Multimodal Diamond Inclusion Detection and Clarity Grading via Enhanced YOLOv7 with ECA and ASFF Modules

Xiaobing Huo
Department of Geology, Hebei Vocational College of Resources and Environment, Shijiazhuang City, Hebei Province, 050091, China
E-mail: xiaobinghuo55@163.com

Keywords: YOLOv7 model, multimodal input, inclusion detection, clarity grading model, grading rule fusion

Received: November 1, 2025

The detection of micro-inclusions and the representation of interpretable results compatible with gemological standards are two major bottlenecks to the automation of diamond clarity grading. In this work, we propose a novel solution using an enhanced YOLOv7 model in a multimodal framework to overcome the above bottlenecks. Specifically, our contributions are threefold: Firstly, we improve the original YOLOv7 by incorporating the Efficient Channel Attention (ECA) mechanism to enhance the extracted fine-grained features and the Adaptively Spatial Feature Fusion (ASFF) module for capturing more robust multi-scale representations; secondly, we build a three-channel input consisting of optical grayscale, gray-level co-occurrence matrix (GLCM) texture linked with the optical properties of inclusions, and morphological operation-enhanced images. Thirdly, we design a traceable grading system by combining the XGBoost classifier with programmable GIA (Gemological Institute of America) rules. Our method achieves 91.3% mAP@0.5 on the Roboflow Diamond Inclusion dataset, outperforming the baseline YOLOv7 by 9.2%. The clarity grading performance on this dataset attains an accuracy of 86.7%, a Kappa coefficient of 0.82, and a weighted F1-score of 0.87, resulting in high consistency with human expert evaluations. Ablation experiments confirm that the proposed components all make individual and complementary contributions. This work represents a significant advance towards the automatic, accurate, and interpretable grading of diamonds, and it creates a practical tool for use within the jewelry industry.

Povzetek: Raziskava predstavlja razložljiv multimodalni sistem na osnovi izboljšanega YOLOv7 za samodejno zaznavanje inkluzij in ocenjevanje čistosti diamantov, ki dosega visoko točnost in dobro skladnost z gemološkimi standardi.

1 Introduction

One of the main factors that influences the value of a diamond is its clarity, which is essentially determined by the number, size, type, and position of the internal inclusions. (Barrie, Teuthof, & Eaton-Magaña, 2021). It is a very slow and expensive procedure, and the outcome can also be influenced by the subjective experience of the person performing it. To ensure the correctness and reproducibility of the results, strict training, multi-expert verification, and standardization of the processes are necessary (Eaton-Magaña, Hardman, & Odake, 2024). In recent years, deep learning technology has provided a new path for the automated detection of diamond inclusions. However, existing methods primarily focus on target positioning and do not fully integrate the core attribute of diamond clarity(Kigo, Omondi, & Omolo, 2023). They ignore the interpretability conversion from “detection” to “grading”, making the model output difficult for the industry to adopt (Hardman et al., 2024; Luo, Nelson, Ardon, & Breeding, 2021). In particular, under the dual constraints of insufficient accuracy in detecting micro-inclusions and opaque logic for determining clarity grades, constructing an automatic grading system with high accuracy and reliability still faces severe challenges. Typical inclusions in diamonds include four microscopic defects: pinpoints, clouds, feathers, and crystals (D'Haenens-Johansson, Butler, & Katrusha, 2022; Ma et al., 2022). These inclusions' number, size, type, spatial distribution, and relative position form the basis for determining clarity grading (Nelson & Reinitz, 2024). Especially in critical grades such as SI1/SI2, slight differences can cause a jump in the clarity grade of a gemstone (Gao et al., 2023). For example, tiny crystals located in the table's center have a greater clarity impact than similar inclusions in the edge area. The center of the table determines the visual center (Pham, Koniuch, Wynne, Brown, & Collins, 2025). Specifically, the defect signal of the middle table is significantly higher than that of the edge (Chepurov et al., 2021). Clustered pinpoints may be judged as “clouds, " notably reducing the grade (Eaton-Magaña et al., 2024). Therefore, precise detection alone is not enough to support reliable grading. Further modeling of inclusions' visual significance and spatial semantics is necessary to achieve intelligent decision-making aligned with GIA rules.

Although existing research has made some progress in automated diamond evaluation, there are still obvious deficiencies in the combination of defect recognition and GIA rules. To systematically identify the research gaps,

Table 1: Summary of related work on automated diamond inclusion detection and grading

Model / Approach	Dataset	Inclusion Types Detected	Key Metrics	Main Limitations
CBAM-ResNet50-FPN (Wenqian, Fengxia, Quanbin, & Qinghai, 2024)	Proprietary	Pinpoints, Crystals	Accuracy (Detection)	Focuses only on detection; no mapping to GIA grades.
VGG-16 (Jianxin et al., 2020)	Proprietary	General Inclusions	Accuracy (Detection)	Fails to quantify size, density, and spatial distribution.
Clarity & Cut Grading (Eaton-Magana, Ardon, Breeding, & Shigley, 2020)	Market /Proprietary	N/A	Accuracy (Grade)	Ignores detailed visual performance and inclusion characteristics.

we summarize key existing approaches and their limitations in Table 1.

As elucidated in Table 1, current research has fundamental flaws in three aspects: a single feature expression dimension (relying solely on RGB images (Tsai, D'Haenens-Johansson, Smith, Zhou, & Xu, 2024)), opaque decision logic, and insufficient adaptability to complex GIA industry rules, leading to outputs that are out of line with standards.

These mentioned limitations are overcome by designing the present work to achieve the following three core research objectives: 1) To significantly raise the micro-inclusion detection accuracy, in particular those smaller than 5 pixels, which are crucial for distinguishing between high clarity grades. 2) The interpretability and GIA-consistency of the grading results are guaranteed, with a traceable decision path that aligns with expert standards. 3) Enhancement of the model's robustness across varying imaging conditions will make it practical for application in the real world.

Our work performs theoretical innovations and experimental validation to achieve these objectives. In this paper, our core methodology is embodied by the proposed hierarchical, decoupled detection-analysis-decision architecture. Multimodal input in this approach combines GLCM texture features at the feature encoding level with a morphological closed-operation enhancement map. In this light, weak inclusion signals are detected according to the mechanisms of light scattering and absorption. At the network structure level, an ECA channel attention mechanism is embedded in the YOLOv7 backbone network to further improve the small object feature response, while the ASFF module can realize adaptive fusion of multi-scale features at the decision-making level. Key features such as location, area ratio, and density extracted from the instances of detected inclusions are fed into an XGBoost classifier integrated with a configurable GIA rule-based inference engine to ensure interpretability and consistency of the results with the rules. Experimental study on the Roboflow Diamond Inclusion dataset verifies that the proposed system realizes a detection mAP@0.5 of 91.3%, clarity grading accuracy of 86.7%, and a Kappa coefficient of 0.82, thereby outperforming significantly the baseline YOLOv7 as well as other mainstream detection models. This is in accordance with the effectiveness and necessity for automated diamond clarity assessment, including multimodal feature fusion and structural improvement.

2 Methods

2.1 Problem formulation and decoupled objectives

The automatic diamond clarity grading task is decomposed into two successive objectives: Inclusion Detection and Clarity Grading, reflecting the process of a gemological expert first locating inclusions and then synthesizing their attributes into a final grade.

Inclusion Detection: An improved YOLOv7 model is trained to locate all the inclusions, each defined by a bounding box and class label, optimizing a composite detection loss jointly refining localization, classification, and confidence.

Grading for Clarity: The findings from detection are converted into a feature vector that encapsulates major GIA criteria, including area ratio, spatial distribution, density, and inclusion type combination.

2.2 Diamond dataset construction

The Roboflow Diamond Inclusion dataset (https://universe.roboflow.com/diamond-classification-data/detect-inclusions/dataset/1) has been utilized to build uniform input data collection from the train/valid/test splits. The dataset represents four different types of microinclusions: pinpoints, clouds, feathers, and crystals.

All images underwent a consistent preprocessing pipeline prior to model input. (1) Automatic white balancing to correct for color temperature variations; (2) Luminance normalization to reduce inter-image intensity variance; (3) Resizing to 640×640 pixels using bilinear interpolation.

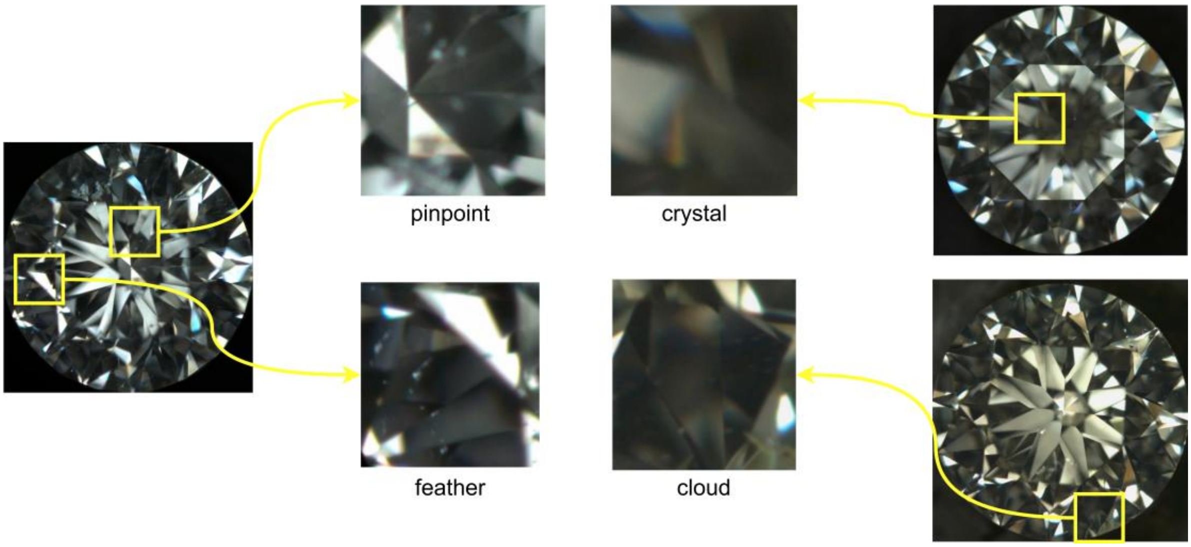

Figure 1: Dataset sample (containing different inclusion types)

The examples in Figure 1 show that pinpoint inclusions are tiny dots (<5 pixels in diameter), feathers with feather-like fractures, geometric crystal inclusions, and diffuse, hazy cloud-like inclusions.

