I. APPENDIX

A. Appendix A: The Proof of Theorem 1

Theorem 1. *Let Z_{ce} be the feature representation learned by minimizing only the cross-entropy loss, and Z_{sieve} be the representation learned by Sieve with a self-supervised contrastive loss weight $\lambda > 0$. Under the Information Bottleneck framework, as the compression coefficient $\beta \to \infty$, Z_{ce} converges to a minimal sufficient statistic of Y, resulting in the complete loss of detection-relevant information, i.e., $I(Z_{ce}; T|Y) = 0$. Conversely, the optimization objective of Sieve guarantees a strictly positive lower bound for detection information, i.e., $I(Z_{sieve}; T|Y) > 0$.*

Proof. Our goal is to mathematically demonstrate that when optimizing solely for classification performance and model compression, non-semantic details in the features (captured by $X|Y$) are forced to zero.

The training process of deep neural networks can be viewed as optimizing the Information Bottleneck objective: maximizing predictive power while minimizing feature complexity. Its Lagrangian form is given by:

$$\min_{\theta} \mathcal{L}_{IB} = -I(Z;Y) + \beta I(Z;X), \tag{1}$$

where β is the Lagrange multiplier. As $\beta \to \infty$ (pursuing extreme feature compression while maintaining classification accuracy), the model tends to learn the **Minimal Sufficient Statistic** of Y [1], denoted as Z_{ce}.

By definition, Z_{ce} must satisfy two conditions simultaneously:

1) **Sufficiency:** The feature contains all information about the label.
$$I(Z_{ce};Y) = I(X;Y). \tag{2}$$
2) **Minimality:** The feature contains the minimum amount of input information.
$$Z_{ce} = \arg\min_{Z} I(Z;X) \quad \text{s.t. } I(Z;Y) = I(X;Y). \tag{3}$$

Using the Chain Rule of mutual information, we decompose the total information $I(Z;X)$ contained in feature Z into two parts [1]:

$$I(Z;X) = \underbrace{I(Z;Y)}_{\text{Semantic Information}} + \underbrace{I(Z;X|Y)}_{\text{Residual Structural Information}}, \tag{4}$$

This decomposition holds based on the Markov chain $Y \to X \to Z$, which implies that given X, Z is conditionally independent of Y, i.e., $I(Z;Y|X) = 0$.

Substituting this decomposition into the minimality objective:

$$\begin{aligned} \min_{Z} I(Z;X) &= \min_{Z} \left[I(Z;Y) + I(Z;X|Y) \right] \\ &= I(X;Y) + \min_{Z} I(Z;X|Y), \end{aligned} \tag{5}$$

where $I(Z;Y)$ is fixed at $I(X;Y)$.

Since mutual information is non-negative, the only solution to minimize the above equation is:

$$\lim_{\beta \to \infty} I(Z_{ce};X|Y) = 0. \tag{6}$$

This proves that in the limit, the features Z_{ce} learned by a pure classification model contain no information about the input X other than the semantics of label Y.

The detection task aims to distinguish between known and unknown samples. The label T is an attribute of the input X but is not fully explained by the semantic label Y (e.g., unknown samples might share similar semantic features with known samples but differ in background distribution). Feature extraction follows the Markov chain: $T \to X \to Z$. Conditioned on the semantic label Y, this Markov chain still holds:

$$T|Y \to X|Y \to Z|Y. \tag{7}$$

According to the conditional form of the Data Processing Inequality (DPI), the information about T contained in feature Z cannot exceed the information contained in input X:

$$I(Z;T|Y) \le I(Z;X|Y). \tag{8}$$

Combining the conclusion $I(Z_{ce};X|Y) = 0$, with the non-negativity of mutual information:

$$0 \le I(Z_{ce};T|Y) \le I(Z_{ce};X|Y) = 0. \tag{9}$$

By the Squeeze Theorem, we necessarily obtain:

$$I(Z_{ce};T|Y) = 0. \tag{10}$$

Theoretically, the pure classification model Z_{ce} loses all non-semantic cues required to distinguish unknown attacks, leading to detection failure.

