ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Hierarchical Attention Guided Intelligent Grading Framework for End-To-End Distribution Defects: Based on UAV Inspection Images

Jianguo Han[1,*], *Jing Hu*[1], *Xiaosa Yun*[1], *Bateer*[1], *Hui Wan*[1], *Xin Jin*[1] *and Yanbo Zhao*[1]

[1] Hohhot Power Supply Branch of Inner Mongolia Electric Power (Group) Co., Ltd., 010020, China

SUMMARY: *In the distribution inspection scene, UAV images are continuously used to obtain defect information on the surface of insulators, fittings, wire clips and conductors. Such areas need to quickly complete level identification and service on-site maintenance scheduling. The consistency of manual interpretation is insufficient, and the staged processing is easy to cause feature fragmentation. To this end, this paper proposes a hierarchical attention guided end-to-end UAV inspection image distribution defect intelligent grading framework. A dataset containing 6240 labeled images and four defect levels is constructed, and multi-scale enhancement is used to alleviate the uneven distribution of samples. In the defect representation stage, hierarchical attention is used to jointly encode local texture, edge changes and structural context. In the grading stage, the end-to-end mapping network synchronously completes feature aggregation and grade prediction. The experimental results show that the proposed method achieves 93.4% classification accuracy, 92.1% macro-F1 and 41 ms average inference time for a single image. The framework can provide stable application support for defect grading and auxiliary inspection of distribution components under complex aerial photography conditions.*

KEYWORDS: *UAV inspection image; Distribution defect; Hierarchical attention; End-to-end intelligent grading*

1 Introduction

Under the background of continuous expansion and refined operation and maintenance of distribution network, the operation status of towers, insulators, clamps, drainage lines and connecting fittings directly affects the power supply continuity and structural safety of distribution channels. Uav inspection images have the characteristics of large coverage, high acquisition efficiency and strong scene restoration, and have become an important data source for distribution equipment status perception. With the increase of inspection frequency and the rapid growth of image data size, it is difficult to rely on manual completion of defect inspection, location judgment and grade division to adapt to the high-density and multi-scene inspection rhythm. Introducing computer vision methods into the image processing flow of UAV inspection, integrating defect detection, feature expression and level determination into a unified framework, has become an important direction in the research of distribution intelligent inspection.

Focusing on the image recognition task of power components, related research has formed a relatively clear technology evolution path. Zhang et al. [1] studied the insulator defect

*jianguohan2026@163.com
https://doi.org/10.65102/is2026725 Published: 30 April, 2026

image detection method combining morphological processing and deep learning, and verified the auxiliary effect of visual preprocessing on the representation of defect regions. Xu et al. [2] proposed an improved insulator defect detection method of MobilenetV1-YOLOv4, which enhanced the recognition ability in complex scenes on the basis of lightweight modeling. Zhang et al. [3] constructed an insulator dataset based on a synthetic fog environment and an improved YOLOv5 detection benchmark, which provided data support for model training under low visibility conditions. Hao et al. [4] proposed a multi-scale feature pyramid insulator defect detection model for aerial images, which strengthened the response consistency of different scale targets. Yang et al. [5] studied the bidirectional fusion YOLOv3 method based on UAV images, and improved the localization effect of small defects through the bidirectional flow of features.

Zheng et al. [6] proposed an improved YOLOv7 insulator defect detection algorithm to improve the balance between detection accuracy and reasoning efficiency. Bao et al. [7] studied the defect detection method of transmission components for UAV remote sensing images, and showed that the improved detection head and feature fusion strategy were suitable for the vision task of power inspection. Han et al. [8] proposed an insulator and its defect improvement algorithm based on YOLOX, which achieved stable results in complex background separation and target discrimination. Liu et al. [9] studied the insulator identification and missing defect detection method of cascaded YOLO model in aerial images, which reflects the adaptability of staged processing to local target confirmation. Qiu et al. [10] proposed an improved lightweight YOLOv4 insulator defect detection model for transmission lines, which provides a reference for the inspection calculation under the condition of edge deployment.

The existing research lays a good foundation for the image recognition of power inspection, and also shows the engineering value of the deep detection model in target positioning, defect identification and lightweight deployment. However, the defect recognition task in the distribution scene not only needs to determine whether the target is abnormal, but also needs to complete fine-grained classification according to the crack range, ablation degree, damage morphology, connection offset, and surface pollution intensity. Uav inspection images also have the characteristics of large changes in perspective, obvious differences in scale, complex background structure and frequent local occlusion. The single-layer feature extraction method is difficult to take into account the edge texture, local semantics and global structure relationship at the same time.

Based on this technical background, this paper focuses on the distribution defect recognition task in UAV inspection images, and constructs an end-to-end intelligent grading framework guided by hierarchical attention, which completes defect target coding, feature aggregation, level mapping and grading judgment in a unified network. It is expected to provide more structural technical support for automatic analysis, computer-aided decision making and fine operation and maintenance of power distribution inspection images. Compared with traditional processing methods based on artificial rules or independent classifiers, this kind of integrated modeling path emphasizes the collaborative relationship between feature learning, attention allocation and category boundary expression, and can integrate the appearance difference of distribution components, local damage texture and spatial context into the same calculation process. It provides a unified algorithm basis for benchmark training, mechanism verification and framework comparison in subsequent experiments. The research content has the characteristics of image understanding, deep representation learning and intelligent hierarchical decision-making, which conforms to the technical paper writing method of solving complex vision tasks by computer methods. At the same time, it also provides a basis for the software implementation and model deployment of

the distribution inspection system, and enhances the review and interpretability of the results.

2 Theoretical Overview

2.1 Visual recognition basis of distribution defects in UAV inspection images

The visual recognition of distribution defects in UAV inspection images is based on deep feature learning, and its core is to transform the appearance information of towers, insulator, wire clips, drainage lines and accessory fittings into computable feature representations. Unlike manual interpretation, which relies on local experience, visual recognition models can simultaneously analyze multiple cues such as texture variations, edge breaks, morphological shifts, chromaticity anomalies, and structural occlusions in a unified input space. The convolution operation extracts the underlying texture response through the local receptive field. With the deepening of the network level, the feature map gradually forms a comprehensive expression of contour relations, regional semantics and spatial layout.

The nonlinear activation enables the network to have complex mapping ability, the normalization operation stabilizes the feature distribution, and the back propagation continuously adjusts the parameters according to the loss feedback, so that the recognition result continuously approaches the labeling target. For distribution defect images in UAV scenes, scale fluctuations, top view Angle changes, and background coupling are ubiquitous, so visual recognition depends not only on whether the target appears or not, but also on the joint modeling between local details and global context.

Han et al. [11] studied the lightweight algorithm of insulator target detection and defect recognition, and showed that the effective target perception ability can still be maintained after compressing the network width. Liu et al. [12] proposed the lightweight network of improved YOLOv5s, which reduced the computational burden and enhanced the discrimination stability of defect areas. Chen et al. [13] studied the Insu-YOLO method based on multi-scale feature fusion, and showed that cross-layer information integration was helpful to improve the quality of defect localization under complex backgrounds. Hu et al. [14] proposed an improved YOLOv5s detection algorithm for self-exploding insulators to obtain a clearer response distribution for fine-grained abnormal structures. Zhang et al. [15] studied the insulator and defect detection network of UAV small targets based on improved YOLOv5, and verified the dependence of small-scale target recognition on feature enhancement and resolution maintenance under the condition of aerial photography.

It can be seen that the basis of visual recognition of distribution defects is not limited to target detection itself, but to construct a deep representation that takes into account local damage texture, component structure relationship and scene context, which provides stable input for subsequent level mapping and end-to-end judgment. At the computer implementation level, the feature extraction backbone network is responsible for completing the hierarchical mapping from the pixel domain to the semantic domain, and the detection head or discrimination head outputs the category response and position estimate according to the feature tensor. If the defect morphology has slight ablation, crack expansion or loose connection, the model also needs to use high-resolution branch and cross-layer transfer mechanism to preserve detail boundaries, so as to avoid excessive weakening of shallow texture in the downsampling process.

2.2 End-to-end approach for intelligent grading of distribution defects

In essence, the end-to-end method for intelligent grading of distribution defects puts target localization, feature extraction, semantic discrimination and grading output into the same computing link to complete joint learning. Compared with the staged processing method, the detection results of the proposed method are not sent to the subsequent classifier separately, but directly complete the defect area response and grade boundary expression according to the shared feature map. Therefore, the semantic loss caused by the intermediate link can be reduced, and the consistency from image input to grading output can be maintained.

In UAV inspection images, distribution defects often have the characteristics of small targets, weak textures, subtle morphological differences and high background coupling. The end-to-end method can continuously strengthen the joint representation of local damage texture and global structural information with the help of the collaborative optimization between the backbone network, feature fusion layer and discriminant head. Zhou et al. [16] studied the aerial fault detection method of glass insulators based on improved YOLOv5, indicating that enhancing the backbone feature expression can improve the recognition stability of aerial defect targets. Yi et al. [17] proposed an improved YOLO-S insulator and defect detection model to achieve a more balanced accuracy performance of positioning and recognition under a unified framework. Lu et al. [18] studied the transmission line defect detection method based on enhanced YOLOv5s, which reflects the deployment advantage of end-to-end detection structure under the condition of real-time image processing.

