mathematics

MDPI

Article

CAGNet: A Multi-Scale Convolutional Attention Method for Glass Detection Based on Transformer

Xiaohang Hu [1], Rui Gao [1], Seungjun Yang [2] and Kyungeun Cho [3,*]

[1] Department of Multimedia Engineering, Dongguk University, 30, Pildongro-1-gil, Jung-gu, Seoul 04620, Republic of Korea; 2020126642@dgu.ac.kr (X.H.); gaorui@dongguk.edu (R.G.)
[2] Electronics and Telecommunications Research Institute, 218 Gajeong-ro, Yuseong-gu, Daejeon 34129, Republic of Korea; sjyang@etri.re.kr
[3] Division of AI Software Convergence, Dongguk University, 30, Pildongro-1-gil, Jung-gu, Seoul 04620, Republic of Korea
* Correspondence: cke@dongguk.edu

Abstract: Glass plays a vital role in several fields, making its accurate detection crucial. Proper detection prevents misjudgments, reduces noise from reflections, and ensures optimal performance in other computer vision tasks. However, the prevalent usage of glass in daily applications poses unique challenges for computer vision. This study introduces a novel convolutional attention glass segmentation network (CAGNet) predicated on a transformer architecture customized for image glass detection. Based on the foundation of our prior study, CAGNet minimizes the number of training cycles and iterations, resulting in enhanced performance and efficiency. CAGNet is built upon the strategic design and integration of two types of convolutional attention mechanisms coupled with a transformer head applied for comprehensive feature analysis and fusion. To further augment segmentation precision, the network incorporates a custom edge-weighting scheme to optimize glass detection within images. Comparative studies and rigorous testing demonstrate that CAGNet outperforms several leading methodologies in glass detection, exhibiting robustness across a diverse range of conditions. Specifically, the IOU metric improves by 0.26% compared to that in our previous study and presents a 0.92% enhancement over those of other state-of-the-art methods.

Keywords: convolutional attention; transformer; feature analyze; semantic segmentation

MSC: 68T07; 68U10; 68T45

Citation: 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. https://doi.org/10.3390/math11194084

Academic Editors: Konstantin Kozlov and Jakub Nalepa

Received: 21 August 2023
Revised: 15 September 2023
Accepted: 25 September 2023
Published: 26 September 2023

1. Introduction

Glass is ubiquitous in various fields, such as architecture, appliances, decorations, and furniture, providing functional and aesthetic benefits. However, its prevalent usage poses unique challenges for computer vision. The inherent properties of glass, such as reflection, refraction, and transparency, significantly increase the complexity of achieving high-precision detection, with the diversity of glass types, from flat and non-double-sided frosted glass to colored decorative window glass, further complicating this problem. Each type of glass displays unique characteristics in different environments and changes its appearance according to the surrounding conditions and lighting, thereby increasing the difficulty of detection.

Accurate glass detection is pivotal in diverse applications. In autonomous vehicles and robotics, correctly identifying glass objects can prevent misjudgments that cause collisions. For three-dimensional (3D) point-cloud reconstruction, identifying and preprocessing the glass locations in two-dimensional (2D) images can significantly minimize the noise caused by glass reflection and refraction, thereby enhancing the overall quality of the reconstructed scenes [1,2]. Additionally, advancements in computer vision have been evident in areas such as semantic segmentation [3–6], object detection [7–9], and image

recognition [8,10], with notable contributions from attention mechanisms [3], pyramid structures [5,7], and transformer architectures [8], depth estimation algorithms [11,12], and salient object detection methods [13–15].

Despite significant progress in collecting vast image data containing glass and glass-like objects [16], detecting glass from a single 2D RGB image remains a formidable challenge. Traditional methods typically rely on depth maps (RGB-D) [17], thermal maps (RGB-T) [18], and polarization maps (RGB-P) [19] to identify glass in images. While utilizing additional data indeed aids models in learning more regarding glass features for detection, such data are often hard to come by in many scenarios. Furthermore, when only a single RGB image is available, these methods often underperform and fail to accurately pinpoint the location of the glass. The imperative of using a single RGB image for glass segmentation is primarily rooted in its practicality and ubiquity. In several real-world applications, such as drones, mobile robots, or some low-cost surveillance systems, only standard RGB cameras might be equipped. This makes acquiring data from additional sensors challenging, and researching algorithms for single RGB images can significantly reduce data acquisition costs and enhance its universality.

To address these issues, a convolutional attention glass segmentation network (CAGNet) based on a transformer structure has been proposed. This study builds upon our previously published work in [20], optimizes its structure, designs two novel convolutional attention mechanisms, and flexibly uses the feature integration and analysis capabilities of the transformer head to play key roles in various modules. The proposed design achieved good efficiency and performance for glass detection in single RGB images, thereby improving the robustness and accuracy of the task.

This study makes the following key contributions:

- A backbone feature analysis (BFA) module was engineered to amalgamate feature information from different layers of the backbone network. Furthermore, a global context (GC) module was incorporated to substantially mitigate the problem of overfitting;
- A multi-level convolutional attention module (MCA), which includes the custom-designed multi-scale convolutional attention module and a convolutional self-attention (CSA) module, was developed. In combination with the transformer head, these methods provide comprehensive feature analysis and fusion, offering an enhanced solution for glass detection in single RGB images;
- A cross-modal feature-analysis fusion module (CFAF) was constructed to fuse and analyze the features, with a custom edge-weighting scheme incorporated into the network. This innovation improves the accuracy of the segmentation process, thereby enabling more refined image-based glass detection.

2. Related Works

Conventional glass detection tasks primarily involve segmenting glass regions in images using semantic segmentation methods. This process entails distinguishing the glass and background as distinct labels, which enables the model to learn the glass features for detection purposes.

Semantic segmentation associates each pixel in an image with a specific category, thus forming the foundation for glass detection; this subfield labels the glass regions and backgrounds in the images. Initial advancements were achieved using fully convolutional networks, which introduced the techniques of feature map fusion and stitching [21,22]. Badrinarayanan et al. [23] further advanced this technology by proposing an optimized encoder-decoder structure utilizing a maximum pooling index for up-sampling. Chen et al. [24–26] enhanced this method by combining a null-pyramid pooling method with the pyramid pooling module of Zhao et al. [5], expanding the receptive field and refining the boundaries through dilated convolution. He et al. [4] conducted binary segmentation on a fast R-CNN [27] and introduced bilinear interpolation for feature up-sampling, which improved the segmentation accuracy.

Mei et al. [16] contributed significantly to glass detection by introducing the glass detection dataset (GDD) and a novel method for extracting and fusing features from different layers of a backbone network. Cao et al. [28] and Hao et al. [29] focused on enhancing boundary discrimination and modeling global shape boundaries.

Recently, transformers have gained widespread attention in the field of computer vision, offering innovative methods for semantic segmentation and glass detection tasks. Wang et al. [7,30] designed a transformer-based backbone characterized by a pyramidal structure of attention layers to optimize the detection of multi-scale features while conserving computational resources. Similarly, Guo et al. [31] designed a convolutional attention module that significantly reduced the requirement for computational resources while maintaining high performance compared with self-attention methods.

In glass detection, Xie et al. [32] provided a transparent object dataset containing various types of glasses and designed an encoder-decoder network based on the Vision Transformer (VIT) network [10]. This configuration provides a global receptive field and effectively classifies the glass regions. Zhang et al. [33] further improved transformer-based methods by proposing a deeper encoder-decoder network combined with a small transformer head to prevent overfitting and enhance the performance of glass detection tasks.

In summary, the CNN and transformer techniques have significantly advanced semantic segmentation and glass detection. CNNs excel in feature extraction and fusion, but they may not always capture long-range dependencies owing to their local receptive fields. By contrast, transformers offer innovative solutions for feature analysis and fusion but can be computationally intensive, particularly when applied to spatial data such as images. The proposed CAGNet synergistically integrates the strengths of both CNNs and Transformers, effectively addressing their individual limitations. A key innovation in our approach is the introduction of the multi-receptive field convolutional attention module. Furthermore, we designed the convolutional self-attention (CSA) mechanism. Unlike traditional transformers in NLP tasks that employ linear layers to compute attention, our model leverages the inherent advantages of convolutions in 2D image processing. Specifically, we utilize convolutions to derive the query, key, and value matrices for attention computation. This convolutional approach to attention not only enhances computational efficiency but also ensures a more spatially coherent representation, which is crucial for tasks such as glass detection in RGB images. Consequently, CAGNet emerges as a robust and efficient model, outperforming in the intricate task of detecting glass from single RGB images.

3. Proposed Method

This section introduces CAGNet, which draws upon and refines the ideas of advanced glass detection and semantic segmentation methods. We opted for ResNeXt-101 as the backbone network owing to its capability to provide multiple feature outputs across different scales, which is paramount for our approach that capitalizes on multi-scale features for glass detection. Furthermore, a plethora of state-of-the-art methodologies employ ResNeXt-101 as their foundational network. Utilizing the same backbone ensures a more equitable quantitative and qualitative assessment when juxtaposing our method with these advanced techniques. Initially, images pass through the ResNeXt-101 backbone network to produce four layers of backbone features that are input to the BFA from a previous method [20]. A GC layer is introduced on top for further feature integration and analysis to obtain better multi-scale feature information. Subsequently, an MCA was designed to learn glass-feature information across small and large receptive fields. Finally, a CFAF was constructed for feature fusion and boundary analysis. This module is used for the cross-modal analysis of boundaries and segmentation and fusion of features in various receptive fields. The final label classification for glass detection was accomplished using classification convolution.

Figure 1 illustrates the overall network architecture, comprising the following four components: the backbone, BFA, MCA, and CFAF. ResNeXt-101 served as the backbone, outputting four distinct feature sizes to BFA for backbone-feature context analysis and fusion, as described in Section 3.1. After fusion, the backbone features were transferred to MCA for multi-level receptive field feature analysis, as described in Section 3.2. Subsequently, the features from various receptive field levels were input to CFAF for cross-modal feature analysis fusion, as described in Section 3.3.