Dataset partitioning and clarity grade distribution statistics are shown in Table 2.

Table 2: Dataset partitioning and clarity grade distribution statistics

Dataset Partition	Number of Images	IF	VVS1	VVS2	VS1	VS2	SI1	SI2
Training Set	394	47	57	63	63	60	44	60
Validation Set	100	11	14	16	16	16	11	16
Test Set	125	15	18	20	20	19	14	19
Total	619	77	93	103	103	98	72	98

The dataset in Table 2 contains 619 diamond images, divided into a training set (394 photos), a validation set (100 images), and a test set (125 images) to ensure the reliability of model training and evaluation. Expert appraisers label clarity grades and classify them into seven categories according to GIA standards: IF (Flawless), VVS1/2 (Very Slightly Included), VS1/2 (Slightly Included), and SI1/2 (Slightly Included).

2.3 Multimodal feature enhancement strategy based on optical property analysis

The three-channel multimodal input strategy is grounded in the physics of light-inclusion interaction to overcome dependencies on specific imaging conditions, encoding key optical properties through GLCM texture and morphological enhancement for robust inclusion detection. This study constructs a three-channel multimodal input strategy based on optical property modeling. By applying a gray-level co-occurrence matrix (GLCM) texture feature map and a morphological closing operation enhancement map, it supplements the scattering and absorption difference information that cannot be explicitly expressed by the RGB (red, green, and blue) channels, thereby improving the model's ability to perceive faint inclusions.

The multimodal feature fusion input is shown in Figure 2.

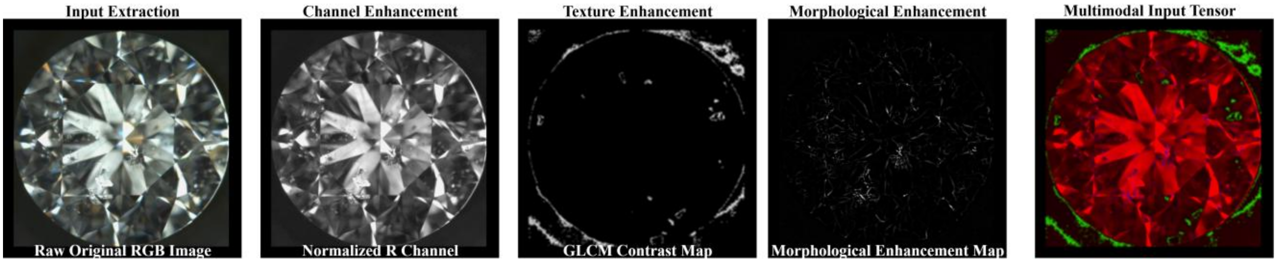

Figure 2: Multimodal feature fusion input

As shown in Figure 2, The normalized R channel is extracted to preserve dominant brightness information.

A Gray-Level Co-occurrence Matrix (GLCM) contrast map T_{GLCM} is computed from the grayscale image. The original RGB image is converted to a grayscale image I_{gray}, and the weighted average method is used:

$$I_{\mathrm{gray}}=0.299R+0.587G+0.114B \quad (1)$$

Formula 1 calculates the retained brightness dominant information as the basis for subsequent texture and morphological analysis.

Based on I_{gray}, we compute the GLCM across four directions (0°, 45°, 90°, 135°) using a 7×7 window and 32 grayscale levels. Among the extracted texture features, the contrast map is selected as the second input channel due to its superior ability to characterize the light scattering heterogeneity caused by refractive index differences between inclusions and the diamond matrix, providing stronger responses to inclusion edges than other descriptors like energy or homogeneity.

A circular structure element with a radius of 3 is used as the structural element, and a closing operation is performed on I_{gray}:

$$I_{\mathrm{close}}=\left(I_{\mathrm{gray}}\oplus B\right)\ominus B \quad (2)$$

In Formula 2, $\oplus$ and $\ominus$ represent dilation and erosion, respectively. The closing operation effectively connects broken edges, fills tiny holes, and enhances the continuity of closed contours.

To enhance the absorption characteristics, I_{close} is differentially processed with the original image:

$$E_{\mathrm{morph}}=\left|I_{\mathrm{close}}\text{-}I_{\mathrm{gray}}\right| \quad (3)$$

According to Formula 3, an edge enhancement map is generated, normalized to the minimum and maximum values, and used as the third channel. The closing operation is chosen because it effectively connects broken edges and fills tiny holes in dark, preserving contiguous inclusion regions's structural integrity while avoiding the risk of removing small inclusions like pinpoints, which opening operations may cause.

The original normalized R channel, GLCM contrast map T_{GLCM}, and a morphological enhancement map E_{morph} are spliced along the channel dimension to form a three-channel input tensor:

$$\mathrm{I}_{\mathrm{multi}}=[R,T_{\mathrm{GLCM}},E_{\mathrm{morph}}]\in\mathrm{R}^{640\times640\times3} \quad (4)$$

The output tensor obtained by Formula 4 enables the network to respond to color, texture heterogeneity, and morphological closure simultaneously at the shallow level, simulating the physical mechanism of the human eye in identifying inclusions by adjusting the lighting angle under a microscope.

2.4 Improved YOLOv7 network structure and clarity grading framework

This paper improves the structure based on the YOLOv7 backbone. The overall network process includes three stages: multi-modal input is used to complete inclusion detection through the improved YOLOv7; the spatial and geometric attributes of the detection results are extracted; the GIA grading rules and XGBoost classifier are integrated to achieve automatic clarity determination. The overall framework of the model is shown in Figure 3.

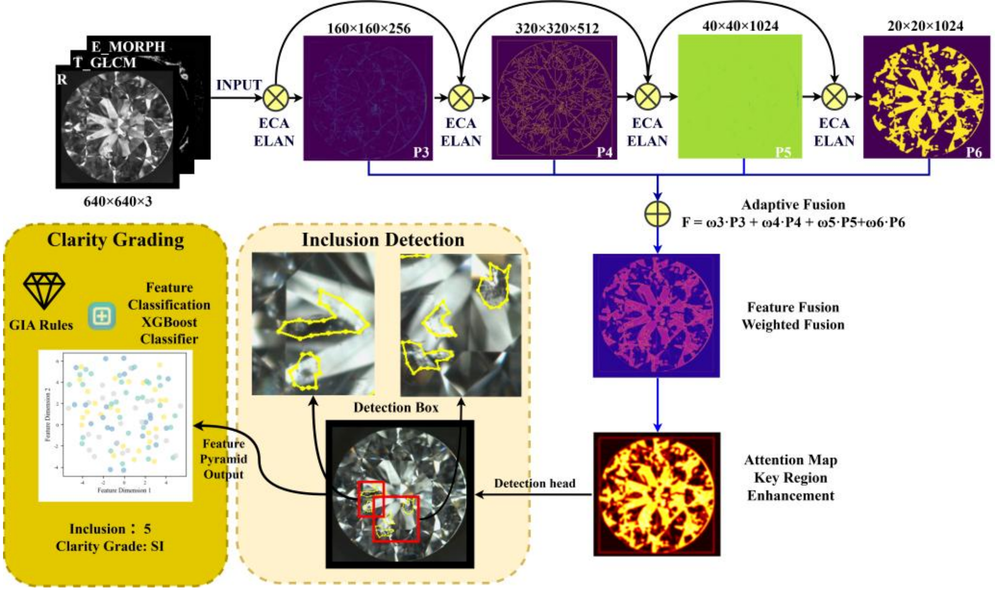

Figure 3: Overall clarity grading framework

Figure 3 Enhanced overall clarity grading framework with detailed tensor dimensions. This module takes the multimodal image as input and produces a set of raw detection outputs, where each detection includes bounding box coordinates and a class label for the inclusion. The second objective, Clarity Grading, leverages the results from the first. It starts with the module Feature Extraction, which translates the raw detections into a structured set of quantitative features: the total inclusion area, spatial distribution, density, and type combinations, among others. These features come together into a single feature vector that encodes the primary criteria for grading. This feature vector then passes through two complementary components for the final decision: the XGBoost Classifier,

which gives a data-driven preliminary grade prediction, and the GIA Rule Engine, where expert-defined logic is applied to ensure that the final grade meets gemological standards. The result of all of the above is the final clarity grade for the diamond.

2.4.1 Improved backbone network

In the backbone network, the ECA attention mechanism addresses the insufficient response of the baseline YOLOv7 ELAN (Extended Efficient Layer Aggregation Network) module to tiny objects.

The integration of ECA modules maintains dimensional compatibility throughout the network. For instance, after each ELAN module, the feature map dimensions remain constant at H×W×C, ensuring seamless integration with the existing PANet structure. By processing through ECA, for example, the P3 feature map remains at 160×160×256, preserving high spatial resolution that is crucial for small inclusion detection.

The ECA module is embedded after the output of each ELAN module, forming a cascaded "convolution-attention" architecture. The feature map $F\in \mathrm{R}^{H\times W\times C}$ is compressed into a channel descriptor $z\in \mathrm{R}^{C}$ through global average pooling; one-dimensional convolution is applied to capture local cross-channel interactions and generate a channel weight vector $a\in \mathrm{R}^{C}$; the weight vector is multiplied with the original feature map to achieve channel recalibration. ECA abandons the fully connected layer and directly uses one-dimensional convolution to achieve cross-channel interactions, reducing the number of parameters $O(C^2)$ from $O(k)$, where k is the adaptively determined convolution kernel size.