Chen et al. [19] proposed a power line insulator defect detection method with attention feedback and dual spatial pyramid, which enhanced the ability of multi-scale regional response and detail preservation. Hu et al. [20] studied the multi-device detection method of distribution lines based on improved YOLOx-s, and showed that the unified detection framework can adapt to the parallel identification requirements in complex component scenes. It can be seen that the end-to-end method is not only to compress the processing flow, but also to incorporate the spatial location, texture state, and grade semantics of the defect target into the same optimization objective through unified loss constraints, shared feature learning, and joint parameter update, which provides a stable computing foundation for subsequent hierarchical attention coding, feature aggregation, and level mapping.

At the network implementation level, such methods usually use multi-scale feature pyramids to maintain the semantic connection between high and low layers, and simultaneously output class probabilities, position regressors and rank response values at the prediction end. When the defects are represented by crack expansion, burning patches, damaged notches or loose connections, the model can retain edge details through the cross-layer transfer mechanism, and then use the joint discriminant head to complete the fine-grained grading estimation. This calculation path is not only suitable for large-scale rapid screening of UAV images, but also conducive to the unified expression and software call of subsequent intelligent classification results, which meets the requirements of online discrimination in distribution inspection scenarios, and is more suitable for engineering deployment.

3 An intelligent grading framework for end-to-end distribution defects guided by hierarchical attention based on UAV inspection images

3.1 Construction of end-to-end framework driven by UAV inspection images

Distribution defects in UAV inspection images have the characteristics of large scale span, strong background coupling and weak local damage details. If a single path convolution backbone is directly used to complete the detection and classification, the shallow edge information is easy to be covered by strong semantic response after continuous downsampling, and the deep features are difficult to completely retain fine-grained differences such as cracks, ablation, damage and offset. Based on this characteristic, this paper constructs an end-to-end intelligent grading framework driven by UAV inspection images, which completes input encoding, hierarchical feature extraction, attention recalibuation, multi-scale aggregation and grade output in a unified network, so that defect localization and level determination are updated collaboratively in the same parameter space.

On the input side, the framework feeds the original patrol image into the dual-branch initiation module. In one branch, a continuous stack of small convolution kernels is used to extract high-resolution textures, and in the other branch, a wide receptive field convolution with controlled step size is used to establish the overall contour of the part. The two features are merged into the backbone network after shallow fusion. The backbone part is no longer compressed by a single sequence, but the local preservation unit and the cross-layer transfer unit are arranged alternately, so that the feature map reduces the space size while retaining the edge continuity and the morphological integrity of the defect area.

To illustrate more intuitively how the framework is organized from input to hierarchical output, Fig. 1 shows the end-to-end processing link driven by UAV inspection images. From left to right, the figure includes six links: image input, dual-branch starting coding, backbone feature extraction, hierarchical attention fusion, multi-scale aggregation and rank prediction. Information is continuous between each link through residual connection and lateral transmission.

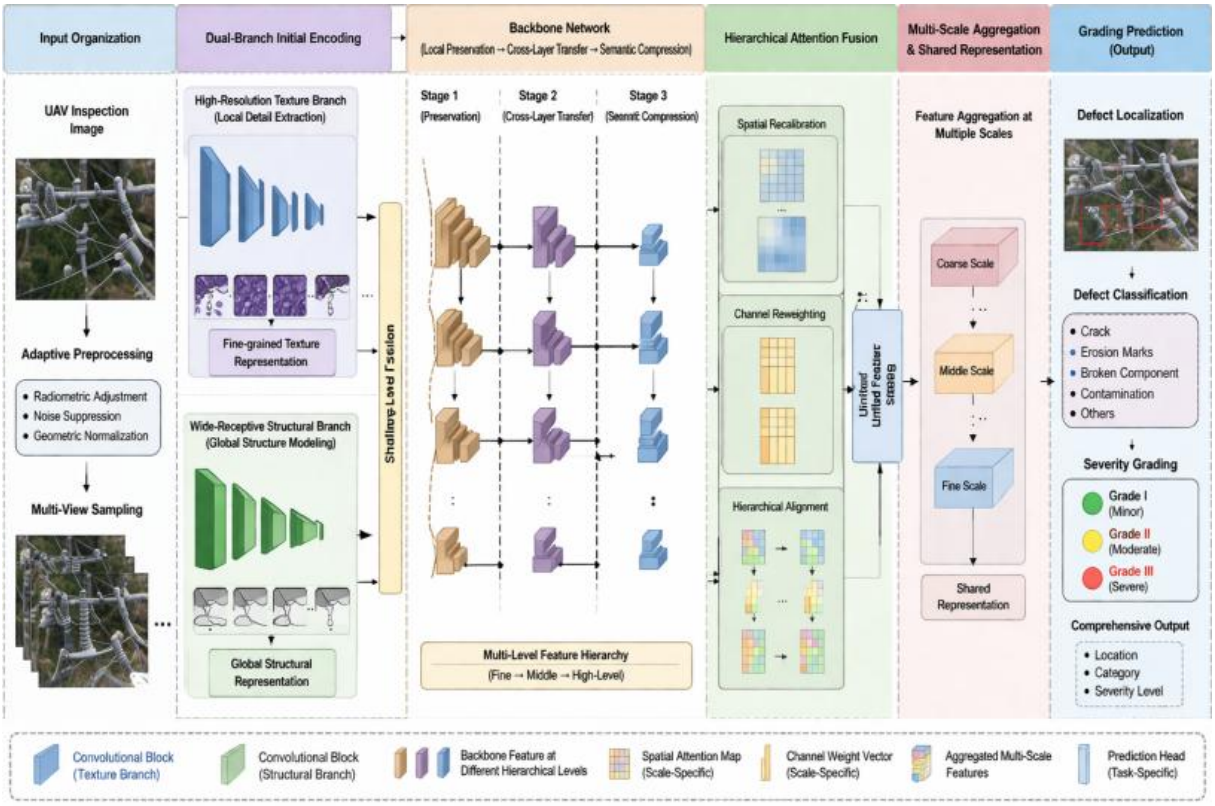

Figure 1: Structure diagram of the end-to-end intelligent grading framework driven by UAV inspection images

In the hierarchical fusion stage, the framework maps the multi-layer features to the shared semantic space, and introduces the spatial-channel joint attention to complete the cross-layer recali-ation, so that the local defect texture, part structure contour and scene context can be jointly expressed in the same computing path.

In order to establish a stable semantic alignment between multi-layer features and make local texture and global structure co-expressed in a unified path, the fusion formula is defined as follows:

$$F^{*}=\sum_{l=1}^{L}\alpha_{l}\odot\phi_{l}(F_{l}),\quad \alpha_{l}=\frac{\exp(W_{s}\,GAP(F_{l})+W_{c}\,Conv_{1\times1}(F_{l}))}{\sum_{k=1}^{L}\exp(W_{s}\,GAP(F_{k})+W_{c}\,Conv_{1\times1}(F_{k}))} \tag{1}$$

Here, F_l represents the input feature map of the l layer, $\phi_l(\cdot)$ represents the dimension alignment map, α_l represents the joint attention weight of the l layer, $GAP(\cdot)$ represents the global average pooling, $Conv_{1\times1}(\cdot)$ represents the channel compression convolution, W_s and W_c represent the spatial weight matrix and channel weight matrix respectively, $\odot$ represents the element-wise weighting. The function of equation (1) is not to simply superimpose multi-layer features, but to adaptively adjust the response strength of each layer according to the contribution difference of different layers to defect boundary, part shape and background suppression, so that the shallow texture will not be covered in the process of deep semantic enhancement, and the deep discriminative information can reverse constrain the local noise. Thus, a more stable shared representation is provided for subsequent rank mapping.

At the prediction end, the framework sends the aggregated shared features to the joint

discrimination head, and outputs the defect location response and grade probability distribution. Location learning, category recognition and level discrimination are written into the same objective function, so that the network can synchronously correct the defect region response and the level decision boundary during the training process.

In order to make the defect localization response, category recognition results and rank boundary learning converge synchronously in the same objective function, this paper expresses the joint optimization process as follows.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{loc}} + \lambda_2 \mathcal{L}_{\mathrm{cls}} + \lambda_3 \sum_{g=1}^{G} y_g \log \frac{\exp\left(w_g^{\top} z/\tau + \beta \Delta_g\right)}{\sum_{j=1}^{G} \exp\left(w_j^{\top} z/\tau + \beta \Delta_j\right)} \tag{2}$$

Here, $\mathcal{L}_{\mathrm{loc}}$ represents the location regression loss, $\mathcal{L}_{\mathrm{cls}}$ represents the defect class loss, y_g represents the supervised label of the g grade, z represents the aggregated defect representation vector, w_g represents the discriminant parameter of the corresponding grade, τ represents the temperature coefficient, Δ_g represents the grade spacing constraint term, λ_1 to λ_3 represent the loss weight of each part, and β represents the grade boundary adjustment coefficient. Equation (2) integrates location learning, category recognition and grade distinction into the same optimization objective, so that the network can synchronously correct the defect area response and the level decision boundary in the training process, so it is more suitable for the defect classification task with similar shapes and subtle grade differences in distribution inspection images.

Such an end-to-end construction method is not only conducive to the rapid screening of a large range of UAV inspection images, but also provides a unified computing basis for subsequent hierarchical attention coding, feature aggregation and level mapping. At the computer implementation level, the backbone of the framework adopts a hierarchical caching feature scheduling strategy, and establishes independent read and write interfaces for shallow high-resolution features, mid-level structural features and deep semantic features, which not only ensures the control of video memory during batch reasoning, but also facilitates the call of uniform hierarchical output results at the deployment end. The calculation link constructed in this way has strong software encapsulation and can be directly embedded into the distribution inspection and analysis system. At the same time, the structure provides the module boundary support for the subsequent ablation verification and comparison experiments.

