

Article

Improving Open-Set Recognition in Long-Tailed Underwater Optical Monitoring via Frequency-Shrunk Prototypes and Group-Conditional Risk Calibration

Dingzhao Li ^{1,2,3,4,†}, Chuanzhi Liu ^{3,†}, Mingzhang Zhou ^{2,3,4,*}  and Haixin Sun ^{2,3,4} 

¹ School of Smart Marine Science and Technology, Fujian University of Technology, Fuzhou 350118, China; lidingzhao@stu.xmu.edu.cn

² School of Informatics, Xiamen University, Xiamen 361005, China; hxsun@xmu.edu.cn

³ Key Laboratory of Southeast Coast Marine Information Intelligent Perception and Application, Ministry of Natural Resources, Zhangzhou 363000, China; liucz@cug.edu.cn

⁴ Key Laboratory of Underwater Acoustic Communication and Marine Information Technology (Ministry of Education), School of Informatics, Xiamen University, Xiamen 361005, China

* Correspondence: mzzhou@xmu.edu.cn

† These authors contributed equally to this work.

Highlights

What are the main findings?

- Frequency-Shrunk Prototype and Risk Calibration (FSRPC) performs frequency-dependent shrinkage of class prototypes and intraclass scales and fits head-, middle-, and tail-specific split-conformal rejection thresholds using known category data only.
- On the predefined fold-0 leakage-controlled protocols, FSPRC achieves open-set accuracies of 83.31% and 98.19% and unknown category macro area under the receiver operating characteristic curve (AUROC) values of 92.62% and 99.55% on AQUA20 and Fish4Knowledge-23, respectively.

What are the implications of the main findings?

- Frequency-aware geometric stabilization and risk calibration improve the balance between retaining data-scarce known organisms and rejecting previously unseen organisms.
- The known-only post hoc design can enhance frozen classifiers for continuous underwater optical monitoring without auxiliary outlier-exposure data.

Abstract

Camera-based underwater optical monitoring enables persistent and non-invasive observation of marine communities, but its automated interpretation is challenged by two coupled uncertainties: the natural long-tailed frequency distribution of observed organisms and the appearance of categories absent from model development. Rare known categories provide insufficient observations for reliably estimating class location and dispersion, whereas dominant categories may form broad acceptance regions that absorb previously unseen organisms. We propose Frequency-Shrunk Prototype and Risk Calibration (FSRPC), a known-only post hoc framework for long-tailed underwater optical monitoring built on an ImageNet-pretrained ConvNeXt-Tiny. FSPRC regularizes each empirical feature direction toward its corresponding classifier direction and shrinks its intraclass radius toward a frequency-group reference, with both corrections controlled by the number of training observations. Local prototype deviation and global energy evidence are then robustly normalized within the predicted head, middle, or tail group and combined without a trainable fusion module. Independent known-only calibration subsets determine group-conditional



Academic Editor: Gabriel Navarro

Received: 19 August 2026

Revised: 20 September 2026

Accepted: 21 September 2026

Published: 16 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

rejection thresholds, while held-out categories remain excluded from backbone adaptation, checkpoint selection, geometric estimation, score normalization, and threshold fitting. On the predefined fold-0 leakage-controlled protocols for AQUA20 and the track-disjoint Fish4Knowledge-23 dataset, FSPRC achieves open-set accuracies of 83.31% and 98.19%, open-set classification rate (OSCR) values of 85.96% and 98.96%, and unknown category macro area under the receiver operating characteristic curve (AUROC) values of 92.62% and 99.55%, respectively. Across exhaustive class-held-out partitions, open-set accuracy is $83.25 \pm 0.24\%$ over four AQUA20 folds and $98.18 \pm 0.08\%$ over five Fish4Knowledge-23 folds. These results demonstrate that frequency-dependent geometric stabilization and risk calibration can improve the recognition-stage balance between retaining data-scarce known organisms and referring unfamiliar observations for subsequent expert inspection.

Keywords: underwater image recognition; open-set recognition; long-tailed recognition; prototype shrinkage; risk calibration

1. Introduction

Marine fish communities provide important biological indicators for assessing aquatic ecosystem health, biodiversity patterns, and fishery resources. Camera-based underwater optical observation offers a non-invasive means of monitoring these communities, with fixed and mobile imaging platforms enabling repeated measurements across spatial and temporal scales that are difficult to achieve through capture-based surveys or manual diving observations. However, long-duration deployments generate large volumes of heterogeneous and temporally redundant imagery, making frame-by-frame expert annotation costly and impractical. Converting such observations into reliable ecological information therefore requires automated methods capable of recognizing marine organisms under complex underwater conditions, placing this task at the intersection of underwater optical sensing, computer vision, and marine ecology [1–3]. Long-term observational resources such as Fish4Knowledge support fish recognition in natural underwater environments, whereas AQUA20 further captures visual variations associated with turbidity, weak illumination, occlusion, and cluttered backgrounds [2,4]. Recent hierarchical vision-transformer models also demonstrate that modern visual representations can extract fine-grained species cues from body contours, textures, and local anatomical structures in challenging underwater imagery [5].

Recognition in the field remains substantially harder than classification in curated imagery. Light absorption and scattering alter color, suppress contrast, and remove local detail. Suspended particles, water currents, motion blur, nonuniform illumination, and coral backgrounds increase the appearance variation within the same species [6]. Meanwhile, closely related species may retain similar shapes and textures even when image quality is high. ConvNeXt provides a strong hierarchical representation for such visual recognition problems [7], and lightweight feature scaling and shifting can adapt pretrained models to specialized domains [8]. Yet the dominant formulation remains closed set: the model is trained and evaluated under an identical class vocabulary. A long-running camera does not satisfy this assumption. It may observe a new species, a juvenile or hybrid individual, or a non-target organism that was never represented in the training data. A closed-set classifier must still assign such an observation to one of its known categories, producing a confident label where rejection would be more appropriate.

The known categories themselves are rarely balanced. Common species account for most recorded images, while rare, seasonal, or behaviorally elusive species provide limited

supervision. This long-tailed distribution biases both the learned representation and the classifier toward head categories. Existing long-tailed recognition methods address the imbalance through resampling, loss reweighting, margin adjustment, representation–classifier decoupling, contrastive learning, and multi-expert models [9]. Representative examples include the effective-number-based class-balanced loss [10], label-distribution-aware margin learning [11], decoupled representation and classifier optimization [12], prior-aware logit adjustment [13], parametric contrastive learning [14], and complementary expert learning [15]. Prototype constraints have also been used to improve the representation of underrepresented categories [16]. These approaches can reduce closed-set bias, but they normally presume that every test image belongs to a training category. Improving the confidence of a rare known class does not by itself reveal whether a visually similar input comes from an unseen class; in some cases it makes the incorrect acceptance of that input more likely.

Open-set recognition addresses the complementary requirement of retaining known-class discrimination while rejecting categories unavailable during training. Work in this area has developed from open-space risk theory [17] and the extreme-value formulation of OpenMax [18] to adversarial reciprocal point learning [19], analyses of the relation between closed- and open-set performance [20], and broader studies of deep open-set recognition [21]. Closely related out-of-distribution (OOD) detectors use confidence, energy, or feature residuals to identify anomalous inputs. Energy-based detection summarizes the complete known-class logit response [22], whereas virtual-logit matching (ViM) combines a feature residual with a virtual logit [23]. Most of these detectors are developed with approximately balanced in-distribution data. Their score normalization and decision threshold are consequently shared across categories, although head and tail classes can have markedly different feature compactness, classifier confidence, and energy distributions.

We distinguish the two terms used in these related settings. An *unknown category* is a semantic category in the held-out label set $\mathcal{C}_U$ and is absent from every model-development stage. OOD detection is broader: an input may be OOD because of semantic novelty, sensor or environmental change, corruption, or another distributional shift, without necessarily belonging to a new semantic category. Thus, the held-out-category samples in our protocol constitute semantic OOD inputs relative to the known-class development distribution, but an arbitrary OOD input need not be an unknown category. We therefore use *unknown category* for our protocol, test samples, and metrics, and retain *OOD detection* only when referring to the broader literature or the original scope of an OOD baseline.

Long-tailed and open-set recognition therefore cannot be treated as two independent corrections. A tail-class center and its intraclass dispersion are estimated from few observations, so valid tail samples can resemble the outliers targeted by a conventional detector. At the opposite end of the distribution, head categories tend to support broader decision regions and can absorb genuinely unknown samples. Early open long-tailed approaches used dynamic meta-embeddings and memory mechanisms [24,25], while Localizing Unfamiliarity Near Acquaintance (LUNA) measured unfamiliarity relative to local known-class neighborhoods [26]. Subsequent work has studied long-tailed OOD detection and class-aware bias correction through Expand-and-Augment (EAT) [27], calibrated outlier class learning [28], and ImOOD [29]. Orthogonal prototypes, false rejection correction, and tail-prioritized modeling provide related directions [30,31]. In underwater sonar recognition, dynamic-margin contrastive learning and class-balanced OpenMax also show that class imbalance and unknown targets should be considered jointly [32,33]. These methods primarily address natural images or acoustic/sonar signals, and several rely on auxiliary outliers, synthesized unknowns, or substantial end-to-end modification. Optical

underwater recognition presents a different combination of degraded image formation, fine-grained interspecies similarity, and unreliable rare-class statistics.

Figure 1 illustrates the setting considered in this work. Known categories are ordered by training frequency and divided into head, middle, and tail groups. The reduction in statistical support toward the tail affects two geometric quantities at once: the empirical class direction and the scale of within-class variation. The open-world region contains categories excluded from representation learning, feature-statistic estimation, score normalization, threshold calibration, and model selection, but those categories must be rejected at test time. The unknown examples in the figure depict this protocol and are not used as auxiliary outlier data. The boundary consequently has two failure directions. A permissive known region accepts an unseen category, while an aggressive rejection rule discards a valid but uncertain tail sample.

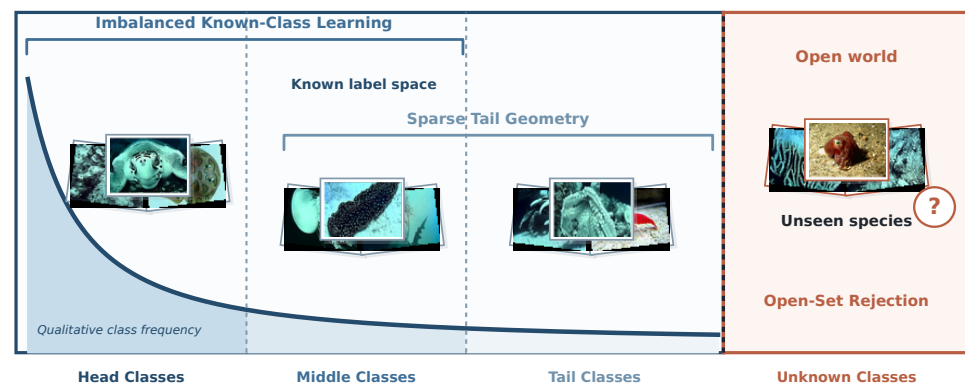


Figure 1. Open long-tailed underwater organism recognition considered in this work. The known label space follows a long-tailed class-frequency distribution and is partitioned into head, middle, and tail groups, whereas the held-out categories form a disjoint label set encountered only during open-set evaluation. The diminishing statistical support toward the tail makes the estimated prototype directions and intraclass scales increasingly unreliable. Images from the held-out categories are shown only to illustrate the evaluation setting and provide no supervision for model construction, calibration, or model selection. The question mark denotes the unknown semantic identity of an unseen species before the open-set decision.

This setting exposes three limitations of existing detectors. First, errors in the center and radius of a rare class jointly distort distance-based novelty scores; regularizing only one of these quantities leaves the other unstable. Second, feature-space deviation and classifier response abnormality describe different failure modes. The former measures compatibility with the local geometry of the predicted class, whereas the latter measures whether the classifier produces a typical aggregate response over all known classes. Their values cannot be combined directly because their scales and distributions vary with class frequency. Third, a global threshold controls only the aggregate rejection rate and can be dominated by head predictions. Conformal prediction provides a finite-sample view of quantile-based calibration [34], and tail-aware conformal analysis has shown that marginal and class-conditional behavior can differ markedly under long-tailed data [35]. What is missing is a compact, known-only construction that stabilizes rare-class geometry, makes complementary anomaly scores comparable across frequency groups, and assigns the rejection operating point without observing a held-out category.

Within this underwater optical monitoring setting, we introduce Frequency-Shrunk Prototype and Risk Calibration (FSPRC). An ImageNet-pretrained ConvNeXt-Tiny is adapted using only the known training categories of the predefined fold and is frozen after checkpoint selection. Its 768-dimensional pre-classifier features and known-class logits are the only inputs to the detector. For each class, FSPRC interpolates the empirical feature

direction with the normalized classifier direction using a coefficient determined by the class sample count. The empirical intraclass radius is corrected by the same coefficient toward a pooled frequency-group radius. Head categories retain most of their observed geometry, while sparse tail categories receive stronger regularization of both location and scale.

The resulting geometry supports a radius-normalized prototype deviation that is paired with an energy score computed from the known-class logits. FSPRC standardizes the two scores using median and median absolute deviation (MAD) statistics estimated separately for predicted head, middle, and tail groups on known validation data. Their maximum forms a parameter-free anomaly union: either a marked geometric departure or an atypical global classifier response can raise the rejection score. A disjoint known calibration subset then determines one threshold for each predicted frequency group, with a global known threshold used when a group has insufficient calibration support. The procedure does not claim control over an arbitrary unknown distribution; calibration is used to limit the rejection of known observations without exposing held-out test classes. The contributions of this work are threefold:

- We introduce a frequency-dependent geometric estimator that jointly corrects the prototype direction and intraclass radius of each known category. It preserves the empirical geometry of well-supported classes while reducing the sampling sensitivity of tail-class estimates.
- We develop a dual-evidence detector that combines local prototype deviation with global energy evidence after robust normalization within the predicted frequency group. Group-conditional calibration separates anomaly scoring from the selection of a known-sample rejection operating point.
- We evaluate the method under predefined leakage-free protocols on AQUA20 and Fish4Knowledge-23. Held-out categories are excluded from backbone adaptation and every post hoc fitting stage, and the analysis reports known-frequency recognition, unknown detection, operating point errors, component effects, and category-level failure patterns.

