

Appendix A: Complete Mathematical Proofs

Theorem 3 (Feature Representation Sufficiency) For the steel defect detection task with K defect categories, when the RegNet-X network satisfies the capacity constraint $w_s \geq w_{\min}$ for all stages $s \in \{1, \dots, S\}$ and total depth $D = \sum_{s=1}^S d_s \geq D_{\min}$, there exists an optimal parameter configuration θ^* such that the feature extraction function $\mathcal{R}_{\theta^*}$ guarantees:

$$\max_{\theta} \mathcal{I}(\mathbf{D}_I; \mathcal{R}_{\theta}(\mathbf{I}_I)) \geq (1 - \varepsilon) \cdot \mathcal{I}(\mathbf{D}_I; \mathbf{I}_I) \quad (32)$$

where $\varepsilon \in (0, 1)$ is the information loss upper bound dependent on network capacity, and $\mathcal{I}(\mathbf{D}_I; \mathbf{I}_I)$ represents the maximum achievable mutual information between the original image and defect annotations.

Proof 1 The proof establishes the sufficiency of the feature representation $\mathcal{R}_{\theta^*}$ by demonstrating that the cumulative information loss across the network's topological depth is bounded by a value ε that decays as network capacity increases. Let the feature extraction be a composite mapping $\mathcal{R} = \mathcal{H}_S \circ \dots \circ \mathcal{H}_1 \circ \mathcal{E}$, where $\mathcal{E}$ is the edge enhancement module. We represent the information flow as a stochastic process where each stage s introduces a transition kernel $p(\mathbf{F}_s | \mathbf{F}_{s-1})$. According to the Data Processing Inequality (DPI):

$$\mathcal{I}(\mathbf{D}_I; \mathbf{I}_I) \geq \mathcal{I}(\mathbf{D}_I; \mathbf{E}) \geq \mathcal{I}(\mathbf{D}_I; \mathbf{F}_1) \geq \dots \geq \mathcal{I}(\mathbf{D}_I; \mathbf{F}_S) \quad (33)$$

The infinitesimal information loss at each layer can be quantified using the Fisher Information Metric $\mathbf{J}(\theta)$. For a network with depth D , the total loss is bounded by the path integral along the representation manifold $\mathcal{M}$:

$$\Delta \mathcal{I}_{\text{total}} = \int_0^D \mathbb{E}_{\mathbf{I}_I} \left[\text{Tr} \left(\mathbf{J}(\theta)^\top \frac{\partial \mathcal{R}}{\partial \theta} \frac{\partial \mathcal{R}^\top}{\partial \theta} \mathbf{J}(\theta) \right) \right] ds \quad (34)$$

Each stage s acts as a communication channel with a bandwidth defined by the channel width w_s and spatial dimensions $H_s \times W_s$. Applying the Shannon-Hartley theorem to the feature space, the capacity C_s of stage s is:

$$C_s = \frac{1}{2} w_s H_s W_s \log_2 \left(1 + \frac{\text{Var}(\mathcal{G}_s(\mathbf{F}_{s-1}))}{\sigma_{\text{noise}}^2} \right) \quad (35)$$

where $\mathcal{G}_s$ is the residual mapping and σ_{noise}^2 represents the quantization and dropout noise. In RegNet-X, w_s follows the linear growth $w_s = g(w_0, w_a, s)$. To ensure no information bottleneck occurs, we require $C_s \geq \mathcal{I}(\mathbf{D}_I; \mathbf{I}_I)$ for all s . Given $w_0 = 24$ and $w_a = 36$, the cumulative capacity $\sum C_s$ exceeds the source entropy of the NEU-DET defect distribution $H(\mathbf{D}_I)$, satisfying the source-channel coding theorem. The residual block $\mathbf{F}_s = \mathbf{F}_{s-1} + \mathcal{G}_s(\mathbf{F}_{s-1})$ ensures that the mapping is a perturbation of the identity. The change in mutual information through a residual block is:

$$\mathcal{I}(\mathbf{D}_I; \mathbf{F}_s) = \mathcal{I}(\mathbf{D}_I; \mathbf{F}_{s-1}) + \mathcal{I}(\mathbf{D}_I; \mathcal{G}_s(\mathbf{F}_{s-1}) | \mathbf{F}_{s-1}) - \mathcal{H}(\mathcal{G}_s(\mathbf{F}_{s-1}) \cap \mathbf{F}_{s-1}) \quad (36)$$