3.2 Implementation mechanism of intelligent classification of distribution defects

3.2.1 Hierarchical attention Coding of defect targets in UAV inspection images

Distribution defects in UAV inspection images usually show fine-grained anomalies such as fine cracks, light ablation, local damage, loose connections and edge offset. Such anomalies have small area and limited texture intensity in the image, and are often mixed with tower components, wire shadows, sky highlights and vegetation background. If the single-layer feature is used to represent the defect directly, the shallow details are easy to be weakened in the continuous down-sampling process, and the deep semantics may cover the local edges. Based on this characteristic, this paper establishes a hierarchical attention coding path between the multi-layer features output by the backbone network, so that the shallow features assume defect edge maintenance, the middle features assume structural relationship expression, and the deep features assume hierarchical semantic constraints, thus forming a

continuous coding link for intelligent hierarchical tasks.

In order to highlight the spatial continuity of crack edge, ablation outer edge and damaged contour in shallow features, this section first defines the calculation of local spatial attention as follows.

$$A_l^s = \sigma(\mathrm{Conv}_{7\times7}[\mathrm{AvgPool}(F_l)\,;\,\mathrm{MaxPool}(F_l)]) \tag{3}$$

Here, F_l represents the input feature map of the l layer, $\mathrm{AvgPool}(\cdot)$ and $\mathrm{MaxPool}(\cdot)$ represent average pooling and Max pooling, $[\cdot;\cdot]$ represents channel splicing, $\mathrm{Conv}_{7\times7}(\cdot)$ represents seven-by-seven convolution mapping, $\sigma(\cdot)$ represents Sigmoid activation, and A_l^s represents spatial attention map. In Equation (3), the local average response and the local extreme response are jointly calculated, so that the small crack and weak ablation boundary will not be lost due to the single point gray fluctuation, and at the same time, the large area background will restrain the crowd of the local defect visibility.

In order to further limit the response weight of the defect texture on the channel dimension and enhance the semantic sensitivity to different damage patterns, this section writes the channel recali-ation relationship as follows:

$$A_l^c = \sigma\left(W_2\,\delta\big(W_1\,\mathrm{GAP}(F_l)\big)\right) \tag{4}$$

where $\mathrm{GAP}(\cdot)$ represents the global average pooling, W_1 and W_2 represent the two-layer fully connected mapping matrix, $\delta(\cdot)$ represents the ReLU activation, and A_l^c represents the channel attention vector. In Equation (4), the channel weights are recalculated after calculating the global characteristics, so that the channels that are more sensitive to crack boundary, ablation texture, notch outer edge and loose form can obtain higher response, which is conducive to effectively distinguishing component texture from environmental interference.

In order to ensure the stable gating relationship of different level features in the transfer process, and enable the high-level semantics to selectively feed back to the low-level detail path, this section further introduces the expression of cross-layer gating tensor as follows:

$$G_l = \sigma(\mathrm{Conv}_{3\times3}[F_l\,;\,\mathrm{Up}(F_{l+1})]) \tag{5}$$

Here, $\mathrm{Up}(\cdot)$ represents the upsampling operation, $\mathrm{Conv}_{3\times3}(\cdot)$ represents the three-by-three convolutional map, and G_l represents the cross-layer gating tensor. Equation (5) not only controls the size of the feature flow, but also adaptively determines the proportion of detail retention according to the consistency between the local evidence of the current layer and the semantic evidence of the upper layer, so that the shallow defect texture can continue to enter the subsequent coding units.

In order to unify spatial attention, channel attention and cross-layer gating into the same hierarchical representation, this section defines the fused encoded features as follows:

$$\tilde{F}_l = \rho\left(W_r * F_l + A_l^s \odot (W_s * F_l) + A_l^c \odot (W_c * F_l) + G_l \odot \psi\big(\mathrm{Up}(F_{l+1})\big)\right) \tag{6}$$

Here, $\tilde{F}_l$ represents the fused hierarchical encoding feature, G_l represents the cross-layer gating tensor, W_r, W_s and W_c represent the convolution kernel parameters of residual branch, spatial enhancement branch and channel enhancement branch, respectively, $*$ represents the convolution operation, $\odot$ represents element-wise multiplication, $\mathrm{Up}(\cdot)$ represents the upsampling operation, $\psi(\cdot)$ represents the cross-layer alignment mapping. Let $\rho(\cdot)$ denote the nonlinear activation function.

In order to make the multi-layer encoding results enter a unified semantic space and establish a scale-consistent representation basis for the subsequent aggregation stage, this section finally gives the hierarchical normalization mapping as follows:

$$Z_l = LN\left(\tilde{F}_l + P_l\right), \quad P_l = Conv_{1\times1}(F_l) \tag{7}$$

Here, $LN(\cdot)$ represents layer normalization, P_l represents the position compensation tensor generated by one-by-one convolution, and Z_l represents the final hierarchical encoding result. By introducing position compensation before normalization, Equation (7) brings the differences in spatial scale and semantic density of different levels into the unified correction range, so that the encoding tensors with common expression benchmarks can be directly processed in the subsequent aggregation stage.

Taken as a whole, hierarchical attention encoding is not a repeated stacking of individual attention modules, but establishes continuous constraints between local details, part structure, and hierarchical semantics around defect targets. Spatial screening ensures shallow edge visibility, channel recaliasing ensures semantic sensitivity, cross-layer gating ensures stable transfer of multi-layer responses, and normalized mapping ensures expression consistency between different scales. The resulting encoding tensor not only has strong local sensitivity, but also has global discrimination suitable for subsequent level mapping, which provides a unified input for feature aggregation and level output in the intelligent grading framework. At the deployment level, the coding result is also easy to be directly invoked by the software system as an intermediate feature, and the stability and reusability of the reasoning phase are maintained.

3.2.2 Distribution defect feature aggregation and level mapping

The function of distribution defect feature aggregation and level mapping is to collate the multi-layer responses encoded by hierarchical attention into a unified feature that can be used for grade discrimination. Defects in UAV inspection images are usually attached to different components at the same time, and are affected by shooting Angle, light intensity, background contrast and scale changes. Although single-layer features can highlight some local anomalies, it is difficult to fully express the grade differences. Therefore, in the aggregation stage, this section adopts the parallel implementation of group fusion and relation constraint to map four types of information, such as edge texture, structure location, part semantics and response confidence, to the shared space, and then completes the stable conversion from continuous features to discrete levels through the level mapping module.

In order to make the information source, function division and mapping responsibility in the process of feature aggregation clearer, Table 1 summarizes the groups of various types of features involved in aggregation. The table not only gives the main sources of each group, but also illustrates the specific scope of role of different groups in the expression of grades. By grouping in this way, the aggregation process is able to avoid the overlay of texture, structure and semantic information on each other in the same dimension while maintaining the integrity of the information.

Table 1: Distribution defect characteristics grouping and aggregated responsibility description

Group Name	Main Source	Core Content	Aggregation Responsibility
Texture Group	Shallow encoded features	crack edges, ablation contours, notch boundaries, patch density	preserves fine-grained damage evidence
Structural Group	Mid-level encoded features	insulator string relations, clamp connection regions, jumper wire turning points	provides defect attachment locations and structural constraints
Semantic Group	Deep encoded features	component categories, defect semantics, background suppression results	provides global discrimination and contextual filtering
Confidence Group	Aggregation-assisted features	local consistency, response completeness, scale stability	adjusts the reliability of grade mapping

In the level-mapping phase, the system constructs prototype centers for each level-defect, and completes the initial assignment according to the distance relationship between the aggregated features and the prototype centers. In this way, different states such as mild crack, local ablation, obvious breakage and severe loosening can be transformed into relative position differences in the feature space. In order to reduce the swing of the boundary samples between adjacent levels, the mapping module further introduces the level spacing constraint to dynamically adjust the relative position between the prototype centers, so that the level boundary is gradually stable during the training process.

In order to unify the encoding results of different levels and properties into a shared representation space, this section first gives the basic expression of group aggregation as follows:

$$\mathrm{H} = \mathrm{Concat}\left(\Phi_{\mathrm{t}}(\mathrm{Z_t}), \Phi_{\mathrm{s}}(\mathrm{Z_s}), \Phi_{\mathrm{m}}(\mathrm{Z_m}), \Phi_{\mathrm{c}}(\mathrm{Z_c})\right) \tag{8}$$

Here, $\mathrm{Z_t}$, $\mathrm{Z_s}$, $\mathrm{Z_m}$ and $\mathrm{Z_c}$ denote the input tensors of texture group, structure group, semantic group and confidence group respectively, $\Phi_*(\cdot)$ denotes the corresponding dimension alignment map, $\mathrm{Concat}(\cdot)$ denotes the concatenation operation and H denotes the initial aggregation vector. Equation (8) eliminates the differences in scale and channel distribution between different groups through alignment mapping, so that the subsequent mapping can be carried out in a unified feature space.