The convolution kernel size k is adaptively determined by the number of channels C , and the calculation formula is:

$$k=\psi(C)=\left/\frac{\log_2(C)}{b}\right/+1 \tag{5}$$

In Formula 5, b is taken as 1.5, and the result is ensured to be an odd number.

2.4.2 Optimized detection head

The ASFF module replaces the fixed upsampling and splicing fusion method in the original PANet of YOLOv7. ASFF applies a learnable spatial weight matrix between feature layers of different scales to achieve adaptive weighted fusion of cross-scale feature maps.

The ASFF fusion logic operates as follows: for each output level i (where $i \in \{3,4,5,6\}$), feature maps from other levels are resized to match the spatial dimensions of level i through bilinear interpolation (for upsampling) or strided convolution (for downsampling). A 1×1 convolution followed by spatial softmax generates attention weight maps α_i^j for each source level j, ensuring that $\Sigma_j \alpha_i^j = 1$ at each spatial position. The final fused feature F_i at level i is computed as $F_i = \Sigma_j(\alpha_i^j \cdot P_j)$, where P_j represents the feature map from level j after spatial alignment.

For the i-th output layer ($i\in\{3,4,5,6\}$, corresponding to feature maps P3-P6), ASFF aligns the higher-level (i+1) and lower-level (i-1) feature maps to the current resolution through differentiable interpolation. A 1×1 convolution is then used to generate a spatial attention weight map, which is then weighted to fuse the four features (P3-P6 feature maps).

The final fused feature is represented as: $F=\omega 3\cdot P3+\omega 4\cdot P4+\omega 5\cdot P5+\omega 6\cdot P6$, where F is the output feature map, and ω is the attention weight for each feature map.

2.4.3 Clarity grading decision

The detection model output includes the class label, confidence score, and bounding box coordinates (x, y, w, h) for each object.

Based on the test results, the key attributes for clarity grading are extracted: (1) inclusion area ratio: the percentage of the union area of all detection frames to the total area of the entire image is calculated; (2) position feature: the image is divided into five areas according to GIA standards: table, crown facets, girdle, pavilion, and culet, and the distribution of the center coordinates of the detection frames in each area is counted; (3) density feature: the number of detection instances per unit area is calculated; (4) type combination feature: the frequency of high-impact categories such as crystals and clouds is counted. All extracted features are normalized by Z-score and fed into the XGBoost classifier as input.

To illustrate the traceable decision pathway, consider a concrete example of how the system grades a diamond:

Step 1: Detection outputs 8 inclusions (3 pinpoints, 2 crystals, 3 clouds) with bounding boxes

Step 2: Feature extraction calculates:

Total area ratio: 0.42%

Table center inclusions: 1 crystal (high impact)

Density: 3.8 inclusions/mm²

High-impact type count: 1 (crystal in table center)

Step 3: XGBoost processes these features and outputs a preliminary grade of VS2 with 87% confidence

Step 4: GIA rule engine applies post-processing:

Checks if any rules mandate grade adjustment

Confirms crystal in table center doesn't exceed VS2 thresholds

Final grade: VS2

The XGBoost classifier uses a gradient boosting tree structure, and the objective function includes a logarithmic loss term and L1 and L2 regularization terms. The 12-dimensional features used systematically quantify the key factors of diamond clarity rating, including position-related features, area, density features, and inclusion type features.

The classification output corresponds to the seven GIA clarity grading system categories. The GIA diamond clarity grading system is ranked from high to low as follows:

(1) IF: No inclusions are visible under a 10x magnifying glass, no visible inclusions, area accounting for 0%, density 0/mm², and no internal features.

(2) VVS1: Inclusions are rare and visible only from certain angles. They contain 1-2 extremely small inclusions and must be located on the girdle or pavilion, occupying less than 0.05% of the area and with a density of less than 0.5/mm². No visible features are allowed in the center of the table.

(3) VVS2: Inclusions are extremely difficult to see but visible from multiple angles. They contain 2-3 tiny inclusions, which may be located on the crown but not in the center of the table. The area accounts for 0.05%-0.1%, with a 0.5-1.0/mm² density. A trace amount of cloudiness is allowed at the girdle.

(4) VS1: Inclusions are difficult to see and require careful observation, containing 3-4 tiny inclusions. One extremely tiny feature is allowed in the center of the table, accounting for 0.1%-0.3% of the area, with a density of 1.0-2.0/mm². The total area of the cloud is <0.05% and does not extend to the surface.

(5) VS2: Inclusions are relatively easy to see and clearly visible, containing 4-6 tiny inclusions. Two tiny features are allowed in the center of the table, accounting for 0.3%-0.5% of the area, with a density of 2.0-3.0/mm². The total area of the cloud is <0.1%, and tiny feather cracks are allowed.

(6) SI1: Inclusions are easily visible, containing 6-8 medium-sized inclusions. Three features are allowed in the center of the table, but the total area is <0.2%, accounting for 0.5%-1.0% of the area, with a density of 3.0-5.0/mm². “Pinpoint groups” (5-8 with a spacing of <20μm) are treated as cloud-like objects.

(7) SI2: Inclusions are easily visible to the naked eye, with >8 medium to large inclusions. The central feature accounts for >0.2% of the table surface, or there is a distinct cloud-like feature, accounting for >1.0% of the area, with a density >5.0/mm². Multiple “pinpoint clusters” are combined, and the cloud-like area is >0.3%.

If cracks extending to the diamond’s surface or noticeable dark crystals are detected, the clarity grade is increased by at least one level. If multiple tiny inclusions appear as “pinpoint clusters” under a 10x microscope, with spacing less than 20 microns, they are combined according to the “clustering” rule and assigned an SI clarity grade. A combination of model predictions and expert rules determines the final clarity grade.

2.5 Model training and optimization configuration

To ensure the reproducibility of all experiments, the random seed was fixed to 42 for PyTorch, NumPy, and the Python built-in random module.

We employed an extensive data augmentation strategy using the Albumentations library to improve model robustness and prevent overfitting. The techniques applied with a probability of 0.5 included:

Horizontal and vertical flips, random rotations within ±15°, and random scaling between 0.8x and 1.2x. HSV color space variations (hue shift ±0.015, saturation ±0.7, value ±0.4), contrast-limited adaptive histogram equalization (CLAHE), Gaussian blur with kernel size up to 5x5, and additive Gaussian noise with variance limit of 0.01. Randomly occluding up to 3% of the image area with 5x5 to 15x15 pixel rectangles to simulate dirt or focus issues on the microscope lens.

During inference, the confidence threshold for accepting a detection was set to 0.5. The Non-Maximum Suppression (NMS) IoU threshold was set to 0.45 to eliminate redundant bounding boxes.

The optimizer uses SGD (Stochastic Gradient Descent) with momentum; the momentum coefficient is set to 0.937; the weight decay coefficient is 1×10⁻⁴; the batch size is 16; the gradient accumulation technique (accumulation step number 4) is used to simulate the equivalent batch size of 64 to avoid optimization instability caused by video memory limitations. The initial learning rate is set to 0.01, and the Cosine annealing scheduling strategy is adopted:

$$\eta_t=\eta_{min}+\frac{1}{2}(\eta_{max}-\eta_{min})(1+\cos\ (\frac{T_{cur}}{T_{max}}\pi)) \tag{6}$$

In Formula 6, η_{min}= 1×10^{-6}, and T_{max}=300 rounds. Linear warmup is performed in the first 5 rounds, and the learning rate is gradually increased from 1×10^{-4} to 0.01 to avoid gradient explosion in the initial stage. Different learning rates are set for different scale detection heads to solve the problem of large differences in inclusion scales: the learning rate coefficient of the shallow feature extraction part (P3) is 0.8, and the deep semantic part (P6) is 0.

The loss function uses a combination of improved CIoU (Complete Intersection over Union) loss and Focal Loss: CIoU is used for positioning loss:

$$\text{CloU}(\text{L}_{loc}=1\text{-}IoU+\frac{\rho^2(b,b^{gt})}{c^2}+\frac{v}{(1\text{-}IoU)+v} \tag{7}$$

In Formula 7, $v=\frac{4}{\pi^2}(\text{arctan}\ \frac{w^{gt}}{\text{h}^{gt}}\text{-arctan}\ \frac{w}{\text{h}})^2$; Focal Loss ($\text{L}_{cls}$=$-\alpha_t(1\text{-}p_t)^{\gamma}\log\ (p_t),\alpha=0.75$ is used for classification loss; binary cross-entropy is used for confidence loss.

The weight ratio of each loss is set to L_{loc}:L_{cls}:L_{obj}=0.05:0.5:1.0. For the positioning loss of slight inclusions (area <36 pixels), an additional 1.5 times weight is added to compensate for sampling bias. The strategy's IoU (Intersection over Union) threshold is set to 0.2. When the IoU between the predicted box and the ground-truth box is <0.2, it is considered background to prevent low-quality predictions from interfering with training.

The performance of the validation set is strictly monitored during the training process. The total number of rounds is 300, and the early stopping mechanism (patience=30) is used. When the validation set mAP@0.5 does not improve for 30 consecutive rounds (improvement <0.1%), the training is terminated, and the best model is retained. A checkpoint is saved every 10 rounds, and the model with the highest validation set mAP@0.5 is finally selected as the best result. During training, gradient clipping (maximum norm 2.0) is implemented to prevent gradient explosion, and EMA (Exponential Moving Average) technology is used to update weights (attenuation coefficient 0.9998) to improve model robustness.

3 Experiments and results

3.1 Experimental environment and setup

The hardware setup features 4×NVIDIA GeForce RTX 4090 GPUs (each with 24GB GDDR6X memory) connected through NVLink 4.0. The CPU is an AMD Ryzen Threadripper 3970X (32 cores, 64 threads) running at 3.7GHz and is paired with 128GB DDR4-3200 ECC memory. The storage configuration includes a 2TB NVMe SSD and an 8TB SATA SSD cache layer. All the experiments take place on the Ubuntu 20.04.6 LTS operating system, kernel version 5.15.0-86-generic, and using a separate Docker container v24.0.6 based on the NVIDIA CUDA 11.7 image.