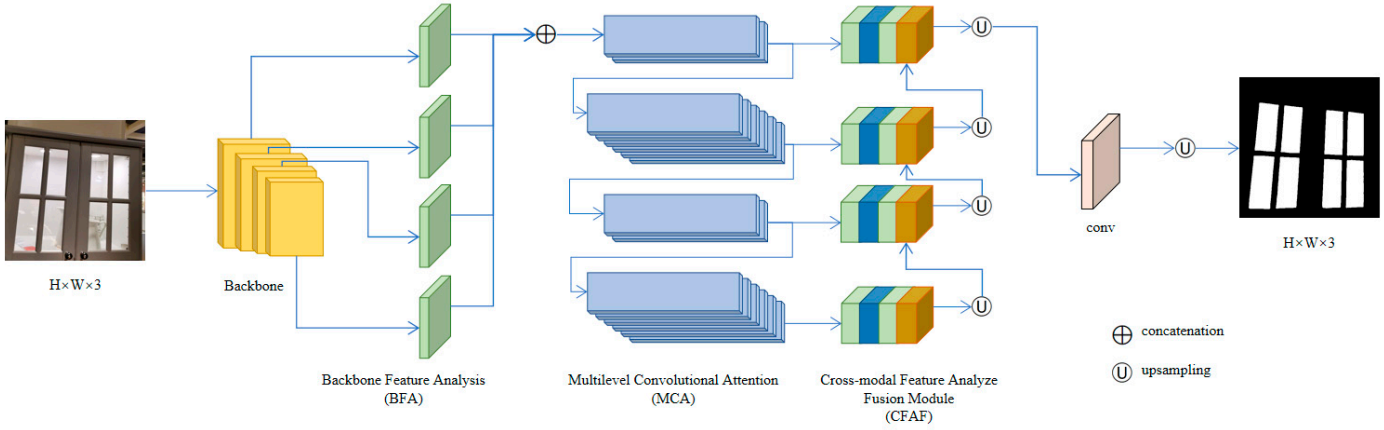

Figure 1. Overall framework structure of CAGNet.

3.1. BFA

In contrast to methods employing convolution for backbone feature extraction or directly feeding features into the network that extinguishes contextual relationships among features, we devised a module for backbone feature analysis and fusion called BFA, as shown in Figure 2. This module accepts the following four different feature sizes output by the backbone network: $(\frac{H}{4}, \frac{W}{4}, 256)$, $(\frac{H}{8}, \frac{W}{8}, 512)$, $(\frac{H}{16}, \frac{W}{16}, 1024)$, and $(\frac{H}{32}, \frac{W}{32}, 2048)$. These features are weighted toward the target area through the multi-head self-attention mechanism of a traditional transformer, thereby focusing on the target feature information [34,35]. Each layer corresponded to the following four output feature sizes: $(\frac{H}{4}, \frac{W}{4}, 256)$, $(\frac{H}{8}, \frac{W}{8}, 512)$, $(\frac{H}{16}, \frac{W}{16}, 1024)$, and $(\frac{H}{32}, \frac{W}{32}, 2048)$. These features were up-sampled and fused through the corresponding GC layer.

The GC layer comprised an adaptive average pooling layer (AAPL) and a fully connected layer (FCL). AAPL adaptively reduces the spatial dimensions of the input feature map to (1, 1), thereby reducing the feature map to a single value. FCL maintains the same number of channels while fusing global context information features [36]. Applying the GC layer in neural networks enhances the model's comprehension of the overall image and utilization of context information [37]. The principle of the GC layer is represented as follows:

$$f_n = conv_{1\times1}(AdaptiveAvgPool2d(F_n)) \quad (1)$$

where "F" represents the output feature of each layer, "n" denotes the corresponding layer, and "f" is the feature processed through GC.

Consequently, BFA concatenates the processed features from the four layers to yield a fused backbone feature.

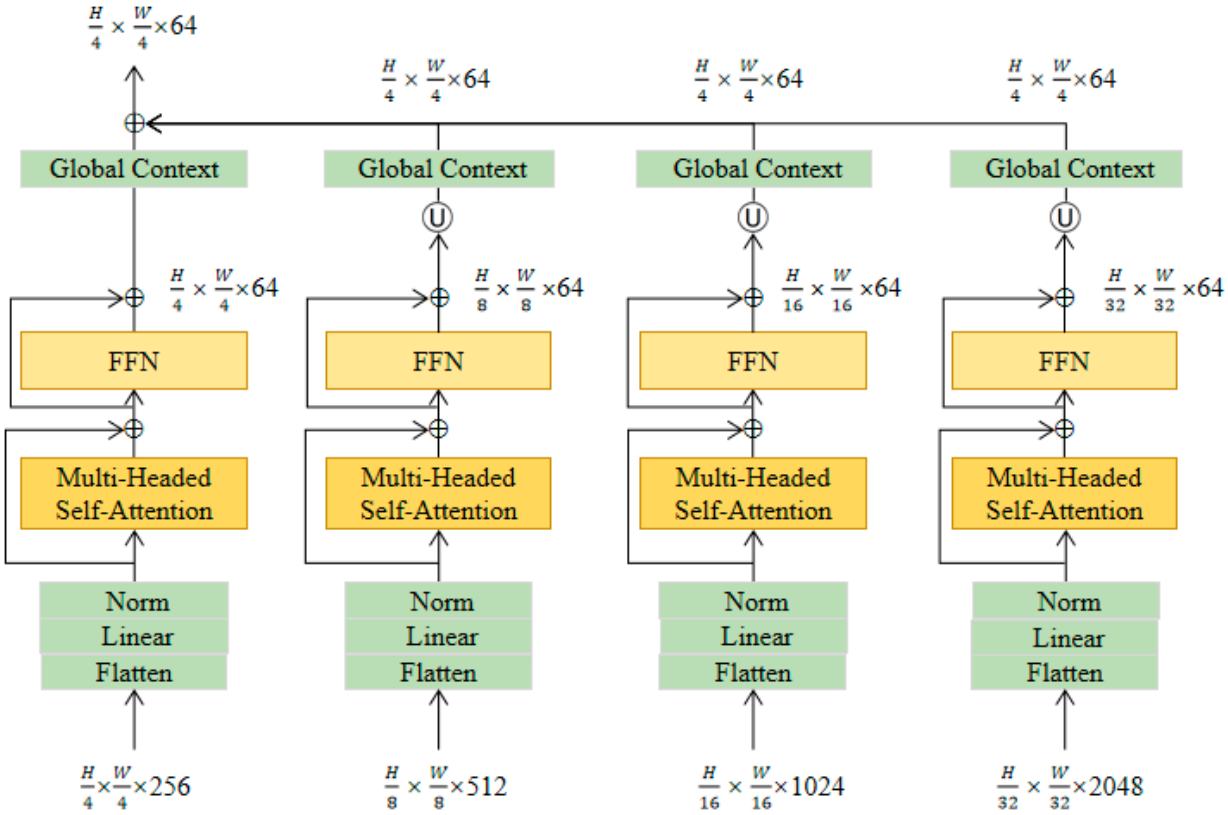

Figure 2. BFA module structure.

3.2. MCA

The features derived from a backbone network often contain vast information that is not always accurate. For instance, given that most glasses are shaped as bars or squares, the network might mistakenly classify objects with square frames, such as door frames or square bookshelves, as target objects. To further analyze and refine these features, the fused features from BFA were fed into the MCA module for multi-level receptive field feature analysis, as shown in Figure 3. This module comprises the following four receptive fields: ($\frac{H}{4}$, $\frac{W}{4}$, 64), ($\frac{H}{8}$, $\frac{W}{8}$, 128), ($\frac{H}{16}$, $\frac{W}{16}$, 256), and ($\frac{H}{32}$, $\frac{W}{32}$, 512), where the three dimensions represent the height, width, and channel number. This design facilitates the model's ability to learn more detailed features in smaller receptive fields and overall glass features in larger receptive fields. Each feature level is progressively transferred, thereby producing its results independently.

To design the MCA module, two convolutional attention mechanisms were incorporated owing to the inherent differences between text and image data [38]. The original transformer attention mechanism was designed for textual data; however, the task involved image processing. Therefore, the first mechanism employs multi-receptive-field convolutional attention as an alternative to the traditional transformer attention mechanism, rendering it more suitable for image data [31]. Instead of linear layers, the second approach utilizes 2D convolutional layers to compute the attention query (Q), key (K), and value (V) values, known as convolutional self-attention (CSA), which tailors the attention computation process to the nuances of image data.