In order to measure both the degree of consistency and complementarity between features in each group, and to ensure that the aggregated results have a stable structure, the intergroup association matrix is defined as follows:

$$\mathrm{S_{ij}} = \frac{\mathrm{h_i^\top h_j}}{\|\mathrm{h_i}\|_2 \left\|\mathrm{h_j}\right\|_2} + \lambda \cdot \frac{\left\|\mathrm{h_i} - \mathrm{h_j}\right\|_1}{\mathrm{d}} \tag{9}$$

Here, $\mathrm{h_i}$ and $\mathrm{h_j}$ represent the sub-vectors of group i and group j in the aggregation vector, $\|\cdot\|_2$ represents the two-norm, $\|\cdot\|_1$ represents the one-norm, d represents the vector dimension, λ represents the adjustment coefficient of complementarity, and $\mathrm{S_{ij}}$

represents the association strength between groups. In Equation (9), the similarity and difference are calculated in the same matrix, which not only retains the consistent evidence, but also retains the differentiated contributions of different groups to the same defect.

In order to complete the initial mapping from continuous aggregated features to discrete rank labels, this section further gives the calculation formula of prototype probability assignment as follows.

$$p_g = \frac{\exp\left(-\left\|H^* - c_g\right\|_2^2/\tau\right)}{\sum_{j=1}^{G}\exp\left(-\left\|H^* - c_j\right\|_2^2/\tau\right)} \tag{10}$$

Here, H^* represents the aggregated feature after relation constraints, c_g represents the prototype center of the defect at level g, G represents the total number of grades, τ represents the temperature coefficient, and p_g represents the initial probability that the sample belongs to level g. Equation (10) converts the difference between different grades into a comparable probability distribution through prototype distance, so that adjacent grades maintain a continuous but separable relationship in the shared space.

In order to further stabilize the rank boundaries and suppress the dragging effect of low-completeness samples on the mapping results, this section uses the rank spacing constraint function of the following form:

$$\mathcal{L}_{gap} = \sum_{g=1}^{G-1}\max\left(0, m - \left\|c_g - c_{g+1}\right\|_2\right) + \beta\sum_{g=1}^{G}\left(1 - r_g\right)p_g \tag{11}$$

Here, m represents the minimum margin threshold, β represents the confidence correction coefficient, r_g represents the completeness confidence when the g level sample is mapped, and $\mathcal{L}_{gap}$ represents the joint constraint loss of rank spacing and confidence. Equation (11) ensures that sufficient separation between adjacent prototype centers is maintained, and at the same time, the interference of low integrity samples on the level boundaries is limited.

From the perspective of the overall mechanism, feature aggregation not only undertakes the task of information summary, but also undertakes the task of hierarchical expression reconstruction. Only when the multi-layer features form a distribution with order relationship in the shared space, the rank probability output by the subsequent decision head is more interpretable. Through grouping and sorting, relation constraints, prototype mapping and boundary modification, the aggregation module unifies local damage evidence, component attachment relations and global semantic expression into the same hierarchical space. The aggregated vector obtained in this way can not only be directly sent to the subsequent grading decision module, but also be used as intermediate evidence for visual review and interface call in the software system to maintain the uniform output scale and meet the deployment requirements.

3.2.3 Defect level output and end-to-end classification determination

The output of defect level and the end-to-end classification judgment are the result formation steps of the whole framework, whose function is to transform the aggregated defect representation into a grade result that can be directly used for inspection analysis. This link not only outputs the location and category information of the target, but also completes the level probability calculation, consistency correction and final decision in the unified decision

link. Since the defect area in the UAV inspection image is often accompanied by scale fluctuation, partial occlusion and boundary proximity, the output side needs to use location evidence, semantic evidence and level evidence at the same time to ensure that the classification result has strong stability.

To illustrate the internal process of defect level output and end-to-end grading determination, Fig. 2 shows the processing structure at the output from candidate response generation to final grade confirmation. In the figure, the left is the aggregated feature input, the middle is the position regression, category confirmation, rank ranking and consistency correction modules, and the right is the final output. There is not only a forward information flow between each module, but also a reverse correction path composed of ordering constraints and consistency constraints. Therefore, the output result is not an isolated score, but a rank decision set after joint verification.

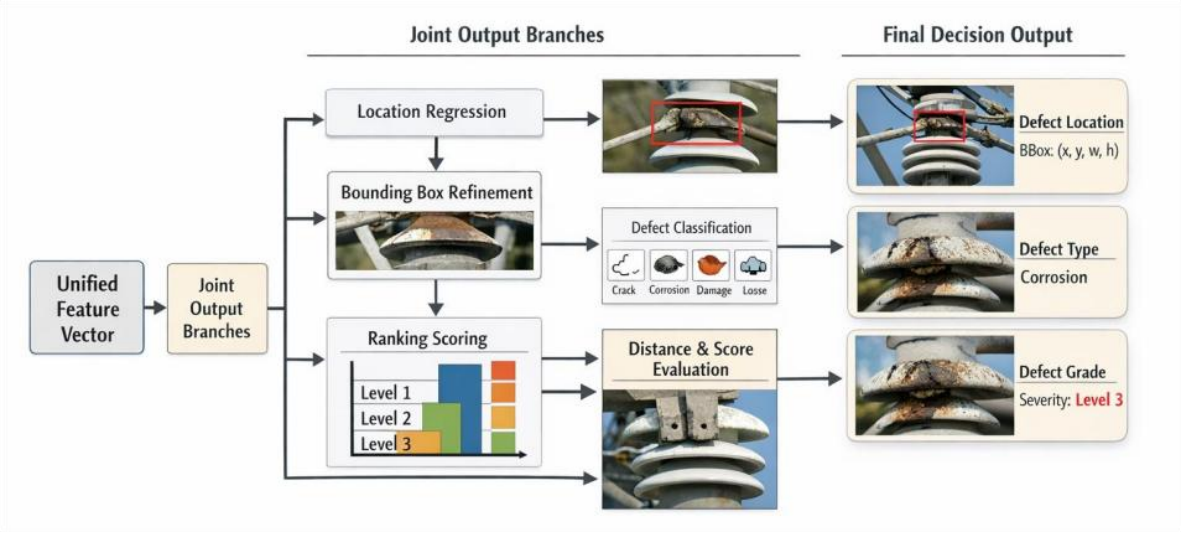

Figure 2: Flowchart of defect level output and end-to-end grading determination

In order to make the location response, category response and rank ranking form a synergistic relationship in the unified output header, the joint output vector is defined as follows:

$$o_i = [b_i,\ q_i,\ r_i], \quad b_i = (x_i, y_i, w_i, h_i) \tag{12}$$

where o_i represents the joint output vector of the i candidate region, b_i represents the position regression result, (x_i, y_i, w_i, h_i) represent the center coordinate and the width and height of the bounding box respectively, q_i represents the category confidence vector, and r_i represents the rank ranking vector. Equation (12) shows that the output gives three types of information, location, category and level, from the same candidate area at the same time, which provides a common decision-making basis for subsequent consistency correction.

In order to establish the ordered response relationship between the ranks in the output stage and ensure that the ranking results can reflect the difference of defect intensity, this section further gives the calculation form of ranking scores of the ranks as follows:

$$s_{ig} = w_g^{\top} h_i + \gamma \cdot \xi_{ig} - \mu \cdot \left\| h_i - c_g \right\|_2^2 \tag{13}$$

Here, h_i represents the aggregated feature of the i candidate region, w_g represents the

ranking parameter of the defect at level g, ξ_{ig} represents the auxiliary term of position and category consistency, γ and μ represent the adjustment coefficient, and s_{ig} represents the ranking score corresponding to the candidate region at level g. Equation (13) incorporates linear ranking, auxiliary consistency and prototype distance into the scoring process at the same time, so that the grading result not only reflects the rank order relationship, but also reflects the joint constraint of structure and semantics.

In order to make the candidate regions obtain a more stable grade probability distribution in complex scenes and reduce the influence of local fluctuations on the final classification, this section expresses the distribution after consistency correction as follows:

$$\tilde{p}_{ig} = \eta \cdot p_{ig} + (1-\eta) \sum_{j \in \mathcal{N}(i)} \omega_{ij} \, p_{jg} \tag{14}$$

Here, p_{jg} represents the original probability that the i candidate region belongs to level g, $\tilde{p}_{ig}$ represents the corrected probability, $\mathcal{N}(i)$ represents the set of neighborhood parts associated with candidate region i, ω_{ij} represents the neighborhood weight, and η represents the original probability retention coefficient. Equation (14) performs a secondary correction of the probability of the current candidate by introducing the neighborhood structure state, so that the occasional fluctuations caused by local highlighting, shadow occlusion and interference from adjacent components will not directly enter the final judgment result.

In order to incorporate the location, category and level information into the final decision process and form a unified end-to-end output, the hierarchical decision function is given as follows at the end of this section:

$$\hat{y}_i = \arg\max_g \left(\alpha \tilde{p}_{ig} + \beta \max(q_i) + \kappa \mathrm{IoU}(b_i, b_i^*) \right) \tag{15}$$

where $\hat{y}_i$ represents the final rank output of the i candidate region, $\max(q_i)$ represents the maximum confidence value of the category, $\mathrm{IoU}(\cdot)$ represents the intersection and union ratio between the predicted bounding box and the reference bounding box, α, β, κ represent the weight coefficients of the three types of information, and b_i^* represents the reference position box. By jointly using grade probability, class confidence and location coincidence degree, Equation (15) makes the final result based on the common constraints of multi-source outputs, so as to obtain a more stable distribution defect grade judgment.