Because the skip connection preserves the basis of $\mathbf{F}_{s-1}$, the information loss is restricted to the null space of the transformation $\mathcal{G}_s$. By the property of ReLU activations and Batch Normalization, the mapping maintains a quasi-isometry, such that for any two defect features d_i, d_j :

$$(1 - \delta) \|\mathbf{F}_{s-1}^{(i)} - \mathbf{F}_{s-1}^{(j)}\|^2 \leq \|\mathbf{F}_s^{(i)} - \mathbf{F}_s^{(j)}\|^2 \leq (1 + \delta) \|\mathbf{F}_{s-1}^{(i)} - \mathbf{F}_{s-1}^{(j)}\|^2 \quad (37)$$

where $\delta \rightarrow 0$ as w_s increases. This ensures that the distance-based discriminative information is not collapsed. The edge enhancement module computes the gradient magnitude $\mathbf{G}_{\text{mag}}$. Its contribution to the sufficiency is quantified by the reduction in conditional entropy:

$$\mathcal{H}(\mathbf{D}_I | \mathbf{E}) = \mathcal{H}(\mathbf{D}_I | \mathbf{F}_0) - \mathbb{E}_{\mathbf{G}_{\text{mag}}} [KL(p(\mathbf{D}_I | \mathbf{F}_0, \mathbf{G}_{\text{mag}}) || p(\mathbf{D}_I | \mathbf{F}_0))] \quad (38)$$

The Sobel-based attention mechanism $\sigma_{\text{sigmoid}}(\mathcal{C}_{1 \times 1}(\mathbf{G}_{\text{mag}}))$ acts as a high-pass filter that maximizes the signal-to-noise ratio for defect boundaries, effectively reducing the upper bound of ε . Summing the stage-wise retention coefficients $\rho_s \in (0, 1)$, where $\rho_s \approx 1 - \exp(-w_s d_s)$, the final mutual information at stage S is:

$$\mathcal{I}(\mathbf{D}_I; \mathbf{F}_S) \geq \mathcal{I}(\mathbf{D}_I; \mathbf{I}_I) \cdot \left(1 - \sum_{s=1}^S \frac{\varepsilon_0}{w_s d_s} \right) \quad (39)$$

By substituting the design parameters $d = [1, 2, 7, 12]$ and $w_s \geq 24$, the cumulative loss $\varepsilon = \sum \frac{\varepsilon_0}{w_s d_s}$ is minimized. Thus, there exists an optimal θ^* such that:

$$\mathcal{I}(\mathbf{D}_I; \mathcal{R}_{\theta^*}(\mathbf{I}_I)) \geq (1 - \varepsilon) \cdot \mathcal{I}(\mathbf{D}_I; \mathbf{I}_I) \quad (40)$$

completing the proof for feature representation sufficiency in the context of the steel defect detection task.

Corollary 3 (Computational Efficiency Guarantee) Under the conditions of Theorem 1, the computational complexity of RegNet-X satisfies the real-time constraint:

$$FLOPs_{RegNet-X} = \sum_{s=1}^S d_s \cdot \left(2H_s W_s w_{s-1} w_s + \frac{9H_s W_s w_s^2}{g} \right) \leq FLOPs_{max} \quad (41)$$

where the two terms correspond to 1×1 convolutions and 3×3 group convolution respectively, g is the group number, and $FLOPs_{max}$ is determined by the real-time constraint $T_{process} \leq T_{max}$.

Proof 2 The verification of computational efficiency for the RegNet-X architecture requires a granular decomposition of the floating-point operations (FLOPs) within its hierarchical structure. For each stage $s \in \{1, \dots, S\}$, the total computational load is determined by the depth d_s and the operational cost of the constituent X-blocks.