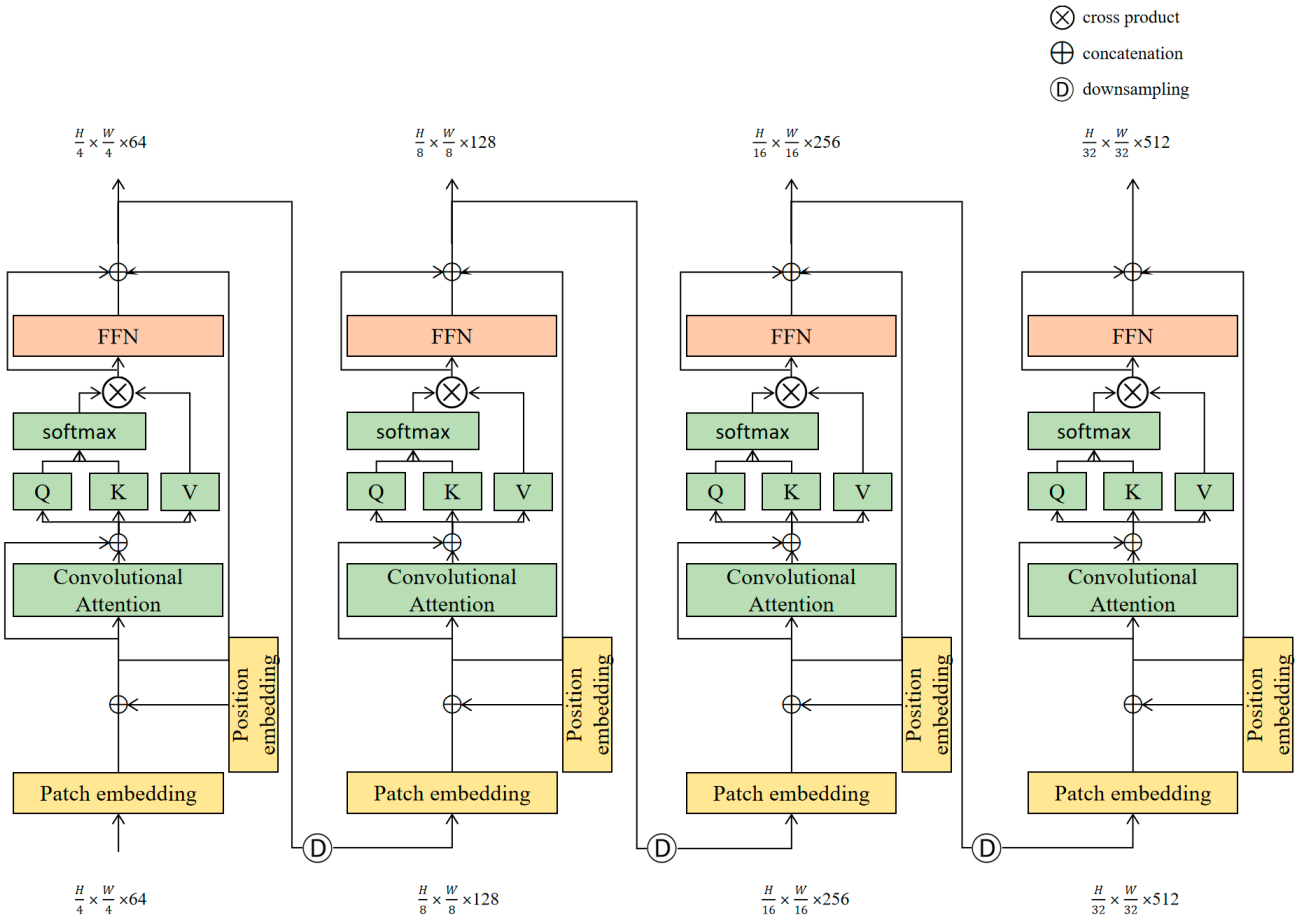

Figure 3. Multi-level convolutional attention module structure.

Figure 4 shows that within the multi-receptive-field convolutional attention module, the proposed design diverges from that in [20]. The convolutional kernels were adjusted to match the nature of the target objects. Initially, a cross-channel linear combination of the input feature information was performed using a 1×1 convolution, which linearly transformed various channels by integrating the information between them. Subsequently, as opposed to a 5×5 convolution, two stacked 3×3 convolutions were used to reduce the number of required parameters. The two sets of convolutional attention sizes, i.e., $(1 \times 7, 7 \times 1)$ and $(3 \times 7, 7 \times 3)$, were incorporated. Band-shaped convolutions can analyze elongated objects and enhance the expressive capability of a network [39]. This design enriches the multi-scale feature information and ensures a broader receptive field, rendering it more attuned to the specific challenges of our task. This process can be expressed as follows:

$$atten = ReLU(Conv_{3\times3}(ReLU(Conv_{3\times3}(F)))) \tag{2}$$

$$atten_1 = Conv_{7\times1}(Conv_{1\times7}(atten)) \tag{3}$$

$$atten_2 = Conv_{7\times3}(Conv_{3\times7}(atten)) \tag{4}$$

where "F" represents the feature obtained after a nonlinear transformation using a $Conv_{1\times1}$; "$atten$" denotes the feature derived after applying a weight to the convolutional feature matrix of size $(3 \times 3, 3 \times 3)$, which subsequently undergoes a nonlinear transformation via the ReLU activation function; "$atten_1$" and "$atten_2$" represent the feature obtained after weighting the convolutional feature matrices of sizes $(1 \times 7, 7 \times 1)$ and $(3 \times 7, 7 \times 3)$, respectively.

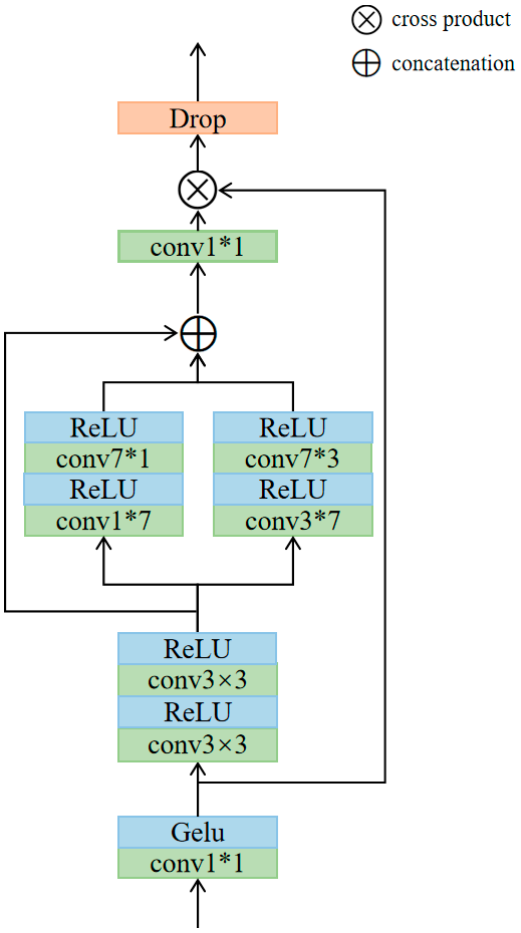

Figure 4. Multi-receptive field convolutional attention module structure.

Subsequently, we followed the approach outlined in reference [31], which involves combining contextual features obtained through additive merging attention mechanisms to derive composite features referred to as "*Attention*". "*Attention*" is multiplied elementwise with the original input feature matrix "*F*", thereby introducing a residual connection. Residual connections alleviate the vanishing gradient problem and enhance the information flow between layers, thereby promoting effective feature reusability and propagation [40]. Furthermore, they improve a network's ability to learn residual mappings and contribute to stable network convergence [41]. The processed features are passed through a dropout layer, yielding the final feature output f as follows:

$$Attention = atten + atten_1 + atten_2 \tag{5}$$

$$f = Drop(F \times Attention) \tag{6}$$

This method minimized the number of parameters and ensured the accuracy of the model. Thereafter, the resulting feature output was fed into the CSA module for further attention analysis. Initially, three 1×1 convolutions were used to obtain the query, key, and value matrices. Following the conventional transformer self-attention mechanism, the dot product of the query and key is computed, and the softmax function is applied to derive the attention matrix $Atten_{conv}$. This matrix is weighted and summed using a value matrix. Subsequently, a 1×1 convolution was used for feature consolidation. After traversing the dropout layer, the processed features are combined with the original output via a skip connection, yielding the final feature representation. The detailed process is as follows:

$$query = Conv_{1\times1}(f) \tag{7}$$

$$key = Conv_{1\times1}(f) \tag{8}$$

$$value = Conv_{1\times1}(f) \tag{9}$$

$$Atten_{conv} = softmax\left(query \times value^{T}\right) \tag{10}$$

$$f_{atten} = F + Drop(conv_{1\times1}(Atten_{conv} \times value)) \tag{11}$$

where "$query$", "key", and "$value$" are the matrices for query, key, and value, respectively, obtained through convolution; "$Atten_{conv}$" represents the attention matrix; "F" denotes the original input; and "f_{atten}" is the final feature obtained.

3.3. CFAF

Figure 5 shows the CFAF module structure, encompassing the four feature analyses and fusion units. Each unit comprises the following four components: a transformer head (TH), cross-modal atrous spatial pyramid pooling module (C-ASPP) [20], transformer conversion head (TCH) [20], and feature fusion block (FFB). Given that the MCA progressively integrates and independently outputs features from each receptive field, features from different levels might exhibit discrepancies, as highlighted in references [42,43]. Directly processing these features could potentially limit the model's performance. To address this, a TH was designed to consolidate these features and mitigate the impact of such discrepancies on the model's efficacy [34,35]. This design aimed to enhance the model's performance by effectively integrating and leveraging features at different levels. C-ASPP performed cross-modal bifurcation of the features into boundary and segmentation features. TCH, equipped with two THs, further weighs and analyzes the boundary and segmentation features. FFB fuses features of the boundary, segmentation, and preceding layer. It weighs and analyzes the fused features again. CFAF accepts four features from the MCA output with dimensions ($\frac{H}{4}$, $\frac{W}{4}$, 64), ($\frac{H}{8}$, $\frac{W}{8}$, 128), ($\frac{H}{16}$, $\frac{W}{16}$, 256), and ($\frac{H}{32}$, $\frac{W}{32}$, 512). Consequently, it outputs a feature of size ($\frac{H}{4}$, $\frac{W}{4}$, 64), which is fed into the classification convolution layer for categorization and up-sampling to obtain the final detection image result.

For CFAF, we adopted the C-ASPP and TCH modules from our previous study [20], which were extensively detailed and experimentally validated. In this study, the final fusion part, i.e., the FFB module, was adjusted, and its structure was optimized to reduce complexity while enhancing computational efficiency.

