Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SICDet: An Oriented Object Detection Model for Complex Ice Crystal Images

Wang Yuhao
Chengdu University of Information Technology
Hu Jinrong
hjr@cuit.edu.cn
Chengdu University of Information Technology
Deng Linjie
Chengdu University of Information Technology
Li Lizhi
Chengdu University of Information Technology

Article

Keywords: ice crystal particles, oriented object detection, cloud microphysics, morphology-aware loss, aggregated ice crystals

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11469187/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

SICDet: An Oriented Object Detection Model for Complex Ice Crystal Images

Wang Yuhao, Hu Jinrong*, Deng Linjie, Li Lizhi
(College of Computer Science, Chengdu University of Information Technology, Chengdu 610225, China; *Corresponding author: hjr@cuit.edu.cn)

Abstract

Ice crystal morphology recognition is fundamental to cloud microphysics. Natural ice crystals exhibit random orientations, large scale variations, and aggregation or adhesion, while image-level single-label classifiers cannot localize their constituent particles. To address these issues, we propose SICDet, an oriented object detection model. SICDet comprises an Orientation-Adaptive Strip Backbone (OASB), a Content-Adaptive Feature Pyramid Network (CA-AFPN), and an Ice-Crystal Morphology-Aware Regression Loss (IC-MARL), which perform multidirectional feature extraction, dynamic scale–detail fusion, and morphology-aware regression weighting, respectively. On NASCENT19, SICDet achieves accuracies of 95.00%, 90.26%, 84.48%, and 78.07% for aggregation-status, basic-habit, microphysical-process, and composite-label recognition. Compared with the YOLO11 oriented detection head, it improves basic-habit, microphysical-process, and composite-label accuracy by 3.54, 5.12, and 2.25 percentage points, respectively. Without fine-tuning on CLOUDLAB, SICDet achieves accuracies of 99.95% for the basic habit of columnar ice crystals and 74.90% for their microphysical processes. The results show that SICDet improves the localization and recognition of constituent particles in complex ice crystal images and generalizes across datasets. It can support automated ice crystal analysis and the interpretation of cloud microphysical processes.

Keywords: ice crystal particles; oriented object detection; cloud microphysics; morphology-aware loss; aggregated ice crystals

1 Introduction

Ice crystals are important carriers of cloud microphysical information. Their habit, spatial orientation, size, and degree of aggregation affect light scattering, depositional growth, particle sedimentation, collision–coalescence, and precipitation formation. Accurate recognition of particle habit and evolutionary state is therefore essential for investigating particle evolution in clouds and processing observations from airborne and balloon-borne imaging probes. Optical array probes and holographic imaging systems have produced large collections of ice crystal images from field campaigns. Manual interpretation, however, is inefficient and sensitive to observer experience [1–3], motivating accurate and robust automated methods.

Automated ice crystal recognition has progressed from handcrafted geometric descriptors to deep feature learning. Early methods classified particles using area, perimeter, aspect ratio, circularity, and related features. Although interpretable, these methods are sensitive to imaging noise, resolution, and observational conditions. More recently, convolutional neural networks (CNNs), transfer learning, and semi-supervised learning have improved feature representation while reducing dependence on large labeled datasets [2–6]. Most of these methods nevertheless treat an entire image as one classification unit and return a single label. They cannot localize multiple constituent particles in an aggregated image or separately describe basic habit and microphysical attributes. IceDetectNet uses oriented bounding boxes to localize constituent particles and multi-label classification to predict their habits and microphysical attributes [7]. It establishes the feasibility of oriented object detection for complex ice crystals, but does not explicitly address random orientation, broad scale variation, or morphological anisotropy.

The available technical routes differ in both output and annotation requirements. Image-level single-label or multi-label classification can identify the overall habit and microphysical attributes of an image, but cannot associate individual particles with spatial locations. Instance segmentation provides precise contours, yet requires pixel-level masks that are costly to annotate in low-contrast images containing adhered particles. Keypoint detection is suited to objects with stable structural landmarks, whereas irregular and aggregated ice crystals have no uniform keypoint definition. Horizontal object detection can localize constituent particles, but boxes around inclined, elongated crystals include substantial background and do not explicitly represent the major-axis direction. Oriented object detection instead uses boxes parameterized by center, scale, and orientation. It can distinguish multiple constituent particles and provide orientation-aligned regional features for habit and microphysical-attribute recognition. Because this study

focuses on joint particle localization and attribute recognition rather than pixel-level contour reconstruction, oriented object detection matches both the task objective and the available annotations.

Recent oriented object detectors have focused on continuous angle representations, orientation-aware feature modeling, and end-to-end architectures. Orientation-aware distance learning and phase-shifting angle coding reduce regression discontinuities caused by angular periodicity and boundary effects [8,9]. Transformer-based oriented detectors and orientation-information integration networks further improve orientation modeling and global relation encoding [10,11]. These methods primarily target regular man-made objects such as aircraft, ships, and buildings in remote-sensing imagery. They do not fully account for the random orientations, imbalanced scale distribution, irregular contours, and adhesion of ice crystals. In particular, fixed-orientation convolutions are poorly matched to random major axes; fixed-weight feature pyramids struggle to preserve small-particle detail while retaining the semantics of large particles; and uniformly weighted oriented-box losses ignore the different sensitivities of objects with different scales and aspect ratios to center and orientation errors.

To address these limitations, we propose SICDet for ice crystal images characterized by orientation dispersion, scale variation, morphological anisotropy, and aggregation or adhesion. The model uses orientation, scale, and anisotropy as a common design basis and modifies three successive stages of an oriented detector. The Orientation-Adaptive Strip Backbone (OASB) extracts orientation and contour features adapted to local structure. The Content-Adaptive Feature Pyramid Network (CA-AFPN) adjusts cross-level fusion according to particle scale and interlevel detail differences. The Ice-Crystal Morphology-Aware Regression Loss (IC-MARL) weights center and orientation errors according to particle scale and anisotropy. The three components operate sequentially in one detection pipeline and are jointly optimized through forward feature propagation and backpropagation. The main contributions are as follows.

(1) We develop OASB. Its Orientation-Adaptive Strip Gating Unit (OASG) uses a Direction-Prior-Guided Selection (DPGS) module to extract double-angle orientation features. It combines horizontal, vertical, and local-context information and dynamically adjusts their contributions according to local orientation and image content, improving the representation of randomly oriented elongated crystals and complex aggregates.

(2) We construct CA-AFPN with a Scale–Detail Guided Fusion (SD Fusion) unit. It predicts cross-level fusion weights from particle-scale information and detail differences between adjacent levels, thereby balancing high-resolution edge information and deep semantic features for small-particle detail preservation and large-aggregate context modeling.

(3) We propose IC-MARL. It decomposes oriented-box regression into center, scale, and orientation terms, applies bounded scale-based weighting to the center error of small crystals, and adjusts orientation-error weights using morphological anisotropy. The resulting constraints reflect how particles of different geometries respond to localization errors.

2 Related Work

2.1 Ice Crystal Image Recognition

Early studies of automated ice crystal recognition relied on handcrafted features such as area, perimeter, circularity, aspect ratio, and texture. These features have clear physical interpretations, but their decision thresholds often require retuning for different instruments, image resolutions, and observation regions. They also provide limited coverage of the continuous morphological changes associated with aging, aggregation, and irregular particles.

With the development of deep learning, CNNs have been used to learn multilevel appearance representations from particle images. Touloupas et al. classified cloud particles recorded by imaging probes and demonstrated the effectiveness of end-to-end feature learning [3]. Liao et al. embedded hypergraph relations in a CNN to model associations among ice crystal habits [4]. Xu et al. used CNNs and transfer learning to improve classification and adaptation across observational conditions [2,5]. Chu et al. introduced contrastive and semi-supervised learning to exploit unlabeled samples and mitigate high annotation costs and class imbalance [6].

These methods generally classify an entire particle image into a single basic-habit or composite category. When an image contains touching, overlapping, or morphologically modified crystals, image-level classification loses the spatial locations of constituent particles and couples basic habit with microphysical process in one label. Zhang et al. proposed IceDetectNet, which combines oriented object detection with multi-label classification to localize and recognize constituent particles in aggregated ice crystals [7]. The method, however, emphasizes particle detection and

label prediction and does not explicitly tailor orientation feature extraction, cross-scale fusion, or regression constraints to random orientation, scale variation, and anisotropy. SICDet instead introduces measured particle characteristics into the backbone, feature pyramid, and oriented-box regression stages.

2.2 Oriented Object Detection

Oriented object detection commonly represents an object with a five-parameter bounding box comprising its center coordinates, width, height, and orientation angle. Compared with horizontal boxes, oriented boxes fit elongated, inclined, and touching objects more tightly and reduce irrelevant background. Early approaches improved localization using rotated anchors, feature alignment, and boundary-aware regression. Yi et al. jointly represented object boundaries and orientations using boundary-aware vectors [12]. Angular periodicity, boundary discontinuity, and orientation ambiguity for near-square objects nevertheless continue to affect training stability.

Recent work has shifted toward continuous orientation representations and end-to-end modeling. Rao et al. measured the geometric discrepancy between predicted and target boxes with an orientation-aware distance, reducing regression jumps at angular boundaries [8]. Yu et al. proposed a phase-shifting coder that maps rotational symmetries of different periods into phase spaces at corresponding frequencies [9]. Zhao et al. developed OrientedFormer, an end-to-end detector combining Gaussian positional encoding, Wasserstein self-attention, and rotated cross-attention [10]. Liu et al. proposed OII, which explicitly integrates orientation information to improve rotated feature representations [11].

These methods mainly address aircraft, ships, vehicles, and buildings in remote-sensing imagery. Ice crystal particles are smaller, cover a wider range of aspect ratios, and have contours that are more susceptible to imaging noise and overlap. Their labels also encode microphysical processes such as aging and aggregation. General oriented detectors therefore require adaptations to ice crystal orientation, scale, and morphology.

For general feature modeling, deformable convolution learns sampling offsets to accommodate geometric deformation [13], whereas the Bidirectional Feature Pyramid Network (BiFPN) performs bidirectional multiscale fusion with learnable weights [14]. Neither explicitly incorporates ice crystal geometry. Among oriented detection frameworks, Oriented R-CNN combines an Oriented RPN with Rotated RoIAlign for efficient proposal generation and orientation-aligned feature extraction [15]. RHINO improves the training of rotated detection Transformers through Hausdorff-distance matching and adaptive query denoising [16]. Strip R-CNN uses sequential orthogonal large-strip convolutions to represent high-aspect-ratio targets [17], but cannot adapt strip-branch contributions to local orientation. IceDetectNet localizes constituent particles with oriented boxes and performs multi-attribute classification, but focuses on its detection and multi-label classification scheme [7]. In contrast, SICDet modifies backbone feature extraction, feature-pyramid fusion, and oriented-box regression for the orientation randomness, scale imbalance, and morphology-dependent localization sensitivity of complex ice crystals.