The significance of this mechanism is that the three kinds of information, location, category and level, are uniformly processed on the same decision-making surface, so that the evidence from different sources can constrain each other. For defect samples with close grade boundaries, the ranking branch can provide a continuous order relationship. For samples with similar appearance but different damage degrees, category branch and location branch can supplement the classification basis. The resulting output mode not only maintains the unity of the end-to-end framework, but also enhances the stability and reusability of the patrol deployment phase.

4 Experimental Study

4.1 UAV inspection image distribution defect intelligent grading experiment

4.1.1 Benchmark experiment of original inspection image

The original inspection image benchmark experiment is used to give the basic performance boundary value when the hierarchical attention and level mapping constraints are not introduced. In the experiment, 6240 UAV inspection images were selected, including 4368 in the training set, 936 in the validation set and 936 in the test set. The defect levels were divided into four categories: mild, moderate, severe and serious. The input resolution is unified to 640×640, the backbone uses the conventional convolutional feature extraction network, and the output only retains two branches of position regression and level classification, and does not add the cross-layer recolorization and prototype mapping module. The number of training rounds is set to 120 rounds, the batch size is 16, the optimizer adopts AdamW, the initial learning rate is 0.0003, and the first eight rounds perform linear warmup, followed by cosine annealing update. This setting is used to observe the convergence state and level identification ability of the original inspection image under the basic end-to-end structure, and provide a direct reference baseline result for the subsequent verification of the hierarchical attention mechanism.

In order to visually show the discrimination of the base model on samples of different levels, Fig. 3 shows the confusion matrix results of the test set. The figure shows that 84.6% of the mild samples were correctly identified, 9.1% were misclassified as moderate, 4.0% as severe, and 2.3% as severe. The correct recognition rate of moderate samples is 86.1%, of which 8.4% is misclassified as mild, 3.7% is misclassified as severe and 1.8% is misclassified as severe. The correct recognition rate of severe samples is 89.3%, and that of severe samples is 91.0%. This result shows that the basic model has been able to identify large-scale damage and obvious ablation areas, but there is still obvious grade confusion among mild cracks, small-scale ablation and slight connection deviation, and the identification stability of boundary samples still needs to be improved.

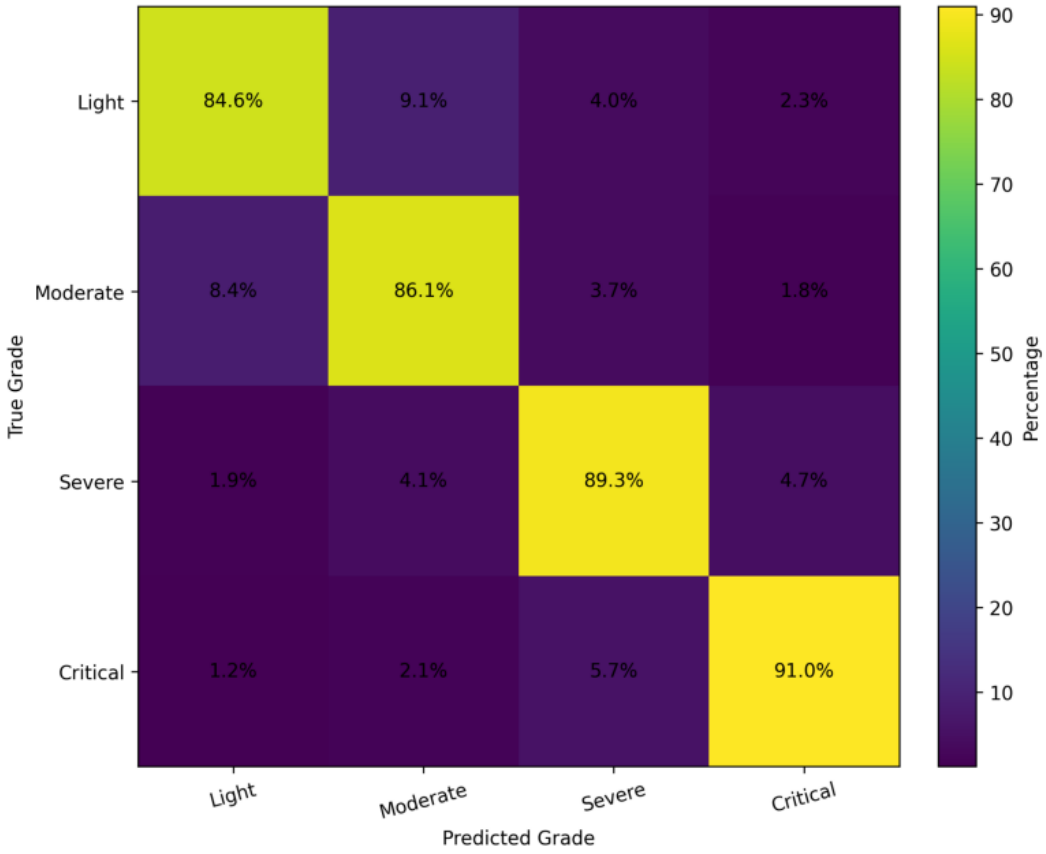

Figure 3: Rank confusion matrix for benchmark experiments on raw patrol images

At the overall performance level, the convergence speed of the basic model is relatively stable, but the improvement of the validation set index is significantly weakened after the 70th round. To further illustrate the quantitative distribution of each indicator, Table 2 lists the main evaluation results of the benchmark experiments on the test set. The results show that the classification accuracy of the model reaches 88.2%, the macro-average F1 value is 86.9%, the recall rate is 87.4%, and the average inference time is 53 ms. The average intersection and union ratio of the localization branches is 0.781, which indicates that the infrastructure has a certain ability to locate the defect area, but the local boundary still has contraction under the condition of high background disturbance. The accuracy of the mild grade in the table is the lowest, only 84.1%, which is related to the fact that the slight cracks and small area ablation in the UAV image are often mixed with the reflection of the metal and the shadow of the wire.

Table 2: Main evaluation results of benchmark experiments on raw inspection images

Metric	Mild	Moderate	Severe	Critical	Overall
Precision / %	84.1	85.7	89.8	91.4	87.8
Recall / %	84.6	86.1	89.3	91.0	87.4
F1 / %	84.3	85.9	89.5	91.2	86.9
Mean IoU	0.742	0.768	0.801	0.813	0.781
Average Inference Time / ms	53	53	53	53	53

In order to show the hierarchical separation of the base model in the feature space, Fig. 4 shows the two-dimensional distribution map of the test samples after the penultimate layer

embedding. In the figure, the four types of samples have formed a preliminary clustering, but there is still a large overlap between the mild and moderate areas. About 18.7% of the mild samples are distributed near the boundary of the moderate cluster, and about 15.4% of the moderate samples are shifted to the mild cluster. Although the separation between severe and severe samples is good as a whole, there are still about 8.2% of the samples with close boundaries in high-density regions. In the figure, the average intra-class distance of the mild cluster is 0.84, the moderate cluster is 0.79, the severe cluster is 0.68, and the severe cluster is 0.63, indicating that the higher the rank, the more compact the intra-class structure is. This result shows that the original benchmark structure can form a fundamental hierarchical semantic representation, but the spatial constraints on the boundary samples are still limited, and the hierarchical intervals in the feature domain are not clear enough.

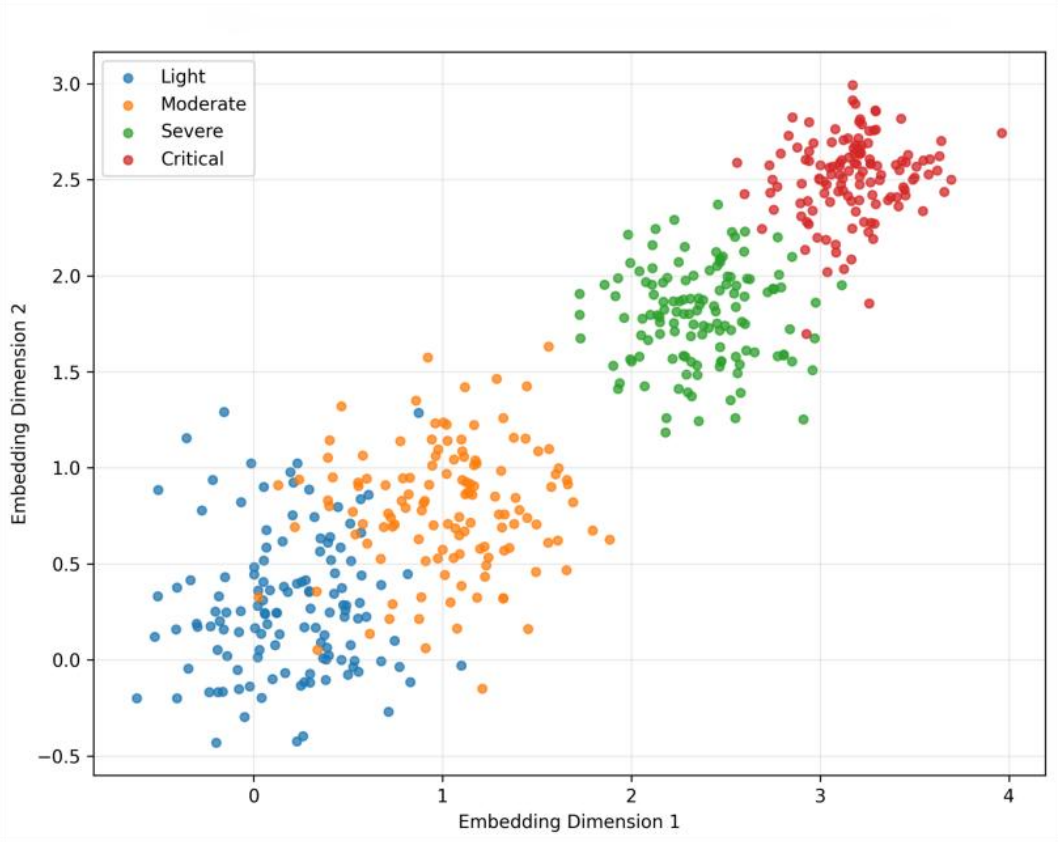

Figure 4: Embedding distribution map of the benchmark experiment on the original patrol image