FFB aimed to establish a multi-level receptive field context association by fusing boundary features with segmentation and the previous layer's features at each level. Before each fusion, the features from the preceding layer must be up-sampled, which may introduce feature loss. To address this, a depth-wise separable convolution was performed for feature analysis post feature-fusion. Such convolution has fewer parameters than traditional convolution operations, offers faster computation, reduces the risk of overfitting, and effectively enhances the generalization capability of the model [44]. As shown in Figure 6, it comprises a depth-wise separable convolution layer and TH. The segmented and boundary features were fused and combined with the previous layer's features. The resulting features were passed to SeparableConv2d to achieve context-related feature fusion across multiple receptive fields. This process can be represented as follows:

$$F = SeparableConv2D(f_s^i + \left(f_s^i \otimes f_b^i\right)) \tag{12}$$

where "F" represents the fused features, "f_b^i" denotes the boundary features, "f_s^i" stands for the segmentation features, and "i" indicates the layer number.

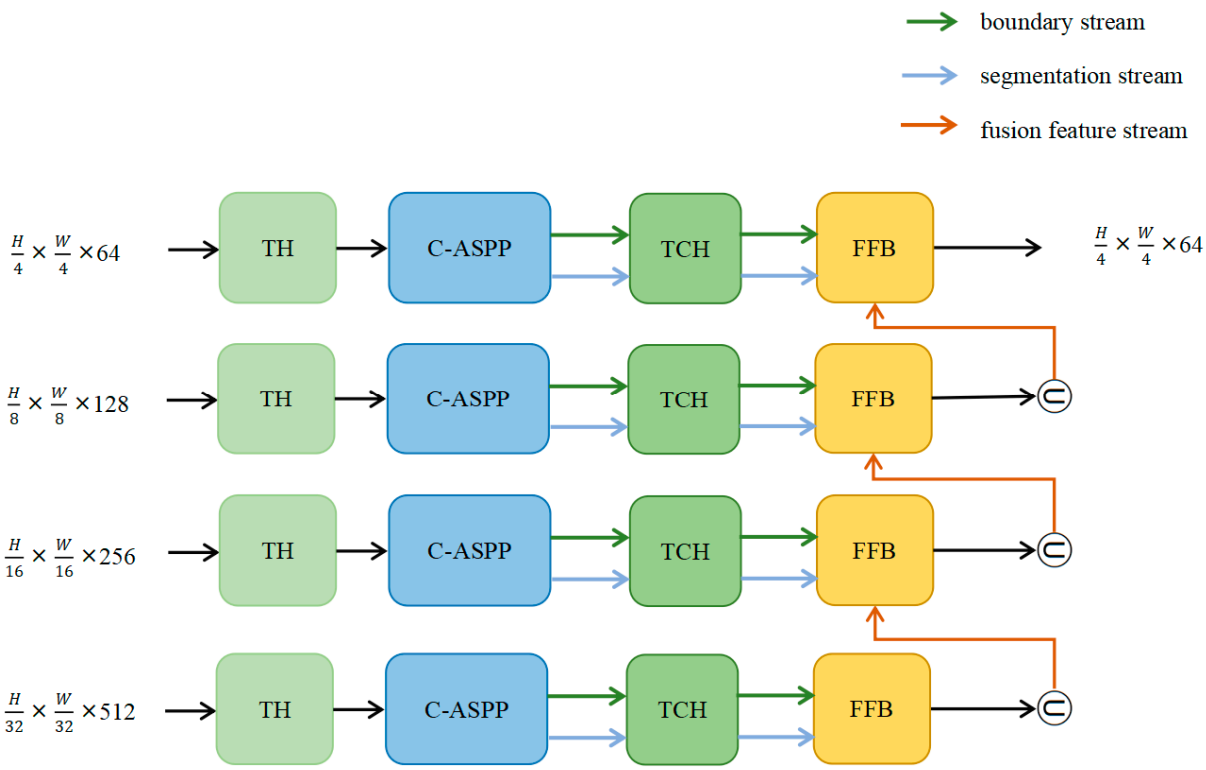

Figure 5. CFAF module structure.

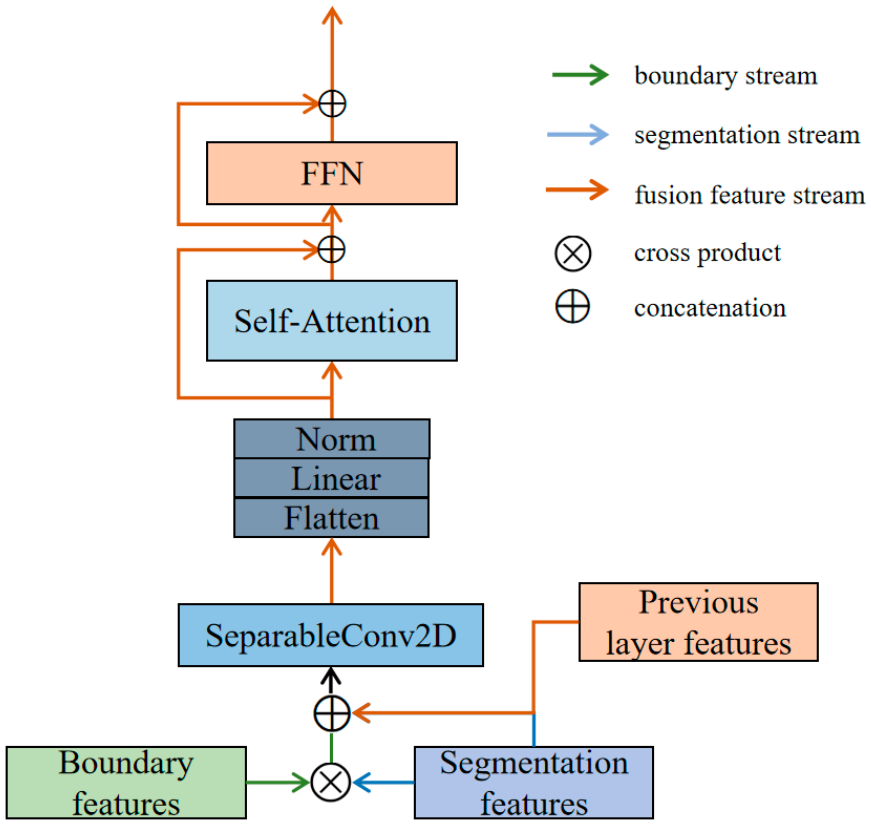

Figure 6. Feature fusion block structure.

The TH described herein shares the same structure as the TH and TCH modules. It applies feature weighting to fused features, thereby minimizing the influence of irrelevant information [34,35]. Subsequently, the features were passed through a feed-forward network (FFN) layer for nonlinear transformation, ensuring that the feature dimensions were aligned with the next layer. This approach enhances the model's expressive capacity and mitigates the risk of overfitting.

3.4. Loss Function

In the proposed network architecture, we primarily employed L1, segmentation, and boundary losses expressed as follows:

$$L_{L1} = L1_{lambda} \times L1_{norm} \tag{13}$$

$$L_{seg} = 1 - \frac{2|P \cap G| + smooth}{|P| + |G| + smooth} \tag{14}$$

$$L_{boundary} = \left(1 - \frac{2|P_b \cap G_b| + smooth}{|P_b| + |G_b| + smooth}\right) \times W \tag{15}$$

$$Loss = L_{seg} + L_{boundary} + L_{L1} \tag{16}$$

where "$L1_{lambda}$" is a hyperparameter set to 0.1 and modulates the weight of $L1$ loss, "$L1_{norm}$" is the summation of the absolute values of all model parameters, "G" represents the ground truth, "G_b" indicates the ground truth for boundaries, "P" represents the prediction result, "P_b" stands for the predicted boundary result, and "W" is a weighting parameter for the boundary loss set to 2.0. The term "smooth" is introduced (assigned a value of 1) to prevent the denominator from vanishing.

This amalgamation of losses ensures the model's proficiency in accurate glass detection, segmentation, and boundary delineation throughout the training phase.

4. Experiment

4.1. Dataset and Settings

4.1.1. Dataset Details

The GDD [16], comprising 2827 indoor and 1089 outdoor images, is a specialized collection tailored for glass detection, encompassing a myriad of everyday scenarios. Adhering to the same data partitioning as GDD, 2980 images were allocated for training, while the remaining 936 images were reserved for testing.

4.1.2. Implementation Details

The proposed network model was constructed using PyTorch 1.8.0, complemented by CUDA 11.3. The training infrastructure employed three RTX A6000 GPUs. The learning rate was initialized to 1×10^{-4}, which subsequently decayed following the poly strategy [45]. The backbone, ResNeXt-101 [46], was pretrained on ImageNet [47]. After training, the learning rate decayed linearly to 1×10^{-6}. The optimizer of choice was AdamW, with ε set to 1×10^{-8} and a weight decay of 1×10^{-4}. The batch size per GPU was configured as eight. After 200 epochs in the GDD dataset [16], convergence was achieved with an average duration of 14 h. The experiments were conducted with an input resolution of 512×512 pixels. Additionally, the dataset's ground truth was binarized and converted into a single-channel image. This approach accentuates the learning focus on the target object labels during training. No data augmentation, online hard-example mining (OHEM), or similar techniques were employed in the experimental tests to ensure a level playing field.