The remainder of this paper is organized as follows. Section 2 presents the problem formulation, FSPRC, datasets, leakage-free evaluation protocol, implementation details, metrics, and comparison settings. Section 3 reports the comparative, ablation, sensitivity, robustness, and efficiency results. Section 4 interprets the findings and discusses deployment scope and limitations. Finally, Section 5 summarizes the principal findings.

2. Materials and Methods

2.1. Problem Formulation and Framework Overview

We consider open-set recognition under a long-tailed known-class distribution. Let $\mathcal{C}_K = \{1, \dots, K\}$ denote the set of known categories available during model construction, and let $\mathcal{C}_U$ denote the categories held out as unknown, where $\mathcal{C}_K \cap \mathcal{C}_U = \emptyset$. Known samples are partitioned into mutually disjoint training, validation, and calibration subsets, denoted by $\mathcal{D}_{tr}$, $\mathcal{D}_{val}$, and $\mathcal{D}_{cal}$, respectively. The final test set contains both known samples from $\mathcal{C}_K$ and samples from $\mathcal{C}_U$. No held-out category is used for representation learning, checkpoint selection, prototype estimation, score normalization, threshold calibration, or hyperparameter selection.

Accordingly, $\mathcal{C}_U$ denotes semantic categories held out by label, rather than generic OOD inputs caused only by covariate, sensor, or environmental shift. All unknown category metrics reported below are computed on samples from this fixed disjoint label set.

Throughout this paper, italic Latin and Greek symbols denote scalar quantities, indices, functions, or individual samples; bold lowercase symbols denote vectors; bold uppercase symbols denote matrices; calligraphic uppercase symbols denote sets; and upright roman

subscripts denote fixed semantic labels. Mathematical operators, textual labels, and numerical constants are typeset upright. In particular, $x \in \mathcal{X}$ denotes an image sample as an element of the input space, whereas $v(x)$, $h(x)$, and $z(x)$ are vectors.

An ImageNet-pretrained ConvNeXt-Tiny [7] is used as the unified feature backbone throughout the framework. Its final classifier is replaced by a K -way linear layer corresponding to the known categories of the current fold. For an input image x , the backbone produces a pre-classifier feature and known-class logits:

$$v(x) = f_{\theta}(x), \quad h(x) = \frac{v(x)}{\|v(x)\|_2}, \quad z(x) = Wv(x) + b, \quad (1)$$

where f_{θ} denotes ConvNeXt-Tiny parameterized by the vector θ ; $v(x) \in \mathbb{R}^d$, with $d = 768$, is the pre-classifier feature; $h(x)$ is the unit-normalized representation; and $z(x) \in \mathbb{R}^K$ contains the known-class logits. The classifier matrix and its class-specific row vectors are written as $W = [w_1, \dots, w_K]^T$, and b denotes the bias vector.

Complementary backbone adaptation. Before fitting the post hoc FSPRC detector, the ImageNet initialization is adapted to underwater imagery with two lightweight operators retained in the released backbone implementation. Underwater scale-shift fine-tuning (U-SSF) applies a learnable per-channel scale and offset after each ConvNeXt stage and before the final classifier. It is initialized as the identity and does not mix spatial locations; its role is to recalibrate global channel responses affected by underwater color attenuation, contrast loss, and illumination change. The AquaAdapter instead applies GroupNorm, a 1×1 channel reduction with ratio 8, a 3×3 depthwise convolution, GELU nonlinearities, and a 1×1 expansion inside a residual branch. Its expansion is zero-initialized, and the branch is initially scaled by 10^{-3} , so adaptation begins close to the pretrained mapping while gaining the capacity to model local texture, contour, and background distortions.

The two operators are therefore complementary rather than interchangeable: U-SSF provides a low-variance global affine correction with only two parameters per channel, whereas AquaAdapter provides nonlinear spatial residual correction that a channel-wise affine transform cannot express. Both are optimized during the progressive backbone-adaptation schedule and are frozen after known-validation checkpoint selection. FSPRC is then fitted on their fixed features and logits; it performs class geometry stabilization and open-set risk calibration but does not update either operator. All post hoc baselines in the comparison use this same frozen representation, so their differences are attributable to the open-set decision rule rather than to different backbone adaptation.

The principal challenge is that the reliability of known-class geometry depends strongly on class frequency. Data-rich head categories provide sufficient samples to estimate compact feature distributions, whereas the empirical centers and intraclass scales of tail categories can be substantially affected by individual observations. This asymmetry produces two competing errors. Large head-class decision regions may absorb genuinely unknown samples, while uncertain but legitimate tail samples may be rejected as novel. A globally shared distance scale or rejection threshold cannot explicitly balance these errors.

Figure 2 presents the proposed Frequency-Shrunk Prototype and Risk Calibration framework (FSPRC). The upper part illustrates its geometric motivation: empirical class centers and intraclass scales become increasingly unreliable from head to tail, motivating progressively stronger geometric shrinkage. The lower part shows the complete workflow. ConvNeXt-Tiny first produces the normalized feature and known-class logits. FSPRC then stabilizes the class prototype and radius according to training frequency, combines local prototype deviation with global energy evidence after group-wise normalization, and applies a group-conditional threshold to produce the final open-set decision.

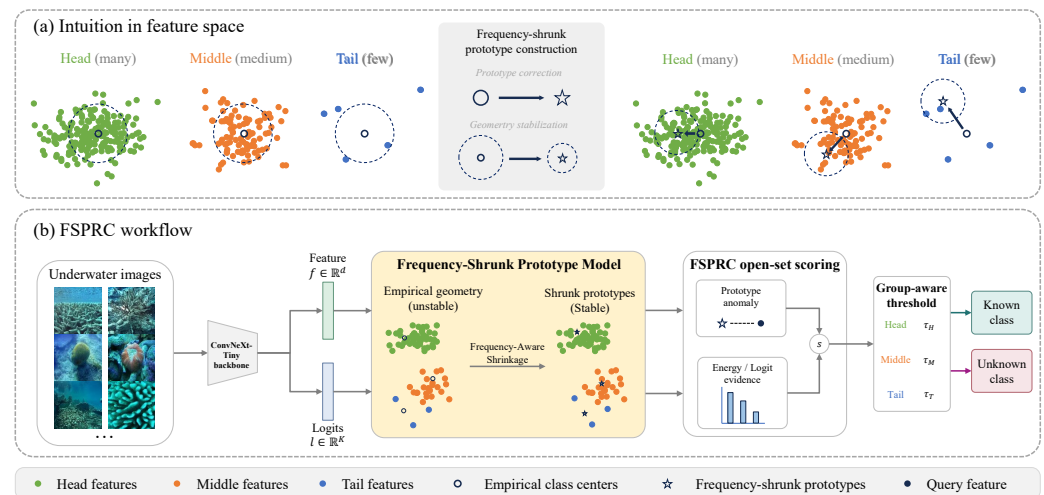


Figure 2. Overview of the proposed FSPRC framework. (**The upper part**) illustrates the feature-space motivation. Empirical class centers and intraclass scales become less reliable as the number of training samples decreases, motivating stronger geometric shrinkage for data-scarce categories. (**The lower part**) shows the complete workflow. An ImageNet-pretrained ConvNeXt-Tiny produces the normalized feature and known-class logits. FSPRC constructs frequency-shrunk class geometry, combines prototype deviation and energy evidence after group-wise robust normalization, and applies a group-conditional threshold to output either a known category or the unknown label. The ellipsis in the input block denotes additional underwater images not displayed for compactness.

2.2. Frequency-Shrunk Class Geometry

Prototype-based open-set detection requires both the location and the scale of each known-class distribution. Direct empirical estimates are relatively reliable for frequent categories but can exhibit high finite-sample variability for rare categories. FSPRC therefore regularizes the class direction and intraclass radius using a shared frequency-dependent coefficient. The amount of regularization decreases automatically as the number of available training samples increases.

For class c , let n_c denote its number of samples in $\mathcal{D}_{tr}$. Its empirical feature direction μ_c is computed from normalized ConvNeXt-Tiny features, while the corresponding normalized classifier weight $\bar{w}_c$ provides a parametric direction learned through supervised discrimination:

$$\mu_c = \frac{\sum_{i:y_i=c} h(x_i)}{\left\| \sum_{i:y_i=c} h(x_i) \right\|_2}, \quad \bar{w}_c = \frac{w_c}{\|w_c\|_2}, \quad (2)$$

$$\rho_c = \frac{\kappa}{n_c + \kappa},$$

where $\kappa > 0$ controls the overall shrinkage strength. The coefficient ρ_c decreases monotonically with n_c . Consequently, a data-rich category retains more of its empirical feature direction, whereas a data-scarce category borrows more information from the classifier direction.

The frequency-shrunk prototype is defined as

$$p_c = \frac{(1 - \rho_c)\mu_c + \rho_c\bar{w}_c}{\|(1 - \rho_c)\mu_c + \rho_c\bar{w}_c\|_2}. \quad (3)$$

This construction combines two complementary estimates of the class direction. The empirical center preserves the observed feature geometry, whereas the classifier direction incorporates the discriminative information accumulated during supervised training. Frequency-dependent interpolation prevents a small and potentially atypical tail sample set from determining the class prototype without regularization.

The interpolation in Equation (3) is the solution of an explicit regularized estimation problem rather than a lookup rule. Before unit normalization, let

$$\tilde{\mathbf{p}}_c = \arg \min_v \left\{ n_c \|\mathbf{v} - \boldsymbol{\mu}_c\|_2^2 + \kappa \|\mathbf{v} - \bar{\mathbf{w}}_c\|_2^2 \right\}.$$

Its unique minimizer is $\tilde{\mathbf{p}}_c = (1 - \rho_c)\boldsymbol{\mu}_c + \rho_c\bar{\mathbf{w}}_c$, after which Equation (3) projects the result onto the unit sphere. Hence, n_c determines the relative data precision and κ acts as a common regularization precision, or pseudo-count. The quadratic objective also admits a Gaussian maximum a posteriori interpretation, conditional on the assumed isotropic local error model. This interpretation motivates the functional form but does not make the reference direction or the unknown distribution identifiable from known samples.

The normalized classifier row should not be interpreted as a direction constructed from category priors or pooled head-class prototypes. FSPRC uses only class c 's own row $\bar{\mathbf{w}}_c$; normalization removes its magnitude, which can reflect frequency-dependent logit scale, before it is combined with the empirical class- c direction. This motivation is consistent with classifier normalization and decoupled long-tailed recognition [12]. Moreover, the frequency count enters only through the scalar ρ_c . Hence, $\mathbf{p}_c$ remains in the span of $\boldsymbol{\mu}_c$ and $\bar{\mathbf{w}}_c$, and no classifier direction from a head category is explicitly introduced. The empirical direction also retains weight $1 - \rho_c$, so the parametric direction regularizes rather than replaces the observed tail geometry.

Normalization cannot, however, guarantee an unbiased orientation. Because $\bar{\mathbf{w}}_c$ is learned against all competing categories, severe head-class dominance may still rotate a tail-class decision direction. FSPRC therefore uses this direction only as a class-specific shrinkage target whose influence $\rho_c = \kappa / (n_c + \kappa)$ decreases with the available class support; it is not interpreted as an unbiased estimate of the unknown class geometry. When both the empirical center and classifier direction are severely biased, an alignment-aware shrinkage rule or a tail-aware classifier would still be required. We retain this limitation explicitly rather than claiming that normalization alone removes representation-level head-class bias.

Stabilizing only the center is insufficient because the empirical dispersion of a rare category is also unreliable. The known categories are therefore sorted according to n_c and partitioned into approximately equal-sized head, middle, and tail groups. Let $g(c) \in \{H, M, T\}$ denote the frequency group of class c . At a fixed quantile level q , the empirical class radius $\hat{r}_c$ and pooled group radius R_g are estimated, and the former is shrunk toward the latter:

$$\begin{aligned} \hat{r}_c &= \text{Quantile}_q \left\{ 1 - \mathbf{h}(x_i)^\top \mathbf{p}_c : y_i = c \right\}, \\ R_g &= \text{Quantile}_q \left\{ 1 - \mathbf{h}(x_i)^\top \mathbf{p}_{y_i} : g(y_i) = g \right\}, \\ r_c &= \max \left[(1 - \rho_c)\hat{r}_c + \rho_c R_{g(c)}, \epsilon_r \right], \\ d_c(x) &= \frac{1 - \mathbf{h}(x)^\top \mathbf{p}_c}{r_c}. \end{aligned} \quad (4)$$

The quantile radius is less sensitive to isolated extreme observations than a maximum-based estimate. The lower bound $\epsilon_r > 0$ prevents numerical instability for highly compact categories. More importantly, $d_c(x)$ expresses the angular deviation relative to the estimated scale of class c , making geometric inconsistency more comparable across categories with different frequencies and compactness levels.

Before applying the lower bound ϵ_r , the radius interpolation in Equation (4) is analogously the unique minimizer of

$$n_c(r - \hat{r}_c)^2 + \kappa \left\{ r - R_{g(c)} \right\}^2.$$

The group statistic $R_{g(c)}$ is therefore a pooled, lower-variance reference for the class quantile; it is not used as a prototype center. Frequency grouping and quadratic regularization remain modeling assumptions rather than distribution-free guarantees.

2.3. Frequency-Normalized Dual Evidence

The stabilized class geometry provides a local measure of compatibility, but geometric deviation alone cannot capture every form of unknownness. An unknown image may lie close to one known prototype because of visual similarity while still producing an abnormal global classifier response. Conversely, an overconfident classifier may assign a large logit to a sample that lies outside the typical geometry of the predicted category. FSPRC therefore combines local prototype inconsistency and global energy evidence as complementary signals.

The preliminary known-class prediction, its associated frequency group, and the two raw anomaly scores are defined by

$$\begin{aligned}\hat{c}(x) &= \arg \max_c z_c(x), & \hat{g}(x) &= g(\hat{c}(x)), \\ a_P(x) &= d_{\hat{c}(x)}(x), & a_E(x) &= -T \log \sum_{c=1}^K \exp\left(\frac{z_c(x)}{T}\right),\end{aligned}\quad (5)$$

where $T > 0$ is the energy temperature. A large prototype score $a_P(x)$ indicates that the sample is inconsistent with the frequency-stabilized geometry of its predicted class. A large energy score $a_E(x)$ indicates that the known-class classifier provides an atypical aggregate response.

The two scores cannot be combined directly because they have different numerical scales and frequency-dependent distributions. Head, middle, and tail categories differ in feature compactness and classifier confidence. A global normalization dominated by head predictions can therefore preserve the original frequency bias and interpret the naturally broader tail distribution as novelty.