Consider a single X-block at stage s with input dimensions $H_s \times W_s$ and input channel width w_{s-1} . The block comprises three primary convolutional operations. The initial 1×1 convolution reduces or maintains the channel dimension with a bottleneck ratio $b = 1$. The number of multiplications and additions for this operation is given by:

$$FLOPs_{conv1} = H_s \cdot W_s \cdot w_{s-1} \cdot w_s \quad (42)$$

The subsequent 3×3 group convolution utilizes a group number g , which partitions the input channels into g independent groups, significantly reducing the dense connectivity. The FLOPs for this group convolution are calculated as:

$$FLOPs_{conv2} = H_s \cdot W_s \cdot w_s \cdot \left(\frac{w_s}{g} \right) \cdot (3 \cdot 3) = \frac{9H_s W_s w_s^2}{g} \quad (43)$$

The final 1×1 projection convolution restores the feature manifold, contributing a complexity of:

$$FLOPs_{conv3} = H_s \cdot W_s \cdot w_s \cdot w_s = H_s W_s w_s^2 \quad (44)$$

Combining these components, the total FLOPs for one X-block can be expressed as:

$$FLOPs_{block} = H_s W_s (w_{s-1} w_s + \frac{9w_s^2}{g} + w_s^2) \approx H_s W_s (2w_{s-1} w_s + \frac{9w_s^2}{g}) \quad (45)$$

The approximation $w_{s-1} w_s + w_s^2 \approx 2w_{s-1} w_s$ holds under the linear parameterization of the RegNet design space where channel increments are controlled. By aggregating the costs across $S = 4$ stages and accounting for the stem layer, the total network complexity is:

$$FLOPs_{RegNet-X} = \sum_{s=1}^S d_s \cdot H_s W_s \cdot \left(2w_{s-1} w_s + \frac{9w_s^2}{g} \right) \quad (46)$$

To evaluate the real-time constraint, the processing time $T_{process}$ is modeled as a function of the total FLOPs and the effective hardware throughput. Given the peak performance R_{peak} and a utilization efficiency factor $\eta_{util} \in (0, 1]$, the execution time is:

$$T_{process} = \frac{FLOPs_{RegNet-X} + FLOPs_{head}}{\eta_{util} \cdot R_{peak}} \leq T_{max} \quad (47)$$

Substituting the design parameters $w_0 = 24, w_a = 36$ and stage depths $[1, 2, 7, 12]$, the total complexity for a 640×640 input approximates 1.6 GFLOPs. On high-performance edge computing platforms (e.g., NVIDIA Jetson), the utilization η_{util} is maximized due to the hardware-friendly channel widths (multiples of 8) and the use of group convolutions which balance the arithmetic intensity and memory bandwidth.

Furthermore, the integration of the edge enhancement module adds negligible overhead $FLOPs_{edge} = 2 \cdot 9 \cdot H_0 W_0 \cdot C_0$, which does not shift the asymptotic complexity. Given that $T_{max} = 10$ ms, the derived $FLOPs_{RegNet-X}$ satisfies the inequality $T_{process} \leq T_{max}$ under standard INT8 or FP16 optimization, ensuring the operational feasibility of the ESAD-Net in high-speed steel rolling production lines.

Theorem 4 (Adaptive Threshold Convergence) For the steel defect detection task with ground truth set $\{g^{(n)}\}_{n=1}^{N_{gt}}$, when the training process reaches statistical equilibrium where the feature representations stabilize, the adaptive threshold $\tau^{(n)}$ converges to a stable value:

$$\lim_{t \rightarrow T_{conv}} |\tau^{(n)}(t) - \tau^{(n)}(t-1)| \leq \epsilon_{conv} \quad (48)$$

where t denotes the training epoch, T_{conv} is the convergence epoch, and ϵ_{conv} is the convergence tolerance. Consequently, the positive sample ratio for each ground truth stabilizes:

$$\pi^{(n)} = \frac{|\mathcal{S}_{pos}^{(n)}|}{|\mathcal{A}_{cand}^{(n)}|} \xrightarrow{t \rightarrow T_{conv}} \pi_{stable}^{(n)} \pm \delta \quad (49)$$

where $\pi_{stable}^{(n)}$ is the converged positive sample ratio and δ is the convergence error bound.

Proof 3 The convergence of the adaptive threshold $\tau^{(n)}$ is established by analyzing the stochastic dynamics of Intersection over Union (IoU) distributions during the optimization of the network parameters θ . Let $a_i \in \mathcal{A}_{cand}^{(n)}$ be a candidate anchor for a ground truth $g^{(n)}$. The IoU at epoch t is a function of the predicted bounding box regression $\Delta b_i(t)$, denoted as $X_i^{(t)} = \text{IoU}(a_i, g^{(n)}; \theta_t)$.