The total loss function of Sieve is $\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{self}$.

- $\mathcal{L}_{ce}$: Minimizing cross-entropy $\iff$ Maximizing $I(Z;Y)$.
- $\mathcal{L}_{self}$: Self-supervised contrastive loss. As proven by [2], minimizing the contrastive loss is equivalent to maximizing the lower bound of the mutual information $I(Z;X)$:
$$\mathcal{L}_{self} \ge \text{const} - I(Z;X), \tag{11}$$

Thus, Sieve's optimization objective can be formalized as finding Z_{sieve}:

$$Z_{sieve} = \arg\max_{Z} \left[I(Z;Y) + \lambda I(Z;X) \right], \quad \lambda > 0. \tag{12}$$

Substituting the chain rule $I(Z;X) = I(Z;Y) + I(Z;X|Y)$ again, we expand Sieve's objective function:

$$\begin{aligned} \mathcal{J}(Z) &= I(Z;Y) + \lambda \left[I(Z;Y) + I(Z;X|Y) \right] \\ &= (1+\lambda) I(Z;Y) + \lambda I(Z;X|Y) \end{aligned} \tag{13}$$

We analyze the behavior of this objective function upon convergence:

1) First Term $(1+\lambda)I(Z;Y)$: The mutual information $I(Z;Y)$ is bounded by the theoretical upper limit of the label entropy $H(Y)$ (i.e., $I(Z;Y) \le H(Y)$). When classification accuracy saturates, the gradient of this term approaches 0.
2) Second Term $\lambda I(Z;X|Y)$: This is the unique term introduced by contrastive learning. As long as $\lambda > 0$, the optimizer will continuously attempt to maximize $I(Z;X|Y)$ to further reduce the total loss.

Since the first term has a ceiling while the second term does not (or its upper bound is far higher than the current value), for the converged Sieve model, there must exist a constant $\epsilon > 0$ such that:

$$I(Z_{sieve}; X|Y) \geq \epsilon. \tag{14}$$

This is referred to as an ϵ-redundant representation in information theory.

Since the upper bound of $I(Z_{sieve}; X|Y)$ is opened, the feasible region for detection information $I(Z_{sieve}; T|Y)$ expands from 0 to the positive real domain:

$$0 < I(Z_{sieve}; T|Y) \leq \epsilon. \tag{15}$$

This implies that Sieve forcibly "backs up" the structural information of input X in the features via contrastive learning. Since cues for detecting unknown attacks (T) are often hidden in this structural information deemed "redundant" by the classification task, Sieve effectively prevents the over-compression of detection-relevant information, thereby providing a solid theoretical guarantee for effective unknown traffic detection. □

B. Appendix B: The Proof of Theorem 2

Theorem 2. *Let $\mathcal{Y}_s$ be the proxy target for self-supervised learning. If $I(\mathcal{Y}_s; Y) = 0$ (the label blindness condition, where the proxy task is statistically independent of the semantic label), then the unsupervised feature Z_{ssl} satisfies $I(Z_{ssl}; Y) = 0$. Sieve breaks this blindness condition by introducing the neighbor-consistency corrected label Y_c, ensuring $I(Z_{sieve}; Y) \gg 0$, thus enabling the detection of Adjacent unknown samples.*

Proof. Based on the logic of Corollary 3.3 (Strict Label Blindness in Filtered Distributions) in the paper [1]. Assume the information contained in input X can be decomposed into two statistically independent components:

$$X = (X_1, X_2), \quad \text{where } X_1 \perp X_2, \tag{16}$$

- X_1 (Proxy Feature): Determines the self-supervised proxy target $\mathcal{Y}_s$. That is, $\mathcal{Y}_s = f_1(X_1)$.
- X_2 (Semantic Feature): Determines the true semantic label Y. That is, $Y = f_2(X_2)$.

