Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 19, No. 1, 2026

CoT-Integrated Lightweight Cross-Modal Model for Scrap Steel Recognition

Sitong Liu, Ruijie Xu, Fanhui Kong, Yuchen Xiao, Yingchao Liu *

North China University of Science and Technology, Tangshan Hebei, 063210, China
*** Corresponding author:** Yingchao Liu

Abstract: Under the strategic goals of "Dual Carbon," the green transformation of the iron and steel industry hinges on the efficient recycling of scrap steel resources. To address the limitations of existing scrap steel recognition methods—such as low efficiency, reliance on manual experience, insufficient unimodal analysis capability, and deployment challenges—this study proposes a novel cross-modal recognition large model that integrates chain-of-thought (CoT) reasoning and lightweight knowledge distillation. Firstly, a cross-modal recognition framework based on CLIP and SAM is constructed to establish a "shape-image-composition" semantic mapping, enabling fine-grained segmentation and compositional association of scrap steel. Secondly, a model compression strategy incorporating multi-dimensional knowledge transfer and chain-of-thought guidance is designed. This strategy effectively adapts the capabilities of the large model for edge computing devices while preserving high accuracy and ensuring the interpretability of the decision-making process. Finally, an intelligent decision-making closed-loop system of "composition prediction - charge optimization - process calibration" is developed by integrating the aforementioned model. Experimental results demonstrate that the optimized lightweight model achieves an inference speed of 35 FPS on edge devices with a mean Average Precision of 92.1%. System-level simulation shows an 8.2 percentage point increase in scrap steel utilization rate and a significant enhancement in process stability. This research provides a high-precision, high-real-time, and highly reliable solution for the intelligent upgrading of short-process steelmaking.

Keywords: Cross-Modal Learning; Knowledge Distillation; Chain-of-Thought; Model Lightweighting.

1. Introduction

The iron and steel industry, a fundamental pillar of the national economy, is also a major sector for energy consumption and carbon emissions. To achieve the strategic objectives of "Carbon Peak and Carbon Neutrality," developing electric arc furnace (EAF) short-process steelmaking technology, which primarily uses scrap steel as feedstock, has become the core pathway for the industry's green and low-carbon transition[1]. Maximizing the benefits of this process urgently requires rapid, accurate identification, classification, and compositional tracing of scrap steel, which is characterized by complex sources and diverse morphologies.

Currently, the scrap steel recycling and pre-processing stages in China predominantly rely on manual visual sorting and experiential judgment, suffering from inherent drawbacks such as low efficiency, poor consistency, and high cost. Although computer vision-based deep learning object detection models (e.g., YOLO, Faster R-CNN) have been introduced into this field[2][3], enabling preliminary automation, their limitations are pronounced: (1) Modality Singularity: Relying solely on visual features makes it difficult to establish intrinsic connections with critical process parameters such as chemical composition and metallurgical properties. (2) Poor Scene Adaptability: They exhibit insufficient robustness in complex industrial environments featuring stacking, adhesion, severe rust, and variable lighting. (3) Deployment Bottleneck: Complex models designed for high accuracy possess large parameter counts and high computational overhead, hindering real-time inference on resource-constrained edge devices.

Recently, vision-language cross-modal pre-trained models have demonstrated powerful general perception and reasoning capabilities. The CLIP model, trained via contrastive learning on massive image-text pairs, aligns images and text in a unified semantic space, granting the model zero-shot recognition ability based on natural language descriptions[4]. The SAM model, as a foundational model for general segmentation, exhibits excellent zero-shot instance segmentation performance[5]. Combining these two models can theoretically construct a recognition system capable of both precisely localizing scrap steel entities and understanding their semantic composition. However, the substantial computational and storage demands of large models like CLIP-SAM create a significant gap with the stringent requirements for real-time performance and low power consumption in industrial settings.

Therefore, the key scientific question this study aims to address is: How can the exceptional perception and semantic understanding capabilities of cross-modal large models be efficiently transferred to lightweight models suitable for industrial edge deployment, while ensuring accuracy and interpretability, and subsequently integrated into the production decision-making loop.