Software Environment Configuration: PyTorch 1.13.1+cu117, TorchVision 0.14.1, CUDA 11.7, cuDNN 8.5.0, OpenCV 4.7.0, scikit-learn 1.2.2, XGBoost 1.7.5, Albumentations 1.3.0, TensorBoard 2.13.0. A fixed random seed (42) was used for all experiments to ensure deterministic and reproducible results.

3.2 Evaluation metrics

The metrics utilized in this study are stated as follows.

Precision: the proportion of samples that are truly positive among all samples detected as positive, calculated as $Precision = TP/(TP+FP)$.

Recall: the proportion of samples that are correctly detected as positive among all true positive samples, calculated as $Recall = TP/(TP+FN)$, where TP, FP, and FN represent true positives, false positives, and false negatives, respectively.

$F\,1score$: the harmonic mean of Precision and Recall, calculated as $F\,1 = 2 \times (Precision \times Recall)/(Precision + Recall)$.

Average Precision (AP): AP is calculated based on the area under the Precision-Recall curve for a single category. The detection results are sorted in descending order of confidence, and the AP is computed using the 11-point interpolation method:

$$AP=\frac{1}{11}\sum_{r\in\{0,0.1,...,1\}}\max_{\tilde{r}:\tilde{r}\geq r}P(\tilde{r}) \tag{8}$$

In Formula 8, the cumulative calculations are $P(k)=\frac{TP(k)}{TP(k)+FP(k)}$ and $R(k)=\frac{TP(k)}{TP(k)+FN(k)}$. The higher the AP value, the higher the detection accuracy.

mAP@0.5: the average AP value. The IoU threshold is fixed at 0.5. The calculation formula is:

$$mAP@0.5=\frac{1}{N}\sum_{i=1}^{N}AP_i \tag{9}$$

In Formula 9, N is the total number of categories, and AP_i is the AP value of the i-th category.

mAP@0.5:0.95: the average AP value is calculated at 10 points with IoU thresholds ranging from 0.5 to 0.95 (step size 0.05):

$$mAP@0.5{:}0.95=\frac{1}{10}\sum_{t=0.5}^{0.95}AP@t \tag{10}$$

In Formula 10, t represents the IoU threshold, and @0.5:0.95 represents the IoU threshold range from 0.5 to 0.95.

Cohen’s Kappa coefficient measures the consistency between the classification results and the manual annotations, considering the influence of random consistency. The calculation formula is:

$$\kappa=\frac{p_o\text{-}p_e}{1\text{-}p_e} \tag{11}$$

In Formula 11, p_o is the observed consistency ratio, and p_e is the random consistency expectation. Kappa > 0.8 indicates firm consistency.

Weighted F1-score: to address the problem of class imbalance, the F1-score is weighted by the support of each class:

$$F1_{weighted}=\sum_{i=1}^{N}w_i\times F1_i \tag{12}$$

In Formula 12, $w_i=\frac{n_i}{\sum_{j=1}^{N}n_j}$, and n_i is the number of samples in class i.

Confusion matrix: recording the correspondence between the prediction of each clarity grade and the true label.

Boundary positioning error: the mean Euclidean distance between the center point of the predicted bounding box and the true bounding box. The calculation formula is:

$$BLE=\frac{1}{K}\sum_{k=1}^{K}\sqrt{(x_k\text{-}\hat{x}_k)^2+(y_k\text{-}\hat{y}_k)^2} \tag{13}$$

In Formula 13, the calculation unit is a pixel.

Model Size: the disk space occupied by the serialized model weights in megabytes (MB).

Inference Time: the average time(ms) required to process a single 640×640 input image and obtain the final detection results, including all pre- and post-processing steps.

To deal with the numerous scales within diamond inclusions, scale-sensitive metrics are also computed: AP_small (area less than 36 pixels), AP_medium (36-144 pixels), and AP_large (more than 144 pixels). To properly evaluate the identification of very small inclusions (less than 10 pixels), the Intersection over Union (IoU) threshold is changed differently: a correct detection is one when the distance between the predicted box and the ground-truth box center is less than 2 pixels and "IoU>0.3".

3.3 Main results and comparisons

3.3.1 Detection performance comparison

To thoroughly analyze the suggested model's sensing abilities, the present work stepwise contrasts it with four representative detection models: YOLOv8-S(Lou et al., 2023), RT-DETR(Real-Time Detection Transformer)-R18(Madan & Reich, 2025), YOLOv7(S. Li, Wang, & Wang, 2023), YOLOv5s (You Only Look Once version 5 small)(Zhan et al., 2022), Faster R-CNN (ResNet50 backbone)(M. H. Li, Yu, Wei, & Chan, 2024), and

EfficientDet-D3(Shin, Lee, & Le, 2025), all of which were run under the same conditions and with the same dataset. Detailed performance measures are given in Table 3.

Table 3 shows each model's comprehensive performance on the Roboflow Diamond Inclusion test set, including core metrics for detection and grading tasks.

Table 3: Comprehensive performance comparison of models

Evaluation Metrics	Proposed Model	YOLOv8-S	RT-DETR-R18	YOLOv7	YOLOv5s	Faster R-CNN (ResNet50)	EfficientDet-D3
mAP@0.5(%)	91.3	85.7	84.2	82.1	83.1	84.7	86.2
mAP@0.5:0.95(%)	68.7	65.9	64.5	62.2	63.5	64.1	67.8
Precision(%)	93.8	78.1	76.8	87.5	88.2	89.3	90.7
Recall(%)	88.5	89.2	87.9	81.3	82.6	83.7	85.4
F1-score(%)	91.1	95.5	94.8	84.3	85.3	86.4	88
AP_small(%)	85.2	90.5	89.1	72.5	72.5	76.8	77.9
AP_medium(%)	92.7	84.9	83.8	85.6	86.4	88.2	89.6
AP_large(%)	95.3	87.6	86.3	93.8	94.1	94.5	94.8
BLE(dpi)	1.83	2.45	2.89	3.72	3.15	3.72	2.91
Grading Accuracy (%)	86.7	80.8	79.9	75.2	78.5	80.1	79.4
Kappa	0.82	0.74	0.72	0.65	0.71	0.73	0.68
Weighted F1(%)	0.87	0.79	0.77	0.74	0.76	0.77	0.72
Model Size(MB)	165.1	22.7	80.5	149.9	27.8	159.3	49.1
Inference Time(ms)	18.5	12.3	16.8	15.2	13.6	24.7	21.4

The improved YOLOv7 model proposed in this paper achieves state-of-the-art performance across virtually all metrics. It attains a mAP@0.5 of 91.3%, a substantial 9.2% increase over the original YOLOv7 baseline, and demonstrates superior overall accuracy with a leading F1-score of 91.1%. Crucially, our model notably outperforms other modern detectors, surpassing YOLOv8-S by significant margins of 5.6% in mAP@0.5 and 7.1% in the critical AP_small metric, while also achieving superior localization accuracy (BLE of 1.83 pixels) over the transformer-based RT-DETR-R18 (2.89 pixels). These results clearly indicate that, for this particular challenge of detecting microinclusions, architectural enhancements (ECA, ASFF) are significantly more effective than the general improvement of state-of-the-art architectures.

From the deployment viewpoint, our model exhibits an excellent balance between accuracy and computational complexity. While keeping real-time performance, with a model size of 165.1 MB and an inference time of 18.5 ms per image, it yields the best accuracy among all alternatives. Although YOLOv8-S offers the fastest inference of 12.3 ms and a smallest footprint of 22.7 MB, it sacrifices significant detection performance, with 85.7% mAP@0.5. The transformer-based RT-DETR-R18 provides an excellent trade-off for moderate size (80.5 MB) and speed (16.8 ms), but has poor overall accuracy of 84.2% mAP@0.5.

This research mainly compares the improved YOLOv7 model with the six mainstream object detection models, according to the proposal. In order to ensure fairness while comparing the results, all models are trained and tested under the same experimental conditions. Figure 4 illustrates the P-R curves for each model.

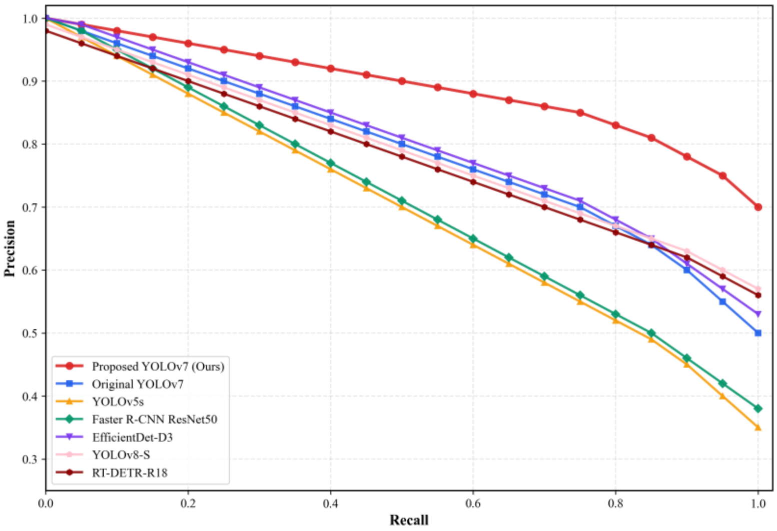

Figure 4: P-R curves for each model

Figure 4: The P-R curves show that the proposed model maintains high precision, ⩾0.70, in the high recall range, Recall ⩾ 0.80, where the baseline YOLOv7 is significantly lowered. Concretely, with a recall of 0.9, the proposed model keeps a precision of 0.77, much higher than YOLOv7's 0.60. This indicates that the proposed improved method significantly strengthens the capability of the model to distinguish complex samples. Hence, it can keep a low false positive rate, especially under high detection requirements.

The improved model also demonstrates significant advantages in detecting inclusions of varying scales. The AP_small metric reaches 85.2%, a 12.7% improvement over YOLOv7 (72.5%), significantly outperforming other compared models. This demonstrates that the ECA attention mechanism and the multimodal input strategy are vital in feature enhancement of tiny inclusions. In the case of medium-sized inclusions, the AP_medium metric is as high as 92.7%, which is by far the best result among the methods compared. AP_large for all models is at the same level; thus, the detection of large-scale inclusions can be considered solved. The variations between the models slightly indicate differences in identifying minor defects and a few medium ones.

