, the authors of the raw RGB-T glass dataset also implemented the registration and alignment of the two-modal images, we used the raw RGB-T glass dataset, the enhanced RGB-T glass dataset, and the ideal RGB-T glass dataset to observe the model's performance across these three datasets. Consistent with the setup in [22], we used 4427 image pairs for training and 1124 image pairs for testing in the RGB-T glass detection task.

5.1.2. Evaluation Metrics

To evaluate the performance of our method against other comparative methods on the datasets, we selected five commonly used metrics to assess the model's prediction results: mean absolute error (MAE) [84], balanced error rate (BER) [85], weighted F-measure (W_F) [86], intersection over union (IoU), and S-measure (S_m) [87]. For MAE and BER, smaller values indicate better model performance, while for W_F, IoU, and S_m, larger values indicate better performance. Additionally, we also calculated precision–recall (PR) curves and F-measure curves to observe the overall performance of the model. For the PR curve, the closer the curve is to the point $(1, 1)$, the better the model's performance. For the F-measure curve, the closer the value is to 1 at different thresholds, the better the model's performance.

5.1.3. Implementation Details

All experiments were conducted on a workstation equipped with an Nvidia RTX 4090 GPU, running CUDA 11.3 and PyTorch 1.11.0. The backbone of our network was initialized with weights pre-trained on the ImageNet dataset [88]. The input image resolution for both training and testing was set to 352×352, with the number of training epochs set to 100. The batch size during training was 6. We used stochastic gradient descent (SGD) [89] to optimize the model parameters, with an initial learning rate of 0.001, which decayed to 0.0001 after 20 epochs and to 0.00001 after 50 epochs. The momentum was set to 0.9, and the weight decay to 0.0005. For the other comparison methods selected in our experiments,

we retrained and tested all methods using the same dataset and data partitioning scheme, ensuring that the parameter settings for all network models remained consistent with those in the original papers to ensure fairness in the experiments. For both the RGB-D mirror detection task and the RGB-T glass detection task, we trained all methods on three different datasets—raw datasets, enhanced datasets, and ideal datasets—to validate the effectiveness of the proposed improvements in the modality data for images containing mirrors/glasses.

5.2. Comparison with State-of-the-Art Methods

5.2.1. Comparison Methods

To demonstrate the effectiveness of the proposed method, we compared the DWMFNet with 18 recent advanced methods in the field across the 6 datasets mentioned. For the RGB-D mirror detection task, the selected methods include MoADNet [90], TBINet [91], SwinNet [92], AFNet [93], HINet [94], HiDANet [95], SAINet [96], MAGNet [97], and HFILNet [98]. For the RGB-T glass detection task, the selected methods include OSRNet [71], CSRNet [99], DCNet [100], LSNet [101], E2Net [102], MGAINet [103], SwinMCNet [104], LAFB [105], and ConTriNet [106].

5.2.2. Quantitative Comparison

For the RGB-D mirror detection task, Table 1 presents the quantitative results of our method and other state-of-the-art (SOTA) methods on the three datasets we provided. Compared to the SOTA methods, our DWMFNet achieves satisfying performance. On the raw RGB-D mirror dataset, our method achieves the second-best results, with only a 0.9%, 1.7%, and 1.8% difference in the S_m, W_F, and IoU metrics compared to the best method, TBINet. On the enhanced RGB-D mirror dataset, our method achieves the best results, outperforming the second-best method, SwinNet, by 1.7%, 4.8%, and 4.6% in the S_m, W_F, and IoU metrics. On the ideal RGB-D mirror dataset, our method achieves the second best results, with only a 0.8% and 1.3% difference in the W_F and IoU metrics compared to the best method, TBINet. Compared to the raw RGB-D mirror dataset, most models performed better on the enhanced RGB-D mirror and ideal RGB-D mirror datasets. For example, MoADNet, which performed the worst on the raw RGB-D mirror dataset, achieved improvements of 38.4%, 40.2%, 80.0%, 85.4%, and 63.4% in MAE, S_m, W_F, IoU, and BER metrics on the enhanced RGB-D mirror dataset. On the ideal RGB-D mirror dataset, it showed improvements of 58.1%, 40.2%, 80.0%, 85.4%, and 80.0% in the same metrics. For HINet, HiDANet, and HFILNet, whose performance did not improve on the enhanced RGB-D mirror dataset, this may be due to the distribution differences in our enhanced RGB-D mirror dataset, which is a secondary-generated dataset. When encountering models with weak generalization ability, the model performance may degrade. However, the stronger performance exhibited by most models on the enhanced and ideal RGB-D mirror datasets proves that our data preprocessing method positively impacts the model's performance in RGB and depth mirror detection tasks.

For the RGB-T glass detection task, Table 2 presents the quantitative results of our method and other SOTA methods on the three datasets. Compared to the SOTA methods, our DWMFNet achieves the best performance on both the enhanced RGB-T glass dataset and the ideal RGB-T glass dataset. On the raw RGB-T glass dataset, our method achieves the second-best result, with only a 1.0% and 1.3% difference in W_F and IoU metrics compared to the best method, LAFB. Additionally, except for ConTriNet, which may have experienced performance degradation on the enhanced RGB-T glass dataset due to model generalization reasons, all other models performed better on the enhanced RGB-T glass dataset and the ideal RGB-T glass dataset. For example, CSRNet, which performed the worst on the raw RGB-T glass dataset, showed improvements of 58.8%, 31.1%, 29.0%, 26.2%,

and 47.1% in MAE, S_m, W_F, IoU, and BER metrics on the enhanced RGB-T glass dataset, respectively. On the ideal RGB-T glass dataset, these improvements were 95.1%, 56.2%, 48.8%, 61.1%, and 77.0%, respectively. This demonstrates that our data preprocessing method significantly enhances the performance of models in RGB and thermal modality-based glass detection tasks. At the same time, it can be observed that most models exhibit strong stability on the RGB-T glass dataset.

Table 1. Quantitative comparison results of different methods on the test sets of the raw, enhanced, and ideal RGB-D mirror detection datasets. All methods are retrained on the corresponding datasets. The best results on the raw dataset are highlighted in green, the best results on the enhanced dataset are highlighted in blue, and the best results on the ideal dataset are highlighted in red. The **bolded** values in parentheses represent the optimization values of each model's performance on the respective datasets compared to the results on the raw dataset.