Since $X_1 \perp X_2$, and $\mathcal{Y}_s, Y$ are deterministic functions of them respectively, hence:

$$I(\mathcal{Y}_s; Y) = 0. \tag{17}$$

This is the **Label Blindness Condition**.

The goal of self-supervised learning (e.g., contrastive learning) is to maximize $I(Z; \mathcal{Y}_s)$ while minimizing $I(Z; X)$ (information bottleneck). Its optimal solution Z_{ssl} is the minimal sufficient statistic with respect to $\mathcal{Y}_s$. This implies:

$$I(Z_{ssl}; X|\mathcal{Y}_s) = 0. \tag{18}$$

That is, Z_{ssl} discards all information in X irrelevant to $\mathcal{Y}_s$.

We calculate $I(Z_{ssl}; Y)$. Since Y is determined entirely by X_2, and $X_2 \perp X_1$ (which implies $X_2 \perp \mathcal{Y}_s$), the information contained in Y falls entirely within the "residual information irrelevant to $\mathcal{Y}_s$".

Expanding $I(Z_{ssl}; X)$ using the chain rule of mutual information:

$$\begin{aligned} I(Z_{ssl}; X) &= I(Z_{ssl}; X_1, X_2) \\ &= I(Z_{ssl}; X_1) + I(Z_{ssl}; X_2|X_1). \end{aligned} \tag{19}$$

Since Z_{ssl} is the minimal sufficient statistic for $\mathcal{Y}_s$ (determined by X_1), it encodes only X_1 and not the independent component X_2. Mathematically, from $I(Z_{ssl}; X|\mathcal{Y}_s) = 0$ and the fact that $\mathcal{Y}_s$ is a function of X_1, it follows that Z_{ssl} is independent of X_2:

$$I(Z_{ssl}; X_2) = 0. \tag{20}$$

According to the Data Processing Inequality ($Y \leftarrow X_2 \leftarrow X \rightarrow Z_{ssl}$):

$$I(Z_{ssl}; Y) \leq I(Z_{ssl}; X_2) = 0, \tag{21}$$

hence

$$I(Z_{ssl}; Y) = 0. \tag{22}$$

Therefore, under the label blindness condition, pure self-supervised features Z_{ssl} contain no semantic label information.

Sieve defines a filtering function: keep only samples with K-nearest neighbor voting consistency $c_i > \xi$.

$$Y_c = \{y_i \mid c_i \geq \xi\}, \tag{23}$$

where y_i is the original (possibly noisy) label.

According to the Manifold Smoothness Assumption [3]: samples sufficiently close in feature space have a probability approaching 1 of sharing the same true label Y.

$$P\left(Y(x) = Y(x') \mid \|f(x) - f(x')\| < \epsilon\right) \to 1. \tag{24}$$

When $c_i \to 1$ (i.e., all K neighbors have label y_i), it implies the sample lies in a high-density class center region rather than a noise-prone decision boundary. At this point, the posterior probability that y_i equals the true label Y is maximized:

$$P(Y = Y_c) = 1 - P_e, \tag{25}$$

where P_e is a minimal error rate.

We view Y as being transmitted from Y_c through a noisy channel. According to Fano's Inequality, the upper bound on the conditional entropy (uncertainty) of Y given observation Y_c is:

$$H(Y|Y_c) \leq H_b(P_e) + P_e \log(|\mathcal{Y}| - 1), \tag{26}$$

- $H_b(P_e)$ is the binary entropy function: $-P_e \log P_e - (1 - P_e) \log(1 - P_e)$.
- $|\mathcal{Y}|$ is the total number of classes.

When the filtering threshold ξ is high (set to 1.0 in Sieve), the error rate $P_e \to 0$. At this point, $H_b(P_e) \to 0$ and $P_e \log(|\mathcal{Y}| - 1) \to 0$. Therefore:

$$H(Y|Y_c) \approx 0. \tag{27}$$

$$I(Y_c; Y) = H(Y) - H(Y|Y_c) \approx H(Y). \tag{28}$$

This proves that the corrected labels Y_c retain almost all information of the true labels Y (i.e., mutual information is maximized).