2.3 Oriented Bounding-Box Regression and Geometric Losses

Conventional object detectors often use L1 or Smooth L1 losses to regress the center coordinates, width, height, and orientation angle separately. This formulation is simple but treats the parameters of an oriented box as independent variables and does not capture the overall geometric relation between predicted and target boxes. Parameter-wise regression can also produce numerical discontinuities when an angle crosses the boundary of its defined interval.

To address these issues, Yang et al. represented oriented boxes as two-dimensional Gaussian distributions and used Kullback–Leibler divergence to jointly measure center, scale, and orientation discrepancies [18,19]. KFIoU subsequently formulated an oriented-box loss from overlap relations between Gaussian distributions and improved localization for high-aspect-ratio objects [20]. Compared with coordinate-wise regression, Gaussian representations place the center, scale, and orientation in one geometric space and reduce discontinuities caused by angular periodicity.

Existing geometric losses usually apply the same regression weights to objects of different scales and shapes. They describe generic distributional discrepancies between oriented boxes but rarely consider the scale imbalance and morphological anisotropy of ice crystals. Small crystals are more sensitive to center displacement, whereas elongated crystals are more sensitive to orientation errors. IC-MARL therefore decomposes oriented-box regression into center, scale, and orientation terms and adjusts the center and orientation weights using target scale and anisotropy.

3 Proposed Method

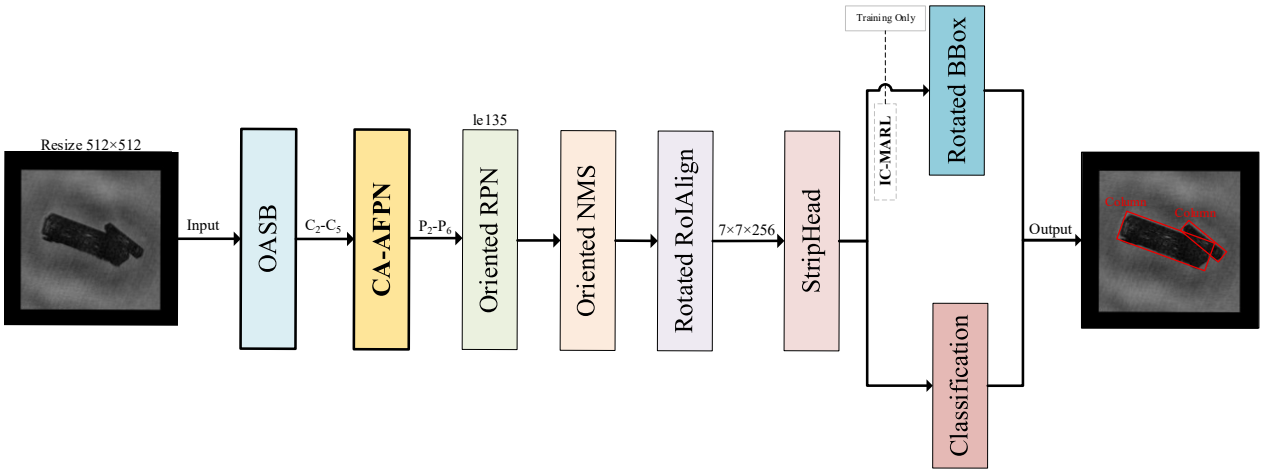


Fig. 1 Overall architecture of SICDet

Figure 1 shows the overall SICDet framework. The input image first passes through OASB to produce four-stage hierarchical features containing orientation and contour information. CA-AFPN then fuses high-resolution details and deep semantics from adjacent levels to form P2–P6 multiscale features. Oriented RPN generates oriented proposals at each pyramid level. After filtering, the proposals are processed by Rotated RoIAlign [15] to obtain orientation-aligned regional features. StripHead [17] performs classification and oriented-box regression, and IC-MARL adjusts the contributions of center, scale, and orientation errors according to target scale and anisotropy. OASB, CA-AFPN, and IC-MARL thus operate at the feature-extraction, cross-level fusion, and regression-constraint stages and are trained end to end to localize and recognize constituent particles.

3.1 Quantitative Characterization of Complex Ice Crystals

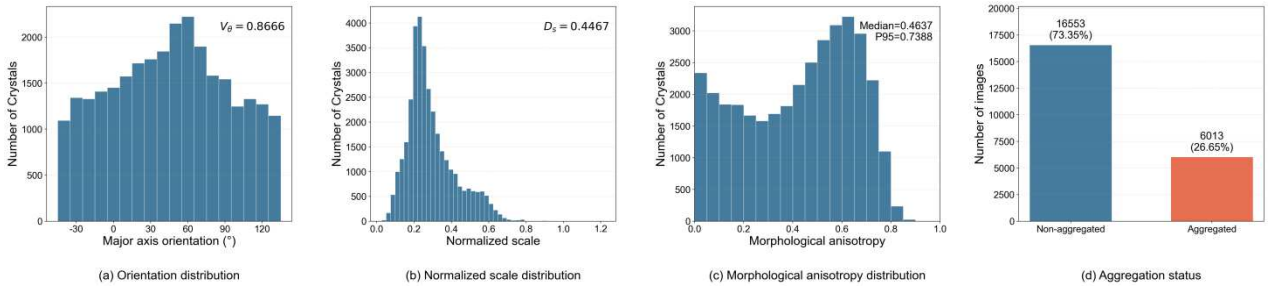


Fig. 2 Characteristics of ice crystals in NASCENT19. (a) Major axis orientation distribution. (b) Normalized scale distribution. (c) Morphological anisotropy distribution. (d) Image-level aggregation status

In this study, a complex ice crystal image is not a new class in the dataset. Rather, it is an image exhibiting one or more of the following properties: dispersed orientations, substantial scale variation, pronounced morphological anisotropy, and aggregation or adhesion. We quantified these factors for 35,193 annotated instances in 22,566 NASCENT19 images. Orientation, scale, and anisotropy were measured at the object-instance level, whereas aggregation status and the number of constituent particles were measured at the image level.

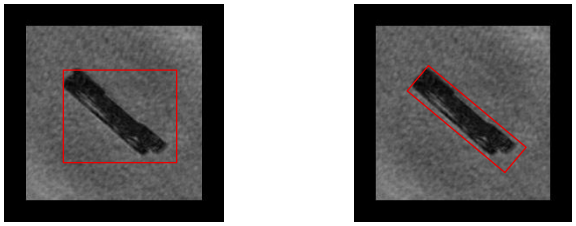
Because an oriented box has a 180° periodicity, we use the axial circular variance $V_\theta = 1 - \left|\frac{1}{N}\sum_{t=1}^{N} \exp(\mathrm{j}2\theta_t)\right|$ to quantify orientation dispersion. After excluding small and irregular particles whose orientations are ambiguous, the axial circular variance is 0.8666 [Fig. 2(a)]. This value indicates substantial dispersion of the major-axis direction and shows why fixed-orientation feature extractors are poorly suited to elongated crystals with different orientations. Random orientations also cause horizontal bounding boxes to include considerable irrelevant background. Oriented boxes fit inclined crystals more closely (Fig. 3), reduce background interference, and provide a more accurate geometric representation of each constituent particle's orientation and position.

For oriented bounding box t, the normalized scale $\bar{s}_t = \sqrt{w_t h_t/(WH)}$ describes its relative size, and $D_s = \sigma(\bar{s})/\mu(\bar{s})$ measures scale variation. In Fig. 2(b), the coefficient of variation is 0.4467 and the 5th–95th percentile

interval is 0.1227–0.5637. The dataset therefore contains both small crystals and large aggregates, requiring dynamic coordination of high-resolution details and deep semantic information.

Morphological anisotropy is defined by $a_t = |w_t - h_t|/(w_t + h_t + \varepsilon)$; values closer to 1 indicate a more elongated target. In Fig. 2(c), the median and 95th percentile are 0.4637 and 0.7388, respectively. A substantial fraction of the particles thus have elongated structures and are sensitive to major-axis features and orientation errors in the oriented box.

Figure 2(d) shows the image-level distribution of aggregation status. The dataset contains 6,013 aggregated images, accounting for 26.65% of all images, with an average of 3.10 constituent particles per aggregated image. Particles in aggregated samples frequently touch or partially occlude one another. The detector must therefore localize each constituent particle while modeling both local structure and contextual relations.



(a)Ice crystal particles in horizontal bounding boxes

(b)Ice crystal particles in oriented bounding boxes

Fig. 3 Ice crystal particles enclosed by different types of bounding boxes. (a) Horizontal bounding box. (b) Oriented bounding box

3.2 Ice Crystal Geometry and Overall Design

SICDet addresses the four factors at different stages of the detection pipeline. Orientation dispersion is handled by local orientation estimation and dynamic branch selection in OASB. Scale variation is addressed by scale–detail fusion in CA-AFPN, while the normalized scale of each object also weights the center term in IC-MARL. Morphological anisotropy motivates strip-structure extraction in OASB and explicitly weights the orientation term in IC-MARL. For aggregation and adhesion, multiple oriented boxes localize constituent particles separately, while the local-context branch in OASG and multiscale fusion in CA-AFPN supplement complex local structure and context.

The three components are connected sequentially. Hierarchical features from OASB provide the input to cross-level fusion in CA-AFPN. The resulting multiscale features support proposal generation and oriented-box prediction, while regression gradients from IC-MARL propagate backward through the detection head and feature pyramid to OASB. The components therefore modify feature extraction, cross-scale fusion, and regression constraints within one jointly optimized detector.

3.3 Orientation-Adaptive Strip Backbone

As shown in Section 2.1, ice crystal particles exhibit marked anisotropy and random orientations. Columns, column–plate crystals, and some aggregates contain elongated major axes, branches, and adhesion structures, whereas aged and irregular crystals have complex local contours. Conventional square convolutions use limited isotropic receptive fields. They capture general texture but are inefficient for long, slender structures and locally varying orientations. Existing strip convolutions enlarge the receptive field along selected axes, yet they fuse horizontal and vertical responses with fixed contributions. Unlike deformable convolution, which adapts to shape through free sampling offsets [13], OASB explicitly uses orientation-specific strip branches to emphasize major axes and elongated structures.

We therefore construct OASB from strip convolutions with orientation-prior estimation and dynamic multibranch gating. As shown in Fig. 4, each OASB block contains a multilayer perceptron (MLP) branch and an OASG branch in parallel. The branches model channel information and spatial structure, respectively, and update the features through layer scaling and residual connections. Their parallel arrangement allows both information types to be modeled independently from the same input.