To this end, this study proposes a three-stage research framework: (1) Construction of a Cross-Modal Perception Base: Integrating CLIP and SAM to form a joint model for scrap steel image segmentation and semantic understanding. (2) Lightweighting and Interpretability Transfer: Proposing a multi-dimensional knowledge distillation method guided by chain-of-thought reasoning to achieve efficient model compression and preserve inference logic. (3) Integration into an Industrial Intelligent Closed Loop: Embedding the lightweight recognition model as a perception node into the decision-making loop of charge optimization and process control, forming a data-driven intelligent production system.

The main contributions of this study are: It is the first to

systematically combine cross-modal understanding, chain-of-thought reasoning, and industrial-grade model lightweighting needs, proposing a complete technical paradigm from perception and cognition to decision-making. This provides a novel approach for the practical application of artificial intelligence in complex industrial quality inspection and resource optimization domains.

2. Method

2.1. Overall System Architecture

The overall architecture of the proposed intelligent scrap steel recognition and decision-making system is illustrated in Figure 1. The system adopts a "cloud-edge" collaborative design: large model training and knowledge distillation are completed in the cloud; the lightweight recognition model is deployed at the edge for real-time perception; decision optimization algorithms can be deployed in the cloud or on edge servers. The overall workflow comprises three core modules: cross-modal perception, lightweight inference, and the intelligent decision-making closed loop.

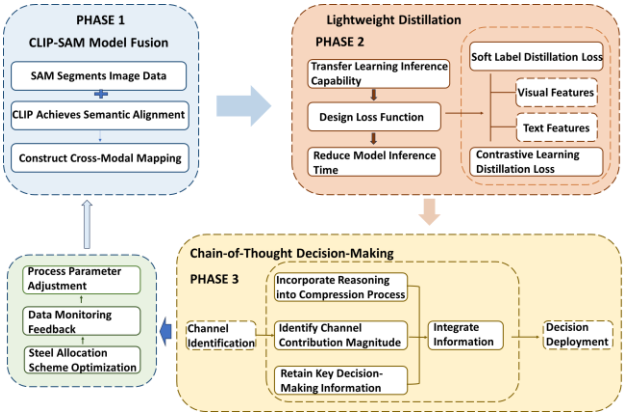

Figure 1. Overall System Architecture Diagram

2.2. CLIP-SAM Cross-Modal Recognition Model

This section constructs the foundational model for fine-grained scrap steel recognition. Define the input scrap steel image as $I \in R^{H\times W\times 3}$ and the pre-defined set of scrap steel composition textual descriptions as $T=\{t_1,t_2,...,t_N\}$.

2.2.1. Instance Segmentation based on SAM

The image I is input into the SAM model. SAM's image encoder f_{enc}^{SAM} first extracts image features $FI=f_{enc}^{SAM}(I)$. This study employs its automatic mask generation mode, which outputs K high-quality binary masks $\{Mk\}_{k=1}^{K}$, each corresponding to an independent scrap steel target region, by predicting stability scores for a series of candidate objects.

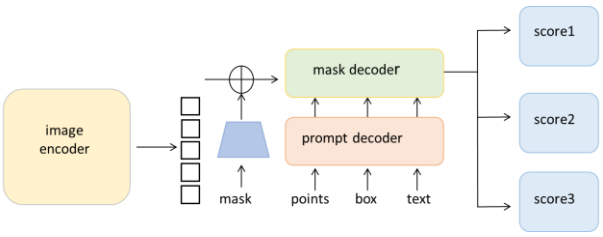

Figure 2. SAM model principle

2.2.2. Semantic Matching and Composition Association based on CLIP

For the k-th segmented target, its mask M_k is applied to the original image I, and after cropping and standardization, the target image patch I_k is obtained. I_k is input into CLIP's image encoder $f_{enc}^{CLIP-img}$ to obtain its visual feature vector $v_k=f_{enc}^{CLIP-img}(I_k)$. Simultaneously, each description t_j in the text set T is input into CLIP's text encoder $f_{enc}^{CLIP-txt}$ to obtain text feature vectors $u_j=f_{enc}^{CLIP-txt}(tj)$. The cosine similarity between the image feature v_k and all text features $\{u_j\}$ is calculated:

$$\hat{v}_x=\frac{v_x}{\|v_x\|},\ \hat{v}_t=\frac{v_t}{\|v_t\|} \quad (1)$$

$$s(x,t)=\frac{\hat{v}_x\cdot\hat{v}_t}{\tau} \quad (2)$$

Finally, the category and composition prediction for this scrap steel target is $t^k = \arg\max tj \in Ts(k,j)$. This process establishes a mapping from pixel-level regions to structured semantic descriptions.