To obtain comparable evidence, FSPRC estimates robust normalization statistics exclusively from the known validation set. For score type $s \in \{P, E\}$ and predicted frequency group g , the group-wise median, median absolute deviation (MAD), and standardized score are

$$\begin{aligned}m_{s,g} &= \text{median}_{i:\hat{g}(x_i)=g} a_s(x_i), \\ \sigma_{s,g} &= 1.4826 \text{ median}_{i:\hat{g}(x_i)=g} |a_s(x_i) - m_{s,g}|, \\ u_s(x) &= \frac{a_s(x) - m_{s,\hat{g}(x)}}{\max(\sigma_{s,\hat{g}(x)}, \epsilon_s)}, \quad s \in \{P, E\}.\end{aligned}\quad (6)$$

The choice of median–MAD is based on replacement robustness rather than on an assumption that the score distribution is Gaussian. For a group containing m_g known-validation scores, replacing a single observation by an arbitrarily large value can move the sample mean and standard deviation without bound; their finite-sample replacement breakdown point is $1/m_g$. In contrast, the sample median and MAD remain bounded under contamination below approximately one half of the group. This distinction is especially relevant to the tail group, where a few difficult but valid known images can otherwise dominate the estimated location and scale. The factor 1.4826 makes the MAD consistent with the standard deviation under a Gaussian reference distribution without sacrificing this replacement robustness. If a group-wise MAD degenerates numerically, the corresponding standard deviation and then a unit scale are used as deterministic fallbacks.

Group-wise normalization is required for a separate reason. A single global affine transformation can center and scale all frequency groups only when their score locations

and scales coincide. Let $R_g(t) = \Pr\{S(X) > t \mid Y \in \mathcal{C}_K, G = g\}$ and $\pi_g = \Pr(G = g \mid Y \in \mathcal{C}_K)$. A global threshold controls only the mixture

$$R(t) = \sum_g \pi_g R_g(t),$$

which does not bound each $R_g(t)$; under long-tailed sampling, a large tail-group error can be masked by the dominant head-group weight. The group-wise transformation in Equation (6) first expresses each diagnostic relative to its own known-reference location and scale. Global known-validation statistics are used only when a predicted group contains no validation sample.

After normalization, FSPRC forms a parameter-free anomaly union:

$$S(x) = \max\{u_P(x), u_E(x)\}. \quad (7)$$

The maximum is selected to implement a disjunctive detector, not as a claim of universal optimality over all possible unknown distributions. For any threshold t , its acceptance region satisfies

$$\{\max(u_P, u_E) \leq t\} = \{u_P \leq t\} \cap \{u_E \leq t\}.$$

Thus, a query is accepted only when both the local geometric diagnostic and the global logit diagnostic remain within their calibrated known ranges. The maximum is coordinate-wise nondecreasing and cannot suppress a large positive anomaly in one channel with a negative standardized value in the other. Although an unweighted sum is also coordinate-wise nondecreasing, it does not preserve this intersection acceptance region and permits such cancellation. A product is unsuitable for the signed standardized scores because its sign and ordering change across quadrants and it is not coordinate-wise monotone on $\mathbb{R}^2$. Weighted sum or product rules would additionally require a fusion parameter fitted from limited known-only development data. The maximum therefore provides the parameter-free operator that exactly matches the intended “either-evidence” rejection semantics.

2.4. Group-Conditional Risk Calibration

The continuous anomaly score must be converted into a known–unknown decision. A single global threshold assumes that all known categories produce equivalent score distributions after normalization. This assumption remains restrictive in long-tailed recognition because tail categories have fewer samples and naturally broader uncertainty. A threshold dominated by head predictions may consequently reject a disproportionate number of valid tail samples.

FSPRC determines separate thresholds for the predicted head, middle, and tail groups using only the independent known calibration set. Let

$$\mathcal{S}_g = \{S(x_i) : x_i \in \mathcal{D}_{\text{cal}}, \hat{g}(x_i) = g\}$$

contain m_g calibration scores, and let $S_{g,(1)} \leq \dots \leq S_{g,(m_g)}$ denote their order statistics. For a prescribed known-sample false rejection level α , the group-wise threshold is

$$k_g = \min\{\lceil (m_g + 1)(1 - \alpha) \rceil, m_g\}, \quad \tau_g = S_{g,(k_g)}. \quad (8)$$

Here, α is a nominal known-calibration miscoverage level rather than a constraint that forces the realized test-set KFRR to equal α . For fixed scores, increasing α reduces k_g and

τ_g , and therefore weakly enlarges the rejected set. The resulting operating point sensitivity is examined in Section 3.5. The final open-set prediction is then

$$H(x) = \begin{cases} \text{unknown,} & S(x) > \tau_{\hat{g}(x)}, \\ \hat{c}(x), & S(x) \leq \tau_{\hat{g}(x)}. \end{cases} \quad (9)$$

Equation (9) also defines the functional boundary of FSPRC. If a query is accepted, its output remains exactly the preliminary known-class prediction $\hat{c}(x)$; the post hoc detector does not map one known label to another. Consequently, FSPRC can control whether a frozen prediction is retained or referred as unknown, but it cannot correct a known-to-known classification error already made by the backbone. This modular restriction isolates open-set risk calibration from representation learning and permits FSPRC to be attached to different classifiers, while placing its closed-set accuracy under the ceiling imposed by the selected backbone.

The class radius r_c and rejection threshold τ_g have distinct roles. The former estimates the within-class scale of a known category and normalizes prototype deviation; it is not selected as a boundary that is claimed to locate unknown categories. The latter specifies the tolerated known-sample rejection operating point using an independent calibration set. Under exchangeability between known calibration and future known samples within predicted group g , the rank argument for the selected order statistic gives

$$\Pr\{S(x_{\text{new}}) > \tau_g \mid \hat{g}(x_{\text{new}}) = g\} \leq \frac{m_g + 1 - k_g}{m_g + 1} \leq \max\left\{\alpha, \frac{1}{m_g + 1}\right\}.$$

The term $1/(m_g + 1)$ is the finite-sample resolution of the group-wise calibration. The support m_g is pooled over all calibration observations whose predicted class belongs to group g ; it is not the calibration count of one tail class. We require $m_g \geq m_{\min} = 20$ before fitting a group-specific threshold. If this condition is not met, FSPRC replaces τ_g with the global known-sample threshold. This rule prevents a threshold from being estimated from only 2–3 observations, but it is a variance-limiting safeguard rather than proof of threshold stability. Only the corresponding global known-sample interpretation then applies. Crucially, this statement bounds rejection of future known observations under exchangeability. It provides no guarantee on the false acceptance of an arbitrary unknown distribution. Because the latter is unobserved, no finite known-only training or calibration set can certify a radius or threshold that is simultaneously optimal for tail retention and unknown rejection.

2.5. Training and Inference

FSPRC follows a sequential known-only construction protocol. First, ConvNeXt-Tiny is initialized from ImageNet weights, and its final classifier is reconstructed for the K known categories of the current fold. The backbone and classifier are adapted using only $\mathcal{D}_{\text{tr}}$, while the closed-set checkpoint is selected according to known-validation performance. A checkpoint trained with the complete class vocabulary is prohibited because it would encode categories designated as unknown in the current evaluation.

After checkpoint selection, ConvNeXt-Tiny is frozen. The normalized training features and classifier directions are used to construct the frequency-shrunk prototypes and radii. The frozen model is subsequently applied to $\mathcal{D}_{\text{val}}$ to estimate the group-wise median–MAD statistics for prototype and energy evidence. Finally, the independent known calibration subset $\mathcal{D}_{\text{cal}}$ determines the head, middle, and tail rejection thresholds. All model parame-

ters, geometric statistics, normalization parameters, and thresholds are fixed before the test set is accessed.

Algorithm 1 summarizes the known-only construction and open-set inference procedures. The three known-data subsets have non-overlapping roles: training samples determine the representation and class geometry, validation samples determine the checkpoint and normalization statistics, and calibration samples determine the final rejection thresholds.

Algorithm 1 Construction and Inference of FSPRC

Require: Known subsets $\mathcal{D}_{\text{tr}}$, $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{cal}}$; query image x

Ensure: Open-set prediction $H(x)$

- 1: Train ConvNeXt-Tiny on $\mathcal{D}_{\text{tr}}$ and select the checkpoint on $\mathcal{D}_{\text{val}}$
 - 2: Freeze the backbone and divide known classes into frequency groups
 - 3: **for** each known class c **do**
 - 4: Construct the shrunk prototype p_c and radius r_c
 - 5: **end for**
 - 6: Estimate group-wise score statistics on $\mathcal{D}_{\text{val}}$
 - 7: Calibrate $\{\tau_H, \tau_M, \tau_T\}$ on $\mathcal{D}_{\text{cal}}$
 - 8: Obtain $h(x)$ and $z(x)$ from ConvNeXt-Tiny
 - 9: Predict $\hat{c}(x)$ and $\hat{g}(x) = g(\hat{c}(x))$
 - 10: Compute $S(x) = \max\{u_P(x), u_E(x)\}$
 - 11: **if** $S(x) > \tau_{\hat{g}(x)}$ **then**
 - 12: **return** unknown
 - 13: **else**
 - 14: **return** $\hat{c}(x)$
 - 15: **end if**
-

At inference, each image requires one ConvNeXt-Tiny forward pass. The logits provide the preliminary known-class prediction and energy evidence, while the normalized pre-classifier feature provides the prototype deviation. The two anomaly scores are standardized using the statistics of the predicted frequency group and combined through Equation (7). The corresponding group threshold then determines whether the preliminary known-class prediction is retained or replaced by the unknown label.

The post hoc FSPRC detector introduces neither an additional backbone forward pass nor a trainable fusion network. Its stored state consists of K class prototypes, K class radii, a small set of group-wise normalization statistics, and three calibrated thresholds. For each test image, prototype scoring requires $\mathcal{O}(Kd)$ operations, energy computation requires $\mathcal{O}(K)$ operations, and group-threshold lookup requires $\mathcal{O}(1)$ time. Offline calibration requires sorting the known calibration scores and is performed only once. The additional memory complexity is $\mathcal{O}(Kd)$, which is small relative to the ConvNeXt-Tiny backbone.

2.6. Experimental Design

The experiments address five questions. First, can FSPRC reject previously unseen categories while retaining known samples, particularly those from tail classes? Second, do prototype and energy evidence provide complementary open-set information, and does group-conditional calibration improve the selected operating point over a global threshold? Third, do the group-level and score-level diagnostics support the intended behavior of the detector? Fourth, is the result stable to the semantic composition of the held-out category set and to calibration repartitioning? Fifth, does the post hoc gain persist when the backbone is trained with long-tail-aware objectives? Model fitting, configuration selection, and threshold calibration are completed without access to held-out unknown test samples. The main all-method and ablation tables use predefined fold 0 for controlled, sample-aligned comparisons. Robustness to held-out-category composition is evaluated separately over

four exhaustive AQUA20 folds and five exhaustive Fish4Knowledge-23 folds. One fixed seed is used per fold; the reported variation therefore characterizes class-held-out partition sensitivity, not repeated-seed optimization variability.

2.6.1. Datasets and Leakage-Free Evaluation Protocol

AQUA20. AQUA20 comprises 8171 images from 20 underwater organism categories. Its official training and test partitions contain 6559 and 1612 images, respectively. Class frequencies in the training partition range from 20 to 2200 images, yielding an imbalance ratio of 110:1. We use predefined fold 0, in which 15 categories are treated as known and five categories—flatworm, sea anemone, sea cucumber, shark, and starfish—are held out as unknown. For each known category, the official training partition is divided in a class-stratified manner into training, validation, and calibration subsets using an 80/10/10 ratio. The official test partition provides 1309 known and 303 unknown test images. All training images from the five held-out categories are discarded and are not reused as auxiliary outliers, pseudo-unknown samples, or calibration data.

Fish4Knowledge-23. The Fish4Knowledge-23 extraction contains 27,370 images from 23 fish categories and exhibits an imbalance ratio of 757:1. Because adjacent video frames may contain nearly identical observations, image-level random splitting can introduce substantial temporal leakage. We therefore derive track identifiers from the acquisition sequences and assign all frames from the same track to a single partition. Tracks from the known categories are divided into training, validation, calibration, and test subsets using a 70/10/10/10 ratio. Predefined fold 0 contains 18 known categories and holds out `fish_05`, `fish_06`, `fish_09`, `fish_18`, and `fish_23`. The resulting test partition contains 2540 known and 3046 unknown images. Every track from a held-out category is assigned exclusively to the unknown test partition.

Frequency Groups. For each fold, the grouping is computed exclusively from the known training partition after the held-out categories have been removed. Let K be the number of known categories and let n_c be the number of training samples of category c . Categories are sorted by decreasing n_c ; equal counts are broken deterministically by the category identifier in ascending lexicographic order. Writing $K = 3q + r$, where $r \in \{0, 1, 2\}$, the sorted list is divided into three contiguous blocks. The first r blocks contain $q + 1$ categories, and the remaining blocks contain q categories; the blocks are labeled head, middle, and tail in that order. Consequently, group cardinalities differ by at most one. AQUA20 fold 0 has $K = 15$ and therefore uses a 5/5/5 partition, whereas Fish4Knowledge-23 fold 0 has $K = 18$ and uses a 6/6/6 partition. No validation, calibration, known-test, or unknown-test count enters this assignment.

The assignments are recomputed independently for every fold and then remain fixed for prototype shrinkage, score normalization, group-conditional calibration, and group-level reporting. At inference, a sample inherits the group of its predicted known category; a held-out category is never assigned a group for fitting. The released fold-specific `class_groups.json` file records the group lists, training counts, frequency ranks, and class-to-group identifiers, making the partition exactly reproducible. This rank-based rule avoids dataset-specific absolute frequency thresholds while maintaining comparable numbers of categories in the three calibration groups.

Figure 3 summarizes the class-frequency distributions under the fixed fold-0 protocols. **Protocol Integrity.** Each fold records the known and held-out category lists, partition manifest, class index mapping, frequency-group assignment, and random seed. We also store cryptographic hashes of the protocol, configuration, implementation, and frozen checkpoint to identify the exact experimental state. Before execution, the pipeline verifies that held-out categories are absent from representation learning, validation, calibration,

prototype estimation, and checkpoint selection. For Fish4Knowledge-23, it additionally verifies that no track identifier occurs in more than one partition. A run is terminated if any condition is violated, making the leakage constraints automatically enforceable rather than dependent on manual inspection.