As a matter of fact, the proposed model goes beyond expectations in terms of localization accuracy as well, with a boundary localization error (BLE) of merely 1.83 pixels, which is substantially lower than that of YOLOv7, Faster R-CNN (all 3.72 pixels), and YOLOv5s (3.15 pixels). This means that the upgraded model is capable of more exact localization of inclusions, thus being able to assess the spatial distribution of inclusions effectively for the following grading process.

3.3.2 Comparison of grading performance

This research uses a two-stage grading strategy based on XGBoost and GIA rules for diamond clarity grading. This method has been the main factor for a significant improvement in the accuracy and reliability of the grading. Table 2 indicates that the upgraded model attains an accuracy of grading of 86.7%, a Kappa coefficient of 0.82, and a weighted F1-score of 0.87. These indicators are at a level far beyond those of other end-to-end detection models that directly produce grading results. The baseline YOLOv7 model, on the other hand, can achieve only a grading accuracy of 75.2% and a Kappa of 0.65; thus, its detection results cannot be used directly for practical grading tasks.

The proposed method is still leading in all consistency metrics when compared with YOLOv5s (grading accuracy 78.5%, Kappa 0.71), Faster R-CNN (80.1%, Kappa 0.73), and EfficientDet-D3 (79.4%, Kappa 0.68). It is worth noting that the mAP@0.5:0.95 of EfficientDet-D3 (67.8%) is quite close to that of the proposed model (68.7%), but its grading Kappa coefficient is only 0.68, which means that a single detection model cannot be sufficiently flexible to GIA's complex grading rules, and also it cannot be adequately adaptable to the internal feature representation of diamond optical properties and grading logic.

Experimental results show that the two-stage grading architecture proposed in this study can better integrate expert prior knowledge and machine vision output. It fully utilizes the inclusion properties (such as size, quantity, position, etc.) provided by the detection model and integrates GIA rules through XGBoost, thereby improving the accuracy of grading decisions.

3.4 Ablation experiments

This study designs ablation experiments to test the performance of different module combinations on the test set. Table 4 shows detailed results of seven configurations compared to the original model in the ablation experiments, including the baseline model (YOLOv7) and variants that gradually added improved components. All experiments are repeated three times under the same conditions, and the average is taken.

Table 4: Ablation experiment analysis

Configuration	mAP @0.5 (%)	mAP @0.5: 0.95(%)	Precision(%)	Recall (%)	F1-score (%)	AP_ Small (%)	AP_ medium(%)	AP_ large(%)	Grading Accuracy (%)	Kappa
Baseline (Original YOLOv7)	82.1	62.2	87.5	81.3	84.3	72.5	85.6	93.8	75.2	0.65
+ECA	86.2	64.5	90.1	83.2	86.5	78.9	87.3	94.2	78.9	0.71
+ASFF	85.9	64.2	89.7	83.8	86.6	76.3	88.7	94.5	79.5	0.72
+Multimodal Input	87.8	66.1	91.2	84.5	87.7	79.6	89.3	94.7	81.8	0.75
+ECA+ASFF	89.5	67	92.4	85.9	89	82.7	90.8	95	83.6	0.77
+ECA+Multimodal Input	90.1	67.5	92.8	86.3	89.4	84.1	91.5	95.1	84.9	0.79
+ASFF+Multimodal Input	89.7	67.3	92.5	86.1	89.2	83.5	91.2	95	84.3	0.78
Full Model	91.3	68.7	93.8	88.5	91.1	85.2	92.7	95.3	86.7	0.82

Table 4 shows that applying the ECA attention mechanism alone improves mAP@0.5 by 4.1 percentage points to 86.2% and mAP@0.5:0.95 by 2.3 percentage points to 64.5%, validating ECA's enhancement of small inclusion features. AP_small increases from 72.5% to 78.9%, a 6.4 percentage point increase, confirming the effectiveness of the channel attention mechanism in enhancing weak inclusion signals. ECA improves the Kappa coefficient from 0.65 to 0.71 in the grading task, indicating that more precise detection results provide a more reliable foundation for subsequent grading.

When only ASFF is used without the original PANet detection head, the mAP@0.5 is raised by 3.8 percentage points to 85.9%, and the AP_medium and AP_large also go up by 3.1% and 0.7%, respectively, which clearly indicates that ASFF is more effective in detecting medium and large inclusions. The adaptive fusion of multi-scale features in ASFF can efficiently handle the large-scale variations of diamond inclusions, especially for cloud-like and crystalline inclusions (AP_medium went up from 85.6% to 88.7%).

By only using the multimodal input strategy, the largest increase in mAP@0.5 is achieved (+5.7% to 87.8%), basically for the detection of very small inclusions, where AP_small goes up by 7.1%. This confirms that the analysis of optical properties is a very efficient method and that the multimodal input can capture the scattering and absorption differences of very small inclusions that an RGB image cannot.

A primary finding is the synergistic effect between the modules: the combined usage of ECA and ASFF results in a substantial synergistic gain, obtaining an mAP@0.5 of 89.5%, which is a considerable increment beyond the separate usage of ECA (86.2%) and ASFF (85.9%). This proves the complementary effects of the two-feature extraction and fusion changes. The synergy here is most convincing in identifying slight inclusions, with an AP_small of 82.7%, thus exemplifying that the enriched feature channels of ECA and the spatial fusion, which ASFF optimizes, lead to the detection of the weakest signals.

Table 5: Sensitivity analysis

Configuration	mAP@0.5 (%)	AP_small (%)	Recall<5px (%)
Baseline (Original YOLOv7)	82.1	72.5	70.2
Multimodal Input - GLCM 5×5	86.5	77.8	80.1
Multimodal Input - GLCM 7×7	87.8	79.6	84.3
Multimodal Input - GLCM 9×9	87.1	78.9	82.7
Morphology Radius 2	87.0	78.5	81.9
Morphology Radius 3	87.8	79.6	84.3
Morphology Radius 4	87.3	79.0	83.2
Full Model (Proposed)	91.3	85.2	88.5

The ablation studies in table 5 further detail the performance impact of these parameter choices. Evaluation of GLCM window sizes and structuring element radii on a validation set containing sub-5-pixel inclusions confirmed that a 7×7 window and a radius of 3 pixels provided the optimal balance, achieving the highest recall rate (84.3%) for tiny inclusions.

The complete model (ECA + ASFF + multimodal input) yields the highest performance: mAP@0.5 is 91.3%, which is 9.2% higher than the baseline; mAP@0.5:0.95 is 68.7%, a 6.5% increment; the grading accuracy is 86.7%, with a Kappa of 0.82. The complete model achieves higher values than any other metrics to which it has been compared, thereby showing that each component is indispensable and that they are also complementary to each other.

3.5 Robustness testing

The robustness test procedure involves a controlled variable method where three typical interferences are applied to the 125 images of the test set: Gaussian noise, brightness change (±20%), and resolution reduction (downsampling to 320×320). Each transformation is performed five times, and the average value is used to get the reliability of the result. The Gaussian noise experiment is conducted with the `skimage.util.random_noise` function, where `mode='gaussian'` and `var=0.0001` (corresponding to "σ=0.01") are used, and the noise is added directly to the normalized image tensor. The brightness change experiment is performed by `RandomBrightnessContrast` from the `albumentations` library, where the brightness factor is set to a single value within the range [0.8, 1.2]. The resolution reduction experiment is done by `cv2.resize` that uses bilinear interpolation to downsample the image from 640×640 to 320×320 and adjusts the anchor box size of the detection head accordingly.

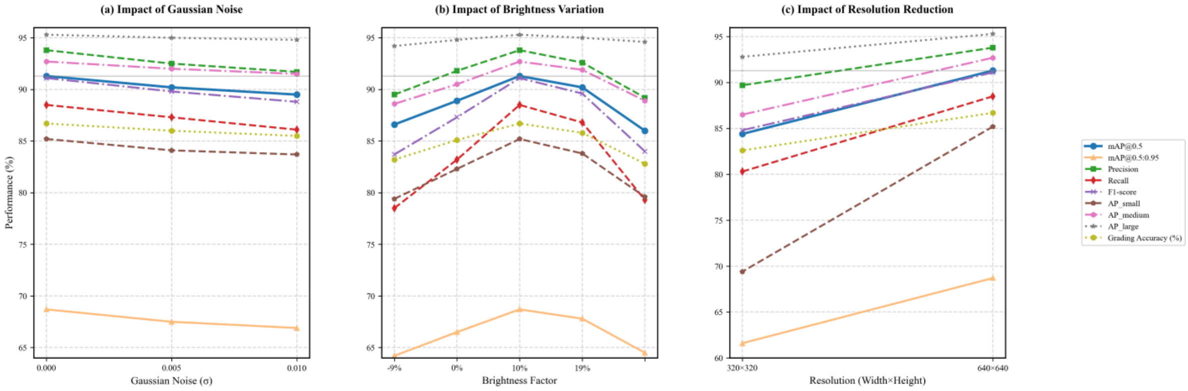

Figure 5: Performance curves under interference

Figure 5 illustrates that the performance decrease in the Gaussian noise experiment is almost linearly negatively correlated with the noise level, indicating that the model is robust to random noise. The model's mAP@0.5 decreases by only 1.8 percentage points to 89.5%, and its grading accuracy decreases by 1.2 percentage points to 85.5%. The study reveals that the model is quite noise-insensitive, which can be attributed to the intrinsic robustness of the GLCM texture features in the multimodal input to Gaussian noise.