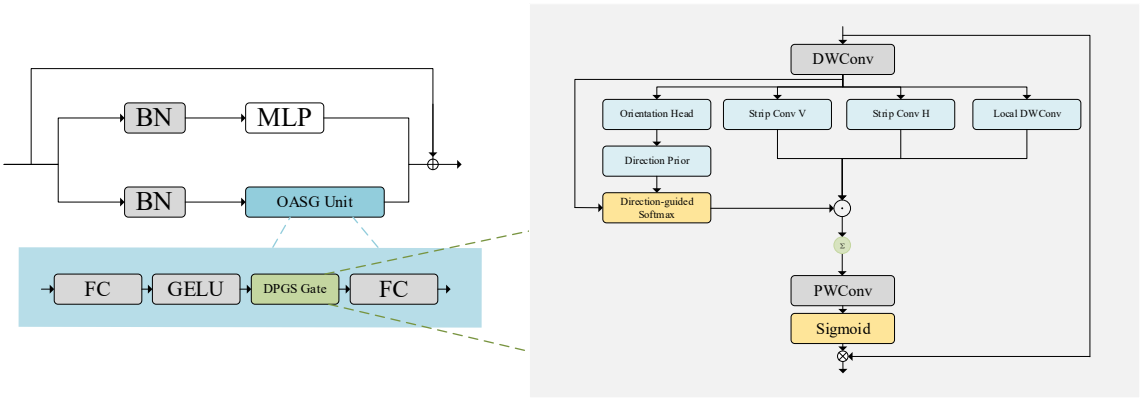


Fig. 4 Structures of the OASB block and OASG. Left, parallel MLP and OASG branches with residual connections. Right, orientation-guided fusion of horizontal strip, vertical strip, and local-context features

Given an input feature ($\boldsymbol{X} \in R^{C\times H\times W}$), the output of the basic block is

$$\boldsymbol{Y} = \boldsymbol{X} + \gamma_{\mathrm{m}}\mathcal{M}(\mathrm{BN}(\boldsymbol{X})) + \gamma_{\mathrm{o}}\mathcal{O}(\mathrm{BN}(\boldsymbol{X})) \tag{1}$$

where $\mathcal{M}(\cdot)$ denotes the MLP branch, $\mathcal{O}(\cdot)$ denotes OASG, and γ_{m} and γ_{o} are learnable layer-scale coefficients.

OASG first applies pointwise convolution (PWConv) and a Gaussian error linear unit (GELU) for channel mapping. An orientation-prior-guided Softmax gate then models spatial structure. A final pointwise convolution restores the channel representation, which is fused with the input through a residual connection.

Within DPGS, a 5 × 5 depthwise convolution (DWConv) first extracts local base features:

$$\boldsymbol{F}_{\mathrm{b}} = \mathrm{DWConv}_{5\times5}(\boldsymbol{X}) \tag{2}$$

Three parallel branches then model horizontal structure, vertical structure, and local context:

$$\boldsymbol{F}_{\mathrm{h}} = \mathrm{DWConv}_{1\times k}(\boldsymbol{F}_{\mathrm{b}}) \tag{3}$$

$$\boldsymbol{F}_{\mathrm{v}} = \mathrm{DWConv}_{k\times1}(\boldsymbol{F}_{\mathrm{b}}) \tag{4}$$

$$\boldsymbol{F}_{\mathrm{l}} = \mathrm{DWConv}_{3\times3}(\boldsymbol{F}_{\mathrm{b}}) \tag{5}$$

Here, $\boldsymbol{F}_{\mathrm{h}}$ and $\boldsymbol{F}_{\mathrm{v}}$ capture long-range horizontal and vertical structures, respectively, while $\boldsymbol{F}_{\mathrm{l}}$ supplements non-axis-aligned, irregular, and local contour information. To avoid fixed-ratio fusion, the orientation-estimation head generates a normalized two-dimensional orientation descriptor from the base features:

$$[\boldsymbol{d}_{\mathrm{c}}, \boldsymbol{d}_{\mathrm{s}}] = \mathrm{Norm}(\mathrm{H}_{\theta}(\boldsymbol{F}_{\mathrm{b}})) \tag{6}$$

where $\boldsymbol{d}_{\mathrm{c}}$ and $\boldsymbol{d}_{\mathrm{s}}$ represent the cosine and sine responses in the double-angle orientation space. The double-angle form accounts for the periodicity of oriented targets. Horizontal, vertical, and non-axis-aligned structural priors are then constructed as

$$\boldsymbol{p}_{\mathrm{h}} = \frac{1+\boldsymbol{d}_{\mathrm{c}}}{2}, \boldsymbol{p}_{\mathrm{v}} = \frac{1-\boldsymbol{d}_{\mathrm{c}}}{2}, \boldsymbol{p}_{\mathrm{l}} = |\boldsymbol{d}_{\mathrm{s}}| \tag{7}$$

To explain how the three branches represent arbitrary orientations, let the ideal outputs of the orientation-estimation head be $d_c = \cos 2\theta$ and $d_s = \sin 2\theta$. The orientation priors in Eq. (7) can then be written as $p_h = \cos^2\theta, p_v = \sin^2\theta, p_l = |\sin 2\theta|$。. OASG does not make a discrete choice between the horizontal and vertical branches. Instead, it continuously adjusts all three branch contributions according to local orientation. When θ approaches 0° or 90°, the horizontal or vertical strip branch receives a stronger prior. When θ approaches 45° or 135°, both axial branches contribute and the non-axis-aligned prior p_l becomes large, increasing the contribution of the local-context branch to oblique contours. The final branch weights also depend on image content at each spatial location, allowing the orientation response to vary continuously with local crystal orientation. Combining the priors with content-gating scores from the base features, Softmax produces three spatially adaptive weights:

$$[\boldsymbol{w}_{\mathrm{h}}, \boldsymbol{w}_{\mathrm{v}}, \boldsymbol{w}_{\mathrm{l}}] = \mathrm{Softmax}\left[H_{\mathrm{g}}(\boldsymbol{F}_{\mathrm{b}}) + \lambda_{\mathrm{p}}\log(\boldsymbol{P}+\varepsilon)\right] \tag{8}$$

where $\boldsymbol{P} = [\boldsymbol{p}_{\mathrm{h}}, \boldsymbol{p}_{\mathrm{v}}, \boldsymbol{p}_{\mathrm{l}}]$, $H_{\mathrm{g}}(\cdot)$ denotes the weight-prediction function. λ_{p} controls the influence of the orientation prior on branch-weight prediction and is set to $\lambda_{\mathrm{p}} = 0.5$ in all experiments; $\varepsilon = 10^{-6}$ is a numerical stability term. The value was selected in preliminary validation experiments. At every spatial location, the three weights satisfy

$$\boldsymbol{w}_{\mathrm{h}} + \boldsymbol{w}_{\mathrm{v}} + \boldsymbol{w}_{\mathrm{l}} = 1 \tag{9}$$

The final structural feature is

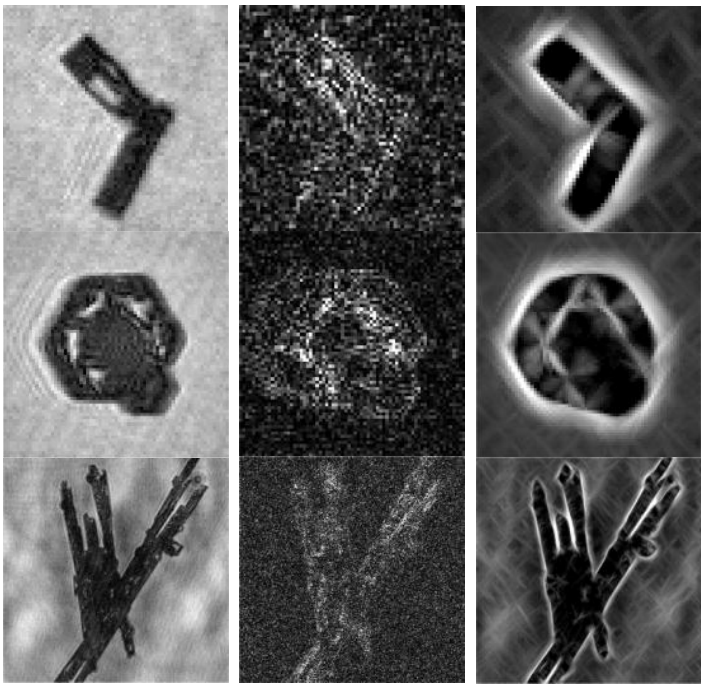
$$\boldsymbol{F}_{\mathrm{fuse}} = \boldsymbol{w}_{\mathrm{h}} \odot \boldsymbol{F}_{\mathrm{h}} + \boldsymbol{w}_{\mathrm{v}} \odot \boldsymbol{F}_{\mathrm{v}} + \boldsymbol{w}_{\mathrm{l}} \odot \boldsymbol{F}_{\mathrm{l}} \tag{10}$$

where $\odot$ denotes element-wise multiplication. The fused feature is processed by pointwise convolution and a sigmoid function to generate a spatial gating map that modulates the input:

$$\boldsymbol{F}_{\mathrm{out}} = \boldsymbol{X} \odot \mathrm{Sigmoid}[\mathrm{PWConv}(\boldsymbol{F}_{\mathrm{fuse}})] \tag{11}$$

The horizontal and vertical strip branches provide long-range structural information along two orthogonal directions. The local-context branch supplements non-axis-aligned edges, branch intersections, and irregular contours. For an oblique elongated crystal, OASG combines both axial branches using spatially varying continuous weights and increases the contribution of the local-context branch. Stacked OASG units progressively aggregate local responses distributed along the oblique major axis. OASG therefore does not rotate kernels to arbitrary angles. Instead, it approximates continuously oriented structures through orientation priors and content-aware gating without enumerating many directional branches.

Figure 5 compares the input image and feature activations produced by conventional convolution and OASG. Conventional convolution yields relatively scattered responses, whereas OASG concentrates high activations around the major axes, edges, and branches of aggregated crystals. Orientation gating strengthens continuous responses along elongated structures. For irregular and aggregated targets, the local-context branch preserves adhesion regions and complex contours.



(a)Original ice crystal image (b)Feature activation map from conventional convolution (c)Feature activation map from OASG

Fig. 5 Feature activation maps extracted by different operators. (a) Original images. (b) Activations from conventional convolution. (c) Activations from OASG

3.4 Content-Adaptive Feature Pyramid Network

The statistics in Section 2.1 show substantial scale variation among ice crystals. Small particles depend on high-resolution edges and local textures, whereas large particles and aggregates require stronger deep semantic information. A conventional feature pyramid network (FPN) fuses adjacent levels by fixed element-wise addition [21], so their relative contributions do not vary with the input. BiFPN uses learnable weights for bidirectional multiscale fusion [14]. CA-AFPN instead predicts spatially varying fusion weights from the scale and detail information of the current features.