Sieve's loss function includes $\mathcal{L}_{ce}(Z, Y_c)$. Minimizing the cross-entropy loss $\mathcal{L}_{ce}(Z, Y_c) = -\mathbb{E}[\log P(Y_c|Z)]$ is equivalent to maximizing the lower bound of the mutual information $I(Z; Y_c)$. Assuming model convergence and sufficient capacity:

$$I(Z_{sieve}; Y_c) \to H(Y_c) \tag{29}$$

(i.e., the model has learned to predict the corrected labels).

We derive the lower bound for $I(Z; Y)$. Consider two expansions of the joint mutual information $I(Z; Y, Y_c)$:

$$I(Z; Y, Y_c) = I(Z; Y_c) + I(Z; Y|Y_c), \tag{30}$$

$$I(Z; Y, Y_c) = I(Z; Y) + I(Z; Y_c|Y), \tag{31}$$

Combining the two equations:

$$I(Z; Y) = I(Z; Y_c) + I(Z; Y|Y_c) - I(Z; Y_c|Y). \tag{32}$$

- Term 1: $I(Z; Y_c)$. This term is maximized by the optimization objective and is large.
- Term 2: $I(Z; Y|Y_c)$. By non-negativity of mutual information, this term ≥ 0.
- Term 3: $I(Z; Y_c|Y)$. This is the information Z contains about Y_c given the true label Y. Since Y_c is a (denoised) observation of Y, and Z is generated from X, this term is bounded by the noise entropy of Y_c itself:

$$I(Z; Y_c|Y) \leq H(Y_c|Y), \tag{33}$$

$H(Y_c|Y)$ represents the "uncertainty of the corrected label given the true label". Since the correction process is highly rigorous (filtering out most noise), Y_c is almost a deterministic function of Y (extremely low noise), so $H(Y_c|Y)$ is very small.

Substituting these relations:

$$\begin{aligned} I(Z_{sieve}; Y) &\geq I(Z_{sieve}; Y_c) - H(Y_c|Y) \\ &\approx H(Y_c) - 0 \\ &\approx H(Y). \end{aligned} \tag{34}$$

Therefore, $I(Z_{sieve}; Y)$ is significantly greater than 0 and approaches the label entropy $H(Y)$.

In the unknown traffic detection task:

- Known samples x_{in} and unknown samples x_{out} have overlapping distributions on X_1, i.e., $P(X_1|in) \approx P(X_1|out)$.
- The distinction lies only in X_2, i.e., $Y(x_{in}) \in \mathcal{C}_{in}$ while $Y(x_{out}) \notin \mathcal{C}_{in}$.

Comparison Analysis:

- **Self-Supervised Learning:** Given the theoretical bound $I(Z_{\text{ssl}}; Y) = 0$, the model encodes solely X_1 while discarding the semantic variations in X_2. Consequently, the representations become indistinguishable, i.e., $Z(x_{\text{in}}) \approx Z(x_{\text{out}})$, leading to detection failure.
- **Sieve (Ours):** In contrast, as derived in our proposition, $I(Z_{\text{sieve}}; Y) \gg 0$, ensuring the effective encoding of X_2. This forces $Z(x_{\text{in}})$ to form compact clusters around ID semantic centers, while $Z(x_{\text{out}})$ is pushed away from these clusters.

Therefore, Sieve successfully breaks the label blindness condition, making unknown traffic samples separable in the feature space. □

C. Appendix C: The Proof of Theorem 3

Theorem 3. *Let Z follow a class-conditional distribution with covariance matrix Σ_k. The self-supervised optimization objective of Sieve ($\mathcal{L}_{self}$) is mathematically equivalent to minimizing the trace of the intra-class covariance matrix Σ_k, i.e., $Tr(\Sigma_k) \to 0$. According to the Arithmetic Mean-Geometric Mean (AM-GM) inequality, this implies that the determinant $\det(\Sigma_k) \to 0$, indicating that the feature distribution becomes extremely compact geometrically. This compactness causes the eigenvalue amplification factor of the Mahalanobis distance metric to tend towards infinity, thereby strictly minimizing the Open Set Risk.*

Proof. To quantify the safety of the detector in unknown environments, we cite the definition from [4].