The performance curves in Figure 5 exemplify a nonlinear decrease in model performance as the interference intensity increases. The brightness variation experiment has a relatively slow performance degradation rate that stays almost constant within the ±10% range (from 0% to +10% brightness, mAP@0.5 drops by 1.1%). A comparison of +20% and +10% brightness shows a significant decrease of mAP@0.5 by 4.2 percentage points; thus, the model is barely tolerant to extreme brightness variations. A drastic brightness change affects the model. The value of mAP@0.5 at -20% brightness is 86.6% after a drop of 4.7 percentage points, and at +20% brightness, it is 86.0% after a decline of 5.3 percentage points. The classification accuracy drops to 83.2% and 82.8%, respectively. The model's recall in dark areas (-20%) is only 78.5%, 10% lower than the recall of the improved model (88.5%), thus the brightness sensitivity issue has been addressed.

Resolution testing reveals that performance gradually deteriorates with resolution, and AP_small decreases very noticeably. When the resolution is halved, the mAP@0.5 drops by 6.9 percentage points to 84.4%, with AP_small going down heavily by 15.8% to 69.4%, whereas AP_large is only reduced by 2.5 percentage points to 92.8%.

Test results show that the improved model maintains good practicality under typical interference conditions, with manageable performance degradation under Gaussian noise and moderate brightness variations. Even when the resolution is reduced to 320×320, it still maintains an mAP@0.5 of 84.4%, meeting the practical requirements of most diamond detection equipment.

4 Discussion

4.1 Physical causes of detection errors

This paper presents actual cases of false negatives and false positives in Figure 6 to further explore the root causes of detection errors.

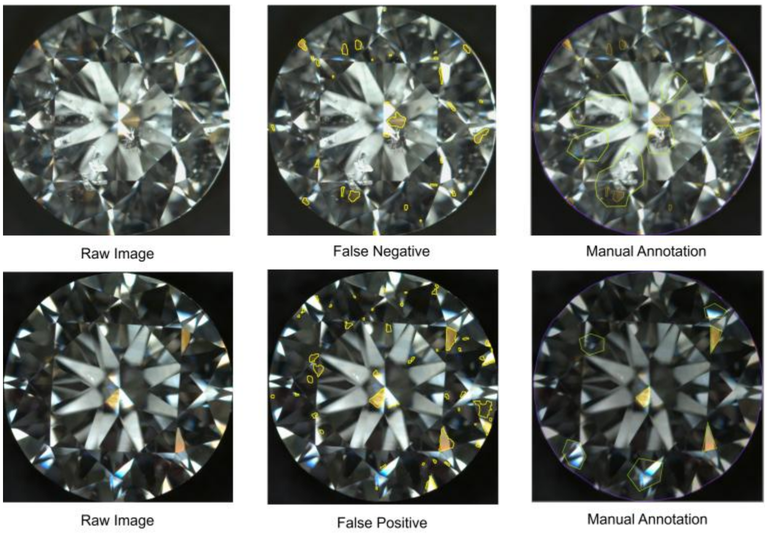

Figure 6: Visualization of detection error cases (false negatives and false positives)

Figure 6 shows that the false negatives are primarily concentrated in the highly reflective region at the junction of the diamond's crown and pavilion. These false negatives are likely due to a physical optical phenomenon: when illumination light is incident at a specific angle, the diamond's crystal faces produce strong specular reflections, with an intensity comparable to the scattered signal from weak inclusions, making it difficult for the model to distinguish them. False positives also mainly occur in areas with dense inclusions. When multiple tiny inclusions are arranged linearly in a local area, irregular fluctuations in mirror reflection near high-reflection areas may be misjudged as tiny inclusions. That suggests a pattern morphological closing operation that gives some reflection suppression but is not sufficient for complex optical scenarios.

Enhancements to the attention mechanism should aim at teaching the model to inhibit attention in high-reflection regions unless supported by complementary inclusion signatures across multiple feature scales. Geometric constraints can be brought into the post-processing refinement based on the facet structure of diamonds to suppress false positives in optically complicated areas, leveraging the successful integration of GIA rules shown (von Rueden et al., 2021).

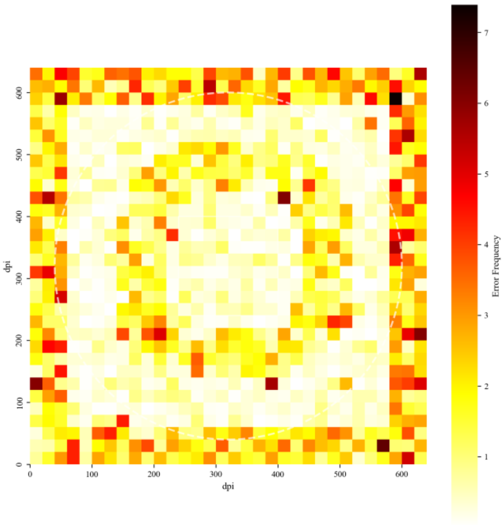

Figure 7: Spatial error distribution heat map

The spatial distribution of detection errors is visualized through a heat map in Fig. 7, which indicates that error frequencies are concentrated along the edges of diamonds (600-640 pixels) and the borders between crowns and pavilions (180-320 pixels from center). It follows that this spatial clustering of errors provides hard evidence for model sensitivity to optical physics rather than architectural deficiencies, hence the necessity for physics-informed methods in computer vision. Indeed, the coincidence of hotspots of false negatives with geometric edges of diamonds evidences that purely RGB-based methods cannot sort out optical ambiguities (Hao et al., 2025), hence integrating polarization imaging (Huang et al., 2023) or multi-angle illumination techniques is unavoidable for capturing additional physical properties.

Misjudgments for "clustered micro-points" could be further overcome with advanced morphological and depth feature extraction (Zhang & Xie, 2024), by better discerning discrete, point-like structures from continuous linear formations, especially through graph-based analyses of the inclusion spatial relationships.

We therefore propose the use of spatially-weighted loss functions to prioritize error correction in high-reflection zones identified from heat map analysis, along with reflection-aware data augmentation using generative adversarial networks that simulate challenging optical conditions at edge and junction regions. Furthermore, we recommend integrating polarization-sensitive features to leverage differential polarization properties between inclusions and diamond matrix (Huang et al., 2023), potentially through dedicated feature extraction branches.

4.2 Decision path tracing of grading deviations

Table 6 presents the confusion matrix between the model output and the expert rating statistically derived from 125 image examples in the test set.

Table 6: Confusion matrix

Predicted \ Actual	IF	VVS1	VVS2	VS1	VS2	SI1	SI2	Accuracy
IF	100	0	0	0	0	0	0	100
VVS1	0	93.5	3.2	2.1	1.2	0	0	93.5
VVS2	0	4.1	91.8	2.7	1.4	0	0	91.8
VS1	0	1.8	4.3	90.3	3.6	0	0	90.3
VS2	0	0	2.7	5.2	92.1	0	0	92.1
SI1	0	0	0	0	0	81.3	18.7	81.3
SI2	0	0	0	0	0	13.5	86.5	86.5

The confusion matrix evaluation in Table 6 indicates that the model is very effective in IF to VS2 grades, where the correct classification rates are over 90% in most cases. IF grades are differentiated correctly, with the following accuracy levels: 93.5% for VVS1, 91.8% for VVS2, 90.3% for VS1, and 92.1% for VS2. This indicates the model's strength in clear grades, as it can find small or scarcely distributed inclusions. On the other hand, the model severely confuses the SI1 and SI2 grades, mixing SI1 with SI2 in 18.7% of cases and SI2 with SI1 in 13.5% of instances.

GIA's definition of "high density" is based on a detailed visual inspection carried out by experts, who use their experience, the context, and the overall visual impression. It is a subjective and hard-to-measure process. Nevertheless, "density" in the model is an objective mathematical calculation. So, suppose the calculated result is near the SI1/SI2 decision boundary.

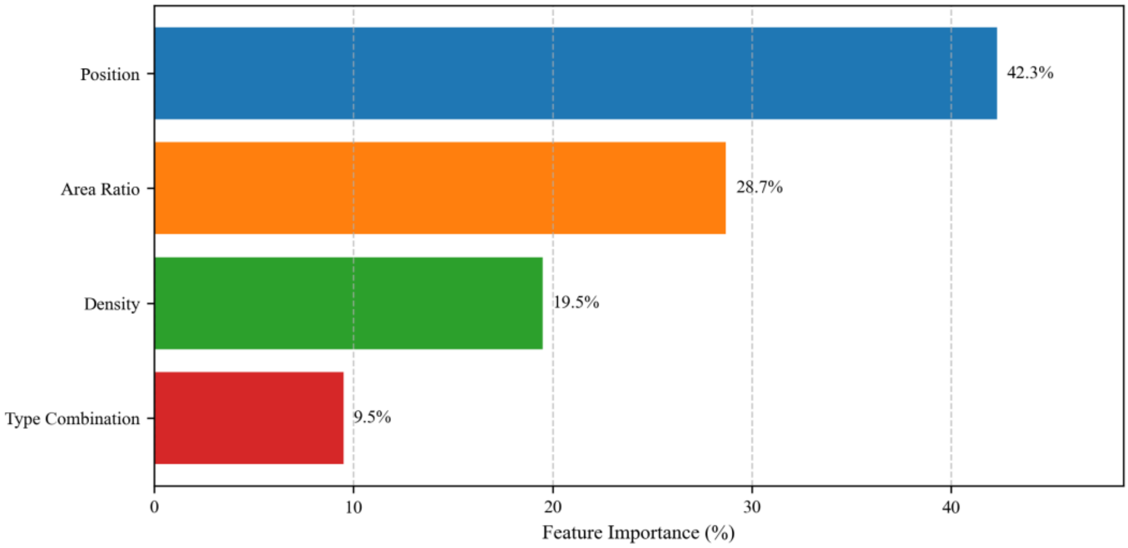

Figure 8: Feature importance of the XGBoost classifier

The examination of feature significance of the XGBoost classifier (Figure 8) indicates that the "position" attribute has the largest influence, accounting for 42.3%, followed by "area share" (28.7%), "density" (19.5%), and "type combination" (9.5%). Such a correlation of the weight values with the GIA clarity grading rules proves the model's decision logic is interpretable. Inclusions that are closer to the center of the table affect clarity more.