CA-AFPN contains an SD Fusion unit, as shown in Fig. 6(a). OASB outputs C = {C2, C3, C4, C5}. A 1 × 1 lateral convolution standardizes the channel dimension at each level, producing L = {L2, L3, L4, L5}. The highest level L5

forms P5. Features are then successively upsampled and fused with the current lateral feature in the order P5 → P4 → P3 → P2, and P5 is downsampled to produce P6. The final set P = {P2, P3, P4, P5, P6} is passed to the Oriented RPN.

Figure 6(b) shows the SD Fusion unit at level i. The higher-level pyramid feature $\boldsymbol{P}_{i+1}$ is first upsampled to the spatial resolution of the current lateral feature $\boldsymbol{L}_i$, and the two are concatenated along the channel dimension. The scale-prior and detail-prior branches then produce the scale descriptor $\boldsymbol{S}_i$ and detail descriptor $\boldsymbol{D}_i$, respectively:

$$\boldsymbol{U}_i = \mathrm{Up}(\boldsymbol{P}_{i+1}) \tag{12}$$

$$\boldsymbol{Z}_i = \mathrm{Concat}(\boldsymbol{L}_i, \boldsymbol{U}_i) \tag{13}$$

$$\boldsymbol{S}_i = \mathrm{H_s}(\boldsymbol{Z}_i) \tag{14}$$

$$\boldsymbol{D}_i = \mathrm{H_d}(|\boldsymbol{L}_i - \boldsymbol{U}_i|) \tag{15}$$

Here, $\boldsymbol{U}_i$ is the upsampled higher-level feature, $\boldsymbol{Z}_i$ is the aligned adjacent-level feature pair, and $\mathrm{H_s}(\cdot)$ and $\mathrm{H_d}(\cdot)$ denote the lightweight scale-prior and detail-prior extraction branches, respectively.

For implementation details, let both the current lateral feature and the upsampled feature contain C channels, that is, $L_i, U_i \in \mathbb{R}^{C\times H_i\times W_i}$. Their concatenation Z_i has $2C$ channels. We set $C = 256$ and use bilinear interpolation for upsampling. The scale-prior branch $H_s(\cdot)$ comprises a 5×5 depthwise convolution and a 1×1 pointwise convolution. The 5×5 depthwise convolution has 512 input and output channels with 512 groups. The subsequent 1×1 pointwise convolution reduces the channel dimension from 512 to 64, namely $C_s = 64$, and aggregates local scale and contextual information with a larger receptive field. The detail-prior branch $H_d(\cdot)$ takes the absolute cross-level difference $|L_i - U_i|$ as input and contains a 3×3 depthwise convolution followed by a 1×1 pointwise convolution. The 3×3 depthwise convolution has 256 input and output channels with 256 groups. The 1×1 pointwise convolution then reduces the channel dimension to 64, namely $C_d = 64$, to extract edge, texture, and cross-level inconsistency cues. All convolutions in both branches use a stride of 1 and padding that preserves spatial resolution, followed by group normalization with 16 groups and GELU activation.

The original fusion feature, scale prior, and detail prior are then passed to the weight-prediction network, which generates normalized spatial fusion weights through Softmax:

$$[\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i] = \mathrm{Softmax}\Big(\mathrm{G}\big(\mathrm{Concat}(\boldsymbol{Z}_i, \boldsymbol{S}_i, \boldsymbol{D}_i)\big)\Big) \tag{16}$$

The weight-prediction network $G(\cdot)$ receives the concatenation of Z_i, S_i, and D_i, with $2C + C_s + C_d = 640$ input channels. It contains a 3×3 convolution followed by a 1×1 convolution. The 3×3 convolution reduces the channel dimension from 640 to 64 and uses a stride of 1, padding of 1, group normalization with 16 groups, and GELU activation. The final 1×1 convolution outputs two channels without normalization or activation. Softmax is then applied across the two fusion branches to obtain $\alpha_i, \beta_i \in \mathbb{R}^{1\times H_i\times W_i}$. After broadcasting along the channel dimension, the weights are applied to L_i and U_i, respectively. CA-AFPN places one SD Fusion unit at each of the P_4、P_3、P_2 levels, and the parameters are not shared across levels.

$$\boldsymbol{\alpha}_i + \boldsymbol{\beta}_i = 1 \tag{17}$$

where $\mathrm{G}(\cdot)$ is the lightweight convolutional weight-prediction network, and $\boldsymbol{\alpha}_i$ and $\boldsymbol{\beta}_i$ are the spatial weights of the current lateral feature and the upsampled higher-level feature, respectively. The two weights are normalized at each spatial location and broadcast along the channel dimension. The output at level i is

$$\boldsymbol{P}_i = \mathrm{Conv}_{3\times 3}(\boldsymbol{\alpha}_i \odot \boldsymbol{L}_i + \boldsymbol{\beta}_i \odot \boldsymbol{U}_i) \tag{18}$$

where $\odot$ denotes element-wise multiplication. Unlike fixed 1:1 addition or weight prediction directly from concatenated features, SD Fusion explicitly incorporates scale and detail priors. The detail prior preserves more high-resolution information around small particles and contour edges, while the scale prior provides stronger context and semantics for large particles and aggregates. Adjacent-level contributions can therefore vary with target scale, edge structure, and local content.

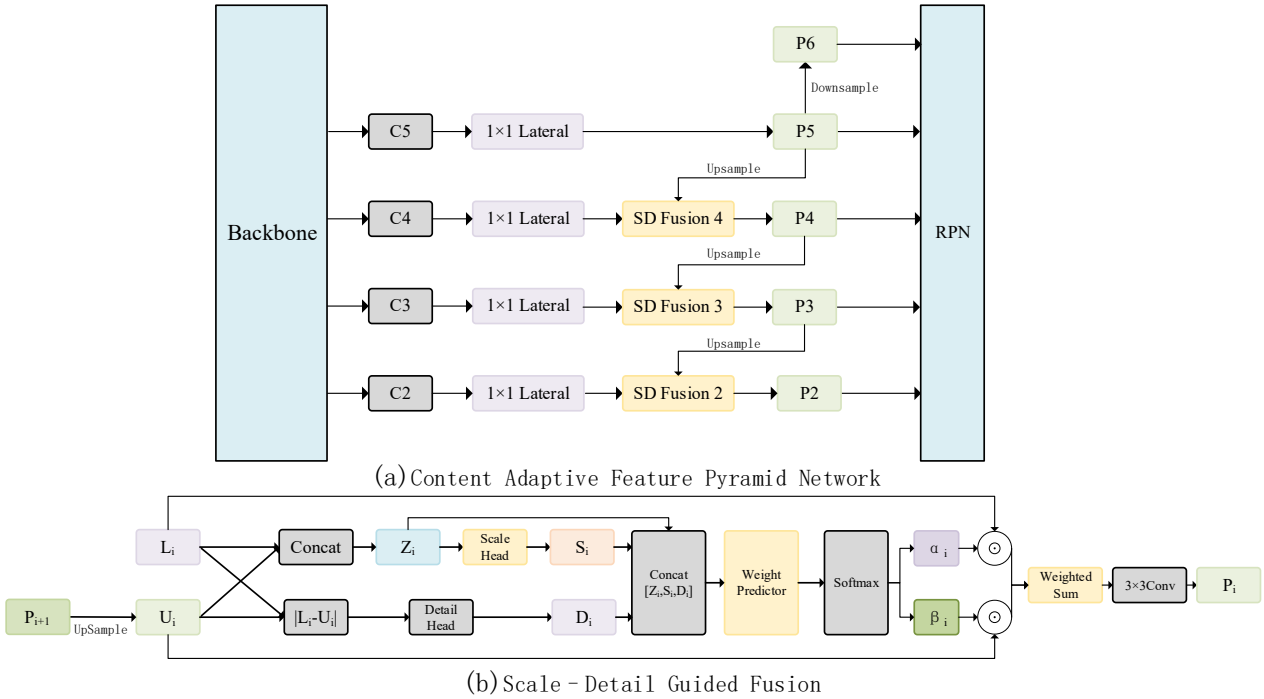


Fig. 6 Architecture of CA-AFPN. (a) Multiscale feature pyramid with top-down fusion. (b) Scale–Detail Guided Fusion unit, which predicts spatially varying weights to combine adjacent-level features

3.5 Ice-Crystal Morphology-Aware Oriented Bounding-Box Regression Loss

Oriented-box regression directly affects the localization of constituent particles. Smooth L1 and L1 losses commonly optimize the center coordinates, width, height, and angle independently [22], although these parameters jointly determine the geometry of the predicted and target boxes. For small or high-aspect-ratio crystals, a slight center displacement or orientation error can cause a substantial decrease in overlap. Angular periodicity can also introduce numerical discontinuities at interval boundaries.

Motivated by Gaussian modeling of oriented boxes [18], we propose the Ice Crystal Morphology-Aware Regression Loss (IC-MARL). The regression error is explicitly decomposed into center-position, scale, and orientation terms, which are weighted according to the normalized scale and morphological anisotropy of the target crystal. Specifically, an oriented bounding box B = (x, y, w, h, θ) is represented as a two-dimensional Gaussian distribution, and center displacement is normalized by the covariance of the target box.

$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{19}$$

The mean vector and covariance matrix are defined as

$$\boldsymbol{\mu} = \begin{bmatrix} x \\ y \end{bmatrix}, \boldsymbol{\Sigma} = \boldsymbol{R}(\theta) \begin{bmatrix} (w/2)^2 & 0 \\ 0 & (h/2)^2 \end{bmatrix} \boldsymbol{R}^{\mathrm{T}}(\theta) \tag{20}$$

$$\boldsymbol{R}(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{21}$$

Under a consistent angle convention for oriented boxes, the regression error is decomposed into center-position, scale, and orientation terms.

The center-position error is measured by the Mahalanobis distance normalized by the covariance of the target box:

$$\mathcal{L}_{\mathrm{c}} = \frac{1}{2}(\boldsymbol{\mu}_{\mathrm{p}} - \boldsymbol{\mu}_{\mathrm{t}})^{\mathrm{T}}(\boldsymbol{\Sigma}_{\mathrm{t}} + \varepsilon \boldsymbol{I})^{-1}(\boldsymbol{\mu}_{\mathrm{p}} - \boldsymbol{\mu}_{\mathrm{t}}) \tag{22}$$

where p and t denote the predicted and ground-truth boxes, respectively, and ε is a stability constant that prevents a singular covariance matrix. This term normalizes center displacement according to the target scale and orientation and is therefore more suitable than direct coordinate differences for crystals with different scales and aspect ratios.