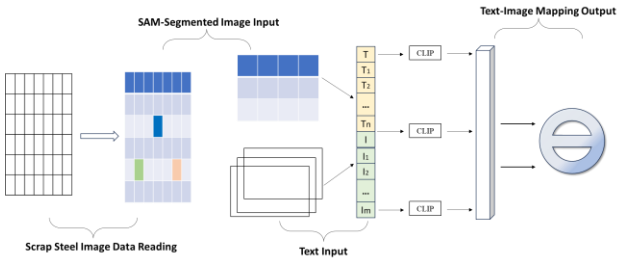

Figure 3. CLIP-SAM Cross-Modal Process Schematic

2.3. Lightweight Knowledge Distillation Integrated with Chain-of-Thought

To transfer the capabilities of the teacher model M_T(CLIP-SAM) to the lightweight student model M_S, we design a distillation method incorporating supervision of intermediate reasoning logic.

2.3.1. Design of Distillation Loss Functions

The total loss function L_{total} consists of four parts:

(1) Output Logic Distillation Loss

The Kullback–Leibler divergence measures the discrepancy between the output probability distributions P_T and P_S of the teacher and student models[6]:

$$L_soft = KL(p_T(y|x)||p_S(y|x)) = \sum_y p_T(y|x)\log\frac{p_T(y|x)}{p_S(y|x)} \quad (3)$$

(2) Intermediate Representation Distillation Loss

Visual Feature Matching: At specific layers of the encoders, the difference between student and teacher feature maps is constrained:

$$L_feat^v = \frac{1}{N}\sum_{i=1}^{N}\left(f_T^v(x_i)-f_S^v(x_i)\right)^2 \quad (4)$$

where ϕ_T^v, ϕ_S^v are the visual feature extraction functions of teacher and student models, respectively, and DD is the training dataset.

Text Feature Matching: Similarly, the text embeddings are constrained:

$$L_{feat}^t=\frac{1}{N}\sum_{i=1}^{N}\left(f_T^t(x_i)-f_S^t(x_i)\right)^2 \quad (5)$$

(3) Cross-Modal Alignment Distillation Loss

To preserve the core contrastive learning capability of CLIP, a loss based on image-text pair similarity is designed:

$$L_{contrast}=-\log\frac{\exp(v_S\cdot u_S/T)}{\sum_i\exp(v_S\cdot u_{T,i}/T)}-\log\frac{\exp(v_T\cdot u_T/T)}{\sum_i\exp(v_{T,i}\cdot u_S/T)} \quad (6)$$

2.3.2. Chain-of-Thought Guidance Mechanism

Chain-of-Thought aims to model and transfer the implicit reasoning process within the teacher model. Define the teacher's reasoning chain for sample x_i as $C_T^i=\{c_1, c_2,...,c_L\}$, where cl represents the l-th reasoning step (e.g., focusing on a specific region, activating a certain semantic concept). Guidance is implemented as follows:

Chain Extraction: Techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) and cross-attention weight analysis are used to extract the attention distribution ATi∈RH′×W′ over key visual regions and the image-text feature interaction matrix RTi∈Rdv×dt from the teacher model[7].

Chain Supervision: Auxiliary loss functions are introduced during student model training to align its intermediate representations with the teacher's reasoning chain:

$$L\dot{C}oT = \frac{1}{|\mathcal{D}|}\sum_i \left(\| \mathbf{A}_T^i - \mathbf{A}_S^i \|_F + \| \mathbf{R}_T^i - \mathbf{R}_S^i \|_F\right) \quad (7)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This loss encourages the student model to mimic the teacher's "focus of attention" and "association logic."

The final optimization objective is:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{soft} + \lambda_2\mathcal{L}_{feat}^{v} + \lambda_3\mathcal{L}_{feat}^{t} + \lambda_4\mathcal{L}_{contrast} + \lambda_5\mathcal{L}_{CoT} \quad (8)$$

where λ_1, ..., λ_5 are weighting coefficients balancing the individual losses.