4.2. Evaluation Metrics

During evaluation, the quartet of semantic segmentation metrics discussed in [48] were implemented to rigorously assess the efficacy of glass detection based on the following metrics: (1) Intersection over union (IoU): it calculates the ratio of the area of overlap between the predicted and ground truth regions to that of their union, robustly measuring the congruence between the predicted segmentation and actual ground truth; (2) F-measure (Fβ) [49]: defined as the harmonic mean of precision and recall, it comprehensively assesses the model's precision (correct positive predictions) against its recall (true positive rate); (3) Mean absolute error (MAE): it quantifies the average absolute differences between the predicted and actual values, clearly indicating the deviation of the model from the ground truth; (4) Balanced error rate (BER) [50]: it computes the average error rates across classes, ensuring that the error of each class is weighted equally, thereby providing a balanced view of the model's performance across diverse classes. These metrics can be expressed as follows:

$$IoU = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W}(G(i,j) * P(i,j))}{\sum_{i=1}^{H}\sum_{j=1}^{W}(G(i,j) + P(i,j) - G(i,j) * P(i,j))} \tag{17}$$

$$F_{\beta}^{w} = (1+\beta^2)\frac{Precision^{w} \cdot Recall^{w}}{\beta^2 \cdot Precision^{w} + Recall^{w}} \tag{18}$$

$$MAE = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}|P(i,j) - G(i,j)| \tag{19}$$

$$BER = 100 \times (1 - \frac{1}{2}(\frac{TP}{N_P} + \frac{TN}{N_n})) \tag{20}$$

4.3. Comparison with Existing Methods

During evaluation, the performance of TGSNet was benchmarked against 14 algorithms associated with our study area. These methods comprise semantic segmentation approaches, such as PSPNet [5], DANet [3], CCNet [51], and FaPN [52]; salient object detection techniques, including EGNet [53] and F3Net [54]; transparent object segmentation methods, namely Trans2Seg [32] and Trans4Trans [33]; the mirror segmentation method, MirrorNet [55], and glass segmentation techniques, including GDNet [16], GSD [56], EBLNet [29], and PGSNet [48]. Additionally, TGDNet from our previous study was considered [20]. To ensure a fair comparison, each method was trained on the GDD dataset without data augmentation. For methods using publicly accessible codes, the recommended optimal parameters were followed. For those without an available code, we relied on the values provided in [48]. All evaluations were performed using the same protocol.

Table 1 comparatively evaluates CAGNet against other prominent methodologies using the GDD dataset. CAGNet evidently exhibits a superior IoU score, which is a crucial metric in semantic segmentation that quantifies the congruence between the predicted and ground truth segmentations. The IoU score of CAGNet is superior to that of TGSNet by a margin of 0.26%, and its F_{β}^{w} is augmented by 0.04%. When benchmarked against leading-edge glass detection paradigms, such as PGSNet [48], the IoU score of our proposed method exhibits an enhancement of 0.92%, and its F_{β}^{w} surpasses PGSNet [48] by 1.1%. These metrics highlight the nuanced capabilities of CAGNet in discerning intricate glass patterns, validating its robustness and precision in various scenarios.

Table 1. Quantitative comparison of the proposed network, CAGNet, with semantic segmentation algorithms, salient object detection algorithms, transparent object segmentation algorithms, mirror segmentation algorithms, and glass detection algorithms on the GDD dataset. The best and second-best results are highlighted in red and blue, respectively. "↑" indicates that higher values are better, "↓" indicates that lower values are better.

Method	Published Journals	Backbone	GDD [16]			
			IoU↑	F_β^w↑	MAE↓	BER↓
PSPNet [5]	CVPR'17	ResNet-50	84.06	0.867	0.084	8.79
DANet [3]	CVPR'19	ResNet-50	84.15	0.864	0.089	8.96
CCNet [51]	ICCV'19	ResNet-50	84.29	0.867	0.085	8.63
FaPN [52]	ICCV'21	ResNet-101	86.65	0.887	0.062	5.69
EGNet [53]	ICCV'19	ResNet-50	85.05	0.870	0.083	7.43
F3Net [54]	AAAI'20	ResNet-50	84.79	0.870	0.082	7.38
MirrorNet [55]	ICCV'19	ResNeXt-101	85.07	0.866	0.083	7.67
Trans2seg [32]	IJCAI'21	ResNet-50	84.41	0.872	0.078	7.36
Trans4Trans [33]	ICCVW'21	PVT-Medium	84.94	0.878	0.076	6.86
GSD [56]	CVPR'21	ResNeXt-101	87.53	0.895	0.066	5.90
GDNet [16]	CVPR'20	ResNeXt-101	87.63	0.898	0.063	5.62
EBLNet [29]	CVPR'21	ResNeXt-101	84.98	0.879	0.076	7.24
PGSNet [48]	TIP'22	ResNeXt-101	87.81	0.901	0.062	5.56
TGSNet [20]	MDPI	ResNeXt-101	88.47	0.908	0.058	5.70
CAGNet (our)	\	ResNeXt-101	88.73	0.913	0.054	5.51

As shown in Figure 7, a qualitative comparison was performed between TGSNet and six state-of-the-art methodologies dedicated to glass and transparent object segmentation. CAGNet consistently outperformed other methods in detecting broken glass regions (rows 2 and 5), multiple glass areas (rows 2–7), expansive glass zones (rows 1, 4, and 7), and segmenting glass regions under outdoor natural illumination (rows 2 and 4). The remaining scenarios depict segmentation within the interior under various lighting conditions. Compared to other methodologies, CAGNet exhibits minimal false detections and smoother edges and ensures the integrity of the segmented glass regions. This superior performance can be attributed to the multi-scale convolutional attention modules embedded within the CAGNet, which encapsulate these features and are further complemented by utilizing THs that weigh the features and effectively filter extraneous information [34,35]. For instance, in the resulting image in row 6, the other methods misclassified the ceiling or ghost as glass. Contrastingly, the proposed approach accurately discerned the glass and mitigated background interference owing to its nuanced feature analysis across different receptive fields, which is crucial for preserving the intricate details, as shown in rows 2, 3, and 4.

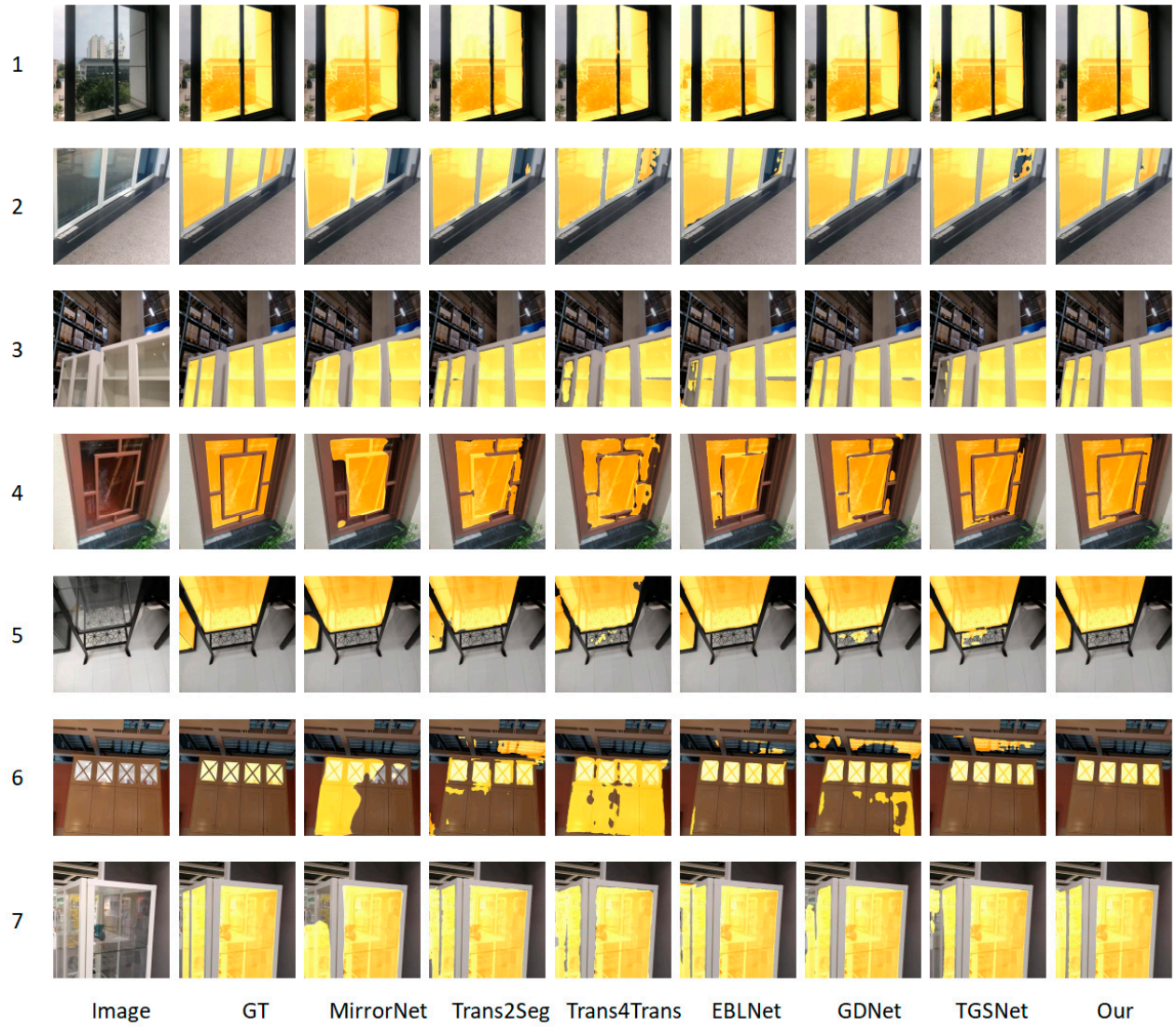

Figure 7. Comparison results of six state-of-the-art glass detection methods and the proposed CAGNet.

We conducted additional experiments to specifically address the challenges posed by varying lighting conditions, as illustrated in Figure 8. In the context of these experiments, daytime scenarios were characterized by ample lighting, whereas nighttime scenarios represented dimly lit environments. Photographs were captured at the same location under both of these conditions. Due to equipment constraints, maintaining the exact positioning was not feasible, but the images were shot from the same vantage point. Ground-truth annotations were provided for these images. As depicted in Figure 8, the proposed method consistently detected glass objects both indoors and outdoors, largely unaffected by the lighting conditions. This underscores the efficiency and robustness of CAGNet across diverse lighting scenarios, and it's evident that the evaluation metrics do vary based on these conditions, highlighting the model's adaptability.