To reduce the effect of absolute target size on width and height regression, the scale term uses logarithmic scale ratios:

$$\mathcal{L}_{\mathrm{s}} = \rho\left(\ln\frac{w_{\mathrm{p}}+\varepsilon}{w_{\mathrm{t}}+\varepsilon}\right) + \rho\left(\ln\frac{h_{\mathrm{p}}+\varepsilon}{h_{\mathrm{t}}+\varepsilon}\right) \tag{23}$$

where $\rho(\cdot)$ is the Smooth L1 function. Logarithmic scale ratios jointly constrain predicted width and height while reducing the absolute numerical advantage of large targets in the loss.

To account for angular periodicity, the orientation loss is defined as

$$\mathcal{L}_{\theta} = 1 - \cos\left[2(\theta_{\mathrm{p}} - \theta_{\mathrm{t}})\right] \tag{24}$$

The double-angle form makes the loss continuous under the π period and avoids large numerical differences between geometrically similar orientations on opposite sides of an angular boundary.

To analyze the relation between target scale and sensitivity to center displacement, assume that the predicted and target boxes have the same width, height, and orientation and differ only in center position. Rotate the displacement into the local coordinate system of the target box and denote it by $\boldsymbol{\delta}_{\mathrm{c}}{}' = (\delta_{\mathrm{u}}, \delta_{\mathrm{v}})^{\mathrm{T}}$. When $|\delta_{\mathrm{u}}| < w_{\mathrm{t}}$ and $|\delta_{\mathrm{v}}| < h_{\mathrm{t}}$, the intersection over union of the two oriented boxes is

$$\mathrm{IoU} = \frac{(w_{\mathrm{t}} - |\delta_{\mathrm{u}}|)(h_{\mathrm{t}} - |\delta_{\mathrm{v}}|)}{2w_{\mathrm{t}}h_{\mathrm{t}} - (w_{\mathrm{t}} - |\delta_{\mathrm{u}}|)(h_{\mathrm{t}} - |\delta_{\mathrm{v}}|)} \tag{25}$$

When the center displacement is small relative to the target dimensions, its first-order approximation is

$$1 - \mathrm{IoU} \approx 2\left(\frac{|\delta_{\mathrm{u}}|}{w_{\mathrm{t}}} + \frac{|\delta_{\mathrm{v}}|}{h_{\mathrm{t}}}\right) \tag{26}$$

Thus, for the same absolute center displacement, a smaller target width or height produces a larger relative displacement and a greater reduction in oriented-box overlap. For example, after image resizing, a one-pixel center shift has a much larger relative effect on a crystal spanning only a few pixels than on a large aggregate.

Using the target covariance defined in Eq. (20) and ignoring the numerical stability term, the center-position loss in Eq. (22) can be written in the local coordinate system of the target box as

$$\mathcal{L}_{\mathrm{c}} = 2\left(\frac{\delta_{\mathrm{u}}^2}{w_{\mathrm{t}}^2} + \frac{\delta_{\mathrm{v}}^2}{h_{\mathrm{t}}^2}\right) \tag{27}$$

Thus, $\mathcal{L}_{\mathrm{c}}$ already normalizes center displacement by target width and height. If the target is uniformly scaled by a factor of κ, then, under the same absolute displacement, $\mathcal{L}_{\mathrm{c}}$ increases approximately with $1/\kappa^2$. This behavior is consistent with the IoU relation above and provides a geometric basis for the center-position sensitivity of small crystals.

To analyze the relation between morphological anisotropy and orientation error, assume that the predicted and target boxes have the same center, width, and height but differ by a small orientation error $\Delta\theta$. According to Eq. (20), the covariance matrix of the oriented box is

$$\boldsymbol{\Sigma}(\theta) = \boldsymbol{R}(\theta)\begin{bmatrix} w_{\mathrm{t}}^2/4 & 0 \\ 0 & h_{\mathrm{t}}^2/4 \end{bmatrix}\boldsymbol{R}^{\mathrm{T}}(\theta) \tag{28}$$

After a first-order perturbation of the orientation angle, the normalized covariance change is approximated by

$$\frac{\|\boldsymbol{\Sigma}(\theta + \Delta\theta) - \boldsymbol{\Sigma}(\theta)\|_{\mathrm{F}}}{\mathrm{tr}(\boldsymbol{\Sigma})} \approx \sqrt{2}\frac{|w_{\mathrm{t}}^2 - h_{\mathrm{t}}^2|}{w_{\mathrm{t}}^2 + h_{\mathrm{t}}^2}|\Delta\theta| \tag{29}$$

Let the aspect ratio of the oriented box be

$$r = \frac{\max(w_{\mathrm{t}}, h_{\mathrm{t}})}{\min(w_{\mathrm{t}}, h_{\mathrm{t}})}, r \geq 1 \tag{30}$$

The orientation-error sensitivity coefficient is then

$$\frac{r^2 - 1}{r^2 + 1} \tag{31}$$

Its derivative with respect to r is

$$\frac{\mathrm{d}}{\mathrm{d}r}\left(\frac{r^2 - 1}{r^2 + 1}\right) = \frac{4r}{(r^2 + 1)^2} > 0 \tag{32}$$

The sensitivity of an oriented box to orientation error therefore increases monotonically with aspect ratio. When $r = 1$, the target is nearly square and rotation produces little covariance change. As the target becomes more elongated, the same orientation error produces a larger geometric discrepancy. This behavior is consistent with columnar ice crystals: their major-axis direction is well defined, and even a small angular error can produce appreciable lateral displacement at both ends. Near-square particles have weaker orientation identifiability.

Because ice crystal morphologies differ in their sensitivity to regression errors, the center and orientation terms are adaptively weighted according to target scale and aspect ratio. The normalized scale and morphological anisotropy of the target box are defined as

$$\bar{s}_{\mathrm{t}} = \sqrt{\frac{w_{\mathrm{t}} h_{\mathrm{t}}}{WH}}, a_{\mathrm{t}} = \frac{|w_{\mathrm{t}} - h_{\mathrm{t}}|}{w_{\mathrm{t}} + h_{\mathrm{t}} + \varepsilon} \tag{33}$$

where W and H are the width and height of the input image. We next relate the morphological-anisotropy coefficient to the orientation-error sensitivity coefficient above. Ignoring the numerical stability term ε,

$$a_{\mathrm{t}} = \frac{r-1}{r+1} \tag{34}$$

we obtain

$$\frac{r^2-1}{r^2+1} = \frac{2a_{\mathrm{t}}}{1+a_{\mathrm{t}}^2}, \frac{\partial}{\partial a_{\mathrm{t}}}\left(\frac{2a_{\mathrm{t}}}{1+a_{\mathrm{t}}^2}\right) = \frac{2(1-a_{\mathrm{t}}^2)}{(1+a_{\mathrm{t}}^2)^2} > 0 \tag{35}$$

Because $0 \le a_{\mathrm{t}} < 1$, the relation above shows that a_t is in a strictly monotonic correspondence with the sensitivity of an oriented box to orientation error. Using a_t to adjust the orientation-loss weight is therefore not solely an empirical choice; it is a bounded monotonic parameterization of aspect-ratio-dependent sensitivity. The center and orientation weights are defined as

$$\lambda_{\mathrm{c}} = 1 + \beta \exp\left(-\frac{\bar{s}_{\mathrm{t}}}{\tau}\right), \lambda_{\theta} = 1 + \alpha a_{\mathrm{t}} \tag{36}$$

where α, β, and τ are hyperparameters. They are fixed at α = 1.0, β = 0.5, and τ = 0.10, respectively, in all experiments. α controls the maximum increase in the orientation-term weight, β controls the maximum increase in the center-term weight, and τ controls how quickly the center weight decays as target scale increases. The Mahalanobis distance in Eq. (22) is already the main scale-normalization mechanism for center displacement. Introducing λ_{c} does not assume that every small target must receive a larger center loss; it applies a bounded adjustment to the contribution of small crystals to the total regression loss. From Eq. (36),

$$1 < \lambda_{\mathrm{c}} \le 1 + \beta, \frac{\partial \lambda_{\mathrm{c}}}{\partial \bar{s}_{\mathrm{t}}} = -\frac{\beta}{\tau} \exp\left(-\frac{\bar{s}_{\mathrm{t}}}{\tau}\right) < 0 \tag{37}$$

Accordingly, λ_{c} decreases smoothly as normalized scale increases, and the upper bound $1 + \beta$ prevents unbounded amplification for extremely small targets. This weight is a limited supplement to the scale-normalized center loss, and its effectiveness is evaluated in the ablation study in Table 13. High-aspect-ratio crystals such as columns have larger a_{t} values and thus receive larger orientation-error weights. For near-square particles, rotation has less effect on geometric overlap, so the orientation weight is reduced.

The final ice-crystal morphology-aware oriented bounding-box regression loss is

$$\mathcal{L}_{\mathrm{IC-MARL}} = \log[1 + \lambda_{\mathrm{c}}\mathcal{L}_{\mathrm{c}} + \lambda_{\mathrm{s}}\mathcal{L}_{\mathrm{s}} + \lambda_{\theta}\mathcal{L}_{\theta}] \tag{38}$$

where λ_{s} is the scale-loss weight and is set to 1 by default. The outer $\log[1 + x]$ smoothly compresses large regression errors, limiting the effect of outliers on the gradient and improving training stability. IC-MARL adaptively adjusts the center, scale, and orientation contributions according to particle size and anisotropy, aligning oriented-box regression with the geometry of complex ice crystals.

4 Experiments and Results

4.1 Datasets

NASCENT19 was collected in 2019 during the Ny-Ålesund Aerosol Cloud Experiment (NASCENT) [23]. A holographic imager mounted on a tethered-balloon system recorded ice crystal particles under natural conditions. As

listed in Table 1, the dataset defines seven basic habits: column, plate, lollipop, column–plate, irregular, droplet, and small. Each particle is also assigned one of four microphysical-process labels (Table 2): pristine, aged, aggregated, or aged and aggregated. The aged label indicates that a crystal has undergone processes such as riming, melting, or sublimation; a rimed crystal, for example, often develops smooth protrusions along its edges. Aggregation indicates that two or more adhered constituent crystals are identifiable in the same image. Each image label combines a basic habit with a microphysical process. Although seven habits and four processes yield 28 possible combinations, several do not occur in the observations. For example, small particles are too small for reliable interpretation of their microphysical evolution, and irregular particles are generally not treated as pristine habits. The dataset therefore contains 19 composite labels and 22,566 images. We randomly split the data before augmentation into 18,799 training images (83.31%) and 3,767 validation images (16.69%). No image appears in both subsets, and random augmentation is applied only to training data. Stratification by the 19 composite labels keeps class proportions as similar as possible across the two subsets.