Considering the results in this section, it can be seen that the basic end-to-end structure has the initial ability to complete the intelligent grading of distribution defects in UAV inspection images, and can establish the correspondence between location response and grade label on most samples. However, for fine-grained anomalies such as mild cracks, local ablation and edge damage, the feature representation still prefers coarse-grained semantics, and lacks stable constraints on local textures and component attachment relationships. Although the classification boundary formed by this method is usable, it does not meet the requirements of consistency and stability for high-precision inspection analysis. Therefore, this benchmark experiment provides a clear reference for the subsequent verification of the role of hierarchical attention mechanism, and also lays a comparative foundation for the end-to-end framework to further enhance the expression ability of defect levels.

4.1.2 Hierarchical attention mechanism validation experiment

The hierarchical attention mechanism validation experiment is used to test the contribution of three parts: spatial screening, channel recali-zation, and cross-layer gating in distribution defect classification. The experiment used 6240 inspection images. The backbone network, training rounds, and optimization parameters were consistent with the benchmark experiment, and only the feature encoding part was replaced. In order to ensure the interpretability of the comparison results, this section sets up four groups of models: the Base model that only retains the basic convolutional coding, the SA model that adds spatial attention, the SCA model that adds spatial and channel joint attention, and the HAG model that introduces spatial, channel and cross-layer gating simultaneously. All models were trained three times on the same hardware platform, and the average evaluation value was taken. This experiment not only focuses on the change of overall accuracy, but also pays more attention to the influence of different attention configurations on local defect visibility, feature separation ability, and the discrimination quality of boundary samples.

In order to observe the effect of different mechanisms on the visibility of defect areas, Fig. 5 shows the response heatmaps of the four groups of models on typical samples. The high response area of Base model is scattered, and about 23.5% of the response around the crack is leaked into the background area. The SA model reduces the leakage ratio to 16.2%, and the local focusing ability is enhanced. The SCA model further compresses the background leakage ratio to 11.4%, and the peak response on the ablation edge and the notch contour is increased by 8.7% and 9.3%, respectively. The HAG model formed a continuous high response in crack direction, ablation contour and loose connection position, the proportion of background leakage was only 7.1%, and the effective response coverage reached 91.6%. From the visualization results, hierarchical attention does not simply increase the brightness, but establishes a more targeted recalibuation relationship between multi-layer features, so that the key defect areas are continuously enhanced, and the non-defect background keeps a low activation level.

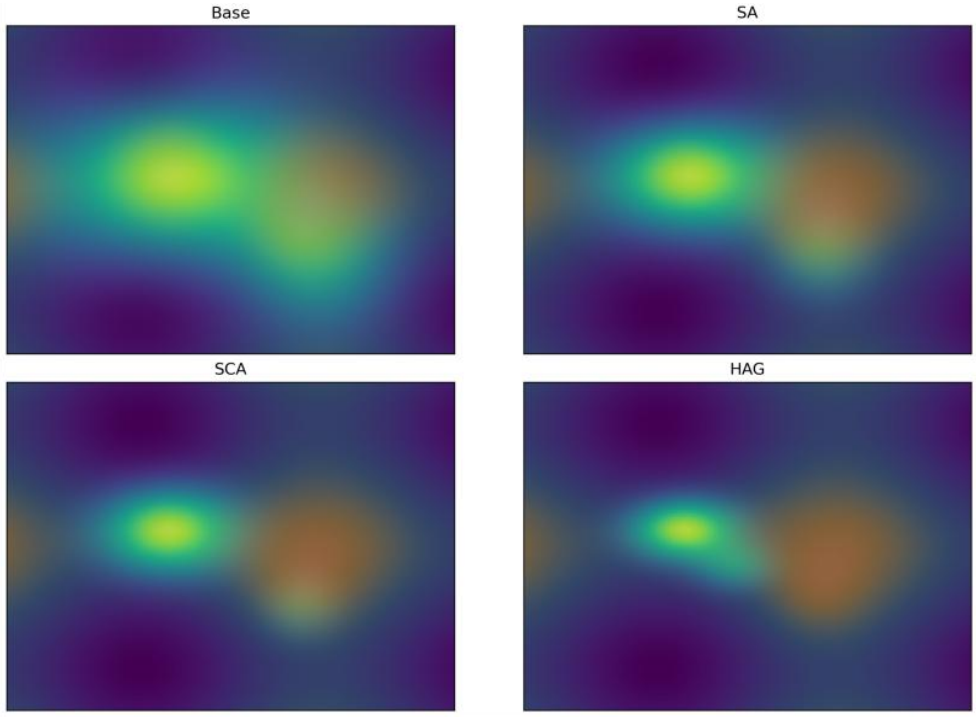

Figure 5: Response heatmap of the hierarchical attention mechanism validation experiment

To quantify the impact of different attention modules on the hierarchical performance, Table 3 lists the core metrics of the four groups of models on the test set. The results show that the classification accuracy of the Base model is 88.2%, and the macro-average F1 value is 86.9%. SA model increased to 89.7% and 88.8% respectively. SCA model further increased to 91.1% and 90.2%; The HAG model achieves 92.8% and 91.6%. In terms of inference efficiency, the HAG model takes 43 ms on average, which is only 10 ms more than the Base model, but the recognition stability of boundary samples is significantly higher. On the mild and moderate samples, the recall rates of the HAG model reached 90.4% and 91.3%, respectively, 5.8 percentage points and 5.2 percentage points higher than that of the Base model. On severe and severe samples, the F1 value increases by 2.1 percentage points and 1.8 percentage points, respectively, indicating that cross-layer gating produces a stable gain for both small-scale defects and high-level anomalies.

Table 3: Comparison of results from the hierarchical attention mechanism validation experiment

Model	Grading Accuracy / %	Macro-F1 / %	Recall / %	Mean IoU	Average Inference Time / ms
Base	88.2	86.9	87.4	0.781	53
SA	89.7	88.8	89.1	0.796	48
SCA	91.1	90.2	90.4	0.808	46
HAG	92.8	91.6	91.9	0.824	43

To further illustrate the separation state of the feature space under different attention configurations, Fig. 6 shows the embedding distributions of the four groups of models on the test samples. In the figure, the average inter-class distance of the four types of samples in the Base model is 1.14, the SA model is increased to 1.29, the SCA model is 1.37, and the HAG model is 1.52. At the same time, the average intra-class distance of the HAG model decreases from 0.74 of the Base model to 0.58, and the intra-class compactness is improved by about 21.6%. Among them, the overlap rate between mild and moderate clusters decreased from 17.8% in Base model to 9.6% in HAG model, and the overlap rate between severe and severe clusters decreased from 10.3% to 5.1%. This shows that the hierarchical attention mechanism not only enhances the local visibility, but also establishes a clearer hierarchical interval in the feature space, so that the boundary samples can enter the subsequent decision module with a more stable distribution.