Table 7: Classification indicators for each category

Grade	Accuracy	Kappa	Weighted F1-score	F1-score
IF	100.00%	0.98	0.12000	1
VVS1	93.50%	0.94	0.12528	0.87
VVS2	91.80%	0.92	0.13600	0.85
VS1	90.30%	0.91	0.13440	0.84
VS2	92.10%	0.91	0.13072	0.86
SI1	81.30%	0.82	0.08512	0.76
SI2	86.50%	0.70	0.11400	0.75
Overall	-	-	0.84552	-

The weighted F1-score shown in Table 7 is 0.84552, slightly lower than the original test set of 0.87, mainly because of the lower F1-scores for the SI1/SI2 classes (0.76 and 0.75, respectively). Thus, this evidence shows the differences in performance when there is a class imbalance. It is hard to tell apart SI1/SI2 samples, which considerably affects the weighted metrics. The model's classification performance across different clarity grades shows a transparent gradient. The model's strong reliability in grading high-quality diamonds is demonstrated by extremely high accuracy and Kappa coefficients (0.91-0.98) for high clarity grades (IF to VS2).

4.3 Quantitative verification of module synergy

This section provides a quantitative verification of the synergistic interactions between the ECA, ASFF, and multimodal input modules through ablation studies. To quantitatively validate the interactions between modules, we analyzed the super-additive effect based on mAP@0.5, with results summarized in Table 8.

Table 8: Analysis of module synergy based on mAP@0.5 improvement

Configuration	Performance Gain (ΔmAP@0.5)	Theoretical Additive Gain	Observed - Additive
+ECA	+4.1%	—	—
+ASFF	+3.8%	—	—
+Multimodal	+5.7%	—	—
ECA + ASFF	+7.4%	4.1% + 3.8% = +7.9%	-0.5%
Full Model	+9.2%	4.1% + 3.8% + 5.7% = +13.6%	-4.4%

Note: Theoretical Additive Gain is the sum of individual modules' improvements. A positive "Observed - Additive" value indicates a super-additive (synergistic) effect.

The cooperative effect of the modules is further demonstrated by performance improvements across different inclusion sizes (Table 8). While the two-module combination shows a near-additive effect, the Full Model achieves a robust gain of +9.2%, significantly outperforming any single module. ECA enhances feature channels for small objects, ASFF improves multi-scale fusion, and multimodal input provides physically-grounded features. Together, they form a system that transcends the sum of its individual parts.

4.4 Interpretation of performance gains and comparison with SOTA

The overall significant improvements in mAP@0.5 by +9.2% and, more notably, AP_small by +12.7% above the state-of-the-art baseline YOLOv7 can be attributed to the synergistic interplay of our core contributions. The multimodal input strategy directly addresses the fundamental limitation of relying solely on RGB data by explicitly encoding texture and morphological information.

Our results confirm its critical role in handling the diverse sizes of diamond inclusions, from tiny pinpoints to larger crystals and feathers. While recent detectors like YOLOv8-S and RT-DETR also employ advanced feature fusion techniques, our tailored combination of multimodal input with ECA and ASFF specifically for the optical properties of diamond inclusions yields superior performance on this specialized task.

4.5 Limitations and error analysis: the optical physics bottleneck

Despite the high general performance, the error analysis in Figures 6 and 7 highlights a more fundamental problem: the performance of the model is intrinsically bounded by the physics of optics. The concentration of false negatives and false positives in high-reflection zones is not a weakness in the network architecture, but rather a result of the nature of the imaging. In these regions, specular reflections give rise to signals that may overwhelm or appear like the scattered light from true inclusions-a problem no degree of network tuning can totally avoid using conventional brightfield microscopy alone.

While robust to mild Gaussian noise, the current model does not have an adaptive mechanism for dynamic recalibration of perception amidst such severe interference with spatially-variant characteristics.

4.6 A control-theoretic perspective on robustness and grading consistency

Inspired by adaptive fuzzy control (Boulkroune, Zouari, & Boubellouta, 2025) and robust neural adaptive control (Farouk Zouari, Saad, & Benrejeb, 2012),, one of the ways a future iteration of this work may improve upon is by having a feedback loop such that the confidence of the grading decision feeds back to either fine-tune the feature extraction or the detection thresholds in cases deemed ambiguous. If the grading module was unsure between SI1 and SI2, for example, it would request a reanalysis of the image using a different sensitivity setting or a focused examination of the density calculation for "clustered micro-dots," just as an adaptive controller adjusts its parameters in order to maintain stability and performance in the presence of disturbances.

4.7 Impact and application in automated jewelry appraisal

The primary application and significant impact of this work lie in the automation of jewelry appraisal and quality control. The model's grading accuracy is 86.7%, with a very high Kappa coefficient of 0.82; thus, it would be a good candidate for high-throughput, preliminary grading to reduce human workload drastically. In fact, the interpretability afforded by the XGBoost-GIA rule fusion is not merely an academic exercise; it is a critical feature toward industry adoption. Appraisers can trust the system because they can trace the logic behind each grade, understanding whether a diamond was downgraded due to a large inclusion in the table center or a high density of pinpoints.

5 Conclusions

5.1 Experimental summary

This research enhances the detection precision of tiny inclusions in diamonds by upgrading YOLOv7 and incorporating multimodal features based on the optical properties. The clarity grading model merges XGBoost with GIA rules, resulting in a highly interpretable automatic grading system. The mAP@0.5 on the Roboflow dataset reaches 91.3%. XGBoost classifier fusion with GIA rules is used for the test results, which are automatically graded, thus achieving an accuracy of 86.7% and a Kappa coefficient of 0.82, thereby simultaneously attaining high precision and high interpretability. The researchers discovered that the primary sources are the model errors due to the optical physical aspects of the highly reflective areas and the insufficient quantification of the density of 'clustered micro-dots'.

5.2 Outlooks

The diamond inclusion detection and clarity grading model still have several limitations, which suggest a direction for future research.

Enhanced Robustness Through Adaptive Control Principles. Future work will explore adaptive control paradigms to dynamically adjust the model in response to input uncertainty. Specifically, adaptive fuzzy control (Boulkroune, Zouari, et al., 2025) and neural adaptive control frameworks(Farouk Zouari et al., 2012) could be employed to weight multimodal features or fine-tune model parameters dynamically, much like these controllers handle uncertainties in complex nonlinear systems(Boulkroune, Boubellouta, et al., 2025; Boulkroune, Hamel, Zouari, Boukabou, & Ibeas, 2017) . For more deterministic optimization under known operational constraints, nonlinear optimal control strategies(Rigatos et al., 2023) could be adapted to manage the trade-offs between detection sensitivity and specificity. Furthermore, integrating adaptive backstepping control methodologies(F. Zouari, Saad, & Benrejeb, 2013) could provide a structured approach to robustly handle the cascaded uncertainties inherent in our detection-then-grading pipeline.

To enhance trust and adoption in high-stakes applications, we will implement specific Explainable AI (XAI) techniques (Dwivedi et al., 2023; K. Li, Zhang, Li, Li, & Fu, 2023). Grad-CAM++ visualizations of image regions most influential in detecting subtle inclusions will be provided for the detection module. For the grading decision, SHAP analysis will quantify the feature contributions to the final clarity grade and provide structured grading reports mirroring the expert's reasoning. This is of great importance to clarify borderline cases, such as SI1/SI2.

Long-tail inclusion distributions need to be addressed by expanding the model's scope through few-shot learning and generative data augmentation for rare types, such as laser drilling (Anjomani & Ardon, 2022) and feather crack healing (Wallace, Plank, Bodnar, Gaetani, & Shea, 2021). Advanced texture analysis will be explored to better quantify the density of clustered inclusions, such as fractal dimension (Liu et al., 2024), local binary patterns (Chow & Reyes-Aldasoro, 2021), and graph neural networks (Zhu, Liu, Yao, & Fischer, 2021) for modeling complex spatial distributions.

References

[1] Anjomani, N., & Ardon, T. (2022). Unusual Laser Drill Holes in a Laboratory-Grown Diamond. In (Vol. 58, pp. 56-57): GEMOLOGICAL INST AMER 5345 ARMADA DR, CARLSBAD, CA 92008 USA.

[2] Barrie, E., Teuthof, S., & Eaton-Magaña, S. (2021). HPHT-Processed DIAMOND with Counterfeit GIA Inscription. *Gems & Gemology, 57*(3), 258.

[3] Boulkroune, A., Boubellouta, A., Bouzeriba, A., & Zouari, F. (2025). Practical Finite-Time Fuzzy Synchronization of Chaotic Systems with Non-Integer Orders: Two Chattering-Free Approaches. *Journal of Systems Science and Systems Engineering, 34*(3), 334-359. Retrieved from https://doi.org/10.1007/s11518-024-5635-7. doi:10.1007/s11518-024-5635-7

[4] Boulkroune, A., Hamel, S., Zouari, F., Boukabou, A., & Ibeas, A. (2017). Output-Feedback Controller Based Projective Lag-Synchronization of Uncertain Chaotic Systems in the Presence of Input Nonlinearities. *Mathematical Problems in Engineering, 2017*(1), 8045803. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1155/2017/8045803. doi:https://doi.org/10.1155/2017/8045803

[5] Boulkroune, A., Zouari, F., & Boubellouta, A. (2025). Adaptive fuzzy control for practical fixed-time synchronization of fractional-order chaotic systems. *Journal of Vibration and Control*, 10775463251320258. Retrieved from https://journals.sagepub.com/doi/abs/10.1177/10775463251320258. doi:10.1177/10775463251320258

[6] Chepurov, A., Sonin, V., Shcheglov, D., Zhimulev, E., Sitnikov, S., Yelisseyev, A., & Chepurov, A. (2021). Surface Porosity of Natural Diamond Crystals after the Catalytic Hydrogenation. *Crystals, 11*(11), 1341. Retrieved from http://dx.doi.org/10.3390/cryst11111341. doi:10.3390/cryst11111341

[7] Chow, B., & Reyes-Aldasoro, C. (2021). Automatic Gemstone Classification Using Computer Vision. *Minerals, 12*(1), 60. Retrieved from http://dx.doi.org/10.3390/min12010060. doi:10.3390/min12010060