Method	Dataset	Evaluation Metrics				
		$MAE\downarrow$	$S_m\uparrow$	$W\_F\uparrow$	$IoU\uparrow$	$BER\downarrow$
MoADNet$_{2022}$ [90]	Raw	0.198	0.419	0.118	7.61	46.17
	Enhanced	0.122 **(−0.076)**	0.701 **(+0.282)**	0.591 **(+0.473)**	52.21 **(+44.60)**	16.88 **(−29.29)**
	Ideal	0.083 **(−0.115)**	0.802 **(+0.383)**	0.743 **(+0.625)**	69.37 **(+61.76)**	9.23 **(−36.94)**
TBINet$_{2022}$ [91]	Raw	0.043	0.858	0.833	77.85	8.26
	Enhanced	0.031 **(−0.012)**	0.873 **(+0.015)**	0.854 **(+0.021)**	80.89 **(+3.04)**	7.51 **(−0.75)**
	Ideal	0.007 **(−0.036)**	0.956 **(+0.097)**	0.968 **(+0.135)**	94.45 **(+16.60)**	2.31 **(−5.95)**
SwinNet$_{2022}$ [92]	Raw	0.047	0.838	0.775	73.63	8.46
	Enhanced	0.029 **(−0.018)**	0.880 **(+0.042)**	0.842 **(+0.067)**	80.32 **(+6.69)**	7.52 **(−0.94)**
	Ideal	0.034 **(−0.013)**	0.895 **(+0.057)**	0.875 **(+0.101)**	81.98 **(+8.35)**	8.27 **(−0.19)**
AFNet$_{2023}$ [93]	Raw	0.054	0.837	0.790	73.05	10.67
	Enhanced	0.046 **(−0.007)**	0.852 **(+0.015)**	0.810 **(+0.020)**	75.55 **(+2.50)**	10.29 **(−0.37)**
	Ideal	0.025 **(−0.028)**	0.908 **(+0.072)**	0.895 **(+0.105)**	84.66 **(+11.61)**	5.91 **(−4.76)**
HINet$_{2023}$ [94]	Raw	0.050	0.840	0.786	72.61	10.97
	Enhanced	0.072 (+0.022)	0.823 (−0.018)	0.723 (−0.063)	66.37 (−6.32)	14.43 (+3.46)
	Ideal	0.015 **(−0.035)**	0.928 **(+0.088)**	0.922 **(+0.136)**	87.95 **(+15.34)**	5.20 **(−5.77)**
HiDANet$_{2023}$ [95]	Raw	0.057	0.823	0.781	72.51	9.78
	Enhanced	0.068 (+0.011)	0.803 (−0.020)	0.732 (−0.049)	68.81 (−3.70)	11.64 (+1.85)
	Ideal	0.017 **(−0.040)**	0.924 **(+0.101)**	0.924 **(+0.143)**	89.46 **(+16.95)**	4.04 **(−5.74)**
SAINet$_{2024}$ [96]	Raw	0.065	0.797	0.740	68.46	12.15
	Enhanced	0.051 **(−0.013)**	0.845 **(+0.048)**	0.818 **(+0.078)**	76.85 **(+8.39)**	9.36 **(−2.79)**
	Ideal	0.035 **(−0.030)**	0.885 **(+0.087)**	0.868 **(+0.127)**	82.32 **(+13.86)**	6.89 **(−5.26)**
MAGNet$_{2024}$ [97]	Raw	0.043	0.844	0.811	75.36	8.93
	Enhanced	0.029 **(−0.014)**	0.877 **(+0.033)**	0.861 **(+0.050)**	81.26 **(+5.90)**	7.27 **(−1.66)**
	Ideal	0.008 **(−0.035)**	0.950 **(+0.106)**	0.960 **(+0.149)**	93.31 **(+17.95)**	2.63 **(−6.30)**
HFILNet$_{2024}$ [98]	Raw	0.118	0.665	0.554	45.13	26.41
	Enhanced	0.139 (+0.021)	0.609 (−0.056)	0.443 (−0.111)	40.37 (−4.76)	26.43 (+0.02)
	Ideal	0.052 **(−0.066)**	0.792 **(+0.126)**	0.727 **(+0.173)**	64.59 **(+19.46)**	17.42 **(−8.99)**
DWMFNet	Raw	0.042	0.850	0.819	76.39	9.06
	Enhanced	0.025 **(−0.017)**	0.895 **(+0.044)**	0.884 **(+0.065)**	84.23 **(+7.84)**	6.24 **(−2.82)**
	Ideal	0.008 **(−0.034)**	0.956 **(+0.105)**	0.960 **(+0.141)**	93.23 **(+16.84)**	2.72 **(−6.34)**

Table 2. Quantitative comparison results of different methods on the test sets of the raw, enhanced, and ideal RGB-D mirror detection datasets. All methods are retrained on the corresponding datasets. The best results on the raw dataset are highlighted in green, the best results on the enhanced dataset are highlighted in blue, and the best results on the ideal dataset are highlighted in red. The **bolded** values in parentheses represent the optimization values of each model's performance on the respective datasets compared to the results on the raw dataset.

Method	Dataset	Evaluation Metrics				
		$MAE\downarrow$	$S_m\uparrow$	$W\_F\uparrow$	$IoU\uparrow$	$BER\downarrow$
$OSRNet_{2022}$ [71]	Raw	0.069	0.829	0.829	79.87	12.29
	Enhanced	0.020 **(−0.049)**	0.903 **(+0.073)**	0.895 **(+0.066)**	88.12 **(+8.25)**	7.56 **(−4.73)**
	Ideal	0.006 **(−0.063)**	0.939 **(+0.110)**	0.921 **(+0.092)**	91.19 **(+11.32)**	5.69 **(−6.60)**
$CSRNet_{2022}$ [99]	Raw	0.265	0.594	0.606	55.11	30.35
	Enhanced	0.109 **(−0.156)**	0.779 **(+0.185)**	0.781 **(+0.176)**	74.65 **(+19.54)**	16.07 **(−14.28)**
	Ideal	0.013 **(−0.252)**	0.928 **(+0.334)**	0.902 **(+0.296)**	88.77 **(+33.66)**	6.99 **(−23.36)**
$DCNet_{2022}$ [100]	Raw	0.065	0.852	0.833	80.90	11.07
	Enhanced	0.024 **(−0.041)**	0.918 **(+0.067)**	0.884 **(+0.051)**	86.87 **(+5.96)**	7.84 **(−3.23)**
	Ideal	0.008 **(−0.057)**	0.950 **(+0.099)**	0.917 **(+0.084)**	90.30 **(+9.40)**	5.65 **(−5.42)**
$LSNet_{2023}$ [101]	Raw	0.059	0.850	0.840	80.96	11.15
	Enhanced	0.031 **(−0.028)**	0.887 **(+0.037)**	0.872 **(+0.031)**	85.07 **(+4.11)**	8.82 **(−2.34)**
	Ideal	0.010 **(−0.048)**	0.939 **(+0.089)**	0.915 **(+0.075)**	90.36 **(+9.40)**	6.03 **(−5.12)**
$E2Net_{2023}$ [102]	Raw	0.073	0.842	0.827	78.12	12.23
	Enhanced	0.019 **(−0.054)**	0.914 **(+0.072)**	0.890 **(+0.064)**	87.22 **(+9.10)**	7.43 **(−4.80)**
	Ideal	0.013 **(−0.060)**	0.935 **(+0.093)**	0.912 **(+0.085)**	89.66 **(+11.54)**	6.04 **(−6.19)**
$MGAINet_{2023}$ [103]	Raw	0.059	0.854	0.842	81.06	11.32
	Enhanced	0.191 (+0.132)	0.795 (−0.059)	0.675 (−0.156)	58.40 (−22.66)	23.00 (+11.68)
	Ideal	0.018 **(−0.041)**	0.928 **(+0.074)**	0.915 **(+0.073)**	90.26 **(+9.20)**	6.63 **(−4.69)**
$SwinMCNet_{2024}$ [104]	Raw	0.053	0.866	0.856	83.33	10.06
	Enhanced	0.025 **(−0.028)**	0.903 **(+0.037)**	0.890 **(+0.034)**	87.23 **(+3.89)**	7.79 **(−2.27)**
	Ideal	0.007 **(−0.046)**	0.938 **(+0.072)**	0.921 **(+0.065)**	91.26 **(+7.93)**	5.45 **(−4.61)**
$LAFB_{2024}$ [105]	Raw	0.039	0.886	0.872	85.01	9.26
	Enhanced	0.016 **(−0.022)**	0.932 **(+0.046)**	0.903 **(+0.031)**	88.05 **(+3.04)**	7.10 **(−2.16)**
	Ideal	0.015 **(−0.024)**	0.963 **(+0.077)**	0.923 **(+0.051)**	91.52 **(+6.51)**	5.96 **(−3.30)**
$ConTriNet_{2024}$ [106]	Raw	0.066	0.851	0.837	80.46	11.04
	Enhanced	0.046 **(−0.021)**	0.899 **(+0.048)**	0.868 **(+0.031)**	85.02 **(+4.56)**	8.38 **(−2.67)**
	Ideal	0.035 **(−0.031)**	0.903 **(+0.052)**	0.877 **(+0.040)**	85.95 **(+5.49)**	8.23 **(−2.81)**
DWMFNet	Raw	0.042	0.894	0.863	83.94	9.56
	Enhanced	0.012 **(−0.030)**	0.943 **(+0.049)**	0.903 **(+0.040)**	88.90 **(+4.96)**	7.09 **(−2.47)**
	Ideal	0.003 **(−0.039)**	0.976 **(+0.082)**	0.924 **(+0.060)**	91.68 **(+7.74)**	5.42 **(−4.14)**