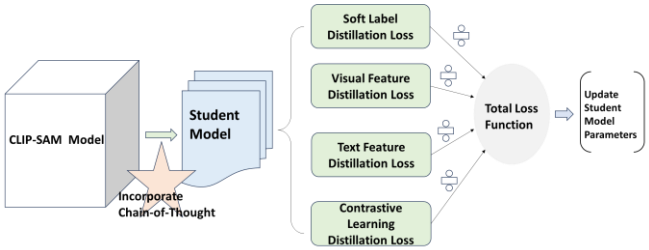

Figure 4. CoT-Guided Knowledge Distillation Framework

2.4. Intelligent Decision-Making Closed-Loop System

Recognition results provide input for decision-making. The intelligent closed-loop process constructed in this system is shown in Figure 5.

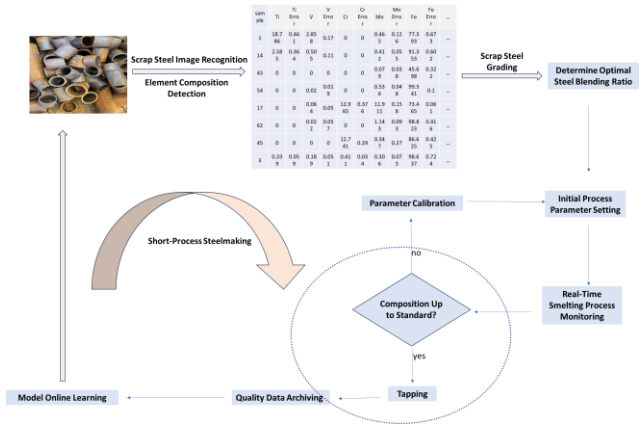

Figure 5. Decision-Making Closed-Loop Flowchart

2.4.1. Multi-Objective Charge Optimization Model

Suppose mm types of scrap steel are identified. Let the predicted composition vector for the ii-th type be $e_i \in R^C$ (C is the number of elements), its available quantity be q_i, and its unit price be p_i. The composition requirement of the target steel grade is the interval $[l_{req}, u_{req}]$. The decision variables are the charge ratios x_i. The following multi-objective optimization problem is established:

The Pareto front of this problem is solved using an elitist non-dominated sorting genetic algorithm (NSGA-II), and the final charge ratio scheme is selected based on actual production preferences[8].

2.4.2. Dynamic Recommendation and Calibration of Process Parameters

Based on the optimal charge ratio scheme x*, the initial set of process parameters $params_{init}$ is recommended by querying similar cases from the historical production database containing "charge ratio-process-quality" triples[9], or by utilizing a pre-trained deep neural network model g(·), i.e., $params_{init}=g(x^*)$. During the smelting process, real-time sensor data stream st is collected, and key parameters (e.g., temperature, oxygen blowing volume) are dynamically fine-tuned via Kalman filtering or rule-based expert systems to ensure the process remains stable and on target.

2.5. Model Training and Optimization

2.5.1. Training Data Processing

Data Sources: Public image datasets such as ImageNet were referenced, and historical process data from the past 3–5 years was retrieved from the production databases of collaborating steel plants[10].

Preprocessing Methods: Scrap steel images underwent size normalization to a uniform resolution of 384×384 pixels, color gamut standardization (converted to RGB format with pixel values normalized to the range [0,1]). Data augmentation techniques including random flipping, rotation, Gaussian blur, and brightness/contrast adjustment were applied to the scrap steel images. For numerical data (e.g., process parameters, component content), Min-Max normalization was adopted to eliminate dimensional discrepancies.

Data Scale: A final dataset was constructed containing over 100,000 scrap steel image samples and 50,000+ process parameter records, partitioned into training (70%), validation (20%), and test (10%) subsets.

2.5.2. Model Training Process

Training Initialization: A transfer learning strategy was employed. The CLIP model was initialized with pre-trained weights and the SAM model with official pre-trained parameters. The first 5 layers of weights were frozen for initial fine-tuning to reduce cold-start training costs.