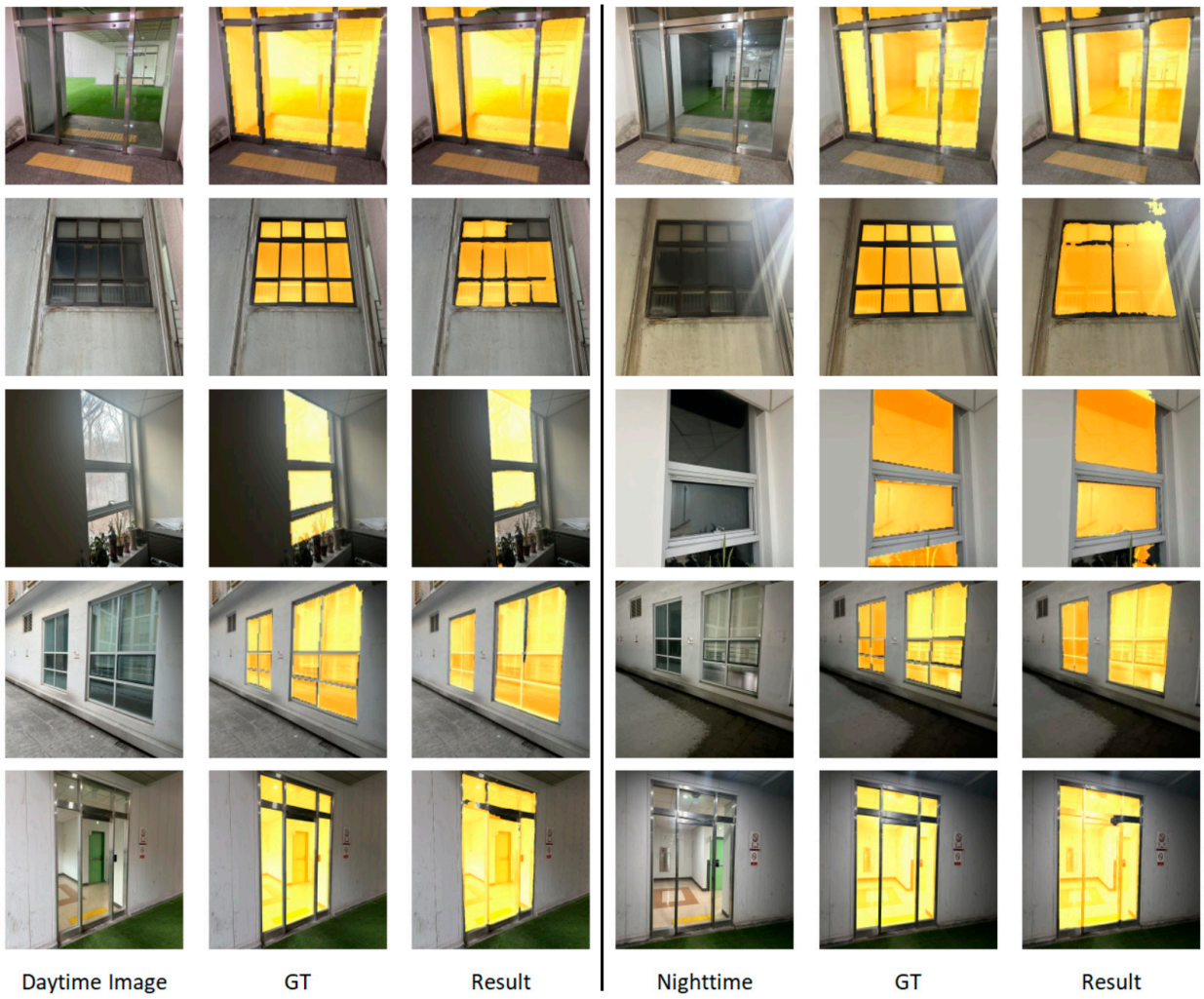

Figure 8. Comparison results of light experiments performed during daytime (left) and nighttime (right).

As shown in Table 2, Compared to TGSNet, the current approach has been significantly improved by optimizing the network structure. As shown in the table, the overall number of parameters in the proposed model was reduced by 36%. Using the same GPU quantity and model, the achieved training convergence speed was 2.6 times faster than that obtained in the previous study. During the testing phase, the average inference time per image was accelerated by 38.4%, and memory consumption during inference was reduced by 48.3%.

Table 2. Comparison of CAGNet with TGSNet in terms of parameter count, GPU usage, training time, number of iterations at convergence, average inference speed per image, and memory consumption. "MParams" represents the number of parameters.

Methods	MParams	GPU	Train Time	Train Epoch	Speed (Per Image)	Memory
TGSNet [20]	185.472	3 * A6000	36 h	500	0.26 s	8921 MiB
CAGNet	**118.734**	3 * A6000	**14 h**	**200**	**0.16 s**	**4610 MiB**

4.4. Ablation Experiments

In this section, we describe three sets of ablation studies. Through various experiments, the effectiveness of BFA, MCA, and CFAF components was validated. The BFA module's performance was compared to that of a CNN in terms of backbone feature extraction.

Additionally, the overall performance was investigated with respect to the influence of GC modules within BFA, the impact of using the CSA module in MCA, and the effect of the feature analysis module in CFAF. The results are presented in Tables 3–5 and Figures 9–11. In the tables, "Networks" and "Backbone" denote the network architecture and backbone network used for training.

Table 3. Results of the ablation experiments for the proposed BFA module, where "conv" denotes the process of gleaning backbone features across varied scales using convolution methods. "↑" indicates that higher values are better, "↓" indicates that lower values are better.

Networks	Backbone	GDD [16]			
		IoU↑	F_β^w↑	MAE↓	BER↓
a. Conv	ResNeXt-101	87.34	0.896	0.068	6.55
b. BFA without GC	ResNeXt-101	88.56	0.907	0.059	5.69
c. BFA with GC	ResNeXt-101	**88.73**	**0.913**	**0.054**	**5.51**

Table 4. Results of the ablation experiments for the MCA module. "↑" indicates that higher values are better, "↓" indicates that lower values are better.

Networks	Backbone	GDD [16]			
		IoU↑	F_β^w↑	MAE↓	BER↓
a. MCA without CSA	ResNeXt-101	88.12	0.905	0.057	5.77
b. MCA with CSA	ResNeXt-101	**88.73**	**0.913**	**0.054**	**5.51**

Table 5. Results of the ablation experiments for the CFAF module. "↑" indicates that higher values are better, "↓" indicates that lower values are better.

Networks	Backbone	GDD [16]			
		IoU↑	F_β^w↑	MAE↓	BER↓
a. CFAF without TH	ResNeXt-101	88.21	0.903	0.062	5.84
b. CFAF with TH	ResNeXt-101	**88.73**	**0.913**	**0.054**	**5.51**

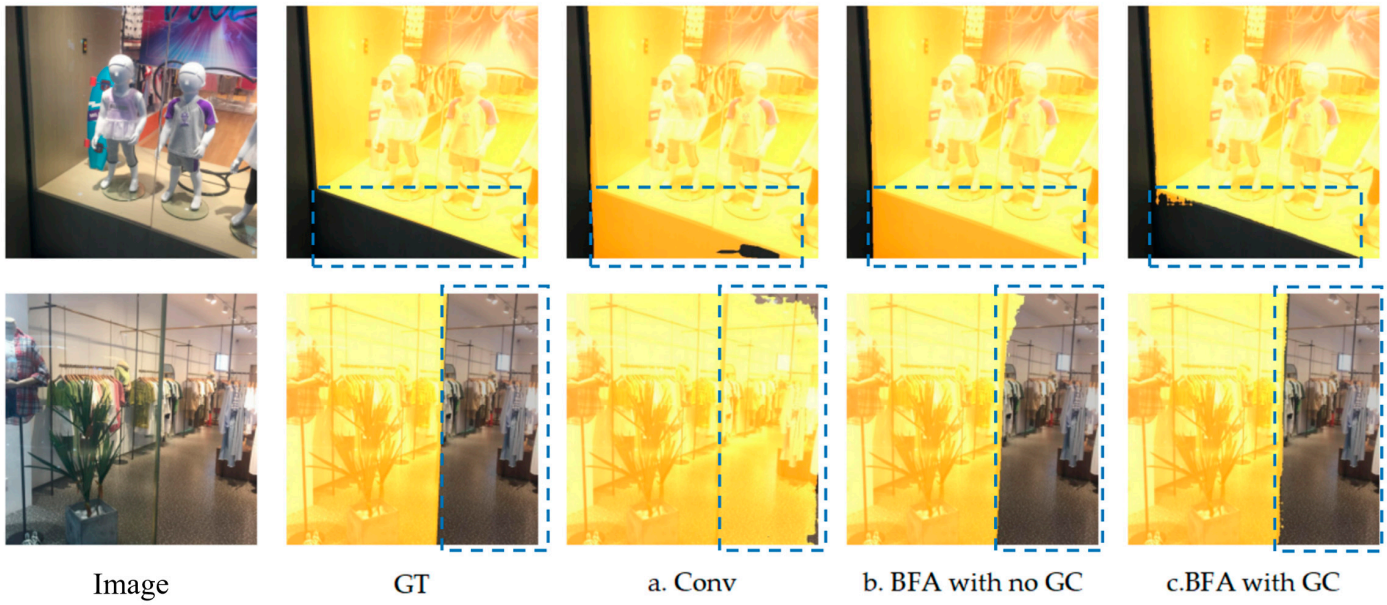

Figure 9. Comparison of results of BFA module ablation experiments using (a) a convolution method, (b) BFA with no GC, and (c) BFA with GC.