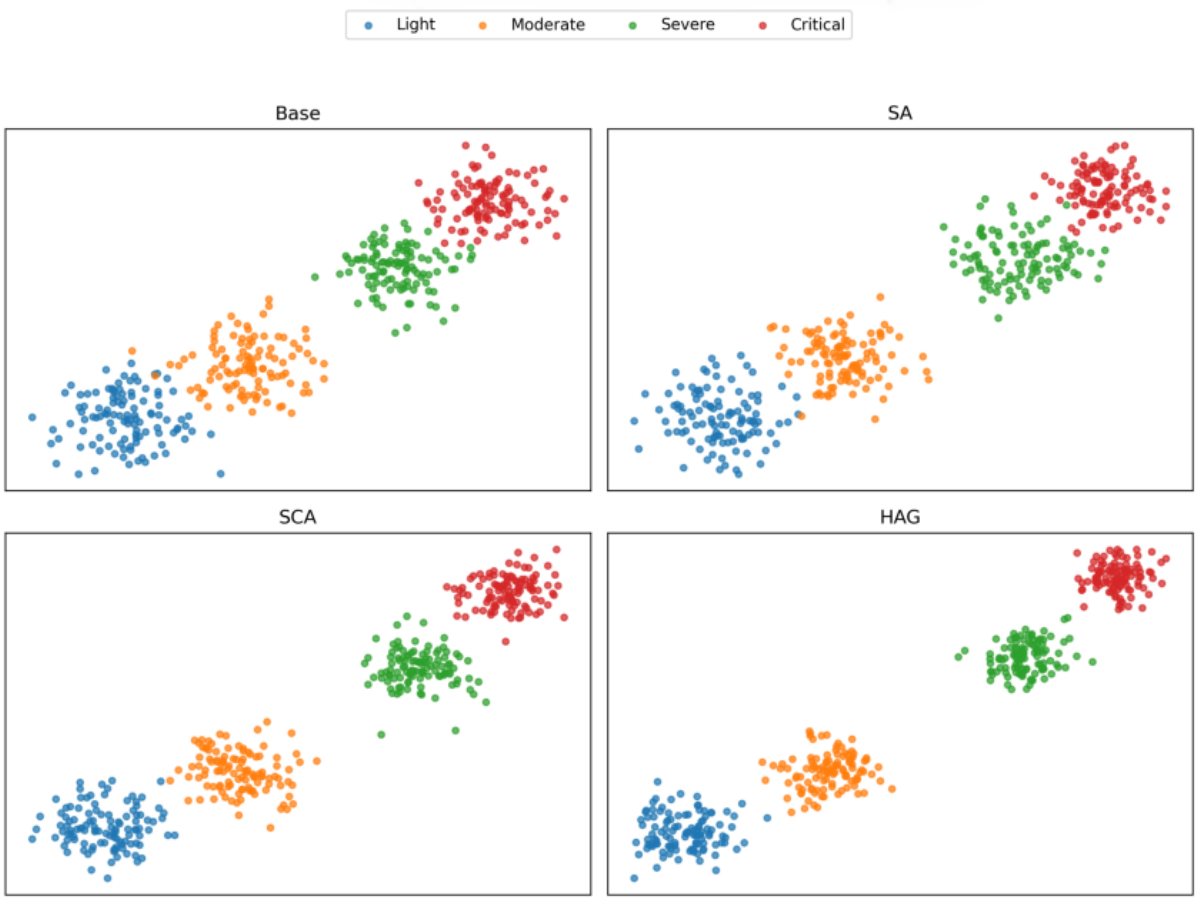

Figure 6: The embedding separation graph of the hierarchical attention mechanism validation experiment

From the experimental results in this section, it can be seen that the three parts of spatial filtering, channel recasting and cross-layer gating do not improve the performance independently of each other, but work together at different levels. Spatial screening improves the visibility of local defects, channel recalitioning strengthens the selection of semantically sensitive channels, and cross-layer gating ensures that shallow textures are not overly weakened in the process of deep semantic enhancement. After the combination of the three, the model shows higher stability under the conditions of boundary samples, fine-grained anomalies and complex background interference. Such mechanism effects show that hierarchical attention is not an additive local enhancement module, but an important part of the intelligent grading framework for distribution defects to determine the quality of representation, and also provides a reliable mechanism basis for subsequent end-to-end framework comparison experiments.

4.1.3 Comparison experiment of end-to-end grading framework

The end-to-end hierarchical framework comparison experiment is used to verify the comprehensive performance of the proposed method under the unified computing link. The experiments compare the HAG-Net proposed in this paper with three comparison models, namely BaseNet using only the conventional convolution backbone, FPN-Net with feature pyramid, and TwoStage-Net using two-stage detection and then performing independent grading. The four groups of models use the same data partition, training rounds, optimizer configuration and input scale. The evaluation metrics include classification accuracy,

macro-average F1, inference time per image and video memory footprint. The experiment not only focuses on the final accuracy, but also focuses on the balance ability between detection, classification and deployment efficiency of the model, to judge whether the framework of this paper is suitable for softwarized calls in actual inspection scenarios.

To show the comprehensive performance of different models on multiple metrics, Fig. 7 shows the normalized radar plots of the four groups of models on the five metrics of accuracy, macro-average F1, recall, intersection over union, and inference time. In the figure, BaseNet has a high score in the speed dimension, but its precision, recall and intersection ratio are significantly lower. FPN-Net improves the intersection and union ratio to 0.807, which is 0.026 higher than that of BaseNet, but the classification accuracy is only increased by 2.2 percentage points. TwoStage-Net has stable classification on severe and severe samples, and the overall accuracy reaches 91.5%, but the average inference time increases to 67 ms. HAG-Net achieves the highest values in the three dimensions of precision, macro-average F1 and recall, which are 93.4%, 92.1% and 92.4%, respectively, and the average IoU reaches 0.836. At the same time, the inference time is controlled at 41 ms. The results show that the proposed method can make full use of the unified feature links to complete defect level expression and boundary determination while maintaining high reasoning efficiency.

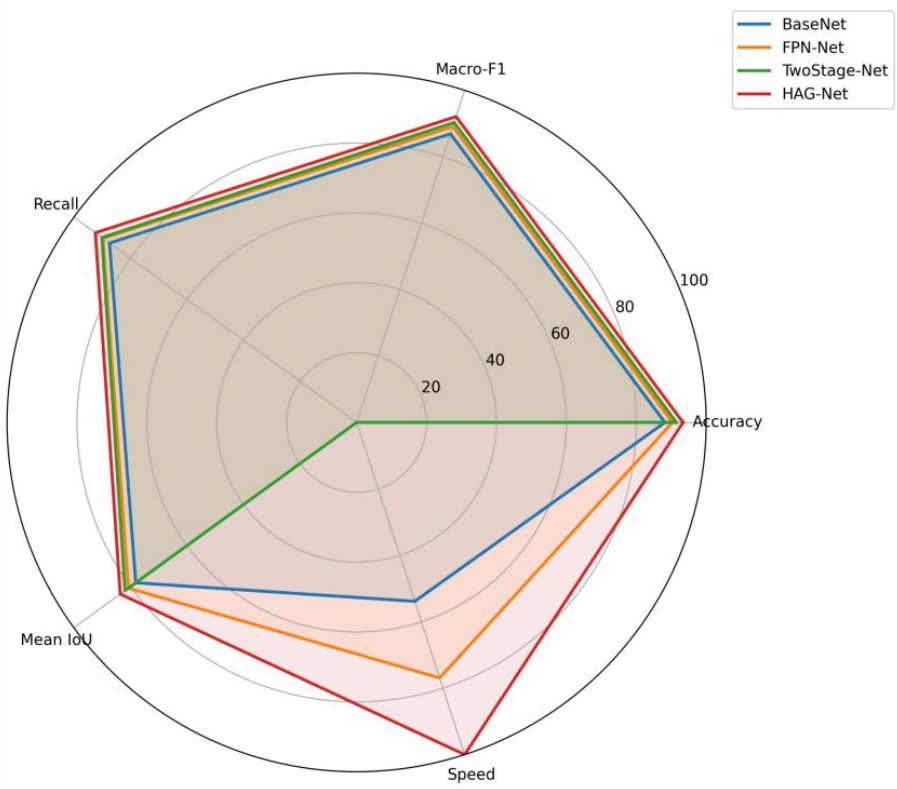

Figure 7: Normalized radar plot of the end-to-end hierarchical framework comparison experiment

To quantitatively compare the core results of each model on the test set, Table 4 lists the overall performance of the four methods. The results show that the classification accuracy of BaseNet is 88.2%, FPN-Net is 90.4%, TwoStage-Net is 91.5%, and this paper HAG-Net reaches 93.4%. Macro average F1 is 86.9%, 89.1%, 90.3% and 92.1%, respectively. In terms of inference speed, the average time of TwoStage-Net is 67 ms, and the proposed method is 41 ms, which is 26 ms less than that of TwoStage-Net. In terms of video memory occupation, HAG-Net is controlled at 3.8 GB, which is lower than 5.1 GB of TwoStage-Net. In the mild

and moderate grades, the recall rates of HAG-Net reach 91.2% and 92.0%, respectively, which are 2.3 percentage points and 2.1 percentage points higher than those of TwoStage-Net, respectively, indicating that the unified end-to-end decision has a better effect on maintaining boundary samples.

Table 4: Overall performance of the end-to-end hierarchical framework comparison experiments

Model	Grading Accuracy / %	Macro-F1 / %	Recall / %	Mean IoU	Inference Time / ms	Memory Usage / GB
BaseNet	88.2	86.9	87.4	0.781	53	3.1
FPN-Net	90.4	89.1	89.7	0.807	47	3.5
TwoStage-Net	91.5	90.3	90.1	0.819	67	5.1
HAG-Net	93.4	92.1	92.4	0.836	41	3.8

To observe how stable the different models are in the output confidence, Fig. 8 shows the boxplots of the confidence distribution of the four groups of methods on the four grade samples. BaseNet has a large box span, with the interquartile range of 0.18 for the mild level and 0.16 for the moderate level. The corresponding values of FPN-Net are reduced to 0.14 and 0.13, indicating that the feature pyramid improves the local stability. The median of TwoStage-Net was higher in severe and severe samples, reaching 0.90 and 0.93, respectively, but the proportion of outliers in mild samples was still 8.6%. The medians of HAG-Net on the four levels reach 0.89, 0.91, 0.94 and 0.96, respectively, and the interquartile ranges shrink to 0.09, 0.08, 0.07 and 0.06, respectively, indicating that its output distribution is more concentrated and the rank confidence is more stable. This result shows that the proposed framework significantly improves the output consistency between samples of different grades while maintaining high accuracy.