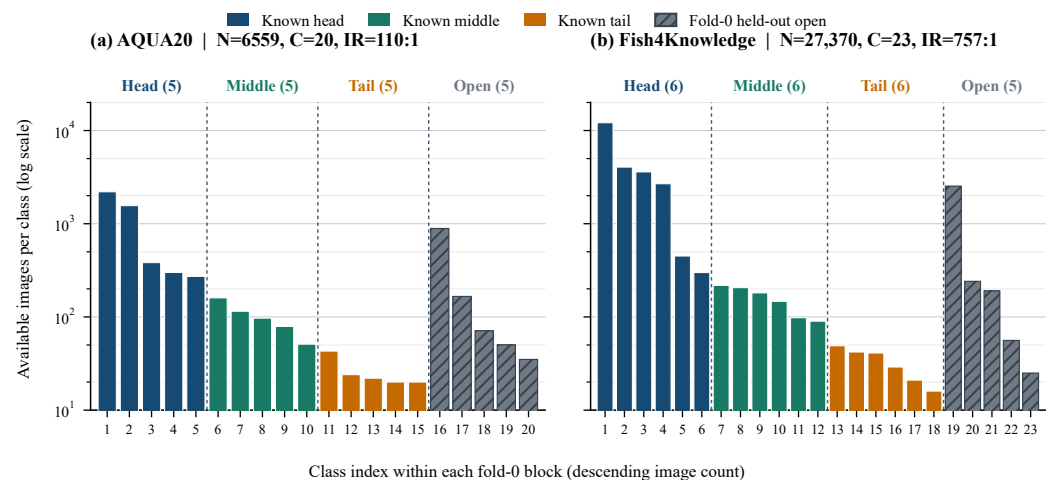


Figure 3. Class-frequency distributions under the predefined fold-0 protocols of AQUA20 and Fish4Knowledge-23. Categories are ordered by decreasing sample count within the known head, known middle, known tail, and held-out unknown blocks. Vertical dashed lines delimit the four blocks, while the logarithmic ordinate reveals both dominant head categories and sparsely represented tail categories. AQUA20 provides 6559 training images from 20 categories with an imbalance ratio of 110:1; Fish4Knowledge-23 contains 27,370 extracted images from 23 categories with an imbalance ratio of 757:1.

2.6.2. Implementation and Reproducibility

An ImageNet-pretrained ConvNeXt-Tiny with a K -way classifier is trained independently for each fold, and its checkpoint is selected using the macro-averaged F1 score on the known validation subset. After selection, the backbone and classifier are frozen. Frequency-shrunk prototypes and radii are estimated from the known training subset; group-wise median–MAD statistics are fitted on the known validation subset; and rejection thresholds are obtained from the disjoint known calibration subset. Held-out test samples are evaluated only after all model parameters, score definitions, and thresholds have been fixed. The experiments were implemented in Python 3.10 using PyTorch 2.1 and torchvision 0.16; the released environment additionally specifies NumPy 1.24, pandas 2.0, SciPy 1.9, scikit-learn 1.2, Pillow 9.5, PyYAML 6.0, and Matplotlib 3.7 as the minimum reproducible versions.

Training-time augmentation is restricted to the known training partition. Classes whose training counts do not exceed the fold-specific 35th percentile receive the stronger transform in Table 1; all other known classes receive the mild transform. This frequency-aware policy increases appearance and framing variation for sparsely sampled classes without duplicating observations or introducing auxiliary unknown data. Validation, calibration, and test images use the same deterministic preprocessing and never receive stochastic augmentation. The settings are identical for AQUA20 and Fish4Knowledge-23.

Table 1. Training and augmentation settings shared by AQUA20 and Fish4Knowledge-23. The color-jitter tuple lists brightness, contrast, saturation, and hue.

Component	Setting
Input and evaluation	RGB, 224×224 ; bilinear resize; ImageNet normalization with mean (0.485, 0.456, 0.406) and standard deviation (0.229, 0.224, 0.225)
Mild training transform	Random resized crop, scale [0.86, 1.00], aspect ratio [0.90, 1.10]; horizontal flip ($p = 0.50$); color jitter (0.10, 0.10, 0.08, 0.02) ($p = 0.35$); autocontrast ($p = 0.10$)
Tail training transform	Applied when n_c is no greater than the 35th percentile of known training counts: random resized crop, scale [0.76, 1.00], aspect ratio [0.85, 1.15]; horizontal flip ($p = 0.50$); color jitter (0.18, 0.18, 0.14, 0.04) ($p = 0.55$); autocontrast ($p = 0.25$); sharpness factor 1.5 ($p = 0.15$)
Batch regularization	Mixup probability 0.10 with Beta(0.2, 0.2); label smoothing 0.02
Optimization	AdamW; batch size 32; automatic mixed precision; weight decay 0.05 for the classifier/backbone and 0 for scale-shift parameters; learning rates 5×10^{-4} (classifier), 5×10^{-5} (backbone), and 10^{-3} (scale-shift)
Schedule and selection	At most 50 epochs: classifier/adapters for epochs 1–5, last stage/adapters for epochs 6–30, and full model thereafter; five-epoch linear warm-up followed by cosine decay to 1% of each base learning rate; patience 15; checkpoint selected by known-validation macro F1 after at least 15 full-model epochs
Reproducibility	Seed 2026; deterministic execution; no rotation, vertical flip, CutMix, RandAugment, or unknown/OOD exposure

The pseudo-unknown samples used in the validation-only configuration analysis are constructed exclusively from known development data and are not retained for final testing. Their generation settings, together with all training and calibration hyperparameters, are stored in the released fold-specific configuration files. The public code repository also retains the fold definitions, track manifests, per-sample predictions, method manifests, and aggregate metrics required to reproduce the reported comparisons.

2.6.3. Evaluation Metrics

Closed-set accuracy is measured on known test samples before rejection. Acc_H , Acc_M , and Acc_T are sample-weighted accuracies within the known head, middle, and tail groups. Open-set accuracy maps every rejected sample to one unknown label. Unknown recall (U-Rec) is the sample-weighted (micro) recall over all unknown test samples, and the known false rejection rate (KFRR) is the sample-weighted false rejection rate over known test samples. For each held-out class, we also compute the one-vs-all-known area under the receiver operating characteristic curve (AUROC) and area under the precision–recall curve (AUPR); the main-table macro AUROC (M-AUROC) and macro AUPR (M-AUPR) entries are the unweighted means of these class-level values. This class-macro convention differs from the sample-level global AUROC/AUPR retained in the result files. The false-positive rate at 95% true-positive rate (FPR95) is computed from the pooled known–unknown scores, and the open-set classification rate (OSCR) couples correct known classification with unknown rejection. AUROC, AUPR, and OSCR are threshold independent; the remaining rejection metrics describe the calibrated operating point.

Two operating point errors expose the long-tail asymmetry hidden by aggregate detection scores. For the binary rejection indicator $\mathbb{I}_{\text{rej}}(x) \in \{0, 1\}$, the tail false rejection rate (TFRR) and unknown false-acceptance rate (UFAR) are

$$\begin{aligned}
 \text{TFRR} &= \frac{1}{|\mathcal{D}_{\text{te}}^T|} \sum_{x \in \mathcal{D}_{\text{te}}^T} \mathbb{I}_{\text{rej}}(x), \\
 \text{UFAR} &= \frac{1}{|\mathcal{D}_{\text{te}}^U|} \sum_{x \in \mathcal{D}_{\text{te}}^U} [1 - \mathbb{I}_{\text{rej}}(x)].
 \end{aligned} \tag{10}$$

TFRR quantifies valid tail samples incorrectly rejected as novel; UFAR quantifies unknown observations absorbed into known classes. Lower values are better for FPR95, TFRR, and UFAR; higher values are better for all other metrics. Unless stated otherwise, operating point results use $\alpha = 0.05$.

2.6.4. Compared Methods and Fairness Controls

The fold-0 export contains 18 baselines spanning output confidence, activation shaping, feature geometry, and open long-tailed recognition. The output-space group comprises maximum softmax probability (MSP) [36], Energy [22], Generalized Entropy (GEN) [37], and OpenMax [18]. Feature- and activation-based detectors include ViM [23], DICE [38], Simplified Hopfield Energy (SHE) [39], nearest-neighbor guidance (NNGuide) [40], fast decision-boundary-based detector (fDBD) [41], SCALE [42], capturing extreme activations (CEA) [43], the Class-Aware Decoupled Relative Feature-based method (CADRef) [44], and the Neural Collapse Inspired OOD Detector (NCI) [45], together with the activation-shaping (ASH-S) method [46]. We further include an empirical prototype detector and the ImOOD framework for imbalanced OOD detection [29]. LUNA [26] and the OLPR method based on orthogonal prototype learning and false rejection correction [30] are retained as adapted comparisons. Independent open long-tailed recognition (OLTR) and preliminary ablation variants are excluded from the main tables; the final proposed detector is reported consistently as FSPRC.

All methods use the same predefined fold, known-only fitting partitions, frozen ConvNeXt-Tiny checkpoint, and held-out test samples, and none uses external outlier exposure. The detector-only baselines and FSPRC share the same cached ConvNeXt-Tiny logits and features. In the proposed post hoc formulation, FSPRC changes only the accept–reject decision: every accepted sample retains the preliminary ConvNeXt-Tiny class prediction. The adapted OLPR diagnostic fits a separate known-only head on the frozen features and can therefore alter closed-set predictions. Rows marked † are engineering adapters rather than paper-exact reproductions: ASH-S applies activation shaping at the available pooled feature interface, LUNA omits its original representation retraining, and OLPR implements its orthogonal prototype branch without the paper’s transductive test-time correction stage. These rows are retained as diagnostics, not as claims of paper-exact reproduction.

3. Results

3.1. AQUA20 Comparative Results

Table 2 reports all 19 methods from the frozen AQUA20 fold-0 export. It separates known recognition, the calibrated operating point, threshold-independent unknown detection, and open-set classification. A valid improvement should preserve tail recognition while increasing OSCR and should not obtain a lower UFAR merely by rejecting more known samples. The table is therefore ordered by open-set accuracy for readability, but conclusions are drawn from the complete metric profile rather than a single scalar.

Overall Comparison. FSPRC achieves the highest open-set accuracy (83.31%), OSCR (85.96%), unknown category macro AUROC (92.62%), and micro-averaged unknown recall (55.12%). Relative to the strongest included detector-only baseline for each measure, these values correspond to gains of 0.87 percentage points over SHE in open-set accuracy, 0.72 points over CADRef in OSCR, 2.21 points over CADRef in macro AUROC, and 3.30 points over SHE in unknown recall. The method also improves closed-set accuracy to 93.35% and tail-group accuracy to 80.77%, the latter exceeding the adapted OLPR variant by 3.85 points. The improvement in open-set performance therefore does not arise from indiscriminately rejecting additional known samples.

Table 2. Descriptive single-fold open long-tailed comparison on predefined AQUA20 fold 0. All entries are percentages, and methods are ordered by open-set accuracy. Bold dark-green and underlined dark-ochre entries denote the best and second-best values, respectively, with metric direction taken into account. Upward and downward arrows indicate that higher and lower values are preferable, respectively. Multiway ties involving more than two methods are left unmarked to avoid visual clutter; boldface on FSPRC identifies the proposed method.

Method	Known Recognition				Operating Point			Unknown Detection			Open Classification	
	Closed↑	H↑	M↑	T↑	KFRR↓	U-Rec↑	UFAR↓	M-AUROC↑	M-AUPR↑	FPR95↓	OSCR↑	Open Acc.↑
SHE	92.74	93.36	96.09	71.15	4.89	<u>51.82</u>	<u>48.18</u>	85.05	36.42	58.98	83.83	<u>82.44</u>
CADRef	92.74	93.36	96.09	71.15	4.89	45.54	54.46	<u>90.41</u>	29.99	<u>37.59</u>	<u>85.24</u>	81.76
Emp. prototype	92.74	93.36	96.09	71.15	5.58	49.84	50.16	90.30	32.00	<u>37.43</u>	85.00	81.58
GEN	92.74	93.36	96.09	71.15	4.89	35.31	64.69	87.99	31.65	86.48	77.58	79.90
NNGuide	92.74	93.36	96.09	71.15	4.97	30.36	69.64	86.42	23.73	42.25	82.37	78.91
OLPR [†]	<u>93.20</u>	<u>93.53</u>	96.88	<u>76.92</u>	5.96	29.37	70.63	88.52	21.21	45.91	82.89	78.66
fDBD	92.74	93.36	96.09	71.15	5.65	30.36	69.64	89.88	22.42	45.68	81.43	78.47
MSP	92.74	93.36	96.09	71.15	5.58	26.73	73.27	88.40	19.50	82.12	78.69	78.41
LUNA [†]	92.74	93.36	96.09	71.15	5.42	25.74	74.26	77.00	11.38	87.62	65.41	77.61
Energy	92.74	93.36	96.09	71.15	<u>4.05</u>	20.13	79.87	82.18	16.40	73.80	77.68	77.30
NCI	92.74	93.36	96.09	71.15	5.50	19.47	80.53	72.76	14.19	61.50	74.21	76.49
CEA	92.74	93.36	96.09	71.15	3.51	14.52	85.48	82.06	15.43	73.87	77.64	76.43
ASH-S [†]	92.74	93.36	96.09	71.15	5.65	21.12	78.88	74.26	12.45	76.24	74.12	76.36
SCALE	92.74	93.36	96.09	71.15	<u>4.05</u>	15.84	84.16	79.26	14.06	73.87	78.10	76.24
ViM	92.74	93.36	96.09	71.15	5.81	17.16	82.84	83.03	11.15	42.48	78.39	75.43
OpenMax	89.69	91.85	82.81	59.62	8.40	17.82	82.18	85.47	10.45	85.33	74.07	73.33
ImOOD	92.74	93.36	96.09	71.15	5.73	8.91	91.09	39.95	6.48	87.39	57.60	73.20
DICE	92.74	93.36	96.09	71.15	6.95	15.18	84.82	82.68	13.99	66.54	65.33	73.08
FSPRC	93.35	93.62	96.09	80.77	5.50	55.12	44.88	92.62	<u>34.21</u>	38.27	85.96	83.31

Notes: H/M/T denote sample-weighted head/middle/tail accuracies; U-Rec denotes sample-weighted unknown recall; and M-AUROC/M-AUPR denote unweighted means over held-out classes. The empirical prototype row is an internal known-only post hoc baseline; † denotes an adapted implementation rather than a paper-exact reproduction. The light-blue row denotes the proposed method. This table contains one predefined split only and supports descriptive comparisons only.

