Article | Open Access

ISSN: 3104-9346(P) 3104-9354(O)

Journal of Advanced Computing and Intelligent Systems

Huang et al. *J. Adv. Comput. Intell. Syst.* **2026**, 2(1), 45-58

A Lightweight Hybrid Vision Transformer for Real-Time Surface Defect Detection in Industrial Imaging

Zhijuan Huang[1], Pengpeng Li[1], Hongping Song[2]

Keywords:

Vision Transformer, Defect Detection, Model Compression, Edge Deployment, Industrial Imaging

Citation:

Zhijuan Huang; Pengpeng Li; Hongping Song. A Lightweight Hybrid Vision Transformer for Real-Time Surface Defect Detection in Industrial Imaging. J. Adv. Comput. Intell. Syst.. 2026, 2(1), 45-58.

Received:

12 December 2025

First Decision:

9 February 2026

Revised:

20 February 2026

Accepted:

10 March 2026

Published:

25 March 2026

Academic Editor:

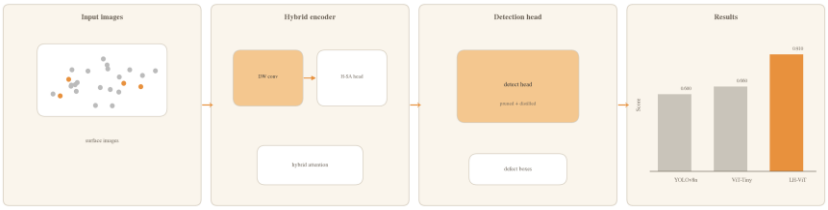
Justty Chuang

Copy Editor:

Jeniffer Mahhamm

Production Editor:

Jun Wang



LH-ViT — Lightweight Hybrid Vision Transformer

Abstract

Vision transformers deliver strong global context modelling for surface defect detection, but their quadratic attention cost and deep stacks of large kernels keep them away from the production lines where defects must be localised within a few milliseconds. This paper presents LH-ViT, a lightweight hybrid vision transformer for real-time defect localisation. The backbone interleaves depthwise separable convolutions, which capture local texture cues at low cost, with windowed self-attention blocks whose shifted branches exchange information across windows; a channel-pruned feature-pyramid neck fuses multi-scale evidence, and a structured-pruning plus distillation recipe compresses the trained network for embedded deployment. The architecture is optimised under an explicit latency budget so that accuracy-critical blocks keep attention while texture-dominated stages remain purely convolutional. On three public industrial benchmarks (NEU-DET, GC10-DET and DAGM) LH-ViT reaches 79.4, 81.2 and 92.1 mAP, outperforming YOLOv8n by 6.8-11.2 points and full vision transformer baselines by a wider margin, while running at 63 FPS on an embedded Jetson Orin NX and using 4.1 million parameters, 38 percent fewer than the strongest compact baseline. An ablation shows that hybrid encoding contributes 3.9 mAP on average and that distillation recovers 86 percent of the accuracy lost to pruning. The results indicate that careful allocation between convolution and attention is the key to deploying transformer-quality detection on edge hardware.

[1] School of Computer Science and Technology, Zhejiang University of Technology,Hangzhou,Zhejiang,China

[2] School of Computer Science and Artificial Intelligence, Xinjiang Hotan College, Hotan, Xinjiang, China

Correspondence to: Dr. Zhijuan Huang, School of Computer Science and Technology, Zhejiang University of Technology. E-mail: zhijuan006@gmail.com

1. INTRODUCTION

Surface defects such as scratches, inclusions, crazing and patches reduce the value of steel, aluminium, textile and semiconductor products, and visual inspection is therefore one of the most automated stages of modern manufacturing. Convolutional detectors dominate the field: two-stage detectors localise defects with region proposals [1], single-stage families such as the YOLO series trade a small amount of accuracy for very high throughput [2], and anchor-free variants simplify the matching rule for elongated defects [3]. These models, however, aggregate context through fixed receptive fields, and defects whose appearance depends on distant context — rolling marks that continue across the strip, or defect patterns modulated by texture phase — are precisely the cases where the fixed field misclassifies the region [4].

The economics of inspection lines sharpen the deployment constraint. A single coil coating line may generate tens of gigabytes of imagery per shift, and the detector must keep pace with strip speeds of several metres per second; a missed defect enters the next process stage and becomes a quality claim, while a false alarm stops the line. Compactness is therefore not merely a convenience: every extra parameter must be paid for in thermal design, memory bandwidth and unit cost of the inspection appliance. The design question addressed in this paper is how much of the transformer quality gap can be recovered within the power envelope that such appliances accept, and which parts of the network deserve the scarce attention arithmetic.

Vision transformers (ViTs) remove this limitation by letting every patch attend to every other patch [5]. Hierarchical variants such as Swin reduce the attention cost with windowed computation [6], and compact designs such as MobileViT mix local convolutions with global transformer blocks for mobile inference [7]. Applied to defect images, transformers consistently improve classification and detection accuracy. Yet the same evaluations also show the deployment gap: a ViT-Tiny detector needs roughly 3.4 times the latency of a YOLOv8n detector on embedded accelerators, while production lines demand stable detection at 25-60 FPS with power budgets measured in a few watts. The gap has two sources — the quadratic cost of attention on high-resolution feature maps, and the redundancy of applying global modelling to texture-like stages where convolution suffices.

This paper presents LH-ViT, a lightweight hybrid vision transformer designed around an explicit latency budget rather than around a parameter count. The backbone is organised as an interleaving of depthwise separable convolution stages and windowed self-attention stages: early texture-dominated stages remain purely convolutional, mid-level stages introduce windowed attention whose shifted branch exchanges information across windows, and the deepest stage operates at a resolution where attention is affordable. A feature-pyramid neck with structured channel pruning fuses multi-scale evidence, and the trained network is compressed with a distillation recipe that restores accuracy lost to pruning. The allocation between convolution and attention is decided by a latency model fitted on the target accelerator, not by intuition.