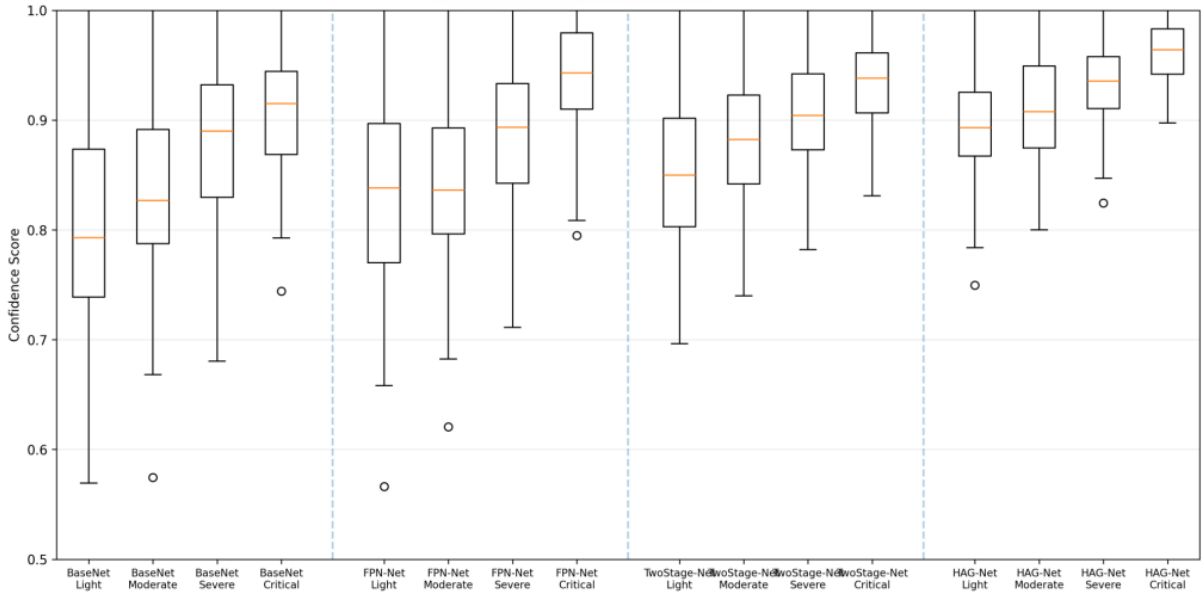

Figure 8: Boxplot of the confidence distribution for the comparison experiment of the end-to-end hierarchical framework

Synthesizing the results in this section, it can be seen that the end-to-end unified modeling not only improves the accuracy of intelligent grading of distribution defects, but also

optimizes the relationship between inference efficiency and video memory occupancy. Compared with the framework that only enhances the detection ability or adopts two-stage separation processing, the performance of the proposed method is more stable on the samples with mild cracks, local ablation and class boundary close to the samples, and the location location, class confirmation and grade determination can be completed simultaneously in a continuous computing link. This structure is not only suitable for large-scale rapid screening of UAV inspection images, but also more suitable for embedded distribution inspection analysis system to perform online reasoning and result review, so it has strong engineering deployment value.

4.2 Discussion

The experimental results show that the distribution defect grading in UAV inspection images does not only depend on whether the defect area is detected, but also depends on whether the features can form a continuous expression between local texture, component structure and grade semantics. The original benchmark experiment has been able to complete the basic identification, but there is still an overlap between mild crack and moderate ablation, indicating that it is difficult to stably maintain the fine-grained boundary by only relying on the conventional convolution response. After the hierarchical attention mechanism is added, the proportion of background leakage is reduced from 23.5% to 7.1%, and the inter-class distance is increased from 1.14 to 1.52, indicating that spatial screening, channel recallation and cross-layer gating jointly improve the visibility of small-scale defects, and also enhance the separation degree of different levels of samples in the shared space. The end-to-end framework comparison experiments further show that the unified modeling not only brings 93.4% classification accuracy and 92.1% macro-average F1, but also controls the inference time in 41 ms, which indicates that the accuracy and deployment efficiency can be balanced after fault localization, feature aggregation and level determination are calculated in the same link. For the distribution inspection system, this structure is more suitable for online screening, result review and interface call, and also provides stable support for subsequent software integration and engineering implementation. At the same time, the recall rate of HAG-Net on mild and moderate grades reaches 91.2% and 92.0%, respectively, which is more stable than the two-stage model, indicating that the level mapping has a direct constraint effect on the boundary samples. It can be seen that attention allocation, feature organization and joint judgment are not isolated modules, but the overall mechanism that determines the reliability of intelligence grading.

5 Summary

In summary, focusing on the intelligent grading task of distribution defects in UAV inspection images, this paper constructs an end-to-end computing framework guided by hierarchical attention, and completes defect localization, feature aggregation, level mapping and level determination in a unified link. Experimental results show that the classification accuracy of the proposed method reaches 93.4%, the macro-average F1 reaches 92.1%, and the average inference time is 41 ms.Compared with the baseline model, the recognition stability of the mild and moderate boundary samples is stronger, which indicates that the hierarchical attention coding and joint decision mechanism can simultaneously enhance the ability of local texture preservation and global semantic expression. At the same time, the method in this paper still has some limitations. The existing data mainly come from specific inspection heights and imaging conditions, and the generalization ability under complex weather, strong

reflection and cross-regional equipment differences still needs to be verified. There is still room for further compression of the hierarchical boundary expression of composite defects and minimal targets. It is also necessary to continue optimizing the video memory occupation and real-time scheduling strategy under the condition of end-side deployment. The follow-up work can be carried out from three aspects. Firstly, the cross-scene inspection samples are extended and more fine-grained annotations are introduced to enhance the adaptation ability of the model to multiple types of distribution components. Secondly, the hierarchical consistency of composite defects and continuous evolution states is improved by combining structural prior and timing inspection information. Thirdly, lightweight compression and reasoning acceleration are carried out for embedded inspection terminals to form an engineering implementation scheme more suitable for online deployment. The proposed method provides a path for intelligent grading of distribution defects in UAV inspection images, and also provides a basis for application and intelligent operation and maintenance analysis.

Funding

This work was supported by Inner Mongolia Electric Power (Group) Co., LTD. 2025 Science and Technology Project-2025-4-1 Research on UAV Distribution Inspection Image Defect Recognition System Based on deep Learning.

References

[1] Zhang Z, Huang S, Li Y, et al. Image detection of insulator defects based on morphological processing and deep learning[J]. Energies, 2022, 15(7): 2465.

[2] Xu S, Deng J, Huang Y, et al. Research on insulator defect detection based on an improved mobilenetv1-yolov4[J]. Entropy, 2022, 24(11): 1588.

[3] Zhang Z D, Zhang B, Lan Z C, et al. FINet: An insulator dataset and detection benchmark based on synthetic fog and improved YOLOv5[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-8.

[4] Hao K, Chen G, Zhao L, et al. An insulator defect detection model in aerial images based on multiscale feature pyramid network[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-12.

[5] Yang Z, Xu Z, Wang Y. Bidirection-fusion-YOLOv3: An improved method for insulator defect detection using UAV image[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-8.

[6] Zheng J, Wu H, Zhang H, et al. Insulator-defect detection algorithm based on improved YOLOv7[J]. Sensors, 2022, 22(22): 8801.

[7] Bao W, Du X, Wang N, et al. A defect detection method based on BC-YOLO for transmission line components in UAV remote sensing images[J]. Remote Sensing, 2022, 14(20): 5176.

[8] Han G, Li T, Li Q, et al. Improved algorithm for insulator and its defect detection based

on YOLOX[J]. Sensors, 2022, 22(16): 6186.

[9] Liu J, Liu C, Wu Y, et al. Insulators' identification and missing defect detection in aerial images based on cascaded YOLO models[J]. Computational Intelligence and Neuroscience, 2022, 2022(1): 7113765.

[10] Qiu Z, Zhu X, Liao C, et al. Detection of transmission line insulator defects based on an improved lightweight YOLOv4 model[J]. Applied Sciences, 2022, 12(3): 1207.

[11] Han G, Zhao L, Li Q, et al. A lightweight algorithm for insulator target detection and defect identification[J]. Sensors, 2023, 23(3): 1216.

[12] Liu C, Yi W, Liu M, et al. A lightweight network based on improved YOLOv5s for insulator defect detection[J]. Electronics, 2023, 12(20): 4292.

[13] Chen Y, Liu H, Chen J, et al. Insu-YOLO: An insulator defect detection algorithm based on multiscale feature fusion[J]. Electronics, 2023, 12(15): 3210.

[14] Hu C, Min S, Liu X, et al. Research on an improved detection algorithm based on yolov5s for power line self-exploding insulators[J]. Electronics, 2023, 12(17): 3675.

[15] Zhang T, Zhang Y, Xin M, et al. A light-weight network for small insulator and defect detection using UAV imaging based on improved YOLOv5[J]. Sensors, 2023, 23(11): 5249.

[16] Zhou M, Li B, Wang J, et al. Fault detection method of glass insulator aerial image based on the improved YOLOv5[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-10.

[17] Yi W, Ma S, Li R. Insulator and defect detection model based on improved YOLO-S[J]. IEEE access, 2023, 11: 93215-93226.

[18] Lu L, Chen Z, Wang R, et al. Yolo-inspection: defect detection method for power transmission lines based on enhanced YOLOv5s[J]. Journal of Real-Time Image Processing, 2023, 20(5): 104.

[19] Chen J, Fu Z, Cheng X, et al. An method for power lines insulator defect detection with attention feedback and double spatial pyramid[J]. Electric Power Systems Research, 2023, 218: 109175.

[20] Hu L, Lu Y, Wang S, et al. Multi-Equipment Detection Method for Distribution Lines Based on Improved YOLOx-s[J]. Computers, Materials & Continua, 2023, 77(3).