Moreover, from the PR curves and F-measure curves in Figures 12 and 13, we observe that nearly all models exhibit improved performance across different thresholds on the enhanced and ideal datasets. This further supports the effectiveness of our mathematical and physical method for enhanced RGB, depth, and thermal modality data, contributing positively to the performance of mirror and glass detection tasks. Moreover, all models achieved unexpectedly good results on the ideal dataset.

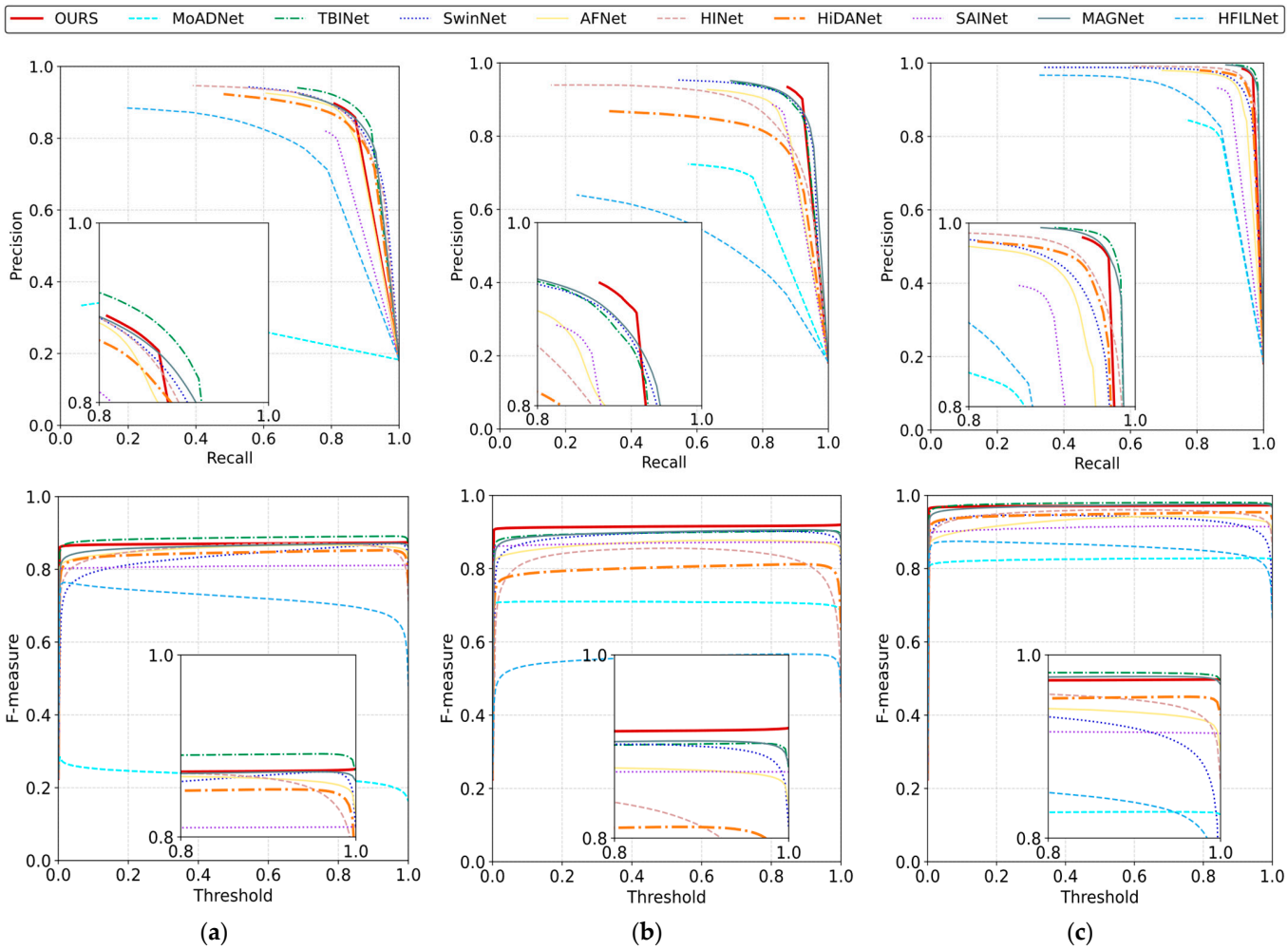


Figure 12. Comparison of PR curves and F-measure curves for DWMFNet (solid red line) and other methods on three different RGB-D mirror detection datasets: (**a**) PR curves and F-measure curves on raw dataset; (**b**) PR curves and F-measure curves on enhanced dataset; (**c**) PR curves and F-measure curves on ideal dataset.

5.2.3. Qualitative Comparison

To visually demonstrate the superiority of our DWMFNet and the enhancement of input modality data through the mathematical and physics method for improving the performance of various models, we present the visual detection results of different methods in Figures 14 and 15. Figure 14 shows the mirror detection results of different models across three datasets. It can be seen that our DWMFNet, when detecting mirror texture details is closer to the ground truth compared to other methods. Additionally, the detection performance of MoADNet improves progressively with our enhanced and ideal datasets, highlighting the effectiveness of these datasets. Similarly, Figure 15 shows that our DWMFNet performs well in detecting smaller glass objects. Moreover, almost all methods achieve more accurate detection results on the enhanced RGB-T glass dataset and the ideal RGB-T glass dataset than on the raw RGB-T glass dataset.