For cross-dataset evaluation, we also use ice crystal images collected by the CLOUDLAB project over the Swiss Plateau [6]. The dataset mainly contains column, aged-column, aggregated-column, and aged-aggregated-column crystals. It is used only for testing and is not included in model training.

Table 1. Sample counts and descriptions of the basic habits in NASCENT19.

Basic habit	Number of images	Description
Column	9857	Columnar ice crystals.
Plate	316	Plate-like crystals with a hexagonal structure and a smooth or irregular surface.
Lollipop	454	Formed when a supercooled droplet collides with and freezes onto a columnar crystal.
Column–plate	3917	Crystals with both columnar and plate-like features, often resembling the letter H; usually formed across column- and plate-growth temperature regimes.
Irregular	1727	Irregular crystals without a distinct basic habit.
Droplet	975	Frozen cloud or drizzle droplets, usually nonspherical or deformed.
Small	1618	Crystals judged visually small by annotators, typically below 75 μm.

Table 2. Sample counts and descriptions of the four microphysical processes in NASCENT19.

Microphysical process	Number of images	Description
Pristine	12380	Crystals that have undergone neither aging nor aggregation.
Aged	4173	Crystals modified by riming or collisions, with smooth protrusions or irregular surfaces.
Aggregated	1197	Two or more mutually adhered ice crystals.
Aged and aggregated	1114	Crystals that have undergone both aging and aggregation.

4.2 Experimental Settings

SICDet is implemented with PyTorch 1.12.1 and MMRotate 0.3.4 [24]. Training uses four NVIDIA GeForce RTX 4090D GPUs with two images per GPU, giving a total batch size of 8. Each experiment runs for 36 epochs. The recorded random seed is 0; deterministic execution and mixed-precision training are disabled. Performance is monitored on the validation set after every epoch, but all reported results use the checkpoint from the final epoch rather than a checkpoint selected by validation performance. Checkpoints are saved after every epoch, with the five most recent retained. Before network input, both image dimensions are expanded by 15%. The added border is filled with pixel value 1, and oriented-box coordinates are translated accordingly so that every box lies fully inside the image. The enlarged image is resized to 512 × 512 by bilinear interpolation, and the boxes are transformed using the corresponding horizontal and vertical scale factors. Horizontal, vertical, and diagonal flips are each applied with probability 0.25. During training, polygon-aware random rotation is applied with probability 0.5; the angle is sampled uniformly from $[-180°, 180°]$, and rotation is performed about the image center. Bilinear interpolation is used for the image affine transformation. Each oriented box is converted to a four-vertex polygon, transformed using the same affine matrix, and refitted as an oriented box under the le135 convention. Images are converted to RGB order, normalized with the ImageNet mean [25] (123.675, 116.28, 103.53) and standard deviation (58.395, 57.12, 57.375), and padded to a multiple of 32. Random augmentation is used only for training. AdamW [26] is used with an initial learning rate of 1×10^{-4} $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 0.05. A step schedule multiplies the learning rate by 0.1 at epochs 27 and 33. The maximum L_2 norm for gradient clipping is 35. Oriented boxes follow the le135 convention with angles in [−45°, 135°). Efficiency is measured on one NVIDIA GeForce RTX 4090D GPU with 512 × 512 inputs and batch size 1. The model runs in inference mode with gradients disabled. After 50 warm-up iterations, 500 consecutive runs are timed using CUDA synchronization. Latency includes the model forward pass and detection postprocessing but excludes image loading. Parameter count is reported in millions (M), and computation is reported in GFLOPs for one 512 × 512 image. All efficiency tests use the same software, hardware, and numerical-precision settings.

4.3 Evaluation Metrics

We use overall accuracy, weighted F1, macro F1, and macro recall. Accuracy is computed from the evaluated instances after task-specific label mapping. For class k, precision and recall are $P_k = TP_k/(TP_k + FP_k)$ and $R_k = TP_k/(TP_k + FN_k)$, and $F1_k = 2P_kR_k/(P_k + R_k)$. Macro F1 and macro recall are arithmetic means across classes, whereas weighted F1 is weighted by class support. Detector predictions are first matched to ground-truth boxes using oriented IoU; unmatched predictions and annotations are retained as false positives and false negatives, respectively. The 19 composite labels are mapped according to each task. Basic-habit evaluation retains only the seven habits and ignores aging and aggregation attributes. Microphysical-process evaluation ignores basic habit and maps labels to pristine, aged, aggregated, and aged-and-aggregated. The aggregated subset contains samples whose ground-truth process is aggregated or aged-and-aggregated, and the non-aggregated subset contains pristine or aged samples. Computational efficiency is measured using the number of parameters (Params), floating-point operations (FLOPs), average per-image latency, and frames per second (FPS). With a batch size of 1, FPS = 1000/Latency, where latency is measured in ms/image.

4.4 Results

4.4.1 Results on NASCENT19

Table 3 summarizes the NASCENT19 results. SICDet obtains accuracies of 95.00%, 90.26%, 84.48%, and 78.07% for aggregation-status, basic-habit, microphysical-process, and composite-label recognition, respectively. For the composite-label task, weighted F1 is 80.24%, whereas macro F1 and macro recall are 54.66% and 58.06%. The gap indicates substantial variation in performance among classes.

Table 3. Summary of SICDet results across recognition tasks.

Task	Subset	Accuracy/%	Weighted F1/%
Composite label	All data (19 classes)	78.07	80.24
Basic habit	All data (7 classes)	90.26	90.24
Basic habit	Non-aggregated particles (7 classes)	89.20	83.60
Basic habit	Aggregated particles (6 classes)	72.60	65.90
Microphysical process	All data (4 classes)	84.48	85.02
Microphysical process	Non-aggregated particles (2 classes)	85.80	87.00
Microphysical process	Aggregated particles (2 classes)	68.90	72.80
Aggregation status	All data (2 classes)	95.00	85.10

For basic-habit recognition, Fig. 7 shows strong performance for column, column–plate, and small particles. These categories have relatively many samples and stable structural characteristics. Irregular crystals and classes with fewer samples are more difficult to recognize.

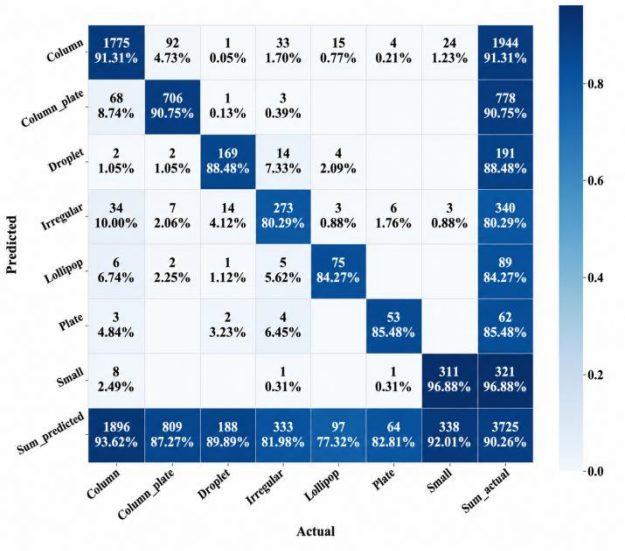


Fig. 7 Confusion matrix of SICDet for basic-habit classification

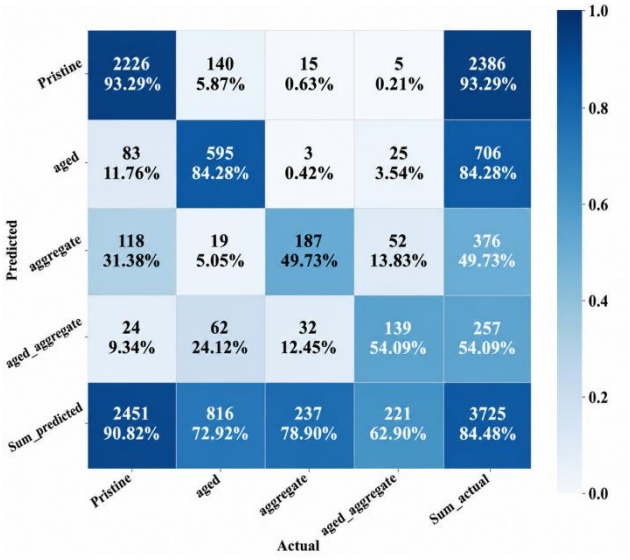


Fig. 8 Confusion matrix of SICDet for microphysical-process classification

4.4.2 Analysis of the Aggregated-Particle Subset

Table 4. Per-class recognition results of SICDet for the 19 composite labels.

Class	Validation samples	Precision/%	Recall/%	F1/%
Column	1088	96.16	92.10	94.08
Column_aged	720	90.21	71.67	79.88
Column_aged_aggregate	159	25.28	42.77	31.78
Column_aggregate	3	11.11	66.67	19.05
Column_plate	701	83.43	84.02	83.72
Column_plate_aggregate	82	25.38	40.24	31.13
Droplet	40	94.29	82.50	88.00
Droplet_aged	96	92.21	73.96	82.08
Droplet_aged_aggregate	45	55.32	57.78	56.52
Droplet_aggregate	13	15.00	23.08	18.18
Irregular	214	96.40	62.62	75.92
Irregular_aggregate	131	31.58	50.38	38.82
Lollipop	83	88.24	72.29	79.47
Lollipop_aggregate	7	4.55	14.29	6.90
Plate	25	37.50	72.00	49.32
Plate_aged	17	91.67	64.71	75.86
Plate_aged_aggregate	18	28.00	38.89	32.56
Plate_aggregate	2	0.00	0.00	0.00
Small	323	97.41	93.19	95.25

To evaluate aggregation and adhesion, we divide the validation set into aggregated and non-aggregated subsets according to the ground-truth microphysical-process labels. As shown in Table 3, basic-habit accuracy and weighted F1 on the aggregated subset are 72.60% and 65.90%, respectively, which are 16.60 and 17.70 percentage points lower than on the non-aggregated subset. Microphysical-process accuracy and weighted F1 are 68.90% and 72.80%, 16.90 and 14.20 percentage points lower, respectively. The 19 composite labels are highly imbalanced. Table 4 shows that Column, Column_aged, Column_plate, and Small account for 75.18% of the validation set and each achieves an F1 score of at least 79.88%; these frequent classes therefore dominate overall accuracy and weighted F1. In contrast, the nine aggregation-related classes have a macro F1 of 26.10%, well below the 80.36% for non-aggregation classes. The low macro F1 is therefore associated not only with the long-tailed label distribution but also with adhesion, partial occlusion, and confusion among similar aggregation-related categories.