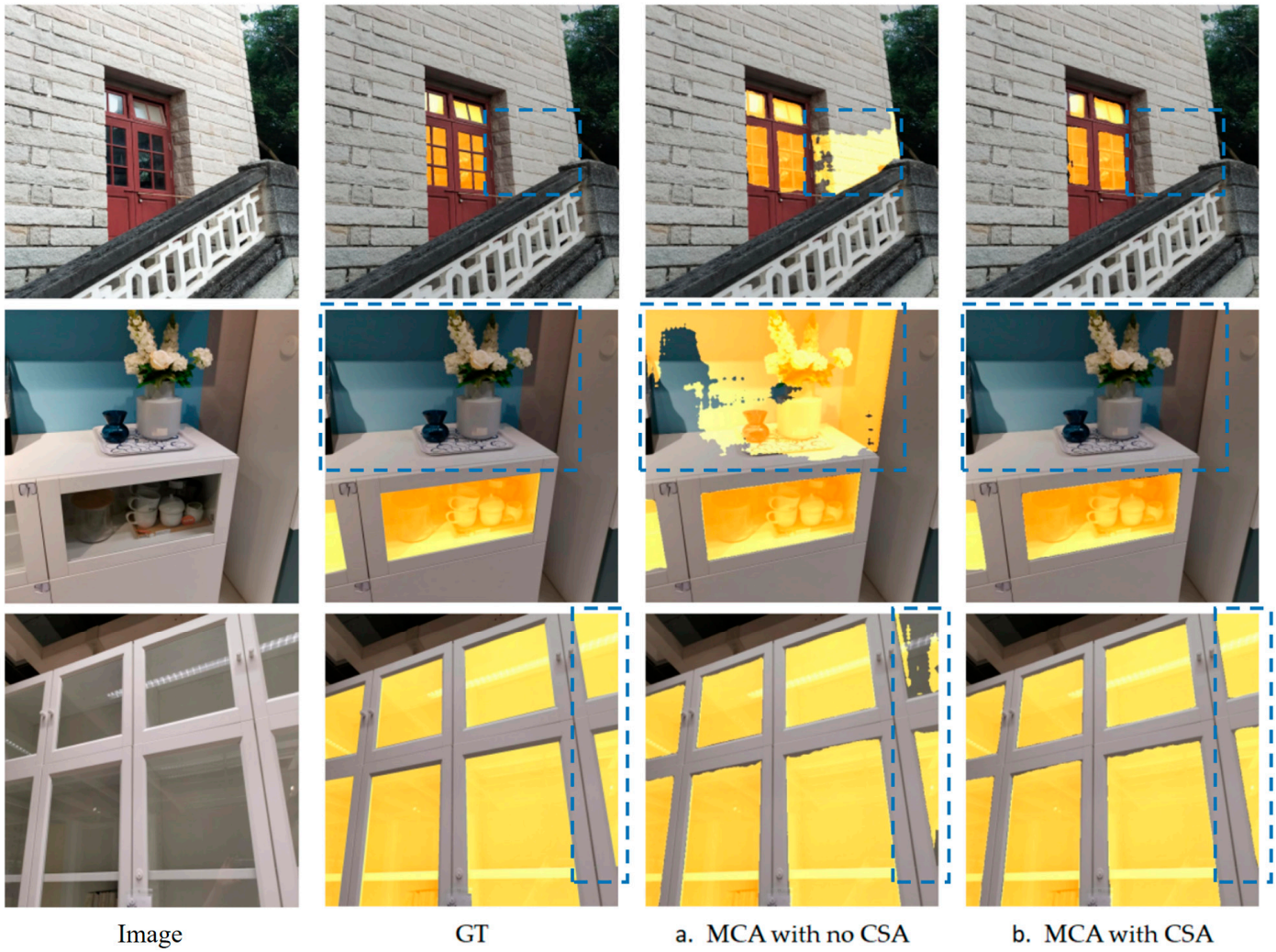

Figure 10. Comparison of results of MCA module ablation experiments. (a) MCA without CSA and (b) using BFA with GC.

4.4.1. Effectiveness of the BFA Module

Herein, the effectiveness of BFA and its GC module is validated. The experiments were divided into the following three groups: (a) convolutional approach, (b) BFA without the GC module, and (c) BFA with the GC module.

Figure 9 shows the limitations of traditional convolutional methods for making accurate predictions while processing the content within frames. Herein, the model may learn erroneous features, misclassifying the content inside a frame as a glass object. Furthermore, this approach fails to detect entire regions, as indicated by the dashed blue box in the figure. The proposed BFA can fuse features across different levels in the absence of a GC module, thereby amplifying the information from the feature backbone. Simultaneously, BFA can filter incorrect features owing to the disparities between different levels [42,43]. BFA enhances the model's performance to some extent; however, instances of misprediction because of overfitting are observed. By introducing the GC module in BFA, the proposed model's performance was significantly boosted, with overfitting issues effectively mitigated.

Table 3 presents a quantitative analysis of the evaluation results from the three experiments. These results unequivocally demonstrate that BFA equipped with the GC module is advantageous for backbone feature extraction and fusion and overcomes the limitation of overfitting. Therefore, the GC module is crucial in enhancing the model's performance, improving feature fusion efficiency, and effectively mitigating overfitting.

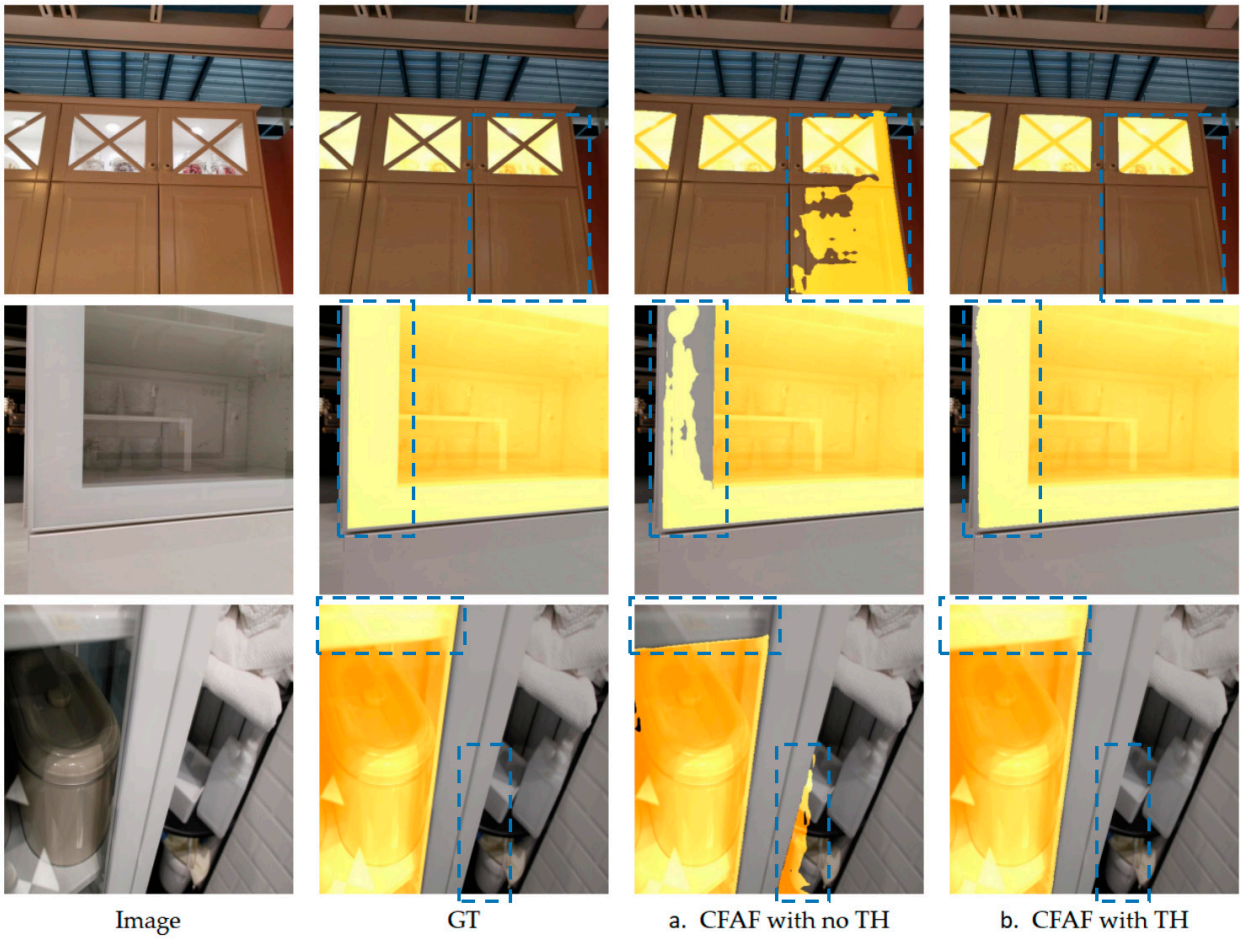

Figure 11. Comparison of the results of CFAF module ablation experiments. (a) CFAF without TH and (b) CFAF with TH.

4.4.2. Effectiveness of the MCA Module

This section presents a scenario wherein the structure remains unchanged except for the MCA. Within the MCA, the multi-scale convolutional attention method and CSA were introduced and quantitatively compared to the scenario without CSA. To validate the effectiveness of the CSA module, the following two sets of experiments were performed: (a) MCA without CSA and (b) MCA with CSA.

Figure 10 shows that in the qualitative experiments without using CSA, the model detected the glass position; however, it struggled to fully distinguish between the target and non-target areas. After incorporating CSA, the model detected the glass region (row 3) more comprehensively and distinguished between the target and non-target areas (blue dashed box in Figure 9) more precisely.

As shown in Table 4, when CSA was introduced into the model, all four evaluation metrics improved significantly. For MCA with CSA, the model's IoU value increased by 0.61% compared to that without CSA. Hence, the introduction of CSA effectively enhanced the feature extraction capability of the model, thereby further boosting its overall performance.