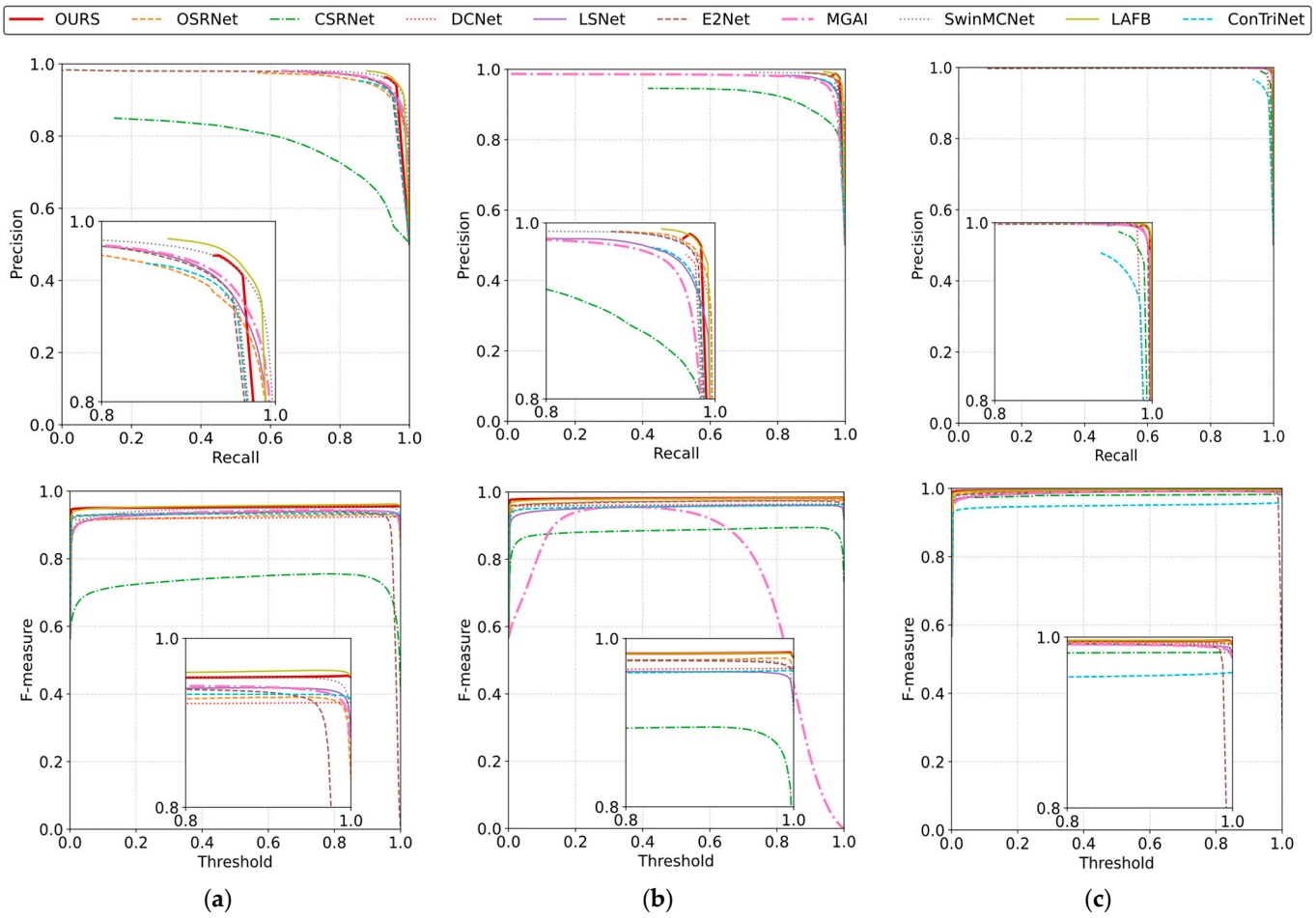


Figure 13. Comparison of PR curves and F-measure curves for DWMFNet (solid red line) and other methods on three different RGB-T glass detection datasets: (**a**) PR curves and F-measure curves on raw dataset; (**b**) PR curves and F-measure curves on enhanced dataset; (**c**) PR curves and F-measure curves on ideal dataset.

5.3. Ablation Study

5.3.1. Ablation Study on Loss Function

To validate the effectiveness of each component in our combined loss function, we conducted ablation experiments on the loss function part of the model. The experimental results are shown in Table 3. All our ablation experiments were conducted using DWMFNet on the enhanced RGB-D mirror dataset. It can be seen that when using our combined loss function l_{total}, the evaluation metrics of the model improve.

Table 3. Ablation experiment of loss function. All experiments are performed on DWMFNet in enhanced RGB-D mirror dataset.

Loss Function		**Evaluation Metrics**				
L_{bce}	L_{iou}	$MAE\downarrow$	$S_m\uparrow$	$W\_F\uparrow$	$IoU\uparrow$	$BER\downarrow$
√	-	0.026	0.890	0.878	83.06	7.03
√	√	0.025	0.895	0.884	84.23	6.24

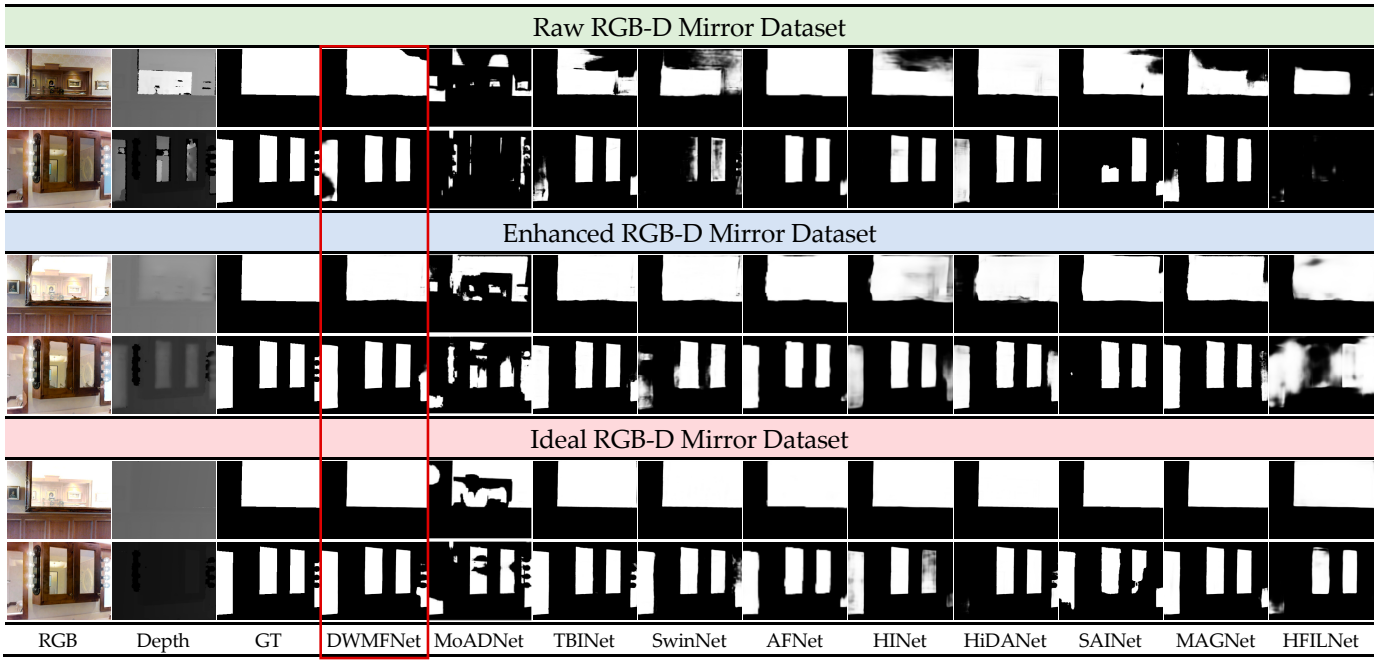


Figure 14. Qualitative comparison results of our model DWMFNet (red bounding boxes) and other methods in raw, enhanced, and ideal RGB-D mirror detection dataset.