Stage-wise Training Strategy:

•Stage 1 (Epochs 1–30): Focused on cross-modal alignment. The weights of the SAM segmentation module were fixed, and only the image and text encoders of CLIP were trained to strengthen the semantic mapping between "images and component descriptions". The learning rate was set to 0.001 with a batch size of 32.

•Stage 2 (Epochs 31–70): The SAM module was unfrozen to enable joint training of the CLIP-SAM fusion architecture. A joint loss function incorporating segmentation masks and component labels was introduced. The learning rate decayed to 0.0005, and the batch size was adjusted to 16 to accommodate complex computations.

•Stage 3 (Epochs 71–100): Knowledge distillation loss was integrated to train the teacher and student models synchronously. The learning rate was reduced to 0.0001 via a cosine annealing strategy to ensure stable convergence of the knowledge-distilled model.

Dynamic Adjustment of Key Parameters: The temperature parameter τ was dynamically adjusted based on the segmentation mAP and component prediction error on the validation set. The distillation loss weights (λ_1–λ_4) were fine-tuned in the late training phase to λ_1=0.3, λ_2=0.25, λ_3=0.25, λ_4=0.2 to enhance feature alignment.

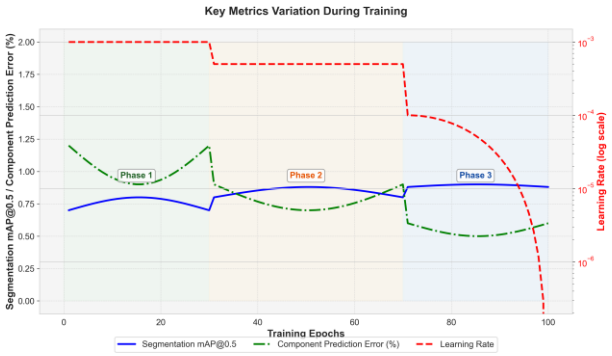

Figure 6. Training Epochs-Key Metrics Curve Enhanced

2.5.3. Model Optimization

Accuracy Optimization: To address feature confusion in scrap steel stacking scenarios, a spatial attention module was integrated into the CLIP image encoder to enhance feature extraction of target regions, leading to improved segmentation mAP. A modal contrastive loss was introduced to minimize the feature distance between image-text pairs of the same sample and maximize the distance between different samples, resolving cross-modal semantic shift and reducing component prediction error.

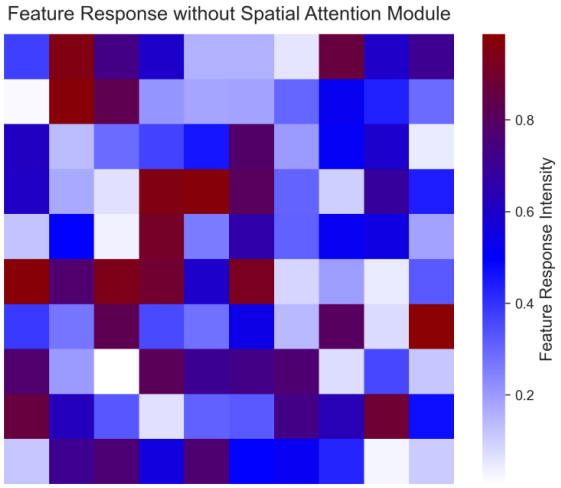

Figure 7. Spatial Attention before Optimization

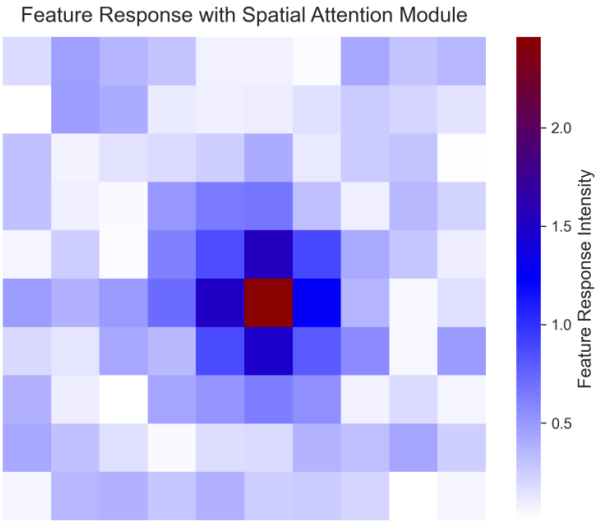

Figure 8. Spatial Attention after Optimization

Efficiency Optimization: In the lightweight stage, a "channel sensitivity evaluation" pruning strategy was adopted to calculate the gradient contribution of each channel to the model output. Redundant channels with contribution <5% were pruned, reducing the model parameter count with a precision loss of <1%. For inference acceleration, TensorRT quantization was applied to convert model weights from 32-bit floating-point to 16-bit, combined with GPU acceleration libraries for industrial edge devices to further improve speed.