The bounding box regression branch is optimized via the GIoU loss $\mathcal{L}_{reg}$, which ensures that for a sufficiently small learning rate η_t , the parameter update $\|\theta_t - \theta_{t-1}\|$ follows a contraction mapping as the gradients $\nabla_{\theta} \mathcal{L}$ vanish near a local optimum. Due to the local Lipschitz continuity of the IoU function with respect to the regression parameters, the variation in individual IoU values is bounded:

$$|X_i^{(t)} - X_i^{(t-1)}| \leq L_{IoU} \cdot \|\theta_t - \theta_{t-1}\| \leq L_{IoU} \cdot \eta_{t-1} \|\nabla_{\theta} \mathcal{L}(\theta_{t-1})\| \quad (50)$$

The adaptive threshold is defined by the first and second statistical moments of the IoU distribution over the candidate set $\mathcal{A}_{cand}^{(n)}$, where $|\mathcal{A}_{cand}^{(n)}| = K$. The mean IoU $\mu_{IoU}^{(n)}(t)$ and the standard deviation $\sigma_{IoU}^{(n)}(t)$ evolve according to:

$$\mu_{IoU}^{(n)}(t) = \frac{1}{K} \sum_{i=1}^K X_i^{(t)}, \quad \sigma_{IoU}^{(n)}(t) = \sqrt{\frac{1}{K} \sum_{i=1}^K (X_i^{(t)} - \mu_{IoU}^{(n)}(t))^2} \quad (51)$$

Applying the triangle inequality to the threshold difference $|\tau^{(n)}(t) - \tau^{(n)}(t-1)|$, where $\tau^{(n)}(t) = \mu_{IoU}^{(n)}(t) + \sigma_{IoU}^{(n)}(t)$, yields:

$$|\tau^{(n)}(t) - \tau^{(n)}(t-1)| \leq |\mu_{IoU}^{(n)}(t) - \mu_{IoU}^{(n)}(t-1)| + |\sigma_{IoU}^{(n)}(t) - \sigma_{IoU}^{(n)}(t-1)| \quad (52)$$

The stability of the mean is directly inherited from the parameter convergence: $|\mu_{IoU}^{(n)}(t) - \mu_{IoU}^{(n)}(t-1)| \leq L_{IoU} \|\theta_t - \theta_{t-1}\|$. For the standard deviation, using the reverse triangle inequality for the L_2 norm, it can be shown that:

$$|\sigma_{IoU}^{(n)}(t) - \sigma_{IoU}^{(n)}(t-1)| \leq \sqrt{\frac{1}{K} \sum_{i=1}^K |(X_i^{(t)} - \mu_t) - (X_i^{(t-1)} - \mu_{t-1})|^2} \leq 2L_{IoU} \|\theta_t - \theta_{t-1}\| \quad (53)$$

Consequently, the total threshold variation is bounded by $|\tau^{(n)}(t) - \tau^{(n)}(t-1)| \leq 3L_{IoU} \|\theta_t - \theta_{t-1}\|$. As $t \rightarrow \infty$ and $\eta_t \rightarrow 0$, the right-hand side converges to zero, satisfying the convergence tolerance ϵ_{conv} at T_{conv} . The stabilization of the positive sample ratio $\pi^{(n)}$ follows from the convergence of the indicator function $\mathbb{I}(X_i^{(t)} \geq \tau^{(n)}(t))$. As the distributions of both the IoU values and the threshold stabilize, the probability of an anchor's status switching between positive and negative categories diminishes. The expected ratio $\mathbb{E}[\pi^{(n)}]$ at equilibrium is:

$$\pi_{stable}^{(n)} = \frac{1}{K} \sum_{i=1}^K \mathbb{P}(X_i^* \geq \tau^{(n)*}) = \int_{\tau^{(n)*}}^1 f_{IoU}^*(x) dx \quad (54)$$

where f_{IoU}^* is the steady-state probability density function of the IoU values for the specific defect geometry. Given the discrete nature of the anchor set, the actual ratio $\pi^{(n)}$ fluctuates within a bound δ determined by the local density of f_{IoU}^* near the threshold $\tau^{(n)*}$. This completes the proof that the ATSS mechanism reaches a statistical equilibrium, ensuring consistent sample supervision for steel defect detection.