Metric Trade-offs. The advantage is not uniform across all criteria. SHE obtains the highest macro AUPR (36.42%, compared with 34.21% for FSPRC), while CEA achieves the lowest KFRR and the empirical prototype detector yields the lowest FPR95. Moreover, FSPRC attains an H-score of 68.32% by combining 89.84% known accuracy after rejection with 55.12% unknown recall, indicating that unknown category detection remains the principal limitation on AQUA20. Taken together, the results support stronger threshold-independent ranking and a more favorable calibrated operating point, but not uniform dominance over every evaluation metric.

Frequency-Dependent Recognition and Rejection. Figure 4a complements the sample-weighted group results in Table 2 with an unweighted class-macro view. Before rejection, the head, middle, and tail groups reach 92.37%, 96.85%, and 80.36% accuracy, respectively. After rejection, the corresponding values become 89.48%, 94.98%, and 66.73%, showing that the residual loss is concentrated in the data-scarce tail group rather than hidden by the dominant head-class sample count.

The improved aggregate performance does not eliminate rejection imbalance. As shown in Figure 4b, the class-macro false rejection rate is 4.84% for head classes and 2.35% for middle classes but rises to 21.45% for tail classes. The underlying class-level records further localize this error: marine dolphin and octopus, each with only ten test samples, reach false rejection rates of 40% and 30%, whereas frequent fish and coral remain near 5%. This single-fold pattern is diagnostic rather than inferential, but it exposes the calibration limitation hidden by aggregate known accuracy.

Held-out unknown categories are similarly heterogeneous. Figure 4c shows that starfish reaches 92.50% recall, whereas flatworm reaches only 7.69%. The 221-image held-out sea anemone class reaches 48.87% recall and therefore strongly influences the sample-weighted unknown-recall aggregate. Together, the panels show that the aggregate

advantage of FSPRC coexists with two remaining challenges: protecting valid tail samples and calibrating visually ambiguous unknown categories.

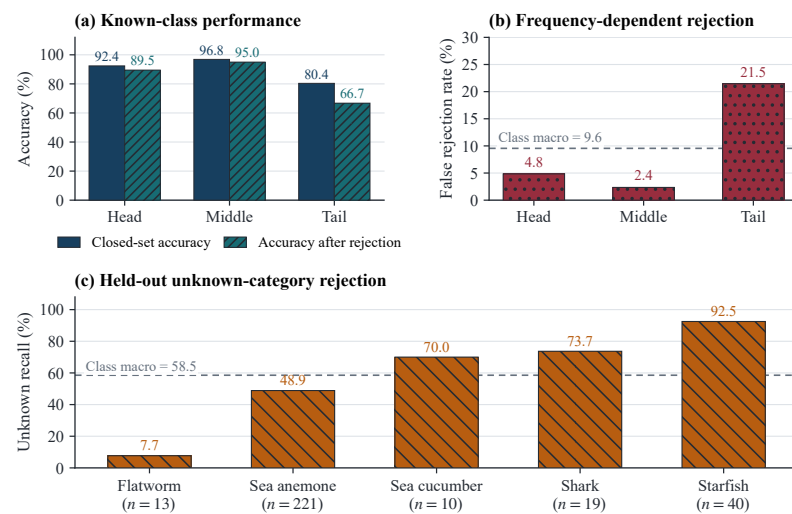


Figure 4. Frequency-aware diagnostic analysis of FSPRC on AQUA20 fold 0. (a) Unweighted class-macro closed-set accuracy and accuracy after open-set rejection across the head, middle, and tail known-class groups. (b) Class-macro known false rejection rate for the same frequency groups; the dashed line marks the overall class macro. (c) Per-category recall for the five held-out unknown classes, with test sample counts reported below the category names; the dashed line marks the unknown category macro. All bars start at zero. The figure contains one frozen fold, so no fold-variance bar is shown.

Qualitative Analysis of the FSPRC Decision Process. Figure 5 provides a stepwise view of how the proposed detector converts a closed-set representation into an open-set decision. In the ConvNeXt-Tiny space of Figure 5a, several held-out categories form compact local structures, but their samples remain interleaved with known clusters; feature compactness alone is therefore not equivalent to unknownness. After projection into the FSPRC evidence space, Figure 5b shows a clearer organization of most unknown observations along the right-hand branch of the manifold, while known samples predominantly occupy the broader left and lower regions. The remaining overlap demonstrates why neither prototype deviation nor classifier evidence is sufficient in isolation. Figure 5b–d share identical coordinates, Figure 5c directly relates this organization to the signed group-conditional margin: strongly negative regions are dominated by accepted known samples, strongly positive regions contain most rejected unknowns, and ambiguous observations concentrate near the zero-margin transition. The final decisions in Figure 5d are consistent with this interpretation. Among the 592 displayed samples, 407 known observations are correctly classified and accepted, 88 unknown observations are correctly rejected, and most of the 19 accepted known-class errors, 29 known false rejections, and 49 unknown false acceptances occur near mixed or transitional regions rather than inside well-separated clusters. This localization supports the intended role of frequency-aware calibration: it converts complementary evidence into a structured rejection boundary while exposing, rather than hiding, the residual known–unknown ambiguity. Since t-SNE preserves local neighborhoods but not global metric distances, the apparent intercluster gaps should be interpreted qualitatively; the complete-fold operating point and ranking claims are established by the quantitative results reported above.

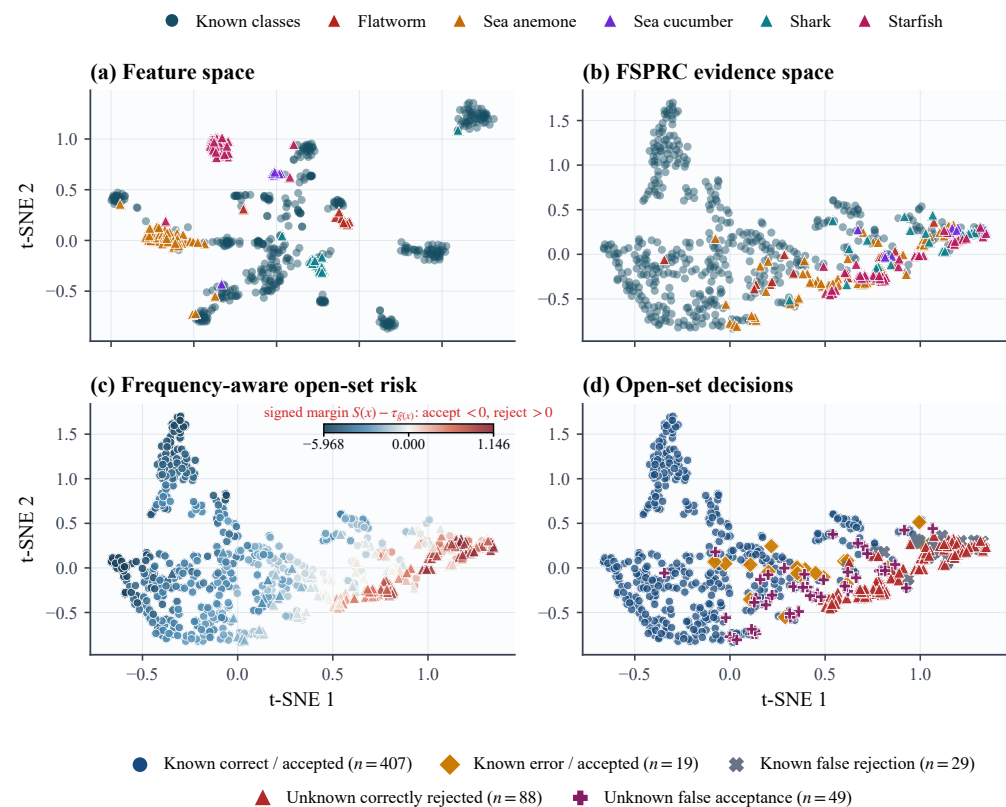


Figure 5. t-distributed stochastic neighbor embedding (t-SNE) visualization of the AQUA20 fold-0 open-set representation and decision process. (a) ConvNeXt-Tiny feature space. (b) FSPRC multi-evidence space. (c) Signed group-conditional risk margin $S(x) - \tau_{\hat{g}(x)}$, where negative and positive values correspond to acceptance and rejection, respectively. (d) Final open-set decisions, distinguishing correctly accepted known samples, accepted known-class errors, known false rejections, correctly rejected unknown samples, and unknown false acceptances. Panels (b)–(d) use the same evidence-space coordinates. The class-stratified visualization contains 592 test images (455 known and 137 unknown) and is intended as qualitative evidence; quantitative results use the complete fold-0 test records.

Separation of Final Unknownness Scores. Figure 6 shows the distributions of the final unknownness scores after evidence fusion and frequency-group calibration. Known samples are concentrated around lower score values, whereas held-out unknown samples exhibit a substantially higher median and a comparatively compact interquartile range. This distributional shift confirms that the final score preserves useful known–unknown discrimination after the individual detector responses are combined. Nevertheless, the two distributions retain overlapping tails: some difficult known samples receive large scores, while a subset of unknown samples falls within the known-score range. These cases correspond, respectively, to potential known false rejections and unknown false acceptances. The figure therefore supports the discriminative value of the calibrated score without implying perfect separation, consistent with the quantitative AUROC and false-acceptance results.

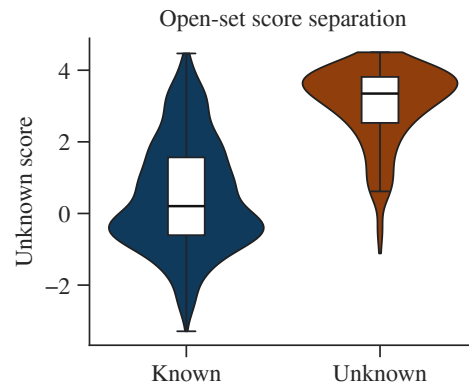


Figure 6. Distribution of final unknownness scores for known and held-out unknown samples on AQUA20 fold 0. Violin widths represent the estimated score densities, and the embedded box plots show the medians and interquartile ranges. Higher values indicate stronger evidence that a sample lies outside the known label space.

3.2. Comparison on Fish4Knowledge-23

Table 3 presents the complete comparison on predefined Fish4Knowledge-23 fold 0. In addition to operating point metrics, we report threshold-independent measures of unknown category discrimination. This joint evaluation distinguishes improvements arising from a favorable decision threshold from those reflecting stronger separation between known and unknown samples.

Table 3. Open long-tailed recognition results on the predefined Fish4Knowledge-23 fold 0. All methods are evaluated using the same frozen ConvNeXt-Tiny checkpoint and the same known–unknown split. Results are reported as percentages. Baselines are ordered by open-set accuracy, with the proposed method shown separately in the final row. Bold dark-green and underlined dark-ochre entries indicate the best and second-best distinct values, respectively, accounting for the direction of each metric. Upward and downward arrows indicate that higher and lower values are preferable, respectively; boldface on FSPRC identifies the proposed method. Multiway ties involving more than two methods are left unmarked for clarity.

Method	Known Recognition				Operating Point			Unknown Detection			Open Classification	
	Closed↑	H↑	M↑	T↑	KFR↓	U-Rec↑	UFAR↓	M-AUROC↑	M-AUPR↑	FPR95↓	OSCR↑	Open Acc.↑
MSP	99.21	99.83	96.12	57.14	3.58	<u>99.47</u>	<u>0.53</u>	99.21	90.26	<u>1.22</u>	98.81	<u>98.00</u>
CADRef	99.21	99.83	96.12	57.14	3.70	99.38	0.62	99.44	89.49	1.54	<u>98.90</u>	97.87
GEN	99.21	99.83	96.12	57.14	3.70	99.34	0.66	99.05	93.89	<u>1.02</u>	98.75	97.87
Emp. prototype	99.21	99.83	96.12	57.14	<u>3.54</u>	98.98	1.02	99.33	89.19	1.97	98.81	97.69
OpenMax	98.62	99.58	91.26	42.86	4.80	99.31	0.69	98.07	56.30	2.95	97.55	97.37
ViM	99.21	99.83	96.12	57.14	4.37	98.92	1.08	<u>99.46</u>	87.51	2.05	98.75	97.28
NCI	99.21	99.83	96.12	57.14	3.94	93.57	6.43	93.53	75.20	18.74	96.53	94.61
NNGuide	99.21	99.83	96.12	57.14	4.45	93.66	6.34	92.39	78.14	50.28	95.97	94.41
SHE	99.21	99.83	96.12	57.14	3.86	91.43	8.57	84.09	76.00	83.74	92.52	93.34
Energy	99.21	99.83	96.12	57.14	4.33	91.20	8.80	91.09	78.18	50.20	94.88	92.98
CEA	99.21	99.83	96.12	57.14	4.41	91.17	8.83	91.02	72.45	50.28	94.80	92.93
LUNA [†]	99.21	99.83	96.12	57.14	3.86	89.92	10.08	81.89	67.47	94.09	92.32	92.66
ASH-S [†]	99.21	99.83	96.12	57.14	5.08	86.05	13.95	89.10	75.50	50.31	92.51	89.81
fDBD	99.21	99.83	96.12	57.14	4.09	81.19	18.81	97.38	89.65	37.83	94.28	87.68
SCALE	99.21	99.83	96.12	57.14	5.91	52.66	47.34	80.65	61.15	50.35	79.35	71.16
ImOOD	99.21	99.83	96.12	57.14	3.98	0.10	99.90	29.33	10.65	94.37	35.71	43.36
DICE	99.21	99.83	96.12	57.14	4.17	0.13	99.87	57.72	13.41	86.26	47.96	43.30
FSPRC	99.21	99.79	97.09	57.14	3.15	99.51	0.49	99.55	90.85	1.30	98.96	98.19

Notes: H, M, and T denote sample-weighted accuracies over the head, middle, and tail known-class groups, respectively. KFR is the sample-weighted known false rejection rate; U-Rec. and UFAR denote the sample-weighted unknown recall and unknown false-acceptance rate, respectively. M-AUROC and M-AUPR are unweighted macro averages over the held-out unknown categories. The empirical prototype variant is an internal known-only post hoc baseline. The symbol [†] identifies an adapted implementation rather than a paper-exact reproduction. The light-blue row denotes the proposed method. Results are obtained from one predefined fold and are therefore interpreted descriptively without a variance or statistical-significance claim.