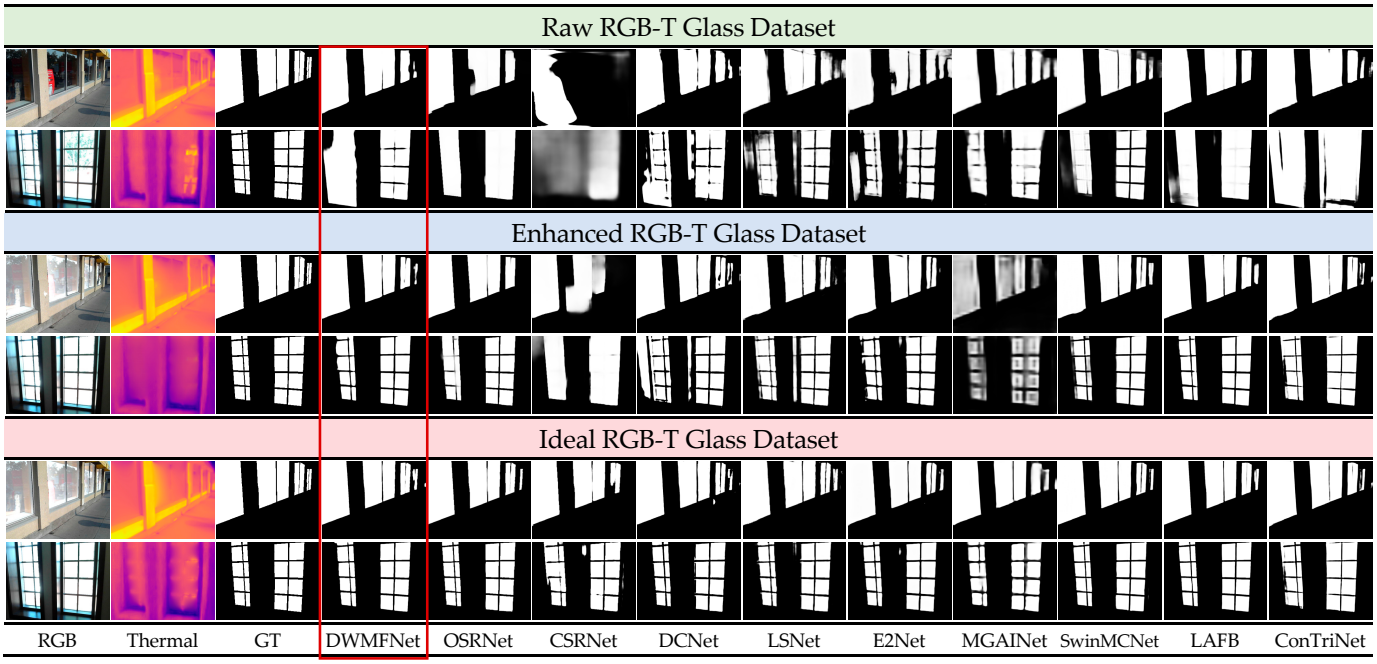


Figure 15. Qualitative comparison results of our model DWMFNet (red bounding boxes) and other methods in raw, enhanced, and ideal RGB-T glass detection dataset.

5.3.2. Ablation Study on Components

Table 4 demonstrates the effectiveness of each component in our DWMFNet model. All our ablation experiments were conducted using DWMFNet on the enhanced RGBD-mirror dataset. For the first ablation experiment, we removed the PFFE, DFFG, and BSSF block from the model. For the decoder, we only used simple summation and upsampling operations to obtain the final prediction results, in order to prove the effectiveness of the BSSF block. It can be observed that each time we add a module to the network, the model's

performance metrics improve. Therefore, it can be concluded that the components used in our network contribute positively to enhancing the network's performance.

Table 4. Ablation experiment of components in our model. All experiments are performed on DWMFNet in enhanced RGB-D mirror dataset.

Modules			Evaluation Metrics				
PFFE	**DFFG**	**BSSF**	***MAE*↓**	***S_m*↑**	***W_F*↑**	***IoU*↑**	***BER*↓**
-	-	-	0.068	0.807	0.774	68.85	14.64
√	-	-	0.066	0.815	0.787	70.40	13.77
√	√	-	0.051	0.828	0.805	72.64	12.89
√	√	√	0.025	0.895	0.884	84.23	6.24

5.3.3. Ablation Study on Input Modal

To demonstrate that combining different modalities for the enhanced RGB-D mirror and enhanced RGB-T glass datasets is beneficial for mirror and glass detection, Tables 5 and 6 present the ablation experiment results obtained by using different modality combinations as inputs to the network. It can be observed that when both input modalities of DWMFNet are depth, the model performs the worst for mirror detection. When the input modalities are changed to RGB only, the detection performance improves significantly. The best performance is achieved when both RGB and depth modalities are combined. From Table 6, it can also be seen that when both input modalities of DWMFNet are thermal, the detection performance for glass is poor. When the input modality is switched to RGB, the performance improves significantly. However, when both modalities are combined, DWMFNet achieves the best performance for mirror detection.

Table 5. Ablation experiment of input modal. All experiments are performed on DWMFNet in enhanced RGB-D mirror dataset.

Input Modal	**Evaluation Metrics**				
	***MAE*↓**	***S_m*↑**	***W_F*↑**	***IoU*↑**	***BER*↓**
Depth-Depth	0.063	0.787	0.726	66.49	13.35
RGB-RGB	0.030	0.885	0.872	82.70	6.72
RGB-Depth	0.025	0.895	0.884	84.23	6.24

Table 6. Ablation experiment of input modal. All experiments are performed on DWMFNet in enhanced RGB-T glass dataset.

Input Modal	**Evaluation Metrics**				
	***MAE*↓**	***S_m*↑**	***W_F*↑**	***IoU*↑**	***BER***
Thermal-Thermal	0.197	0.658	0.659	60.58	24.41
RGB-RGB	0.017	0.937	0.898	88.41	7.03
RGB-Thermal	0.012	0.943	0.903	88.90	7.09

6. Conclusions

In this paper, we mainly propose and investigate the mirror and glass detection problem using mathematical and physical methods for multimodal image enhancement. First, based on the three-point plane determination and the steady-state heat conduction equation for two-dimensional planes, we construct the plane equation and Laplace equation for the mirror surface in depth data and the glass surface in thermal data, completing the reconstruction of depth and thermal images. Next, we propose an RGB enhancement

module to further strengthen the mirror/glass surfaces in RGB images. Based on the raw RGB-D mirror and RGB-T glass datasets, we construct enhanced RGB-D mirror and enhanced RGB-T glass datasets, as well as ideal RGB-D mirror and ideal RGB-T glass datasets. Then, we introduce the prior fusion feature extraction module (PFFE) and deep fusion feature guidance module (DFFG) to extract low-level clue weights and high-level contextual weights from the input data, strengthening the model's global perception of image information. We also incorporate the Mamba module, which efficiently captures long-range dependencies, and propose the bidirectional state space fusion block (BSSF) to complete the fusion of different modality images. Based on this, we present the double weight Mamba fusion network (DWMFNet) for multimodal-based mirror/glass detection tasks. Extensive experiments show that our model achieves satisfactory results on the six datasets mentioned above. Moreover, nearly all models show varying degrees of performance improvement on the enhanced and ideal datasets.

Author Contributions: J.Q.: writing—original draft, visualization, validation, resources, project administration, methodology, formal analysis, data curation; C.J.: validation, writing—review and editing, methodology. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data will be made available by the authors on request.