Figure 9 shows representative results for aggregated particles. When constituent-particle contours remain relatively complete and mutual occlusion is limited, SICDet localizes the particles separately and correctly identifies their basic habits.

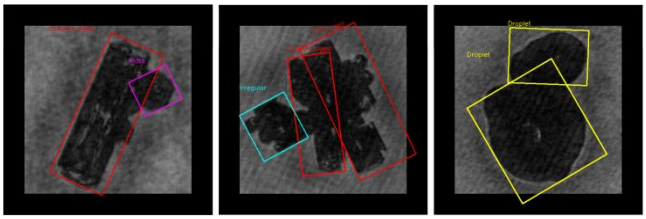


Fig. 9 Representative SICDet results for constituent particles in aggregated ice crystals

4.4.3 Comparison with Image Classification Models

Tables 5–7 compare SICDet with four image classifiers: VGG16, DenseNet169, ResNet152, and ConvNeXt-Tiny [27–30]. SICDet obtains the highest accuracy and weighted F1 under all three label mappings. Unlike image-level classifiers, the detector represents multiple constituent crystals separately instead of forcing the entire image into one visual category, which is particularly important for aggregated-particle images.

Table 5. Comparison of basic-habit classification results.

Model	Accuracy/%	Weighted F1/%	Macro F1/%	Macro Recall/%
SICDet	90.26	90.24	87.27	86.42
VGG16	83.67	83.54	74.26	73.19
DenseNet169	83.70	83.56	75.68	74.24
ResNet152	83.75	83.61	76.53	75.85
ConvNeXt-Tiny	84.34	82.10	67.47	78.13

Table 6. Comparison of microphysical-process classification results.

Model	Accuracy/%	Weighted F1/%	Macro F1/%	Macro Recall/%
SICDet	84.48	85.02	72.35	76.38
VGG16	82.56	82.10	67.14	65.06
DenseNet169	81.90	81.62	67.22	65.76
ResNet152	82.51	82.11	67.48	65.60
ConvNeXt-Tiny	82.61	82.10	67.47	65.18

Table 7. Comparison of composite-label classification results.

Model	Accuracy/%	Weighted F1/%	Macro F1/%	Macro Recall/%
SICDet	78.07	80.24	54.66	58.06
VGG16	77.67	77.17	49.66	49.47
DenseNet169	77.33	76.92	49.90	49.07
ResNet152	77.91	77.42	53.77	54.40
ConvNeXt-Tiny	77.13	77.56	52.77	51.06

4.4.4 Comparison with Oriented Object Detectors

Tables 8–10 compare oriented object detectors. IceDetectNet is the domain-specific model most closely related to this task [7], while RHINO and Strip R-CNN provide relevant architectural baselines [16,17]. Rotated Faster R-CNN, RTMDet, the YOLO11 oriented detection head, and RT-DETR are included as general detection baselines [31–34]. On the composite-label task, SICDet achieves a macro F1 of 54.66%, 6.54 percentage points higher than the YOLO11 oriented detection head. The result indicates better overall class-wise recognition, although aggregation-related minority classes remain challenging.

Table 8. Comparison of oriented object detectors for basic-habit classification.

Model	Accuracy/%	Weighted F1/%	Macro F1/%	Macro Recall/%
SICDet	90.26	90.24	87.27	86.42
IceDetectNet	79.67	79.73	61.55	61.01
RHINO	55.80	58.20	45.53	50.65
Strip R-CNN	85.43	85.91	68.00	65.58
Rotated Faster R-CNN	74.48	75.01	52.47	54.69
RTMDet	82.14	82.65	64.82	63.90
YOLO11 oriented detection head	86.72	87.15	72.84	70.96
RT-DETR	83.08	83.56	66.21	65.37

Table 9. Comparison of oriented object detectors for microphysical-process classification.

Model	Accuracy/%	Weighted F1/%	Macro F1/%	Macro Recall/%
SICDet	84.48	85.02	72.35	76.38
IceDetectNet	60.02	67.48	37.31	45.51
RHINO	14.28	15.53	11.38	28.94

Model	Accuracy/%	Weighted F1/%	Macro F1/%	Macro Recall/%
Strip R-CNN	77.67	79.80	50.43	54.56
Rotated Faster R-CNN	57.15	62.63	33.27	39.65
RTMDet	71.84	74.62	45.73	50.21
YOLO11 oriented detection head	79.36	81.44	55.62	59.18
RT-DETR	73.05	75.91	47.68	51.74

Table 10. Comparison of oriented object detectors for composite-label classification.

Model	Accuracy/%	Weighted F1/%	Macro F1/%	Macro Recall/%
SICDet	78.07	80.24	54.66	58.06
IceDetectNet	57.92	64.10	29.22	32.94
RHINO	10.41	8.62	7.24	16.05
Strip R-CNN	74.65	76.21	46.30	47.10
Rotated Faster R-CNN	55.23	60.75	26.34	30.02
RTMDet	68.42	71.37	38.90	40.88
YOLO11 oriented detection head	75.82	77.43	48.12	50.63
RT-DETR	70.16	72.85	40.75	42.91

4.4.5 Cross-Dataset Evaluation

Figure 10 shows the CLOUDLAB confusion matrices. SICDet achieves 99.95% basic-habit accuracy for columnar ice crystals; only one column is misclassified as a column–plate crystal. Microphysical-process accuracy is 74.90%. The model thus preserves reliable basic-habit recognition for columns across different instruments and scenes and retains some ability to distinguish microphysical processes.

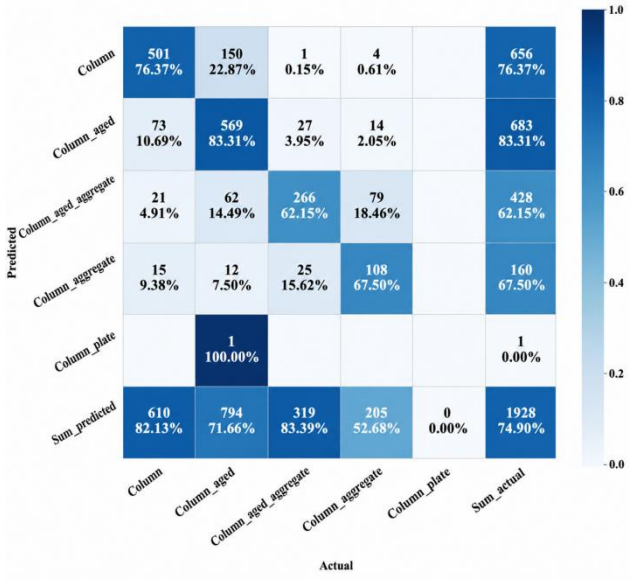


Fig. 10 SICDet confusion matrices on the independent CLOUDLAB dataset

4.4.6 Computational Efficiency

Table 11 compares the parameter count, computational cost, inference latency, and frame rate of the baseline and different component combinations. All models are evaluated with the same hardware and input resolution. IC-MARL is used only during training and adds no network structure, parameters, or computation at inference.

The baseline contains 32.80 M parameters, requires 71.60 GFLOPs, and runs at 70.42 FPS. OASB and CA-AFPN add 2.60 M and 1.90 M parameters and 5.80 and 8.70 GFLOPs, respectively. The complete SICDet contains 37.30 M parameters and requires 86.10 GFLOPs, increases of 13.72% and 20.25% over the baseline, while running at 60.61 FPS. At the same time, composite-label accuracy increases from 72.19% to 78.07%, a gain of 5.88 percentage points. Because IC-MARL is used only in training, it does not affect inference parameters or theoretical computation. SICDet therefore improves recognition with a moderate increase in computational cost.

Table 11. Computational efficiency of different models.

Model	Params/M	FLOPs/G	Latency/(ms·image^{-1})	Frame rate/(frames·s^{-1})	Accuracy/%
Baseline	32.80	71.60	14.20	70.42	72.19
Baseline + OASB	35.40	77.40	15.10	66.23	76.35

Model	Params/M	FLOPs/G	Latency/(ms·image^{-1})	Frame rate/(frames·s^{-1})	Accuracy/%
Baseline + CA-AFPN	34.70	80.30	15.60	64.10	76.98
Baseline + IC-MARL	32.80	71.60	14.20	70.42	75.14
SICDet	37.30	86.10	16.50	60.61	78.07

4.4.7 Ablation Studies

Table 12 reports ablations of OASB, CA-AFPN, and IC-MARL on composite-label classification. The baseline accuracy is 72.19%. Adding OASB, CA-AFPN, or IC-MARL individually increases accuracy to 76.35%, 76.98%, and 75.14%, gains of 4.16, 4.79, and 2.95 percentage points, respectively. Each component therefore improves the baseline independently. CA-AFPN provides the largest individual gain, indicating that content- and scale-dependent fusion of hierarchical features is important for complex ice crystal recognition.

Table 12 also reports all two-component combinations. OASB + CA-AFPN, OASB + IC-MARL, and CA-AFPN + IC-MARL obtain accuracies of 77.52%, 77.18%, and 77.61%, respectively, all above the corresponding single-component models. The components are therefore complementary. Orientation and contour features from OASB provide targeted inputs for cross-level fusion in CA-AFPN and support orientation regression of elongated crystals under IC-MARL. The improved multiscale features from CA-AFPN, in turn, provide a stronger basis for the differentiated regression constraints imposed by IC-MARL.

Integrating all three components yields the highest accuracy, 78.07%, outperforming every single- and two-component configuration. Jointly improving orientation feature extraction, scale–detail fusion, and morphology-aware regression better accommodates variation in crystal orientation, scale, and shape. These results establish the individual contributions and complementarity of the three components.

Table 12. Ablation results for the proposed components.

Model	OASB	CA-AFPN	IC-MARL	Accuracy/%
Baseline	×	×	×	72.19
Baseline + OASB	√	×	×	76.35
Baseline + CA-AFPN	×	√	×	76.98
Baseline + IC-MARL	×	×	√	75.14
Baseline + OASB + CA-AFPN	√	√	×	77.52
Baseline + OASB + IC-MARL	√	×	√	77.18
Baseline + CA-AFPN + IC-MARL	×	√	√	77.61
SICDet	√	√	√	78.07

Table 13. Ablation results for weighting strategies in IC-MARL.