[8] D'Haenens-Johansson, U. F. S., Butler, J. E., & Katrusha, A. N. (2022). Synthesis of Diamonds and Their Identification. *Reviews in Mineralogy and Geochemistry, 88*(1), 689-753. Retrieved from http://dx.doi.org/10.2138/rmg.2022.88.13. doi:10.2138/rmg.2022.88.13

[9] Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., . . . Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Computing Surveys, 55*(9), 1-33. Retrieved from http://dx.doi.org/10.1145/3561048. doi:10.1145/3561048

[10] Eaton-Magana, S., Ardon, T., Breeding, C. M., & Shigley, J. E. (2020). Natural-Color D-to-Z Diamonds: A Crystal-Clear Perspective. *Gems &amp; Gemology, 56*(3), 318-335. Retrieved from http://dx.doi.org/10.5741/gems.56.3.335. doi:10.5741/gems.56.3.335

[11] Eaton-Magaña, S., Hardman, M. F., & Odake, S. (2024). Laboratory-Grown Diamonds: An Update on

Identification and Products Evaluated at GIA. *Gems &amp; Gemology, 60*(2). Retrieved from http://dx.doi.org/10.5741/gems.60.2.146. doi:10.5741/gems.60.2.146

[12] Gao, Y., He, M., Sun, X., Zhen, C., Li, H., Wei, X., & Zhao, Y. (2023). Jedi Spinel from Man Sin, Myanmar: Color, Inclusion, and Chemical Features. *Crystals, 13*(1), 103. Retrieved from http://dx.doi.org/10.3390/cryst13010103. doi:10.3390/cryst13010103

[13] Hao, H., Fang, Y., Diao, Z., Li, X., Chen, L., Wang, Q., . . . Luo, X. (2025). High-accuracy and broadband polarization detection via metasurface vector beam modulation. *Photonics Research, 13*(9), 2487. Retrieved from http://dx.doi.org/10.1364/prj.565136. doi:10.1364/prj.565136

[14] Hardman, M. F., Homkrajae, A., Eaton-Magaña, S., Breeding, C. M., Palke, A. C., & Sun, Z. (2024). Classification of Gem Materials Using Machine Learning. *Gems &amp; Gemology, 60*(3), 306-329. Retrieved from http://dx.doi.org/10.5741/gems.60.3.306. doi:10.5741/gems.60.3.306

[15] Huang, X., Wu, C., Xu, X., Wang, B., Zhang, S., Shen, C., . . . Chang-Hasnain, C. J. (2023). Polarization structured light 3D depth image sensor for scenes with reflective surfaces. *Nature Communications, 14*(1). Retrieved from http://dx.doi.org/10.1038/s41467-023-42678-5. doi:10.1038/s41467-023-42678-5

[16] Jianxin, Y., Xiaoping, L., Bo, W., Lei, Y., Zhen, Z., & Yadong, F. (2020). Detection method based on deep learning for yellow industrial diamond. *金刚石与磨料磨具工程, 40*(6), 13-19.

[17] Kigo, S. N., Omondi, E. O., & Omolo, B. O. (2023). Assessing predictive performance of supervised machine learning algorithms for a diamond pricing model. *Scientific Reports, 13*(1), 17315. Retrieved from https://doi.org/10.1038/s41598-023-44326-w. doi:10.1038/s41598-023-44326-w

[18] Li, K., Zhang, Y., Li, K., Li, Y., & Fu, Y. (2023). Image-Text Embedding Learning via Visual and Textual Semantic Reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(1), 641-656. Retrieved from http://dx.doi.org/10.1109/tpami.2022.3148470. doi:10.1109/tpami.2022.3148470

[19] Li, M. H., Yu, Y., Wei, H., & Chan, T. O. (2024). Classification of the qilou (arcade building) using a robust image processing framework based on the Faster R-CNN with ResNet50. *Journal of Asian Architecture and Building Engineering, 23*(2), 595-612. Retrieved from https://doi.org/10.1080/13467581.2023.2238038. doi:10.1080/13467581.2023.2238038

[20] Li, S., Wang, S., & Wang, P. (2023). A Small Object Detection Algorithm for Traffic Signs Based on Improved YOLOv7. *Sensors, 23*(16), 7145. Retrieved from https://www.mdpi.com/1424-8220/23/16/7145.

[21] Liu, Y., Sun, T., Wu, K., Zhang, H., Zhang, J., Jiang, X., Feng, M. (2024). Fractal-Based Pattern Quantification of Mineral Grains: A Case Study of Yichun Rare-Metal Granite. *Fractal and Fractional, 8*(1), 49. Retrieved from http://dx.doi.org/10.3390/fractalfract8010049. doi:10.3390/fractalfract8010049

[22] Lou, H., Duan, X., Guo, J., Liu, H., Gu, J., Bi, L., & Chen, H. (2023). DC-YOLOv8: Small-Size Object Detection Algorithm Based on Camera Sensor. *Electronics, 12*(10), 2323. Retrieved from https://www.mdpi.com/2079-9292/12/10/2323.

[23] Luo, Y., Nelson, D., Ardon, T., & Breeding, C. M. (2021). Measurement and Characterization of the Effects of Blue Fluorescence on Diamond Appearance. *Gems &amp; Gemology, 57*(2), 102-123. Retrieved from http://dx.doi.org/10.5741/gems.57.2.102. doi:10.5741/gems.57.2.102

[24] Ma, Y., Qiu, Z., Deng, X., Ding, T., Li, H., Lu, T., . . . Wu, J. (2022). Chinese Colorless HPHT Synthetic Diamond Inclusion Features and Identification. *Crystals, 12*(9), 1266. Retrieved from http://dx.doi.org/10.3390/cryst12091266. doi:10.3390/cryst12091266

[25] Madan, M., & Reich, C. (2025). Strengthening Small Object Detection in Adapted RT-DETR Through Robust Enhancements. *Electronics, 14*(19), 3830. Retrieved from https://www.mdpi.com/2079-9292/14/19/3830.

[26] Nelson, D. P., & Reinitz, I. M. (2024). Metrology at GIA. *Gems & Gemology, 60*(4), 596-603.

[27] Pham, S. T., Koniuch, N., Wynne, E., Brown, A., & Collins, S. M. (2025). Microscopic crystallographic analysis of dislocations in molecular crystals. *Nature Materials, 24*(5), 682-687. Retrieved from http://dx.doi.org/10.1038/s41563-025-02138-5. doi:10.1038/s41563-025-02138-5

[28] Rigatos, G., Abbaszadeh, M., Sari, B., Siano, P., Cuccurullo, G., & Zouari, F. (2023). Nonlinear optimal control for a gas compressor driven by an induction motor. *Results in Control and Optimization, 11*, 100226. Retrieved from https://www.sciencedirect.com/science/article/pii/S2666720723000280. doi:https://doi.org/10.1016/j.rico.2023.100226

[29] Shin, S.-P., Lee, S.-Y., & Le, T. H. M. (2025). Feasibility of EfficientDet-D3 for Accurate and Efficient Void Detection in GPR Images. *Infrastructures, 10*(6), 140. Retrieved from https://www.mdpi.com/2412-3811/10/6/140.

[30] Tsai, T.-H., D'Haenens-Johansson, U. F. S., Smith, T., Zhou, C., & Xu, W. (2024). Multi-excitation photoluminescence spectroscopy system for gemstone analysis. *Optics Express, 32*(14), 24839. Retrieved from http://dx.doi.org/10.1364/oe.525832. doi:10.1364/oe.525832

[31] von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., . . . Schuecker, J. (2021).

Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Transactions on Knowledge and Data Engineering*, 1-1. Retrieved from http://dx.doi.org/10.1109/tkde.2021.3079836. doi:10.1109/tkde.2021.3079836

[32] Wallace, P. J., Plank, T., Bodnar, R. J., Gaetani, G. A., & Shea, T. (2021). Olivine-Hosted Melt Inclusions: A Microscopic Perspective on a Complex Magmatic World. *Annual Review of Earth and Planetary Sciences*, *49*(1), 465-494. Retrieved from http://dx.doi.org/10.1146/annurev-earth-082420-060506. doi:10.1146/annurev-earth-082420-060506

[33] Wenqian, F., Fengxia, Z., Quanbin, D., & Qinghai, W. (2024). Diamond particle clarity detection method based on CBAM-ResNet50. *金刚石与磨料磨具工程, 44*(5), 588-598.

[34] Zhan, W., Sun, C., Wang, M., She, J., Zhang, Y., Zhang, Z., & Sun, Y. (2022). An improved Yolov5 real-time detection method for small objects captured by UAV. *Soft Computing*, *26*(1), 361-373. Retrieved from https://doi.org/10.1007/s00500-021-06407-8. doi:10.1007/s00500-021-06407-8

[35] Zhang, S., & Xie, M. (2024). MIPANet: optimizing RGB-D semantic segmentation through multi-modal interaction and pooling attention. *Frontiers in Physics, 12*. Retrieved from http://dx.doi.org/10.3389/fphy.2024.1411559. doi:10.3389/fphy.2024.1411559

[36] Zhu, D., Liu, Y., Yao, X., & Fischer, M. M. (2021). Spatial regression graph convolutional neural networks: A deep learning paradigm for spatial multivariate distributions. *GeoInformatica, 26*(4), 645-676. Retrieved from http://dx.doi.org/10.1007/s10707-021-00454-x. doi:10.1007/s10707-021-00454-x

[37] Zouari, F., Saad, K. B., & Benrejeb, M. (2012). Robust neural adaptive control for a class of uncertain nonlinear complex dynamical multivariable systems. *International Review on Modelling and Simulations, 5*(5), 2075-2103.

[38] Zouari, F., Saad, K. B., & Benrejeb, M. (2013, 18-21 March 2013). *Adaptive backstepping control for a class of uncertain single input single output nonlinear systems.* Paper presented at the 10th International Multi-Conferences on Systems, Signals & Devices 2013 (SSD13).