Acknowledgments: The authors thank the anonymous reviewers and editors for their suggestions and insightful comments on this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Whelan, T.; Goesele, M.; Lovegrove, S.J.; Straub, J.; Green, S.; Szeliski, R.; Butterfield, S.; Verma, S.; Newcombe, R. Reconstructing Scenes with Mirror and Glass Surfaces. *ACM Trans. Graph.* **2018**, *37*, 102. [CrossRef]
2. Xie, E.; Wang, W.; Wang, W.; Ding, M.; Shen, C.; Luo, P. Segmenting Transparent Objects in the Wild. In *Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 696–711.
3. Wang, T.; He, X.; Barnes, N. Glass Object Localization by Joint Inference of Boundary and Depth. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012; pp. 3783–3786.
4. Nuanmeesri, S. Enhanced Hybrid Attention Deep Learning for Avocado Ripeness Classification on Resource Constrained Devices. *Sci. Rep.* **2025**, *15*, 3719. [CrossRef] [PubMed]
5. Nuanmeesri, S. Spectrum-Based Hybrid Deep Learning for Intact Prediction of Postharvest Avocado Ripeness. *IT Prof.* **2024**, *26*, 55–61. [CrossRef]
6. Nuanmeesri, S. The Affordable Virtual Learning Technology of Sea Salt Farming across Multigenerational Users through Improving Fitts' Law. *Sustainability* **2024**, *16*, 7864. [CrossRef]
7. Chan, K.-H.; Im, S.-K. Light-Field Image Super-Resolution with Depth Feature by Multiple-Decouple and Fusion Module. *Electron. Lett.* **2024**, *60*, e13019. [CrossRef]
8. Mayo, B.; Hazan, T.; Tal, A. Visual Navigation with Spatial Attention. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 16893–16902.
9. Liu, Z.; Zhang, Z.; Cao, Y.; Hu, H.; Tong, X. Group-Free 3D Object Detection via Transformers. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 2929–2938.
10. Lin, T.; Zhang, Y.; Zhu, Y. Improved Pseudo-Stereo for Monocular 3D Object Detection in Autonomous Driving. In Proceedings of the 2024 9th International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), Okinawa, Japan, 21–23 November 2024; Volume 9, pp. 100–104.
11. Sun, T.; Zhang, G.; Yang, W.; Xue, J.-H.; Wang, G. TROSD: A New RGB-D Dataset for Transparent and Reflective Object Segmentation in Practice. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 5721–5733. [CrossRef]
12. Liang, Y.; Deng, B.; Liu, W.; Qin, J.; He, S. Monocular Depth Estimation for Glass Walls With Context: A New Dataset and Method. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 15081–15097. [CrossRef] [PubMed]
13. Chang, Q.; Liao, H.; Meng, X.; Xu, S.; Cui, Y. PanoGlassNet: Glass Detection With Panoramic RGB and Intensity Images. *IEEE Trans. Instrum. Meas.* **2024**, *73*, 1–15. [CrossRef]

14. Hu, X.; Gao, R.; Yang, S.; Cho, K. CAGNet: A Multi-Scale Convolutional Attention Method for Glass Detection Based on Transformer. *Mathematics* **2023**, *11*, 4084. [CrossRef]
15. Li, F.; Ma, J.; Tian, Z.; Ge, J.; Liang, H.-N.; Zhang, Y.; Wen, T. Mirror-Yolo: A Novel Attention Focus, Instance Segmentation and Mirror Detection Model. In Proceedings of the 2022 7th International Conference on Frontiers of Signal Processing (ICFSP), Paris, France, 9–11 September 2022; pp. 76–80.
16. He, H.; Li, X.; Cheng, G.; Shi, J.; Tong, Y.; Meng, G.; Prinet, V.; Weng, L. Enhanced Boundary Learning for Glass-like Object Segmentation. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 15839–15848.
17. Xu, Y.; Nagahara, H.; Shimada, A.; Taniguchi, R. TransCut2: Transparent Object Segmentation From a Light-Field Image. *IEEE Trans. Comput. Imaging* **2019**, *5*, 465–477. [CrossRef]
18. Wang, T.; He, X.; Barnes, N. Glass Object Segmentation by Label Transfer on Joint Depth and Appearance Manifolds. In Proceedings of the 2013 IEEE International Conference on Image Processing, Melbourne, Australia, 15–18 September 2013; pp. 2944–2948.
19. Kalra, A.; Taamazyan, V.; Rao, S.K.; Venkataraman, K.; Raskar, R.; Kadambi, A. Deep Polarization Cues for Transparent Object Segmentation. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 8599–8608.
20. Wang, F.; Sheng, J.; Sfarra, S.; Zhou, Y.; Xu, L.; Liu, L.; Chen, M.; Yue, H.; Liu, J. Multimode Infrared Thermal-Wave Imaging in Non-Destructive Testing and Evaluation (NDT&E): Physical Principles, Modulated Waveform, and Excitation Heat Source. *Infrared Phys. Technol.* **2023**, *135*, 104993.
21. Mei, H.; Dong, B.; Dong, W.; Peers, P.; Yang, X.; Zhang, Q.; Wei, X. Depth-Aware Mirror Segmentation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 3043–3052.
22. Huo, D.; Wang, J.; Qian, Y.; Yang, Y.-H. Glass Segmentation With RGB-Thermal Image Pairs. *IEEE Trans. Image Process.* **2023**, *32*, 1911–1926. [CrossRef]
23. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv* **2024**, arXiv:2312.00752. [CrossRef]
24. Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Nießner, M.; Savva, M.; Song, S.; Zeng, A.; Zhang, Y. Matterport3D: Learning from RGB-D Data in Indoor Environments. *arXiv* **2017**, arXiv:1709.06158. [CrossRef]
25. Olson, E. AprilTag: A Robust and Flexible Visual Fiducial System. In Proceedings of the 2011 IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 3400–3407.
26. Yang, X.; Mei, H.; Xu, K.; Wei, X.; Yin, B.; Lau, R. Where Is My Mirror? In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8808–8817.
27. Lin, J.; Wang, G.; Lau, R.W.H. Progressive Mirror Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3694–3702.
28. Guan, H.; Lin, J.; Lau, R.W.H. Learning Semantic Associations for Mirror Detection. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 5931–5940.
29. He, R.; Lin, J.; Lau, R.W.H. Efficient Mirror Detection via Multi-Level Heterogeneous Learning. *arXiv* **2022**, arXiv:2211.15644. [CrossRef]
30. Huang, T.; Dong, B.; Lin, J.; Liu, X.; Lau, R.W.H.; Zuo, W. Symmetry-Aware Transformer-Based Mirror Detection. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, Washington, DC, USA, 7–14 February 2023*; AAAI'23/IAAI'23/EAAI'23; AAAI Press: Washington, DC, USA, 2023.
31. Tan, X.; Lin, J.; Xu, K.; Chen, P.; Ma, L.; Lau, R.W.H. Mirror Detection With the Visual Chirality Cue. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 3492–3504. [CrossRef] [PubMed]
32. Lin, J.; Tan, X.; Lau, R.W.H. Learning to Detect Mirrors from Videos via Dual Correspondences. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 9109–9118.
33. Xu, K.; Siu, T.W.; Lau, R.W. ZOOM: Learning Video Mirror Detection with Extremely-Weak Supervision. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 6315–6323.
34. McHenry, K.; Ponce, J. A Geodesic Active Contour Framework for Finding Glass. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 1, pp. 1038–1044.