Setting	Scale weightingλ_c	Anisotropy weightingλ_θ	Accuracy/%
Fixed weights	×	×	73.58
Scale weighting only	√	×	74.37
Anisotropy weighting only	×	√	74.11
Full IC-MARL	√	√	75.14

To assess the contributions within OASG, we ablate the horizontal branch $\boldsymbol{F}_h$, vertical branch $\boldsymbol{F}_v$, local-context branch $\boldsymbol{F}_l$, and orientation prior separately. When one branch is removed, the others are unchanged and their Softmax weights are renormalized. To remove the orientation prior, we set $\lambda_p = 0$ and predict branch weights only from image content. All configurations use the same data split and training strategy; the results are listed in Table 14.

Table 14. Ablation results for OASG components.

Setting	Horizontal branch$\boldsymbol{F}_h$	Vertical branch$\boldsymbol{F}_v$	Local branch$\boldsymbol{F}_l$	Orientation prior	Accuracy/%
Without horizontal branch	×	√	√	√	75.21
Without vertical branch	√	×	√	√	75.36
Without local branch	√	√	×	√	74.82
Without orientation prior$\lambda_p = 0$	√	√	√	×	75.57
Full OASG	√	√	√	√	76.35

Table 14 shows that the complete OASG achieves the highest accuracy, 76.35%. Removing the horizontal or vertical branch decreases accuracy by 1.14 or 0.99 percentage points, indicating that the two axial strip branches extract elongated structures from complementary directions. Removing the local-context branch reduces accuracy by 1.53 percentage points, confirming its contribution to non-axis-aligned, irregular, and adhered aggregate structures. When

$\lambda_p = 0$, accuracy decreases by 0.78 percentage points. Content-only branch weighting is therefore insufficient, and an explicit orientation prior improves branch selection. The complete OASG combines axial structure, local contours, and orientation priors and consequently performs best.

5 Conclusion

We proposed SICDet, an oriented object detection model for complex ice crystal images characterized by orientation dispersion, scale variation, morphological anisotropy, and aggregation or adhesion. OASB adjusts strip-branch contributions according to local structural orientation. CA-AFPN changes cross-level fusion according to scale information and interlevel detail differences, and IC-MARL weights center and orientation errors using particle scale and anisotropy. The three components operate at successive stages and are jointly optimized through feature propagation and backpropagation. All two-component configurations outperform the corresponding single-component models, and the complete model achieves the highest accuracy, confirming their complementarity. On NASCENT19, SICDet obtains accuracies of 95.00%, 90.26%, 84.48%, and 78.07% for aggregation-status, basic-habit, microphysical-process, and composite-label recognition, respectively. On the independent CLOUDLAB dataset, basic-habit accuracy for columnar ice crystals reaches 99.95%, and microphysical-process accuracy reaches 74.90%. SICDet therefore supports constituent-particle localization and the recognition of basic habits and microphysical processes across datasets. Limitations remain: basic-habit and microphysical-process accuracies on the aggregated subset are 72.60% and 68.90%, respectively, below those on non-aggregated samples. Severe adhesion, partial occlusion, and limited samples for aggregation-related classes continue to affect localization and recognition. Future work will investigate occlusion-aware feature enhancement and class-balanced learning, and will connect particle recognition with cloud microphysical parameter analysis for mixed-phase cloud studies, precipitation-process analysis, and weather-modification applications.

6 Funding

This research received no external funding.

References

[1] Duan, J., Guo, H., Hu, J., et al.: Research progress in observation and shape classification of natural snow and ice crystals. Acta Meteorologica Sinica 81(5), 685–701 (2023). https://doi.org/10.11676/qxxb2023.20220180 (in Chinese)

[2] Xu, Y.-F., Jiao, R.-L., Huang, M.-S.: Classification of ice crystal images from airborne cloud particle imager probe (CPI) using convolutional neural networks (CNN). Journal of Computers 35(3), 159–174 (2024). https://doi.org/10.53106/199115992024063503012

[3] Touloupas, G., Lauber, A., Henneberger, J., et al.: A convolutional neural network for classifying cloud particles recorded by imaging probes. Atmospheric Measurement Techniques 13(5), 2219–2239 (2020). https://doi.org/10.5194/amt-13-2219-2020

[4] Liao, M., Duan, J., Zhang, R., et al.: A hypergraph-embedded convolutional neural network for ice crystal particle habit classification. Intelligent Automation & Soft Computing 29(3), 787–801 (2021). https://doi.org/10.32604/iasc.2021.018190

[5] Xu, Y., Jiao, R., Li, Q., et al.: Classification of cloud particle habits using transfer learning with a deep convolutional neural network. Atmosphere 16(3), 294 (2025). https://doi.org/10.3390/atmos16030294

[6] Chu, Y., Zhang, H., Li, X., et al.: Exploring the effect of training set size and number of categories on ice crystal classification through a contrastive semi-supervised learning algorithm. Atmospheric Measurement Techniques 18(12), 2781–2801 (2025). https://doi.org/10.5194/amt-18-2781-2025

[7] Zhang, H., Li, X., Ramelli, F., et al.: IceDetectNet: a rotated object detection algorithm for classifying components of aggregated ice crystals with a multi-label classification scheme. Atmospheric Measurement Techniques 17(24), 7109–7128 (2024). https://doi.org/10.5194/amt-17-7109-2024

[8] Rao, C., Wang, J., Cheng, G., et al.: Learning orientation-aware distances for oriented object detection. IEEE Transactions on Geoscience and Remote Sensing 61, 1–11 (2023). https://doi.org/10.1109/TGRS.2023.3278933

[9] Yu, Y., Da, F.: Phase-shifting coder: predicting accurate orientation in oriented object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13354–13363. IEEE, Piscataway (2023). https://doi.org/10.1109/CVPR52729.2023.01283

[10] Zhao, J., Ding, Z., Zhou, Y., et al.: OrientedFormer: an end-to-end Transformer-based oriented object detector in remote sensing images. IEEE Transactions on Geoscience and Remote Sensing 62, 1–16 (2024). https://doi.org/10.1109/TGRS.2024.3456240

[11] Liu, Y., Jiang, W.: OII: an orientation information integrating network for oriented object detection in remote sensing images. Remote Sensing 16(5), 731 (2024). https://doi.org/10.3390/rs16050731

[12] Yi, J., Wu, P., Liu, B., et al.: Oriented object detection in aerial images with box boundary-aware vectors. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2150–2159. IEEE, Piscataway (2021). https://doi.org/10.1109/WACV48630.2021.00220

[13] Dai, J., Qi, H., Xiong, Y., et al.: Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 764–773. IEEE, Piscataway (2017). https://doi.org/10.1109/ICCV.2017.89

[14] Tan, M., Pang, R., Le, Q.V.: EfficientDet: scalable and efficient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10781–10790. IEEE, Piscataway (2020). https://doi.org/10.1109/CVPR42600.2020.01079

[15] Xie, X., Cheng, G., Wang, J., et al.: Oriented R-CNN for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3500–3509. IEEE, Piscataway (2021). https://doi.org/10.1109/ICCV48922.2021.00350

[16] Lee, H., Song, M., Koo, J., et al.: Hausdorff distance matching with adaptive query denoising for rotated detection transformer. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1872–1882. IEEE, Piscataway (2025). https://doi.org/10.1109/WACV61041.2025.00189

[17] Yuan, X., Zheng, Z., Li, Y., et al.: Strip R-CNN: large strip convolution for remote sensing object detection. Proceedings of the AAAI Conference on Artificial Intelligence 40(15), 12259–12267 (2026). https://doi.org/10.1609/aaai.v40i15.38217

[18] Yang, X., Yang, X., Yang, J., et al.: Learning high-precision bounding box for rotated object detection via Kullback-Leibler divergence. In: Advances in Neural Information Processing Systems, vol. 34, pp. 18381–18394. Curran Associates, Red Hook (2021). https://proceedings.neurips.cc/paper_files/paper/2021/hash/98f13708210194c475687be6106a3b84-Abstract.html

[19] Kullback, S., Leibler, R.A.: On information and sufficiency. The Annals of Mathematical Statistics 22(1), 79–86 (1951). https://doi.org/10.1214/aoms/1177729694

[20] Yang, X., Zhou, Y., Zhang, G., et al.: The KFIoU loss for rotated object detection. In: Proceedings of the 11th International Conference on Learning Representations (2023). https://openreview.net/forum?id=qUKsCztWlKq

[21] Lin, T.-Y., Dollár, P., Girshick, R., et al.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125. IEEE, Piscataway (2017). https://doi.org/10.1109/CVPR.2017.106

[22] Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448. IEEE, Piscataway (2015). https://doi.org/10.1109/ICCV.2015.169

[23] Pasquier, J.T., David, R.O., Freitas, G., et al.: The Ny-Ålesund Aerosol Cloud Experiment (NASCENT): overview and first results. Bulletin of the American Meteorological Society 103(11), E2533–E2558 (2022). https://doi.org/10.1175/BAMS-D-21-0034.1

[24] Zhou, Y., Yang, X., Zhang, G., et al.: MMRotate: a rotated object detection benchmark using PyTorch. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 7331–7334. ACM, New York (2022). https://doi.org/10.1145/3503161.3548541

[25] Deng, J., Dong, W., Socher, R., et al.: ImageNet: a large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE, Piscataway (2009). https://doi.org/10.1109/CVPR.2009.5206848

[26] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proceedings of the 7th International Conference on Learning Representations (2019). https://openreview.net/forum?id=Bkg6RiCqY7

[27] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations (2015). https://www.robots.ox.ac.uk/~vgg/publications/2015/Simonyan15/

[28] Huang, G., Liu, Z., van der Maaten, L., et al.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708. IEEE, Piscataway (2017). https://doi.org/10.1109/CVPR.2017.243

[29] He, K., Zhang, X., Ren, S., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE, Piscataway (2016). https://doi.org/10.1109/CVPR.2016.90

[30] Liu, Z., Mao, H., Wu, C.-Y., et al.: A ConvNet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986. IEEE, Piscataway (2022). https://doi.org/10.1109/CVPR52688.2022.01167

[31] Ren, S., He, K., Girshick, R., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(6), 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031

[32] Lyu, C., Zhang, W., Huang, H., et al.: RTMDet: an empirical study of designing real-time object detectors. arXiv:2212.07784v2 [preprint] (2022). https://doi.org/10.48550/arXiv.2212.07784

[33] Jocher, G., Qiu, J.: Ultralytics YOLO11, version 11.0.0 [software]. Ultralytics (2024). https://github.com/ultralytics/ultralytics. Accessed 30 August 2026.

[34] Zhao, Y., Lv, W., Xu, S., et al.: DETRs beat YOLOs on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16965–16974. IEEE, Piscataway (2024). https://doi.org/10.1109/CVPR52733.2024.01605