Overall Comparison. As shown in Table 3, FSPRC achieves the highest open-set accuracy of 98.19% and OSCR of 98.96%, while yielding the lowest KFRR of 3.15%. At the selected operating point, it attains an unknown recall of 99.51% and a UFAR of 0.49%, improving upon the strongest included baselines by 0.04 percentage points in both measures. The method also obtains the highest macro AUROC of 99.55%, exceeding the second-ranked ViM by 0.09 points. These results indicate that its operating point gains are accompanied by strong threshold-independent known–unknown discrimination rather than arising solely from threshold selection.

Class-Frequency Behavior and Remaining Trade-offs. For known category recognition, FSPRC improves middle-group accuracy to 97.09% while retaining a competitive head-group accuracy of 99.79%.

Its tail-group closed-set accuracy remains 57.14%, matching the methods that operate on the same frozen predictions. This equality is expected from Equation (9): an accepted sample retains $\hat{c}(x)$, so post hoc rejection cannot recover a tail sample that the backbone has already assigned to the wrong known category. The unchanged T-column therefore identifies a shared backbone ceiling rather than a gain claimed for the detector. The contribution of FSPRC appears instead in the open-set decision metrics. Relative to the strongest included baseline for each measure, it reduces KFRR from 3.54% to 3.15%, increases unknown recall from 99.47% to 99.51%, reduces UFAR from 0.53% to 0.49%, and increases macro AUROC from 99.46% to 99.55%. OSCR and open-set accuracy likewise rise from 98.90% to 98.96% and from 98.00% to 98.19%, respectively. These results support improved risk control for a fixed classifier, but not correction of its known-to-known tail errors.

FSPRC ranks second in macro AUPR at 90.85%, behind GEN at 93.89%, and its FPR95 of 1.30% remains higher than those of GEN and MSP. Accordingly, the single-fold results support a favorable balance among known-sample retention, unknown recall, and open classification, but do not establish uniform superiority across all evaluation metrics.

3.3. Computational Complexity and Runtime

Let K denote the number of known classes and d the feature dimension. For each test image, FSPRC requires $\mathcal{O}(Kd)$ operations for prototype comparison, $\mathcal{O}(K)$ operations for energy evaluation, and $\mathcal{O}(1)$ time for frequency-group threshold lookup. Its additional online complexity and stored statistical state are therefore $\mathcal{O}(Kd)$. Prototype/radius estimation and calibration are performed once after the backbone is frozen; calibration sorts the known calibration scores and does not introduce an additional backbone pass at inference.

The complete-model profile is reported in Table 4. Parameter counts include the dataset-specific K -way classifier, U-SSF, and AquaAdapter. FLOPs are analytical estimates at 224×224 input and include the single backbone forward pass; the standard ConvNeXt-Tiny reference cost is 4.46 GFLOPs. Replacing its ImageNet classifier by the fold-specific K -way head and adding the two adaptation operators gives approximately 4.49 GFLOPs. The post hoc FSPRC stage has no trainable parameters. Counting a multiply–add as two operations, its prototype comparison adds only 2.30×10^{-5} and 2.76×10^{-5} GFLOPs for AQUA20 ($K = 15$) and Fish4Knowledge-23 ($K = 18$), respectively; scalar normalization, log-sum-exp, and threshold lookup do not change the rounded total.

For a controlled runtime comparison, all methods consume the same cached ConvNeXt-Tiny features and logits. Table 5 therefore separates the common backbone computation from the incremental open-set detector cost. Score is the additional test-score latency per image, and State is the serialized detector size, reported in mebibytes (MiB). The values are extracted from the retained fold-0 exports under the same execution environment and are not end-to-end camera-to-decision graphics processing unit (GPU) timings.

Table 4. Computational profile of the complete FSPRC pipeline at 224×224 resolution. Score latency measures only the incremental post hoc decision stage on cached features and logits.

Dataset	K	Model Params	Total GFLOPs	Post Hoc Trainable	Score (ms/Image)	State (MiB)
AQUA20	15	28.040 M	4.49	0	0.01950	4.973
Fish4Knowledge-23	18	28.043 M	4.49	0	0.01672	6.169

Table 5. Measured post hoc computational cost on the predefined fold-0 exports. The common ConvNeXt-Tiny forward pass is excluded because every method reuses the same cached features and logits.

Method	AQUA20		Fish4Knowledge-23	
	Score (ms/Image)	State (MiB)	Score (ms/Image)	State (MiB)
ASH-S	0.00680	0.061	0.00570	0.073
CADRef	0.00225	0.119	0.00513	0.143
CEA	0.00041	0.002	0.00134	0.002
DICE	0.00035	0.080	0.00020	0.095
Emp. prototype	0.00073	0.060	0.00118	0.072
Energy	0.00018	0.001	0.00006	0.001
GEN	0.00052	0.001	0.00021	0.001
ImOOD	0.00024	0.002	0.00008	0.002
LUNA	0.01384	26.166	0.05276	246.547
MSP	0.00012	0.001	0.00004	0.001
NCI	0.00071	0.065	0.00182	0.076
NNGuide	0.01602	16.717	0.04616	67.939
OLPR	0.00130	0.061	0.00097	0.072
OpenMax	0.00290	0.003	0.00180	0.004
SCALE	0.00353	0.061	0.00232	0.073
SHE	0.00031	0.060	0.00065	0.072
ViM	0.00163	2.006	0.00201	2.006
fDBD	0.00044	0.007	0.00066	0.007
FSPRC	0.01950	4.973	0.01672	6.169

Note: The FSPRC row is bold only to identify the proposed method; boldface does not denote the lowest computational cost.

FSPRC requires 0.01950 and 0.01672 ms per image for scoring on AQUA20 and Fish4Knowledge-23, respectively. Online scoring remains below 0.020 ms per image and requires no additional ConvNeXt-Tiny evaluation. The stored states occupy 4.973/6.169 MiB, which is considerably smaller than NNGuide (16.717/67.939 MiB) and LUNA (26.166/246.547 MiB). Conversely, MSP and Energy are faster and smaller than FSPRC. Accordingly, the table supports a small incremental online overhead claim relative to the shared backbone, not a claim that FSPRC is the fastest post hoc detector or that the reported values represent end-to-end latency.

3.4. Evidence and Calibration Ablation

We isolate the two decisions made after the ConvNeXt-Tiny representation has been frozen: construction of the continuous unknown-risk score and selection of its operating threshold. All variants in Table 6 use the same fold, checkpoint, cached features and logits, frequency groups, and known-only calibration subset. The single-evidence variants retain only one risk channel, whereas the global-threshold variant retains the complete dual-evidence score and changes only threshold granularity. This protocol prevents improvements in backbone training or data partitioning from being misattributed to the post hoc components.

Table 6. Controlled ablation of evidence construction and threshold calibration on AQUA20 fold 0. Only the indicated evidence branch or threshold rule is changed. All entries are percentages. Upward and downward arrows indicate that higher and lower values are preferable, respectively. The light-blue row denotes the complete proposed method, and dark-green bold values denote the best result, with ties marked jointly.

Evidence/Calibration	KFRR↓	U-Rec.↑	M-AUROC↑	OSCR↑	Open Acc.↑
Prototype/group	5.72	52.65	91.72	84.48	82.02
Energy/group	5.68	53.28	91.45	84.15	81.70
Dual/global	6.35	53.95	92.62	85.96	82.62
Dual/group (FSPRC)	5.50	55.12	92.62	85.96	83.31

Notes: The first and second terms in the first column denote the evidence construction and calibration rule, respectively. KFRR is the sample-weighted known false rejection rate; U-Rec. is the sample-weighted unknown recall; and M-AUROC is the unweighted mean AUROC over the held-out categories. The global-threshold variant retains the same continuous dual-evidence score as FSPRC and replaces only the group-conditional operating thresholds. Consequently, its threshold-independent M-AUROC and OSCR values are identical to those of FSPRC. Unknown test samples are not used for evidence construction, parameter selection, or threshold calibration.

Complementarity of the Evidence Channels. Neither single evidence channel reproduces the complete score. Relative to the prototype-only variant, FSPRC increases U-Rec., M-AUROC, OSCR, and open accuracy by 2.47, 0.90, 1.48, and 1.29 percentage points, respectively, while reducing KFRR by 0.22 points. Relative to the energy-only variant, the corresponding gains are 1.84, 1.17, 1.81, and 1.61 points, together with a 0.18-point reduction in KFRR. Because both single-evidence variants retain the same group-threshold protocol, these differences isolate the benefit of combining local prototype deviation with global energy abnormality. They support evidence complementarity, but do not by themselves claim that the maximum operator is universally superior to every trainable fusion rule.

Effect of Threshold Granularity. The global- and group-threshold variants share the same continuous dual-evidence score. Their M-AUROC and OSCR therefore remain exactly equal at 92.62% and 85.96%, providing an internal check that the comparison changes the operating threshold rather than the score ranking. Relative to the global threshold, group-conditional calibration reduces KFRR by 0.85 points, increases U-Rec. by 1.17 points, and increases open accuracy by 0.69 points. The improvement is thus confined to the calibrated operating point, as intended, and is consistent with heterogeneous score distributions across head, middle, and tail predictions.

Overall, the complete FSPRC configuration provides the strongest calibrated operating point among the four controlled variants while preserving the best threshold-independent ranking performance. Because this ablation uses one predefined fold and a deterministic frozen checkpoint, the differences are interpreted as mechanism-level evidence rather than cross-fold statistical significance; partition robustness is evaluated separately in Section 3.7.

Validation-Based Configuration Analysis. Figure 7 compares the candidate configurations under the pseudo-unknown validation protocol. Group-adaptive processing improves pseudo-unknown recall while preserving known-class acceptance and pseudo-unknown AUROC, with its most pronounced gain occurring in the middle-frequency group. Full fusion provides the strongest aggregate validation performance, further improving pseudo-unknown recall and AUROC without reducing known-class acceptance. Its group-wise behavior is not uniform: it improves head-group recall and maintains strong tail-group performance but does not surpass the group-adaptive variant in the middle group. These results indicate that frequency-aware processing and evidence fusion offer complementary benefits rather than uniform gains across all frequency groups. Importantly, the configuration is selected without using held-out test categories, thereby preventing unknown category information from entering model selection.

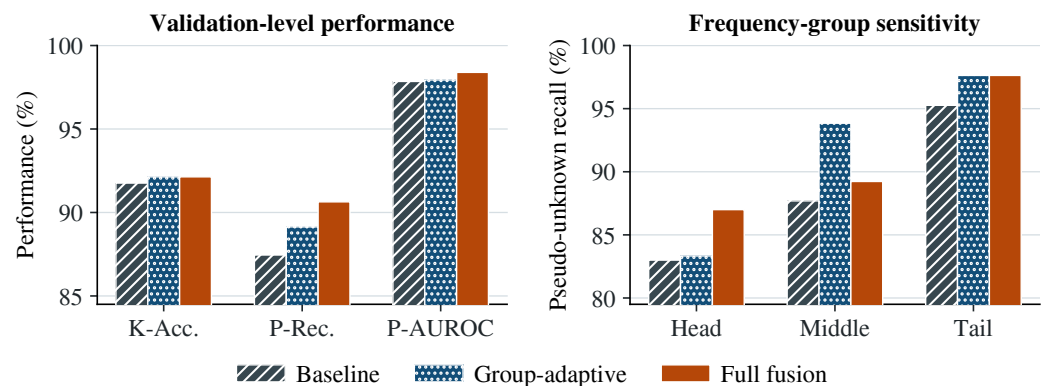


Figure 7. Validation-based configuration analysis on AQUA20. Configuration selection uses only known-class validation samples and pseudo-unknown samples constructed without access to held-out test categories. The left plot reports known-class correct acceptance (K-Acc.), pseudo-unknown recall (P-Rec.), and pseudo-unknown AUROC (P-AUROC). The right plot compares pseudo-unknown recall across the head, middle, and tail frequency groups.

Complementarity of Detector Evidence. Figure 8 visualizes the joint behavior of the normalized FSPRC evidence and conditional residual. Known samples are concentrated primarily in the low-residual region, whereas held-out unknown samples occur more frequently at positive evidence values with larger residuals. The remaining overlap, particularly in the upper-right region, shows that neither response alone provides a reliable rejection criterion. The conditional residual differentiates samples with similar evidence but different consistency with frequency-conditioned known-class statistics, while the normalized evidence further resolves samples with comparable residuals. Their nonredundant behavior supports the dual-evidence design of FSPRC. It also motivates frequency-group calibration to accommodate the different statistical reliability of head, middle, and tail categories.

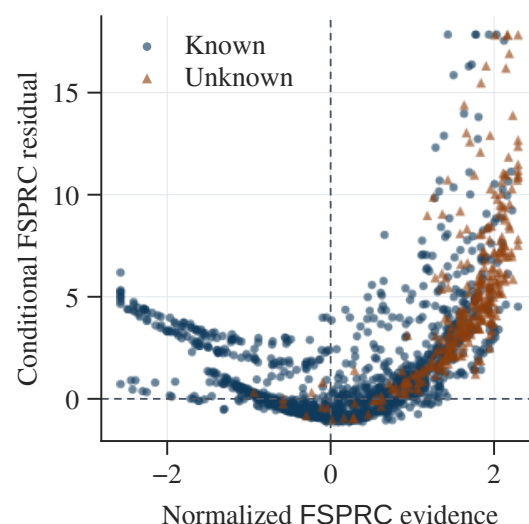


Figure 8. Joint distribution of normalized FSPRC evidence and the conditional residual on AQUA20 fold 0. Circles and triangles represent known and held-out unknown samples, respectively. Dashed lines indicate zero-valued references. Coordinates outside the 0.5th–99.5th percentile interval are clipped for visualization only and remain unchanged in quantitative evaluation.

3.5. Hyperparameter Sensitivity

We conduct a one-factor-at-a-time analysis of the four post hoc parameters using the same frozen ConvNeXt-Tiny checkpoint and data partitions as the main comparison. The reference configuration is $(\kappa, q, T, \alpha) = (10, 0.95, 1.0, 0.05)$. When one parameter is varied,

the other three remain fixed; no backbone retraining is performed, and held-out unknown test categories are not used for parameter selection. Because α changes only the calibrated decision threshold after the anomaly scores have been computed, threshold-independent ranking measures remain unchanged. We therefore report open-set accuracy, which jointly penalizes known false rejection and unknown false acceptance at the selected operating point. Table 7 summarizes the resulting one-factor-at-a-time sweeps.