Definition: The Open Set Risk $R_{\mathcal{O}}$ is defined as the effective measure (Lebesgue measure) of the recognition function f over the open space $\mathcal{O}$ (the space outside the support of the training data):

$$R_{\mathcal{O}}(f) = \frac{\int_{\mathcal{O}} f(z)dz}{\int_{S_o} f(z)dz} \tag{35}$$

- $f(z)$ is the recognition function: if z is classified as a known class (ID), then $f(z) = 1$, otherwise 0.
- S_o represents the entire feature space.

Minimizing $R_{\mathcal{O}}$ necessitates minimizing the volume of the acceptance region of $f(z) = 1$ within the feature space. That is, the more compact the distribution of known classes, the lower the probability of erroneously infringing upon the open space $\mathcal{O}$.

Assume that the features Z extracted by Sieve follow a multivariate Gaussian distribution for class k: $Z|Y = k \sim \mathcal{N}(\mu_k, \Sigma_k)$. According to information theory, the differential entropy of a multivariate Gaussian is given by:

$$H(Z|Y = k) = \frac{1}{2} \ln \left((2\pi e)^d \det(\Sigma_k) \right), \tag{36}$$

where $\det(\Sigma_k)$ is the determinant of the covariance matrix, which is geometrically proportional to the square of the volume of the distribution ellipsoid:

$$\text{Vol(ID Distribution)} \propto \sqrt{\det(\Sigma_k)}. \tag{37}$$

Minimizing the intra-class conditional entropy $H(Z|Y)$ is mathematically strictly equivalent to minimizing the determinant of the covariance matrix $\det(\Sigma_k)$, which in turn is equivalent to minimizing the volume of the ID distribution.

Sieve's loss function includes the self-supervised contrastive loss $\mathcal{L}_{self}$. We need to prove that this loss function possesses the mathematical property of compressing covariance.

Citing the conclusion from [5], the $\mathcal{L}_{self}$ loss asymptotically optimizes two properties: Alignment and Uniformity.

For positive sample pairs in Sieve (samples (z_i, z_j) belonging to the same class k), the Alignment loss term is defined as:

$$\mathcal{L}_{align} \triangleq \mathbb{E}_{(z_i,z_j)\sim P_k}\|z_i - z_j\|^2. \tag{38}$$

We expand the relationship between the expected Euclidean distance between samples and the trace of the covariance matrix:

$$\begin{aligned}\mathbb{E}\|z_i - z_j\|^2 &= \mathbb{E}\|(z_i - \mu_k) - (z_j - \mu_k)\|^2 \\ &= \mathbb{E}\|z_i - \mu_k\|^2 + \mathbb{E}\|z_j - \mu_k\|^2 \\ &\quad - 2\mathbb{E}[(z_i - \mu_k)^T(z_j - \mu_k)] \\ &= \mathrm{Tr}(\Sigma_k) + \mathrm{Tr}(\Sigma_k) - 0 \\ &\quad \text{(assumption of independent sampling)} \\ &= 2\mathrm{Tr}(\Sigma_k),\end{aligned} \tag{39}$$

where $\mathrm{Tr}(\Sigma_k) = \sum_{m=1}^{d} \lambda_m$ is the sum of the eigenvalues of the covariance matrix (total variance).

By minimizing $\mathcal{L}_{self}$, Sieve directly minimizes $\mathcal{L}_{align}$, thereby forcing the trace of the covariance matrix $\mathrm{Tr}(\Sigma_k) \to 0$.