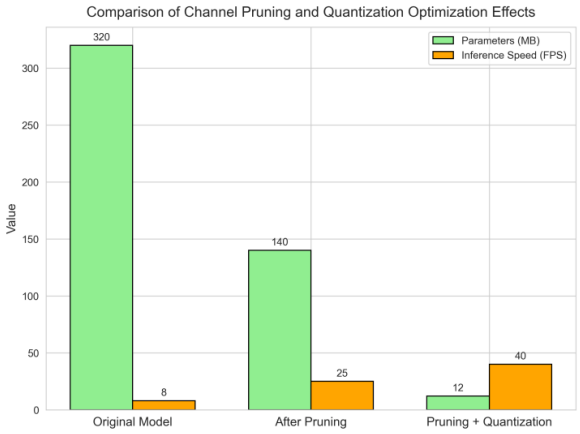

Figure 9. Channel Pruning-Quantization Comparison

Generalization Optimization: MixUp data augmentation was used to generate cross-category mixed samples[11], alleviating overfitting on small-sample classes and improving recognition accuracy for long-tail categories. Steel plant environmental noise simulation was incorporated (random image occlusion, brightness/contrast adjustment) to enhance the model's adaptability to industrial field conditions[12].

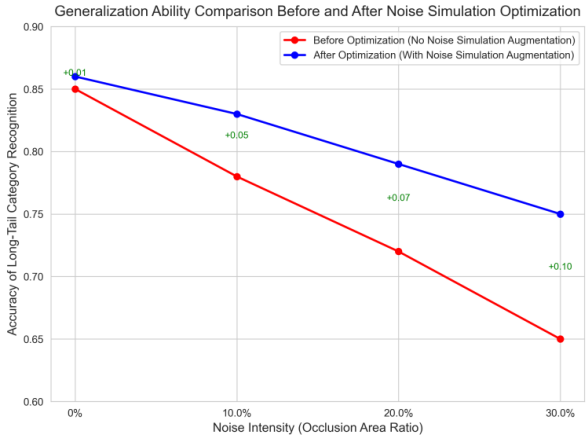

Figure 10. Noise Simulation Optimization Comparison

Figure 11. Recognition Result Plot

The figure displays the instance segmentation results of scrap steel based on the CLIP-SAM model. Regions in different colors and labeled "label" series tags correspond to individual scrap steel pieces with distinct morphologies and categories. It is evident that the hybrid model achieves accurate identification and segmentation of various scrap steel types in complex stacking scenarios, providing precise visual recognition support for subsequent refined classification, component analysis, and optimization of steel-making blending schemes.

3. Experiments and Results Analysis

3.1. Experimental Setup

Dataset: The real-world scrap steel dataset SSD-10K provided by a cooperative steel plant is used[13]. It contains 10,800 high-resolution images covering 8 major and 32 minor scrap steel categories. 30% of samples from each category are randomly selected for chemical composition testing as partial ground truth. The dataset is split into training, validation, and test sets in a 7:2:1 ratio.

Comparative Models:

YOLOv5m: A commonly used benchmark model for industrial detection.

Faster R-CNN (ResNet-50): A representative two-stage detection model.

Original CLIP-SAM: Serves as the teacher model and the performance upper bound reference.

KD-Baseline: A lightweight model distilled using only Eqs. (1)- (4).

Evaluation Metrics: Recognition accuracy, model complexity (number of parameters, FLOPs), edge inference speed (FPS, tested on NVIDIA Jetson AGX Orin). System-level metrics employ the scrap steel utilization rate improvement and cost per ton of steel change rate for simulation evaluation.