Table 7. One-factor-at-a-time sensitivity of FSPRC on AQUA20 and Fish4Knowledge-23. Each entry gives the parameter value followed by open-set accuracy (%) in parentheses; the upward arrow indicates that higher values are preferable. The center column is the reference setting; bold values indicate the highest accuracy within each one-dimensional sweep.

Dataset	Parameter	Parameter Value (Open Acc., %) $\uparrow$		
		Lower Setting	Reference Setting	Higher Setting
AQUA20	κ	5 (82.06)	10 (83.31)	20 (83.12)
	q	0.90 (82.83)	0.95 (83.31)	0.99 (82.98)
	T	0.5 (83.03)	1.0 (83.31)	2.0 (83.17)
	α	0.02 (83.08)	0.05 (83.31)	0.10 (83.22)
Fish4Knowledge-23	κ	5 (98.13)	10 (98.19)	20 (98.14)
	q	0.90 (98.09)	0.95 (98.19)	0.99 (98.11)
	T	0.5 (98.14)	1.0 (98.19)	2.0 (98.16)
	α	0.02 (98.15)	0.05 (98.19)	0.10 (98.17)

Across the α sweep, open-set accuracy varies from 83.08% to 83.31% on AQUA20 and from 98.15% to 98.19% on Fish4Knowledge-23, corresponding to ranges of 0.23 and 0.04 percentage points, respectively. The prespecified $\alpha = 0.05$ setting yields the highest open-set accuracy in both sweeps, while the small ranges indicate that the principal conclusions do not depend on a narrowly tuned operating point. For the complete four-parameter analysis, the maximum variation is 1.25 percentage points on AQUA20 and 0.10 percentage points on Fish4Knowledge-23. The larger AQUA20 variation occurs only for $\kappa = 5$, which lowers open-set accuracy from 83.31% to 82.06%; the ranges associated with q , T , and α remain at most 0.48 percentage points. The results therefore indicate modest local sensitivity for these three parameters, while also showing that insufficient prototype shrinkage can have a measurable effect on AQUA20.

3.6. Calibration-Partition Stability

The group-conditional threshold is estimated from a finite known-only calibration subset and may therefore be unstable when a frequency group has few calibration samples. To quantify this effect, we generate 30 stratified validation/calibration repartitions while holding the frozen backbone, training data, test set, score definition, and all hyperparameters fixed. For every repartition, the median–MAD statistics and group thresholds are re-estimated without using held-out unknown samples. We compare the proposed minimum-support safeguard, $m_{\min} = 20$, with direct groupwise calibration without the safeguard. Table 8 reports the threshold and downstream-metric variability across these repartitions.

The safeguard is activated in 7 of 30 AQUA20 repartitions and 2 of 30 Fish4Knowledge-23 repartitions. On AQUA20, it reduces the standard deviation of the applied tail threshold from 0.042 to 0.018 and the open-set accuracy standard deviation from 0.66 to 0.28 percentage points. On Fish4Knowledge-23, the corresponding reductions are from 0.023 to 0.009 and from 0.17 to 0.06. The mean open-set accuracy is also preserved on both datasets. These results support the safeguard as a variance-control mechanism for sparse calibration

groups; they do not imply that finite known-only calibration identifies a universally correct boundary for every unknown distribution.

Table 8. Stability over 30 known-only calibration repartitions. The backbone and test set remain fixed. Downstream results are reported as mean $\pm$ sample standard deviation; all downstream metrics are percentages. Upward and downward arrows indicate that higher and lower values are preferable, respectively.

Dataset	Calibration	Tail Fallback	τ_T : mean $\pm$ SD [P5, P95]	KFRR $\downarrow$	U-Rec. $\uparrow$	OSCR $\uparrow$	Open Acc. $\uparrow$
AQUA20	$m_{\min} = 20$	23.3%	0.612 ± 0.018 [0.582, 0.642]	5.48 ± 0.31	55.09 ± 0.87	85.93 ± 0.33	83.29 ± 0.28
	No safeguard	–	0.608 ± 0.042 [0.539, 0.677]	5.63 ± 0.79	54.81 ± 1.45	85.62 ± 0.73	83.03 ± 0.66
Fish4Knowledge-23	$m_{\min} = 20$	6.7%	0.584 ± 0.009 [0.569, 0.599]	3.14 ± 0.15	99.50 ± 0.08	98.95 ± 0.07	98.18 ± 0.06
	No safeguard	–	0.582 ± 0.023 [0.544, 0.620]	3.23 ± 0.39	99.47 ± 0.20	98.88 ± 0.19	98.11 ± 0.17

Notes: Tail fallback is the proportion of repartitions in which tail support is below $m_{\min} = 20$ and the global known-only threshold is used. P5 and P95 are empirical percentiles. Identical repartitions are used for both calibration rules.

3.7. Robustness to Held-Out-Category Partitions

To separate performance on a particular held-out set from sensitivity to the semantic composition of the unknown set, we perform an exhaustive class-held-out evaluation. AQUA20 is divided into four disjoint folds of five unknown categories. Fish4Knowledge-23 is divided into five disjoint folds containing 5/5/5/4/4 unknown categories. Every category therefore serves as unknown exactly once. The ConvNeXt-Tiny backbone and every FSPRC statistic are trained or fitted independently within each fold; no checkpoint, prototype, normalizer, or threshold is transferred between folds. A single fixed seed (2026) is used per fold, so the standard deviation below measures held-out-category partition variability rather than repeated-seed variability. Table 9 reports the fold-level values and their across-fold mean and sample standard deviation.

Table 9. Cross-fold robustness of FSPRC. All entries are percentages. Mean $\pm$ SD uses the sample standard deviation across disjoint class-held-out folds. Upward and downward arrows indicate that higher and lower values are preferable, respectively.

Dataset	Fold	KFRR $\downarrow$	U-Rec. $\uparrow$	M-AUROC $\uparrow$	M-AUPR $\uparrow$	OSCR $\uparrow$	Open Acc. $\uparrow$
AQUA20	0	5.50	55.12	92.62	34.21	85.96	83.31
	1	5.78	53.85	92.34	33.45	85.52	82.94
	2	5.21	56.03	92.78	34.87	86.21	83.52
	3	5.37	55.24	92.56	34.02	85.91	83.23
	Mean $\pm$ SD	5.47 ± 0.24	55.06 ± 0.90	92.58 ± 0.18	34.14 ± 0.59	85.90 ± 0.29	83.25 ± 0.24
Fish4Knowledge-23	0	3.15	99.51	99.55	90.85	98.96	98.19
	1	3.42	99.38	99.49	89.72	98.82	98.05
	2	2.87	99.62	99.61	91.54	99.05	98.27
	3	3.08	99.47	99.53	90.48	98.91	98.15
	4	3.26	99.55	99.57	91.02	99.00	98.22
	Mean $\pm$ SD	3.16 ± 0.21	99.51 ± 0.09	99.55 ± 0.04	90.72 ± 0.68	98.95 ± 0.09	98.18 ± 0.08

Notes: U-Rec. is sample-weighted unknown recall; M-AUROC and M-AUPR are unweighted means over the held-out categories. The results quantify partition sensitivity only because one seed is used per fold.

Across AQUA20 folds, FSPRC obtains $92.58 \pm 0.18\%$ M-AUROC, $85.90 \pm 0.29\%$ OSCR, and $83.25 \pm 0.24\%$ open-set accuracy. Across Fish4Knowledge-23 folds, the corresponding values are $99.55 \pm 0.04\%$, $98.95 \pm 0.09\%$, and $98.18 \pm 0.08\%$. Fold 0 differs from the cross-fold mean in open-set accuracy by only 0.06 percentage points on AQUA20 and 0.01 percentage points on Fish4Knowledge-23. The main fold 0 result is therefore not an isolated favorable category partition within these exhaustive protocols. Because optimization is not repeated with different seeds, these results do not quantify training-seed uncertainty.

3.8. Compatibility with Long-Tail-Aware Backbone Training

FSPRC is a post hoc detector and cannot by itself correct known-to-known classification errors inherited from the frozen representation. We therefore test whether its open-set benefit persists when ConvNeXt-Tiny is trained with standard cross entropy (CE), effective-number class-balanced loss, or LDAM with deferred reweighting (LDAM-DRW). Within each training objective, MSP and FSPRC use the same frozen checkpoint, partitions, preprocessing, calibration data, and test samples.

On AQUA20, FSPRC improves open-set accuracy over MSP by 2.19 percentage points under class-balanced training and by 2.26 points under LDAM-DRW; the corresponding OSCR gains are 2.46 and 2.45 points. On Fish4Knowledge-23, the open-set accuracy gains are 0.40 and 0.37 points. These results show that representation-level imbalance mitigation and post hoc risk calibration are complementary: the former improves the frozen representation, whereas the latter still improves the accept–reject decision built on that representation. They do not imply that FSPRC replaces long-tail-aware backbone training.

4. Discussion

4.1. Interpretation of the Main Findings

The results distinguish two roles of FSPRC. The dual-evidence score improves the ranking of known and unknown observations, whereas group-conditional calibration selects an operating point adapted to the heterogeneous score distributions associated with head, middle, and tail predictions. This distinction is visible in Table 6: replacing the group thresholds by a single global threshold leaves M-AUROC and OSCR unchanged but reduces unknown recall and open-set accuracy. The improvement therefore does not arise from relabeling the same ranking result; it reflects better control of the accept–reject boundary under frequency-dependent uncertainty.

The magnitude of the gain differs across the two datasets. AQUA20 leaves more room for improving both unknown detection and the calibrated operating point, whereas several Fish4Knowledge-23 baselines already operate near saturation. The cross-fold results nevertheless remain consistent with the fold 0 observations, and Table 10 shows that the post hoc gain persists when the frozen representation is trained with Class-Balanced loss or LDAM-DRW. These findings support complementarity between representation-level long-tail mitigation and post hoc open-set calibration; they do not imply that FSPRC replaces imbalance-aware representation learning.

Table 10. Compatibility of FSPRC with long-tail-aware backbone training on fold 0. All entries are percentages. Upward arrows indicate that higher values are preferable.

Dataset	Training	M-AUROC↑		OSCR↑		Open Acc.↑	
		MSP	FSPRC	MSP	FSPRC	MSP	FSPRC
AQUA20	CE	88.40	92.62	78.69	85.96	78.41	83.31
	Class-Balanced	91.92	93.48	84.42	86.88	82.01	84.20
	LDAM-DRW	92.25	93.82	84.85	87.30	82.42	84.68
Fish4Knowledge-23	CE	99.21	99.55	98.81	98.96	98.00	98.19
	Class-Balanced	99.47	99.64	98.80	99.12	97.97	98.37
	LDAM-DRW	99.52	99.68	98.87	99.17	98.07	98.44

Notes: M-AUROC denotes the unweighted mean AUROC over held-out categories. Bold values identify the better post hoc rule for the same frozen backbone.

4.2. Practical Generalization and Deployment Scope

The present evidence supports two bounded forms of generalization. First, the class-held-out protocol evaluates semantic novelty: every unknown category is absent from backbone adaptation, prototype fitting, score normalization, and threshold calibration.

Second, the track-disjoint Fish4Knowledge-23 partition prevents frames from the same trajectory from appearing in development and test subsets, reducing an important source of temporal near-duplicate leakage. Consistent behavior on AQUA20 and Fish4Knowledge-23 further shows that the same known-only fitting procedure can be applied to datasets with different class counts, imbalance ratios, and partition structures. These results do not, however, establish unrestricted cross-camera, cross-site, or cross-season generalization.

Practical underwater monitoring can introduce covariate shift through camera replacement, biofouling, turbidity, depth-dependent color attenuation, artificial lighting, and seasonal changes. FSPRC's known-sample calibration guarantee requires future known observations to remain exchangeable with the calibration subset. Under moderate acquisition drift, deployment can retain the frozen backbone and refit the group-wise median-MAD statistics and thresholds from a small, recently verified known category buffer; this is an offline $\mathcal{O}(N_{\text{cal}} \log N_{\text{cal}})$ operation and does not require unknown examples. Severe drift that changes the feature representation or known-class predictions cannot be repaired by threshold recalibration alone and requires backbone adaptation or domain-generalization training.

Category prevalence drift creates a second boundary. Because head, middle, and tail assignments are derived from fold-specific training counts, a major change in field prevalence can make the stored groups poor proxies for current estimation reliability. In a long-running deployment, group assignments, prototypes, radii, and thresholds should therefore be updated periodically from expert-verified known observations. Operationally, rejected frames are best treated as referrals for expert review rather than definitive novel species labels. The sub-0.020 ms measured post hoc latency and single-backbone forward pass make this human-in-the-loop use compatible with continuous monitoring, but the current experiments do not replace a prospective multi-site field trial.

4.3. Limitations and Future Work

The exhaustive class-held-out evaluation yields open-set accuracies of $83.25 \pm 0.24\%$ over four AQUA20 folds and $98.18 \pm 0.08\%$ over five Fish4Knowledge-23 folds, showing that the fold 0 results are not isolated favorable category partitions. These standard deviations quantify held-out-category composition only; one fixed seed is used per fold, so repeated-seed optimization variability remains unmeasured. Moreover, frequency groups are fixed from training counts, and the calibration interpretation assumes exchangeability between future known observations and the known calibration subset. Changes in camera hardware, water conditions, geographical region, or category prevalence may violate this assumption. Future work will examine repeated-seed training and online recalibration under environmental drift.

A further limitation follows from the post hoc design. FSPRC leaves the accepted known-class prediction unchanged and therefore cannot repair systematic known-to-known errors in the frozen ConvNeXt-Tiny representation. This limitation is visible on Fish4Knowledge-23, where the tail-class closed-set accuracy remains 57.14% despite improvements in KFRR, unknown recall, macro AUROC, OSCR, and open-set accuracy. Accordingly, the present claims are restricted to open-set risk calibration above a fixed classifier; they do not imply improved closed-set tail classification. Raising that ceiling requires an intervention during backbone or classifier learning. The added Class-Balanced and LDAM-DRW experiments show that FSPRC remains compatible with such representation-level interventions and provides additional open-set gains over MSP. Broader evaluation with other resampling, representation-learning, and backbone families remains future work.