We need to derive the determinant from the trace. For a d-dimensional positive semi-definite matrix Σ_k, according to the Arithmetic Mean-Geometric Mean (AM-GM) inequality:

$$\frac{1}{d}\mathrm{Tr}(\Sigma_k) = \frac{\sum \lambda_i}{d} \geq \left(\prod \lambda_i\right)^{1/d} = (\det(\Sigma_k))^{1/d}. \tag{40}$$

Rearranging gives:

$$\det(\Sigma_k) \leq \left(\frac{\mathrm{Tr}(\Sigma_k)}{d}\right)^d. \tag{41}$$

Since we have proven that $\mathrm{Tr}(\Sigma_k) \to 0$, by the Squeeze Theorem:

$$\lim_{\text{train}\to\infty} \det(\Sigma_k) = 0. \tag{42}$$

The optimization objective of Sieve causes the feature distribution to become extremely compact geometrically, with its volume tending towards zero.

Sieve employs the Mahalanobis distance as the detection score:

$$D_M(z) = \sqrt{(z - \mu_k)^T \Sigma_k^{-1} (z - \mu_k)}. \tag{43}$$

Perform eigendecomposition on Σ_k: $\Sigma_k = Q\Lambda Q^T$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$. Since $\mathrm{Tr}(\Sigma_k) \to 0$, all eigenvalues $\lambda_i \to 0^+$. The eigenvalues of Σ_k^{-1} are $1/\lambda_i$, and their limit behavior is:

$$\lim_{\lambda_i \to 0} \frac{1}{\lambda_i} = \infty. \tag{44}$$

For any unknown sample z_{out}, assume there is a non-zero deviation from the class center (i.e., $z_{out} \neq \mu_k$). We project the deviation vector $v = z_{out} - \mu_k$ onto the eigenvector basis; there exists at least one component $v_j \neq 0$. The squared Mahalanobis distance is:

$$D_M^2(z_{out}) = \sum_{i=1}^{d} \frac{v_i^2}{\lambda_i} \geq \frac{v_j^2}{\lambda_j}. \tag{45}$$

As $\lambda_j \to 0$:

$$\lim_{\lambda_j \to 0} \frac{v_j^2}{\lambda_j} = \infty. \tag{46}$$

As Sieve optimization proceeds, Σ_k^{-1} constructs an extremely steep potential well. For any non-ID sample, its distance metric is drastically amplified by the reciprocal of the eigenvalues, making it extremely easy to filter out by a threshold.

Returning to the risk definition from [4]. Sieve's detection strategy is based on an acceptance region defined by threshold τ:

$$A_\tau = \{z \mid D_M(z) < \tau\}. \tag{47}$$

This region is an ellipsoid, and its volume $V(A_\tau)$ is proportional to the square root of the determinant of the covariance:

$$V(A_\tau) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} \tau^d \sqrt{\det(\Sigma_k)}. \tag{48}$$

Combining with $\det(\Sigma_k) \to 0$, the volume of the acceptance region tends to zero:

$$\lim_{\text{train}\to\infty} V(A_\tau) = 0. \tag{49}$$

Therefore, in the infinite open space $\mathcal{O}$, the integral term $\int_{\mathcal{O}} f(z)dz$ (the overlap between the acceptance region and the open space) is strictly constrained to 0.

Sieve minimizes the trace of the intra-class covariance matrix through contrastive learning, leading to the collapse of the feature distribution volume. This not only maximizes the separability of known and unknown traffic in the Mahalanobis distance space (signal-to-noise ratio $\to \infty$), but also strictly minimizes the Open Set Risk mathematically. □

REFERENCES

[1] H. Yang, Q. Yu, and T. Desell, "Can we ignore labels in out of distribution detection?" *arXiv preprint arXiv:2504.14704*, 2025.

[2] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[3] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples." *Journal of machine learning research*, vol. 7, no. 11, 2006.

[4] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 7, pp. 1757–1772, 2012.

[5] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *International conference on machine learning*. PMLR, 2020, pp. 9929–9939.