35. Wei, H.; Li, X.; Shi, Y.; You, B.; Xu, Y. Fusing Sonars and LRF Data to Glass Detection for Robotics Navigation. In Proceedings of the 2018 IEEE International Conference on Robotics and Biomimetics (ROBIO), Kuala Lumpur, Malaysia, 12–15 December 2018; pp. 826–831.
36. Tibebu, H.; Roche, J.; De Silva, V.; Kondoz, A. LiDAR-Based Glass Detection for Improved Occupancy Grid Mapping. *Sensors* **2021**, *21*, 2263. [CrossRef] [PubMed]
37. Mei, H.; Yang, X.; Wang, Y.; Liu, Y.; He, S.; Zhang, Q.; Wei, X.; Lau, R.W.H. Don't Hit Me! Glass Detection in Real-World Scenes. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 3684–3693.
38. Lin, J.; He, Z.; Lau, R.W.H. Rich Context Aggregation with Reflection Prior for Glass Surface Detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13410–13419.
39. Zheng, C.; Shi, D.; Yan, X.; Liang, D.; Wei, M.; Yang, X.; Guo, Y.; Xie, H. GlassNet: Label Decoupling-Based Three-Stream Neural Network for Robust Image Glass Detection. *Comput. Graph. Forum* **2022**, *41*, 377–388. [CrossRef]
40. Fan, K.; Wang, C.; Wang, Y.; Wang, C.; Yi, R.; Ma, L. RFENet: Towards Reciprocal Feature Evolution for Glass Segmentation. *arXiv* **2023**, arXiv:2307.06099. [CrossRef]
41. Lin, J.; Yeung, Y.-H.; Lau, R.W.H. Exploiting Semantic Relations for Glass Surface Detection. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022*; NIPS '22; Curran Associates Inc.: Red Hook, NY, USA, 2024.
42. Qi, F.; Tan, X.; Zhang, Z.; Chen, M.; Xie, Y.; Ma, L. Glass Makes Blurs: Learning the Visual Blurriness for Glass Surface Detection. *IEEE Trans. Ind. Inform.* **2024**, *20*, 6631–6641. [CrossRef]
43. Lin, J.; Yeung, Y.-H.; Ye, S.; Lau, R.W.H. Leveraging RGB-D Data with Cross-Modal Context Mining for Glass Surface Detection. *arXiv* **2024**, arXiv:2206.11250. [CrossRef]
44. Mei, H.; Dong, B.; Dong, W.; Yang, J.; Baek, S.-H.; Heide, F.; Peers, P.; Wei, X.; Yang, X. Glass Segmentation Using Intensity and Spectral Polarization Cues. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 12612–12621.
45. Yan, T.; Xu, S.; Huang, H.; Li, H.; Tan, L.; Chang, X.; Lau, R.W.H. NRGlassNet: Glass Surface Detection from Visible and near-Infrared Image Pairs. *Knowl. Based Syst.* **2024**, *294*, 111722. [CrossRef]
46. Li, G.; Wang, Y.; Liu, Z.; Zhang, X.; Zeng, D. RGB-T Semantic Segmentation With Location, Activation, and Sharpening. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 1223–1235. [CrossRef]
47. Wang, K.; Lin, D.; Li, C.; Tu, Z.; Luo, B. Alignment-Free RGBT Salient Object Detection: Semantics-Guided Asymmetric Correlation Network and a Unified Benchmark. *IEEE Trans. Multimed.* **2024**, *26*, 10692–10707. [CrossRef]
48. Fan, D.-P.; Ji, G.-P.; Sun, G.; Cheng, M.-M.; Shen, J.; Shao, L. Camouflaged Object Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June; 2020; pp. 2774–2784.
49. Hazirbas, C.; Ma, L.; Domokos, C.; Cremers, D. FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture. In *Computer Vision—ACCV 2016*; Lai, S.-H., Lepetit, V., Nishino, K., Sato, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 213–228.
50. Ha, Q.; Watanabe, K.; Karasawa, T.; Ushiku, Y.; Harada, T. MFNet: Towards Real-Time Semantic Segmentation for Autonomous Vehicles with Multi-Spectral Scenes. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 5108–5115.
51. Sun, Y.; Zuo, W.; Liu, M. RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes. *IEEE Robot. Autom. Lett.* **2019**, *4*, 2576–2583. [CrossRef]
52. Chen, X.; Lin, K.-Y.; Wang, J.; Wu, W.; Qian, C.; Li, H.; Zeng, G. Bi-Directional Cross-Modality Feature Propagation with Separation-and-Aggregation Gate for RGB-D Semantic Segmentation. In *Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 561–577.
53. Zhou, W.; Yuan, J.; Lei, J.; Luo, T. TSNet: Three-Stream Self-Attention Network for RGB-D Indoor Semantic Segmentation. *IEEE Intell. Syst.* **2021**, *36*, 73–78. [CrossRef]
54. Lan, X.; Gu, X.; Gu, X. MMNet: Multi-Modal Multi-Stage Network for RGB-T Image Semantic Segmentation. *Appl. Intell.* **2022**, *52*, 5817–5829. [CrossRef]
55. Zhao, S.; Zhang, Q. A Feature Divide-and-Conquer Network for RGB-T Semantic Segmentation. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 2892–2905. [CrossRef]
56. Chen, Z.; Cong, R.; Xu, Q.; Huang, Q. DPANet: Depth Potentiality-Aware Gated Attention Network for RGB-D Salient Object Detection. *IEEE Trans. Image Process.* **2021**, *30*, 7012–7024. [CrossRef] [PubMed]
57. Tu, Z.; Xia, T.; Li, C.; Wang, X.; Ma, Y.; Tang, J. RGB-T Image Saliency Detection via Collaborative Graph Learning. *IEEE Trans. Multimed.* **2020**, *22*, 160–173. [CrossRef]

58. Song, K.; Wang, J.; Bao, Y.; Huang, L.; Yan, Y. A Novel Visible-Depth-Thermal Image Dataset of Salient Object Detection for Robotic Visual Perception. *IEEE/ASME Trans. Mechatron.* **2023**, *28*, 1558–1569. [CrossRef]
59. Qiu, J.; Jiang, C.; Wang, H. ETFormer: An Efficient Transformer Based on Multimodal Hybrid Fusion and Representation Learning for RGB-D-T Salient Object Detection. *IEEE Signal Process. Lett.* **2024**, *31*, 2930–2934. [CrossRef]
60. Wang, Q.; Yang, J.; Yu, X.; Wang, F.; Chen, P.; Zheng, F. Depth-Aided Camouflaged Object Detection. In *Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023*; MM '23; Association for Computing Machinery: New York, NY, USA, 2023; pp. 3297–3306.
61. Wang, L.; Yang, J.; Zhang, Y.; Wang, F.; Zheng, F. Depth-Aware Concealed Crop Detection in Dense Agricultural Scenes. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 17201–17211.
62. Liu, X.; Qi, L.; Song, Y.; Wen, Q. Depth Awakens: A Depth-Perceptual Attention Fusion Network for RGB-D Camouflaged Object Detection. *Image Vis. Comput.* **2024**, *143*, 104924. [CrossRef]
63. Song, S.; Lichtenberg, S.P.; Xiao, J. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 567–576.
64. Dai, A.; Chang, A.X.; Savva, M.; Halber, M.; Funkhouser, T.; Nießner, M. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2432–2443.
65. Armeni, I.; Sax, S.; Zamir, A.R.; Savarese, S. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *arXiv* **2017**, arXiv:1702.01105. [CrossRef]
66. Seib, V.; Barthen, A.; Marohn, P.; Paulus, D. Friend or Foe: Exploiting Sensor Failures for Transparent Object Localization and Classification. In Proceedings of the International Conference on Machine Vision, Moscow, Russia, 14–16 September 2016.
67. Sajjan, S.; Moore, M.; Pan, M.; Nagaraja, G.; Lee, J.; Zeng, A.; Song, S. Clear Grasp: 3D Shape Estimation of Transparent Objects for Manipulation. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 3634–3642.
68. Dinh, B.-D.; Nguyen, T.-T.; Tran, T.-T.; Pham, V.-T. 1M Parameters Are Enough? A Lightweight CNN-Based Model for Medical Image Segmentation. *arXiv* **2023**, arXiv:2306.16103. [CrossRef]
69. Hu, Q.; Guo, X. Single Image Reflection Separation via Component Synergy. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 13092–13101.
70. Gao, S.-H.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; Torr, P. Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 652–662. [CrossRef]
71. Huo, F.; Zhu, X.; Zhang, Q.; Liu, Z.; Yu, W. Real-Time One-Stream Semantic-Guided Refinement Network for RGB-Thermal Salient Object Detection. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–12. [CrossRef]
72. Song, K.; Wen, H.; Ji, Y.; Xue, X.; Huang, L.; Yan, Y.; Meng, Q. SIA: RGB-T Salient Object Detection Network with Salient-Illumination Awareness. *Opt. Lasers Eng.* **2024**, *172*, 107842. [CrossRef]
73. Li, Q.; Li, R.; Ji, K.; Dai, W. Kalman Filter and Its Application. In Proceedings of the 2015 8th International Conference on Intelligent Networks and Intelligent Systems (ICINIS), Tianjin, China, 1–3 November 2015; pp. 74–77.
74. Zhou, M.; Li, T.; Qiao, C.; Xie, D.; Wang, G.; Ruan, N.; Mei, L.; Yang, Y. DMM: Disparity-Guided Multispectral Mamba for Oriented Object Detection in Remote Sensing. *arXiv* **2024**, arXiv:2407.08132. [CrossRef]
75. Dong, W.; Zhu, H.; Lin, S.; Luo, X.; Shen, Y.; Liu, X.; Zhang, J.; Guo, G.; Zhang, B. Fusion-Mamba for Cross-Modality Object Detection. *arXiv* **2024**, arXiv:2404.09146. [CrossRef]
76. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. VMamba: Visual State Space Model. *arXiv* **2024**, arXiv:2401.10166. [CrossRef]
77. Zhou, W.; Wu, H.; Jiang, Q. MDNet: Mamba-Effective Diffusion-Distillation Network for RGB-Thermal Urban Dense Prediction. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *2024*, 1. [CrossRef]
78. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.
79. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
80. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks. In *Proceedings of the 29th International Conference on Neural Information Processing Systems—Volume 2, Montreal, QC, Canada, 7–15 December 2015*; NIPS'15; MIT Press: Cambridge, MA, USA, 2015; pp. 2017–2025.
81. Xie, S.; Tu, Z. Holistically-Nested Edge Detection. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1395–1403.
82. de Boer, P.-T.; Kroese, D.P.; Mannor, S.; Rubinstein, R.Y. A Tutorial on the Cross-Entropy Method. *Ann. Oper. Res.* **2005**, *134*, 19–67. [CrossRef]