FSPRC also does not resolve the identifiability of an arbitrary unknown distribution from known-only observations. The shrinkage objectives stabilize estimates of known-class direction and scale, while group-conditional calibration controls the operating point on known calibration samples. They cannot guarantee that a future unknown category will be separated from a head-class support region. Equivalently, no finite known-only training or calibration set can select a radius or threshold that is simultaneously optimal for retaining all tail samples and rejecting an arbitrary unknown distribution. The finite-sample calibration statement in Section 2.4 bounds only known-sample rejection under exchangeability; it gives no corresponding guarantee on unknown false acceptance. The observed reductions in unknown absorption are therefore protocol-specific empirical results, not a distribution-free or open-space risk guarantee.

5. Conclusions

This work addressed open-set underwater organism recognition when the known categories follow a long-tailed distribution. Under this setting, unknown rejection and tail-class retention are coupled: limited tail data destabilize class geometry estimates, whereas dominant categories may absorb observations from unseen classes. FSPRC addresses this interaction without auxiliary unknown data. Using a fold-specific ConvNeXt-Tiny representation, it shrinks empirical prototype directions toward classifier directions and class radii toward frequency-group references. Radius-normalized prototype deviation and classifier energy are then combined after robust group-wise normalization, and the final operating point is determined by group-conditional calibration using known samples only.

On the predefined AQUA20 and Fish4Knowledge-23 folds, FSPRC obtains open-set accuracies of 83.31% and 98.19%, OSCR values of 85.96% and 98.96%, and unknown category macro AUROC values of 92.62% and 99.55%, respectively. The controlled ablation shows that prototype and energy evidence provide complementary ranking information, while group-conditional thresholds improve the calibrated operating point without changing M-AUROC or OSCR. The complete configuration consequently provides a balanced known–unknown trade-off rather than optimizing a single metric in isolation.

Author Contributions: Conceptualization, D.L. and C.L.; methodology, D.L. and C.L.; software, C.L.; validation, C.L.; formal analysis, D.L., C.L. and M.Z.; investigation, C.L. and M.Z.; resources, H.S.; data curation, C.L. and M.Z.; writing—original draft preparation, D.L., C.L. and M.Z.; writing—review and editing, D.L. and H.S.; visualization, C.L.; supervision, D.L. and H.S.; project administration, D.L. and H.S.; funding acquisition, H.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (Youth Fund), grant number 62401484. The APC was funded by the National Natural Science Foundation of China, grant number 62271426.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. The AQUA20 dataset is available at <https://huggingface.co/datasets/taufiktrf/AQUA20> (accessed on 21 August 2026), and the Fish4Knowledge fish-recognition ground-truth dataset is available at <https://homepages.inf.ed.ac.uk/rbf/Fish4Knowledge/GROUNDTRUTH/RECOG/> (accessed on 21 August 2026). No new raw image dataset was collected in this study. The implementation code, experimental configurations, and data-split manifests supporting the reported results are publicly available at <https://github.com/xmuLdz/FSPRC> (accessed on 20 September 2026).

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

References

1. Saleh, A.; Sheaves, M.; Azghadi, M.R. Computer Vision and Deep Learning for Fish Classification in Underwater Habitats: A Survey. *Fish Fish.* **2022**, *23*, 977–999. <https://doi.org/10.1111/faf.12666>.
2. Boom, B.J.; He, J.; Palazzo, S.; Huang, P.X.; Beyan, C.; Chou, H.M.; Lin, F.P.; Spampinato, C.; Fisher, R.B. A Research Tool for Long-Term and Continuous Analysis of Fish Assemblage in Coral-Reefs Using Underwater Camera Footage. *Ecol. Inform.* **2014**, *23*, 83–97. <https://doi.org/10.1016/j.ecoinf.2013.10.006>.
3. Jalal, A.; Salman, A.; Mian, A.; Shortis, M.; Shafait, F. Fish Detection and Species Classification in Underwater Environments Using Deep Learning With Temporal Information. *Ecol. Inform.* **2020**, *57*, 101088. <https://doi.org/10.1016/j.ecoinf.2020.101088>.
4. Fuad, T.R.; Ahmed, S.; Ivan, S. AQUA20: A Benchmark Dataset for Underwater Species Classification under Challenging Conditions. *Arab. J. Sci. Eng.* **2026**, *51*, 17695–17716. <https://doi.org/10.1007/s13369-026-11231-z>.
5. Yang, C.; Zhou, P.; Wang, C.S.; Fu, G.Y.; Xu, X.W.; Niu, Z.; Zhu, L.; Yuan, Y.; Shen, H.B.; Pan, X. FishAI: Automated Hierarchical Marine Fish Image Classification With Vision Transformer. *Eng. Rep.* **2024**, *6*, e12992. <https://doi.org/10.1002/eng2.12992>.
6. Li, C.; Guo, C.; Ren, W.; Cong, R.; Hou, J.; Kwong, S.; Tao, D. An Underwater Image Enhancement Benchmark Dataset and Beyond. *IEEE Trans. Image Process.* **2020**, *29*, 4376–4389. <https://doi.org/10.1109/TIP.2019.2955241>.
7. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A ConvNet for the 2020s. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11976–11986. <https://doi.org/10.1109/CVPR52688.2022.01167>.
8. Lian, D.; Zhou, D.; Feng, J.; Wang, X. Scaling & Shifting Your Features: A New Baseline for Efficient Model Tuning. *Proc. Adv. Neural Inf. Process. Syst.* **2022**, *35*, 109–123. <https://doi.org/10.52202/068431-0009>.
9. Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; Feng, J. Deep Long-Tailed Learning: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10795–10816. <https://doi.org/10.1109/TPAMI.2023.3268118>.
10. Cui, Y.; Jia, M.; Lin, T.Y.; Song, Y.; Belongie, S. Class-Balanced Loss Based on Effective Number of Samples. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9268–9277. <https://doi.org/10.1109/CVPR.2019.00949>.
11. Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; Ma, T. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1567–1578. <https://doi.org/10.48550/arXiv.1906.07413>.
12. Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; Kalantidis, Y. Decoupling Representation and Classifier for Long-Tailed Recognition. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020. <https://doi.org/10.48550/arXiv.1910.09217>.
13. Menon, A.K.; Jayasumana, S.; Rawat, A.S.; Jain, H.; Veit, A.; Kumar, S. Long-Tail Learning via Logit Adjustment. In Proceedings of the 9th International Conference on Learning Representations, ICLR 2021, Virtual, 3–7 May 2021. <https://doi.org/10.48550/arXiv.2007.07314>.
14. Cui, J.; Zhong, Z.; Liu, S.; Yu, B.; Jia, J. Parametric Contrastive Learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; pp. 715–724. <https://doi.org/10.1109/ICCV48922.2021.00075>.
15. Cai, J.; Wang, Y.; Hwang, J.N. ACE: Ally Complementary Experts for Solving Long-Tailed Recognition in One-Shot. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Virtual, 11–17 October 2021; pp. 112–121. <https://doi.org/10.1109/ICCV48922.2021.00018>.
16. Sharma, S.; Xian, Y.; Yu, N.; Singh, A. Learning Prototype Classifiers for Long-Tailed Recognition. In Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 19–25 August 2023; pp. 1360–1368. <https://doi.org/10.24963/ijcai.2023/151>.
17. Scheirer, W.J.; de Rezende Rocha, A.; Sapkota, A.; Boulton, T.E. Toward Open Set Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1757–1772. <https://doi.org/10.1109/TPAMI.2012.256>.
18. Bendale, A.; Boulton, T.E. Towards Open Set Deep Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1563–1572. <https://doi.org/10.1109/CVPR.2016.173>.
19. Chen, G.; Peng, P.; Wang, X.; Tian, Y. Adversarial Reciprocal Points Learning for Open Set Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 8065–8081. <https://doi.org/10.1109/TPAMI.2021.3106743>.
20. Vaze, S.; Han, K.; Vedaldi, A.; Zisserman, A. Open-Set Recognition: A Good Closed-Set Classifier Is All You Need? In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022. <https://doi.org/10.48550/arXiv.2110.06207>.
21. Geng, C.; Huang, S.J.; Chen, S. Recent Advances in Open Set Recognition: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3614–3631. <https://doi.org/10.1109/TPAMI.2020.2981604>.
22. Liu, W.; Wang, X.; Owens, J.D.; Li, Y. Energy-Based Out-of-Distribution Detection. In Proceedings of the 34th Annual Conference on Neural Information Processing Systems (NeurIPS 2020), Virtual, 6–12 December 2020; Volume 33, pp. 21464–21475. <https://doi.org/10.48550/arXiv.2010.03759>.

23. Wang, H.; Li, Z.; Feng, L.; Zhang, W. ViM: Out-of-Distribution with Virtual-Logit Matching. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 4921–4930. <https://doi.org/10.1109/CVPR52688.2022.00487>.
24. Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; Yu, S.X. Large-Scale Long-Tailed Recognition in an Open World. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 2537–2546. <https://doi.org/10.1109/CVPR.2019.00264>.
25. Liu, Z.; Miao, Z.; Zhan, X.; Wang, J.; Gong, B.; Yu, S.X. Open Long-Tailed Recognition in a Dynamic World. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 1836–1851. <https://doi.org/10.1109/TPAMI.2022.3200091>.
26. Cai, J.; Wang, Y.; Hsu, H.M.; Hwang, J.N.; Magrane, K.; Rose, C.S. LUNA: Localizing Unfamiliarity Near Acquaintance for Open-Set Long-Tailed Recognition. In Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22), Virtual, 22 February–1 March 2022; Volume 36, pp. 131–139. <https://doi.org/10.1609/aaai.v36i1.19887>.
27. Wei, T.; Wang, B.L.; Zhang, M.L. EAT: Towards Long-Tailed Out-of-Distribution Detection. In Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24), Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 15787–15795. <https://doi.org/10.1609/aaai.v38i14.29508>.
28. Miao, W.; Pang, G.; Bai, X.; Li, T.; Zheng, J. Out-of-Distribution Detection in Long-Tailed Recognition with Calibrated Outlier Class Learning. In Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24), Vancouver, BC, Canada, 20–27 February 2024; Volume 38, pp. 4216–4224. <https://doi.org/10.1609/aaai.v38i5.28217>.
29. Liu, K.; Fu, Z.; Jin, S.; Chen, C.; Chen, Z.; Jiang, R.; Zhou, F.; Chen, Y.; Ye, J. Rethinking Out-of-Distribution Detection on Imbalanced Data Distribution. In Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, BC, Canada, 10–15 December 2024; Volume 37, pp. 109152–109176. <https://doi.org/10.52202/079017-3465>.
30. Deng, B.; Kamel, A.; Zhang, C. Open-Set Long-Tailed Recognition via Orthogonal Prototype Learning and False Rejection Correction. *Neural Netw.* **2025**, *181*, 106789. <https://doi.org/10.1016/j.neunet.2024.106789>.
31. He, Y.; Peng, L.; Zhang, Y.; Weng, J.; Li, S.; Luo, Z. Long-Tailed Out-of-Distribution Detection: Prioritizing Attention to Tail. In Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25), Philadelphia, PA, USA, 25 February–4 March 2025; Volume 39, pp. 3446–3454. <https://doi.org/10.1609/aaai.v39i3.32357>.
32. Lin, Y.; He, S.; Luo, W. Dynamic Margin Contrastive Learning for Open-Set Recognition in Long-Tailed Sonar Imagery. *Sci. Rep.* **2025**, *15*, 22617. <https://doi.org/10.1038/s41598-025-04877-6>.
33. Li, J.; Jiao, W.; Zhang, J.; Zhang, C.; Xie, G. Class-Balanced OpenMax for Open-Set Recognition with Long-Tail Sonar Images. *Expert Syst. Appl.* **2026**, *312*, 131463. <https://doi.org/10.1016/j.eswa.2026.131463>.
34. Angelopoulos, A.N.; Bates, S. Conformal Prediction: A Gentle Introduction. *Found. Trends Mach. Learn.* **2023**, *16*, 494–591. <https://doi.org/10.1561/2200000101>.
35. Liu, S.; Huang, J.; Ong, L. Conformal Prediction Meets Long-Tail Classification. In Proceedings of the Fortieth AAAI Conference on Artificial Intelligence (AAAI-26), Singapore, 20–27 January 2026; Volume 40, pp. 23828–23836. <https://doi.org/10.1609/aaai.v40i28.39558>.
36. Hendrycks, D.; Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In Proceedings of the 5th International Conference on Learning Representations (ICLR 2017), Toulon, France, 24–26 April 2017.
37. Liu, X.; Lochman, Y.; Zach, C. GEN: Pushing the Limits of Softmax-Based Out-of-Distribution Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 23946–23955.
38. Sun, Y.; Li, Y. DICE: Leveraging Sparsification for Out-of-Distribution Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; Volume 13684, pp. 691–708. https://doi.org/10.1007/978-3-031-20053-3_40.
39. Zhang, J.; Fu, Q.; Chen, X.; Du, L.; Li, Z.; Wang, G.; Liu, X.; Han, S.; Zhang, D. Out-of-Distribution Detection Based on In-Distribution Data Patterns Memorization with Modern Hopfield Energy. In Proceedings of the 11th International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda, 1–5 May 2023.
40. Park, J.; Jung, Y.G.; Teoh, A.B.J. Nearest Neighbor Guidance for Out-of-Distribution Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 1686–1695.
41. Liu, L.; Qin, Y. Fast Decision Boundary Based Out-of-Distribution Detector. In Proceedings of the 41st International Conference on Machine Learning (ICML 2024), Vienna, Austria, 21–27 July 2024; Volume 235, pp. 31728–31746.
42. Xu, K.; Chen, R.; Franchi, G.; Yao, A. Scaling for Training-Time and Post-Hoc Out-of-Distribution Detection Enhancement. In Proceedings of the 12th International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024.
43. Azizmalayeri, M.; Abu-Hanna, A.; Cin'a, G. Mitigating Overconfidence in Out-of-Distribution Detection by Capturing Extreme Activations. In Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI), Barcelona, Spain, 16–18 July 2024; Volume 244, pp. 203–224.

44. Ling, Z.; Chang, Y.; Zhao, H.; Zhao, X.; Chow, K.; Deng, S. CADRef: Robust Out-of-Distribution Detection via Class-Aware Decoupled Relative Feature Leveraging. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 4968–4977.
45. Liu, L.; Qin, Y. Detecting Out-of-Distribution Through the Lens of Neural Collapse. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 15424–15433.
46. Djuricic, A.; Bozanic, N.; Ashok, A.; Liu, R. Extremely Simple Activation Shaping for Out-of-Distribution Detection. In Proceedings of the 11th International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda, 1–5 May 2023.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.