4.4.3. Effectiveness of the CFAF Module

This section validates the efficacy of TH by comparing the results of CFAF with and without TH while keeping the other conditions constant.

Through qualitative analysis, Figure 11 indicates that by employing TH to integrate features from each layer of MCA, the model can effectively filter out larger erroneous regions (as indicated by the blue dashed box in the first row). Furthermore, the introduction

of TH enables the model to classify non-target regions more accurately (as shown by the blue dashed boxes in the second and third rows). Therefore, incorporating TH significantly enhances the model's capability to assimilate features, reduce errors, and improve classification accuracy, thereby optimizing the overall performance of the framework.

As shown in Table 5, the quantitative experimental results indicate that the overall performance of the model is significantly enhanced by the introduction of TH, highlighting the pivotal role of TH in the aggregation and fusion of convolutional features within CFAF. Particularly, by integrating features from each MCA layer, TH effectively filters the errors and elevates classification accuracy, enhancing the model's performance.

5. Conclusions

This study investigated the glass detection subtask within image segmentation, emphasizing the effective extraction and fusion of features to enhance model performance. Through experiments, we identified the limitations of traditional convolution methods in predicting the content within frames, causing the acquisition of incorrect features. Consequently, the BFA approach was introduced to address these limitations. BFA promotes feature fusion across different levels, amplifying information from the backbone feature and filtering inaccuracies owing to interlevel differences. While BFA improved the model's performance, overfitting issues remained. Therefore, to further enhance the model's capabilities and reduce overfitting, a GC module was incorporated into the BFA. GC captures global contextual information, which significantly improves the model's ability to handle complex scenarios. Additionally, the proposed MCA module introduced CSA, which further refined feature extraction and fusion using self-attention. Finally, we integrated TH into CFAF to consolidate features from different levels, thereby mitigating the impact of disparities on the model's performance. The experimental results indicated a marked improvement in overall performance with the inclusion of TH, emphasizing its importance in enhancing performance and reducing overfitting. In conclusion, the proposed BFA, GC, CSA, and TH modules demonstrated significant efficacy in boosting model performance, optimizing feature fusion efficiency, and effectively controlling overfitting. In future endeavors, we aim to investigate the intricacies and potential enhancements of these modules. We will explore the practical applications of this model, particularly in domains requiring precise glass detection, such as autonomous driving and urban planning. Furthermore, we aim to apply our research findings to segmentation tasks involving remote sensing images and 3D point-cloud data.

Author Contributions: Conceptualization, X.H.; Formal analysis, X.H.; Funding acquisition, K.C.; Investigation, R.G.; Methodology, X.H.; Project administration, K.C.; Software, X.H.; Supervision, K.C.; Validation, X.H. and R.G.; Visualization, X.H.; Writing—original draft, X.H.; Writing—review and editing, R.G., S.Y. and K.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government (23ZH1200), the research of the fundamental media contents technologies for hyper-realistic media space and Institute of Information and Communications Technology Planning and Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00254592) grant funded by the Korea government (MSIT).

Data Availability Statement: Publicly available datasets were analyzed in this study. These data can be found here: URL: https://mhaiyang.github.io/CVPR2020_GDNet/index.html (accessed on 20 August 2023.) [16].

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

References

1. Gao, R.; Li, M.; Yang, S.-J.; Cho, K. Reflective Noise Filtering of Large-Scale Point Cloud Using Transformer. *Remote Sens.* **2022**, *14*, 577. [CrossRef]
2. Gao, R.; Park, J.; Hu, X.; Yang, S.; Cho, K. Reflective noise filtering of large-scale point cloud using multi-position LiDAR sensing data. *Remote Sens.* **2021**, *13*, 3058. [CrossRef]
3. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
4. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2961–2969.
5. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
6. Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; Agrawal, A. Context encoding for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7151–7160.
7. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 568–578.
8. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 379–387.
9. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
10. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
11. Liu, F.; Shen, C.; Lin, G. Deep convolutional neural fields for depth estimation from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5162–5170.
12. Zheng, C.; Cham, T.J.; Cai, J. T2net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 767–783.
13. Zhang, L.; Dai, J.; Lu, H.; He, Y.; Wang, G. A bi-directional message passing model for salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1741–1750.
14. Wu, Z.; Su, L.; Huang, Q. Cascaded partial decoder for fast and accurate salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3907–3916.
15. Liu, J.J.; Hou, Q.; Cheng, M.M.; Feng, J.; Jiang, J. A simple pooling-based design for real-time salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June; pp. 3917–3926.
16. Mei, H.; Yang, X.; Wang, Y.; Liu, Y.; He, S.; Zhang, Q.; Wei, W.; Lau, R.W. Don't hit me! glass detection in real-world scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3687–3696.
17. Lin, J.; Yeung, Y.H.; Lau, R.W.H. Depth-aware glass surface detection with cross-modal context mining. *arXiv* **2022**, arXiv:2206.11250.
18. 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]
19. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12622–12631.
20. Hu, X.; Gao, R.; Yang, S.; Cho, K. TGSNet: Multi-Field Feature Fusion for Glass Region Segmentation Using Transformers. *Mathematics* **2023**, *11*, 843. [CrossRef]
21. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
22. Ronneberger, O.; Fischer, P.; Brox, T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*; Springer: Cham, Switzerland, 2015; pp. 234–241.
23. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
24. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.
25. Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef] [PubMed]
26. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.

27. Girshick, R. Fast r-cnn. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1440–1448.
28. Cao, Y.; Zhang, Z.; Xie, E.; Hou, Q.; Zhao, K.; Luo, X.; Tuo, J. FakeMix augmentation improves transparent object detection. *arXiv* **2021**, arXiv:2103.13279.
29. 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 IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15859–15868.
30. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pvt v2: Improved baselines with pyramid vision transformer. *Comput. Vis. Media* **2022**, *8*, 415–424. [CrossRef]
31. Guo, M.H.; Lu, C.Z.; Hou, Q.; Liu, Z.; Cheng, M.M.; Hu, S.M. Segnext: Rethinking convolutional attention design for semantic segmentation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 1140–1156.
32. Xie, E.; Wang, W.; Wang, W.; Sun, P.; Xu, H.; Liang, D.; Luo, P. Segmenting Transparent Objects in the Wild with Transformer. *IJCAI* **2021**, 1194–1200. [CrossRef]
33. Zhang, J.; Yang, K.; Constantinescu, A.; Peng, K.; Muller, K.; Stiefelhagen, R. Trans4Trans: Efficient transformer for transparent object segmentation to help visually impaired people navigate in the real world. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1760–1770.
34. Aboutalebi, H.; Pavlova, M.; Gunraj, H.; Shafiee, M.J.; Sabri, A.; Alaref, A.; Wong, A. MEDUSA: Multi-Scale Encoder-Decoder Self-Attention Deep Neural Network Architecture for Medical Image Analysis. *Front. Med.* **2021**, *8*, 821120. [CrossRef]
35. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.
36. Kang, D.; Dhar, D.; Chan, A. Incorporating side information by adaptive convolution. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 2897–2918. [CrossRef]
37. Shi, Y.; Wang, M.; Chen, S.; Wei, J.; Wang, Z. Transform-based feature map compression for cnn inference. In Proceedings of the 2021 IEEE International Symposium on Circuits and Systems (ISCAS), Daegu, Republic of Korea, 22–28 May 2021; pp. 1–5.
38. Kiyak, E.O.; Cengiz, A.B.; Birant, K.U.; Birant, D. Comparison of image-based and text-based source code classification using deep learning. *SN Comput. Sci.* **2020**, *1*, 266. [CrossRef]
39. Ding, X.; Guo, Y.; Ding, G.; Han, J. Acnet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1911–1920.
40. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
41. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the AAAI-17: Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; p. 31.
42. Xiao, J.; Zhao, T.; Yao, Y.; Yu, Q.; Chen, Y. Context augmentation and feature refinement network for tiny object detection. In Proceedings of the Tenth International Conference on Learning Representations, Virtual, 25–29 April 2021.
43. Chen, L.I.; Jianxun, L.I. Orthogonal Features Extraction Method and Its Application in Convolution Neural Network. *J. Shanghai Jiaotong Univ.* **2021**, *55*, 1320.
44. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
45. Liu, W.; Rabinovich, A.; Berg, A.C. Parsenet: Looking wider to see better. *arXiv* **2015**, arXiv:1506.04579.
46. Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; He, K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1492–1500.
47. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, 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.
48. Yu, L.; Mei, H.; Dong, W.; Wei, Z.; Zhu, L.; Wang, Y.; Yang, X. Progressive Glass Segmentation. *IEEE Trans. Image Process.* **2022**, *31*, 2920–2933. [CrossRef] [PubMed]
49. Margolin, R.; Zelnik-Manor, L.; Tal, A. How to evaluate foreground maps? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 248–255.
50. Nguyen, V.; Yago Vicente, T.F.; Zhao, M.; Hoai, M.; Samaras, D. Shadow detection with conditional generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4510–4518.
51. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 603–612.
52. Huang, S.; Lu, Z.; Cheng, R.; He, C. FaPN: Feature-aligned pyramid network for dense image prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 864–873.
53. Zhao, J.X.; Liu, J.J.; Fan, D.P.; Cao, Y.; Yang, J.; Cheng, M.M. EGNet: Edge guidance network for salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8779–8788.

54. Wei, J.; Wang, S.; Huang, Q. F^3Net: Fusion, feedback and focus for salient object detection. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12321–12328.
55. Yang, X.; Mei, H.; Xu, K.; Wei, X.; Yin, B.; Lau, R.W. Where is my mirror? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8809–8818.
56. Lin, J.; He, Z.; Lau RW, H. Rich context aggregation with reflection prior for glass surface detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13415–13424.