83. Máttyus, G.; Luo, W.; Urtasun, R. DeepRoadMapper: Extracting Road Topology from Aerial Images. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 3458–3466.
84. Perazzi, F.; Krähenbühl, P.; Pritch, Y.; Hornung, A. Saliency Filters: Contrast Based Filtering for Salient Region Detection. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 733–740.
85. Nguyen, V.; Vicente, T.F.Y.; Zhao, M.; Hoai, M.; Samaras, D. Shadow Detection with Conditional Generative Adversarial Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 4520–4528.
86. Margolin, R.; Zelnik-Manor, L.; Tal, A. How to Evaluate Foreground Maps. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 248–255.
87. Cheng, M.-M.; Fan, D.-P. Structure-Measure: A New Way to Evaluate Foreground Maps. *Int. J. Comput. Vis.* **2021**, *129*, 2622–2638. [CrossRef]
88. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F.-F. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
89. Ruder, S. An Overview of Gradient Descent Optimization Algorithms. *arXiv* **2017**, arXiv:1609.04747. [CrossRef]
90. Jin, X.; Yi, K.; Xu, J. MoADNet: Mobile Asymmetric Dual-Stream Networks for Real-Time and Lightweight RGB-D Salient Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 7632–7645. [CrossRef]
91. Wang, Y.; Zhang, Y. Three-Stage Bidirectional Interaction Network for Efficient RGB-D Salient Object Detection. In Proceedings of the Asian Conference on Computer Vision (ACCV), Macao, China, 4–8 December 2022; pp. 3672–3689.
92. Liu, Z.; Tan, Y.; He, Q.; Xiao, Y. SwinNet: Swin Transformer Drives Edge-Aware RGB-D and RGB-T Salient Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 4486–4497. [CrossRef]
93. Chen, T.; Xiao, J.; Hu, X.; Zhang, G.; Wang, S. Adaptive Fusion Network for RGB-D Salient Object Detection. *Neurocomputing* **2023**, *522*, 152–164. [CrossRef]
94. Bi, H.; Wu, R.; Liu, Z.; Zhu, H.; Zhang, C.; Xiang, T.-Z. Cross-Modal Hierarchical Interaction Network for RGB-D Salient Object Detection. *Pattern Recognit.* **2023**, *136*, 109194. [CrossRef]
95. Wu, Z.; Allibert, G.; Meriaudeau, F.; Ma, C.; Demonceaux, C. HiDAnet: RGB-D Salient Object Detection via Hierarchical Depth Awareness. *IEEE Trans. Image Process.* **2023**, *32*, 2160–2173. [CrossRef] [PubMed]
96. Yan, Y.; Jia, X.; Song, K.; Cui, W.; Zhao, Y.; Liu, C.; Guo, J. Specificity Autocorrelation Integration Network for Surface Defect Detection of No-Service Rail. *Opt. Lasers Eng.* **2024**, *172*, 107862. [CrossRef]
97. Zhong, M.; Sun, J.; Ren, P.; Wang, F.; Sun, F. MAGNet: Multi-Scale Awareness and Global Fusion Network for RGB-D Salient Object Detection. *Knowl. Based Syst.* **2024**, *299*, 112126. [CrossRef]
98. Gao, H.; Su, Y.; Wang, F.; Li, H. Heterogeneous Fusion and Integrity Learning Network for RGB-D Salient Object Detection. *ACM Trans. Multimed. Comput. Commun. Appl.* **2024**, *20*, 1–24. [CrossRef]
99. Huo, F.; Zhu, X.; Zhang, L.; Liu, Q.; Shu, Y. Efficient Context-Guided Stacked Refinement Network for RGB-T Salient Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 3111–3124. [CrossRef]
100. Tu, Z.; Li, Z.; Li, C.; Tang, J. Weakly Alignment-Free RGBT Salient Object Detection With Deep Correlation Network. *IEEE Trans. Image Process.* **2022**, *31*, 3752–3764. [CrossRef] [PubMed]
101. Zhou, W.; Zhu, Y.; Lei, J.; Yang, R.; Yu, L. LSNet: Lightweight Spatial Boosting Network for Detecting Salient Objects in RGB-Thermal Images. *IEEE Trans. Image Process.* **2023**, *32*, 1329–1340. [CrossRef] [PubMed]
102. Bi, H.; Zhang, J.; Wu, R.; Tong, Y.; Fu, X.; Shao, K. RGB-T Salient Object Detection via Excavating and Enhancing CNN Features. *Appl. Intell.* **2023**, *53*, 25543–25561. [CrossRef]
103. Song, K.; Huang, L.; Gong, A.; Yan, Y. Multiple Graph Affinity Interactive Network and a Variable Illumination Dataset for RGBT Image Salient Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 3104–3118. [CrossRef]
104. Jiang, X.; Hou, Y.; Tian, H.; Zhu, L. Mirror Complementary Transformer Network for RGB-thermal Salient Object Detection. *IET Comput. Vis.* **2023**, *18*, 15–32. [CrossRef]
105. Wang, K.; Tu, Z.; Li, C.; Zhang, C.; Luo, B. Learning Adaptive Fusion Bank for Multi-Modal Salient Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 7344–7358. [CrossRef]
106. Tang, H.; Li, Z.; Zhang, D.; He, S.; Tang, J. Divide-and-Conquer: Confluent Triple-Flow Network for RGB-T Salient Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *47*, 1958–1974. [CrossRef]