Corollary 4 (Small Target Detection Improvement) Under the convergence conditions of Theorem 2, the ATSS adaptive threshold mechanism provides systematic improvement for small target defect detection compared to fixed threshold methods:

$$\Delta mAP^{small} = mAP_{ATSS}^{small} - mAP_{fixed}^{small} \geq \kappa \cdot \left(\frac{\sigma_{IoU}^{small}}{\mu_{IoU}^{small}} \right) \quad (55)$$

where mAP_{ATSS}^{small} and mAP_{fixed}^{small} are the average precision for small targets using adaptive and fixed thresholds respectively, $\kappa > 0$ is a task-dependent coefficient, and $\sigma_{IoU}^{small} / \mu_{IoU}^{small}$ is the coefficient of variation of IoU distribution for small targets.

Proof 4 The systematic improvement for small target detection is derived by analyzing the sensitivity of the positive sample selection process to target scale under fixed versus adaptive thresholding. For small defects g^{small} with area $A < 32^2$, the IoU distribution $f(x)$ typically exhibits a lower mean μ_{IoU}^{small} and higher relative dispersion compared to larger targets, primarily due to the increased sensitivity of the IoU metric to small spatial pixel shifts.

Under a fixed thresholding strategy (e.g., $\tau_{fixed} = 0.5$), the expected number of positive training samples N_{pos}^{fixed} for a small target is governed by the tail probability of the IoU distribution:

$$N_{pos}^{fixed} = |\mathcal{A}_{cand}| \cdot \int_{\tau_{fixed}}^1 f_{IoU}^{small}(x) dx \quad (56)$$

As the target size decreases, μ_{IoU}^{small} shifts leftward, often resulting in $\tau_{fixed} > \mu_{IoU}^{small} + \sigma_{IoU}^{small}$. In such cases, N_{pos}^{fixed} approaches zero, leading to the vanishing gradient problem in the classification loss $\mathcal{L}_{cls}$ and regression loss $\mathcal{L}_{reg}$, which severely degrades the average precision mAP_{fixed}^{small} .

In contrast, the ATSS mechanism defines the adaptive threshold as $\tau^{(n)} = \mu_{IoU}^{(n)} + \sigma_{IoU}^{(n)}$. This ensures that the selection criterion is statistically normalized to the specific IoU characteristics of small targets. The expected number of positive samples under ATSS remains relatively invariant to scale:

$$N_{pos}^{ATSS} = |\mathcal{A}_{cand}| \cdot \mathbb{P}(X \geq \mu_{IoU}^{small} + \sigma_{IoU}^{small}) \approx |\mathcal{A}_{cand}| \cdot \Phi(-1) \quad (57)$$

where Φ is the cumulative distribution function of the normalized IoU distribution. The increase in the number of effective positive samples $\Delta N_{pos} = N_{pos}^{ATSS} - N_{pos}^{fixed}$ directly enhances the signal-to-noise ratio of the training gradients. By applying the central limit theorem to the gradient updates, the reduction in variance of the learned feature representation $\mathcal{R}_\theta$ is proportional to $\sqrt{\Delta N_{pos}}$. The precision-recall performance improvement ΔmAP^{small} can be modeled by integrating the precision gain over the recall range. Since small targets suffer most from low recall due to sample scarcity, the gain is dominated by the expansion of the recall boundary. The improvement is mathematically linked to the coefficient of variation $CV^{small} = \sigma_{IoU}^{small} / \mu_{IoU}^{small}$. For a distribution with a low mean, a higher CV^{small} indicates a heavier tail relative to the mean, which ATSS exploits by lowering the threshold $\tau^{(n)}$ more aggressively than a fixed threshold would allow. The relationship can be expressed through the sensitivity of the threshold to the distribution spread:

$$\Delta mAP^{small} \approx \int_0^1 \Delta P(R) dR \propto \frac{\partial \pi^{(n)}}{\partial \sigma_{IoU}} \cdot \frac{1}{\mu_{IoU}} \quad (58)$$

By aggregating the task-specific constants into a coefficient κ , which accounts for the network's learning capacity and the defect density in the NEU-DET dataset, we obtain the bound:

$$\Delta mAP^{small} \geq \kappa \cdot \left(\frac{\sigma_{IoU}^{small}}{\mu_{IoU}^{small}} \right) \quad (59)$$

This derivation confirms that the ATSS mechanism provides a necessary statistical correction for small target defects, where the performance gain is intrinsically tied to the relative variance of the anchor-target overlap distribution.