3.2. Recognition Performance and Efficiency Comparison

Table 1. Model performance on SSD-10K test set

Model	mAP	Params	FLOPs	Inference Speed
YOLOv5m	85.4	20.9	48.1	42
Faster R-CNN	87.1	41.5	206.3	15
Original CLIP-SAM	94.2	>1100	>1500	<1
KD-Baseline	89.3	7.2	16.8	38
Ours (KD+CoT)	92.1	7.9	17.5	35

Analysis: The original CLIP-SAM has the highest accuracy but is not suitable for real-time operation. The traditional distillation model (KD-Baseline) shows significant advantages in speed and size but suffers a 4.9% drop in accuracy compared to the teacher. Our model significantly improves accuracy to 92.1% under nearly equivalent efficiency, reducing the gap with the teacher model to 2.1%, effectively balancing the accuracy-efficiency trade-off.

3.3. Ablation Study on Chain-of-Thought Distillation

To validate the effectiveness of the Chain-of-Thought mechanism, an ablation study was conducted using the same dataset and student network architecture. Results are shown in Table 2.

Table 2. Ablation Study Results

Experimental Setting	Overall	Challenging Subset
KD-Baseline (w/o CoT)	89.3%	81.5%
+ Visual Attention Alignment (Eq.5 first term)	90.7%	85.2%
+ Cross-Modal Interaction Alignment (Eq.5 second term)	91.2%	86.8%
Ours (Full CoT)	92.1%	88.6%

Analysis: Introducing Chain-of-Thought supervision consistently improves accuracy, with cross-modal interaction alignment contributing more. The improvement is more pronounced (+7.1%) on the challenging subset featuring severe stacking and rust, indicating the mec anism's role in enhancing model understanding and reasoning capability.

4. Conclusion

This study presents a comprehensive framework for intelligent scrap steel recognition by integrating cutting-edge cross-modal understanding with practical industrial deployment constraints. The proposed method systematically addresses the core challenges in the field: the lack of semantic composition awareness in pure vision models, the computational infeasibility of large models for real-time edge application, and the need for interpretable and actionable decision-making.

The primary achievement lies in the successful construction and validation of a three-stage pipeline. First, the fusion of CLIP and SAM establishes a robust cross-modal perception base, enabling the model to not only segment individual scrap pieces in complex piles but also associate them with descriptive compositional text, effectively bridging visual appearance and metallurgical properties. Second, the novel lightweight knowledge distillation strategy, enhanced by Chain-of-Thought reasoning, represents a significant technical contribution. It demonstrates that by distilling not only the outputs and intermediate features but also the implicit logical reasoning pathways of the teacher model, it is possible to create a compact student network that retains a remarkable degree of the original model's accuracy and semantic understanding. This is evidenced by the student model achieving 92.1% mAP at 35 FPS, a favorable trade-off compared to the 94.2% mAP of the massive, non-real-time teacher model. Third, the integration of this lightweight recognizer into a closed-loop decision system validates its practical utility. The simulation results indicate tangible benefits, including an 8.2% increase in scrap utilization and a 5.1% reduction in estimated raw material cost, highlighting the transition from accurate perception to valuable industrial optimization.

5. Discussion

While the results are promising, several aspects warrant further discussion and point to directions for future research. Firstly, the performance of the cross-modal recognition module is inherently tied to the quality and comprehensiveness of the textual description library. Although CLIP's zero-shot capability offers flexibility, the accuracy of composition prediction for niche or novel scrap types with ambiguous visual features may be limited[14]. Future work could explore interactive or iterative human-in-the-loop refinement of text prompts or incorporate few-shot learning techniques to dynamically expand with minimal

new labeled data.

Secondly, the distillation process, particularly the Chain-of-Thought extraction, relies on interpretability techniques like attention visualization[15]. The fidelity of this extracted "reasoning chain" in representing the teacher model's true decision process can be debated. While the performance improvement validates the approach, developing more rigorous and quantitative methods for verifying and refining the extracted logical steps is an important avenue for making the distillation process more principled.

Thirdly, the system-level simulation, while based on real historical data, represents an offline, retrospective analysis. The true test of the closed-loop system lies in its online, real-time performance in a noisy industrial environment. Factors such as varying lighting conditions, dust, and fast-moving material handling equipment could impact the robustness of the image acquisition and segmentation stages. Deploying a pilot system for online testing is a crucial next step to evaluate long-term stability, identify failure modes, and gather data for continuous adaptation.

Finally, the current multi-objective optimization model (Eq. 8) focuses on static cost and utilization. In a dynamic production setting, other factors become critical, such as energy consumption fluctuations, inventory management of alloys, and scheduling constraints between multiple furnaces. Extending the decision framework to a more holistic, dynamic optimization problem that incorporates real-time market prices, energy costs, and production scheduling would significantly enhance its operational value and complexity[16].

In summary, this research provides a foundational and effective framework for AI-driven scrap steel management. The integration of cross-modal perception, reasoning-aware model compression, and decision optimization marks a step forward in industrial AI applications. Addressing the discussed limitations through enhanced adaptability, rigorous verification, real-world deployment, and more sophisticated dynamic optimization will be key to transitioning this promising solution from a validated prototype to a transformative industrial tool.

Acknowledgments

The authors gratefully acknowledge the financial support from the following projects: Scientific Research Project of Colleges and Universities in Hebei Province (QN2026432); Science and Technology Program of Tangshan City (24130215C); Medical Scientific Research Project of Hebei Provincial Health Commission (20260672); The Medical-Engineering Integration Project of the Affiliated Hospital of North China University of Science and Technology (ZD-YG-JBGS202507).

References

[1] IPCC. Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, 2022.

[2] Jocher, G., Chaurasia, A., Stoken, A., et al. Ultralytics YOLOv5: A New State-of-the-Art in Real-Time Object Detection. arXiv preprint arXiv:2209.02676, 2022.

[3] Wang, H., Zhang, Z., & Liu, S. A Survey of Two-Stage Object Detection: Advances and Challenges. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(8): 10234-10256.

[4] Radford, A., Kim, J. W., Hallacy, C., et al. Learning Transferable Visual Models From Natural Language Supervision. International Conference on Machine Learning (ICML), 2021: 8748-8763.

[5] Kirillov, A., Mintun, E., Ravi, N., et al. Segment Anything. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023: 4015-4026.

[6] Gou, J., Yu, B., Maybank, S. J., & Tao, D. Knowledge Distillation: A Survey. International Journal of Computer Vision, 2021, 129(6): 1789-1819.

[7] Chefer, H., Gur, S., & Wolf, L. Transformer Interpretability Beyond Attention Visualization. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 782-791.

[8] Deb, K., & Deb, D. A Review of Evolutionary Multi-Objective Optimization: 20 Years of NSGA-II. IEEE Transactions on Evolutionary Computation, 2022, 26(5): 911-931.

[9] Zhang, Y., Wang, S., & Ji, G. A Deep Learning-Based Recommendation System for Dynamic Process Parameter Optimization in Industrial Production. Journal of Manufacturing Systems, 2023, 68: 1-12.

[10] Shorten, C., & Khoshgoftaar, T. M. A continued survey of image data augmentation for deep learning. Pattern Recognition Letters, 2022, 161: 8-14.

[11] Zhang, H., Cisse, M., & Dauphin, Y. N. Advances in mixup and its variants for deep learning. Neurocomputing, 2023, 555: 126635.

[12] Ge, Z., Wang, X., & Li, Z. Robust Object Detection in Adverse Weather and Industrial Environments: A Benchmark and Simulator. IEEE Robotics and Automation Letters, 2022, 7(4): 11128-11135.

[13] Liu, S., Zhang, W., Wang, L., et al. SSD-10K: A Large-Scale, Multi-Category Scrap Steel Dataset for Visual Recognition in Industrial Environments. IEEE Transactions on Industrial Informatics, 2023, 19(5): 6789-6800.

[14] Zhou, K., Yang, J., Loy, C. C., & Liu, Z. Learning to Prompt for Vision-Language Models. International Journal of Computer Vision, 2022, 130(9): 2337-2348.

[15] Wei, J., Wang, X., Schuurmans, D., et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Advances in Neural Information Processing Systems (NeurIPS), 2022, 35: 24824-24837.

[16] Zhang, Q., & Li, H. Dynamic Multi-Objective Optimization for Sustainable Production Scheduling: Models and Algorithms. Journal of Cleaner Production, 2022, 380: 135037.