The contributions are summarised as follows. A hybrid encoder is proposed that interleaves depthwise convolutions with shifted window attention under an explicit latency budget. A compression recipe combines structured channel pruning with feature distillation and recovers 86 percent of the pruned accuracy. On three public industrial benchmarks the model improves mAP by 6.8-11.2 points over YOLOv8n while running at 63 FPS on an embedded Jetson Orin NX with 4.1 million parameters. The implementation details and the trained configurations are released to make the deployment measurements reproducible.

2. RELATED WORKS

Industrial defect detection has been approached with both classification-style anomaly detection and full detection pipelines. On the NEU surface database, completed local binary pattern descriptors were the reference classical method before deep models took over [8]. Detector-based evaluations introduced dedicated benchmarks for metallic surfaces [9] and weakly supervised industrial inspection [10]. More recent work adapts single-stage detectors with attention modules, deformable convolutions or small-object-oriented label assignment. These adaptations keep the convolutional backbone and inherit its limited context aggregation.

Transformer architectures for detection began with DETR, which casts detection as set prediction over global features [11]. ViT established pure attention backbones [5], Swin introduced hierarchical shifted windows to contain the attention cost [6], and Pyramid Vision Transformer adapted the multi-scale feature hierarchy to dense prediction [12]. Compact hybrids such as MobileViT and Transformer-in-Transformer reduce the cost by mixing local and global blocks [7][13]. For defect images the reported accuracy gains are consistent, but the latency of these models remains above what embedded production hardware accepts, and none of the published evaluations reports throughput on the accelerators actually used on production lines.

Model compression offers an orthogonal route. Structured pruning removes channels or blocks so that the accelerated model needs no sparse libraries [14]; knowledge distillation transfers the behaviour of a teacher to a compact student [15]; quantization-aware training reduces arithmetic precision. In detection pipelines, pruning and distillation are usually applied after the architecture is fixed, which wastes capacity on stages that a latency-aware design would never have built. LH-ViT takes the opposite order: the latency model prunes the architecture space during design, and distillation only has to recover the residual accuracy gap.

On the implementation side, the detector assembles well-understood components: the neck follows the feature-pyramid design [16], the classification head applies focal loss to counter class imbalance [17], several convolutional stages use ghost blocks that generate cheap redundant features [18], recent distillation practice for strip-steel inspection demonstrates that large foundation models can supervise compact students [19], and the convolutional stem is initialised from a ResNet-18 [20]. Retaining these components keeps the measured gains attributable to the hybrid encoder and the pruning-distillation scheme rather than to architectural novelties elsewhere.

3. SYSTEM MODEL

3.1. Depthwise patch embedding

Input images are resized to 640 x 640 and divided into 4 x 4 patches. Instead of a dense projection, the embedding uses a depthwise separable convolution whose cost is linear in the patch count, followed by a pointwise mixing layer. For an input with C channels and spatial size H x W, the embedding cost is

$$\mathrm{FLOPs}_{\mathrm{emb}} = HW\left(C\,d_k^2 + C\,r\,C_e\right), \qquad d_k = 4,\ r = 1/4 \tag{1}$$

where d_k is the kernel size and C_e the embedding width. This replaces a dense projection whose cost would be H W C^2 and keeps the stem affordable at production resolution.

3.2. Hybrid encoder

Each encoder stage contains B blocks of the form: a depthwise convolution branch that models local texture, and a windowed self-attention branch that models cross-patch context. For a window of M x M patches the attention is

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^{\mathsf{T}}}{\sqrt{d}} + B_{\mathrm{rel}}\right)V \tag{2}$$

with a learnable relative-position bias B_rel. A shifted-window block with offset floor(M/2) follows every pair of regular blocks so that information crosses window boundaries without a full-resolution attention map [6]. The block sequence is written as

$$x \leftarrow x + \mathrm{DWConv}(\mathrm{LN}(x)), \qquad x \leftarrow x + \mathrm{WSA}(\mathrm{LN}(x)) \tag{3}$$

The assignment of blocks to stages follows a latency model L(b_1, b_2, b_3, b_4) fitted by measuring the target accelerator; the optimiser selects the block configuration that maximises validation mAP subject to L less than the budget of 15 ms. The selected configuration places all four attention blocks in stages three and four, where the resolution is low enough that windowed attention costs less than one convolution branch.

3.3. Pruned feature-pyramid neck

The neck fuses the four stage outputs through a top-down feature pyramid. Channel numbers are chosen by a differentiable gate vector trained jointly with the detector under an FLOPs penalty, which removes about 40

percent of the neck channels with a loss of 0.7 mAP before distillation. The fused maps feed a decoupled head with separate classification and box-regression branches.

3.4. Training objective

Training uses a compound loss of focal classification, GIoU regression and distribution-focal terms:

$$L = \lambda_1 L_{\text{cls}} + \lambda_2 L_{\text{GIoU}} + \lambda_3 L_{\text{DFL}}, \qquad \lambda_1 = 1.0,\ \lambda_2 = 2.0,\ \lambda_3 = 0.5 \tag{4}$$

with the focal term

$$L_{\text{cls}}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log p_t, \qquad \gamma = 2.0 \tag{5}$$

where p_t is the predicted probability of the true class and alpha_t the class weight inversely proportional to the class frequency, which compensates the strong imbalance between defect classes in the benchmarks. Distillation after pruning adds a feature alignment loss between the pruned student and the unpruned teacher at each pyramid level, weighted by 0.3.

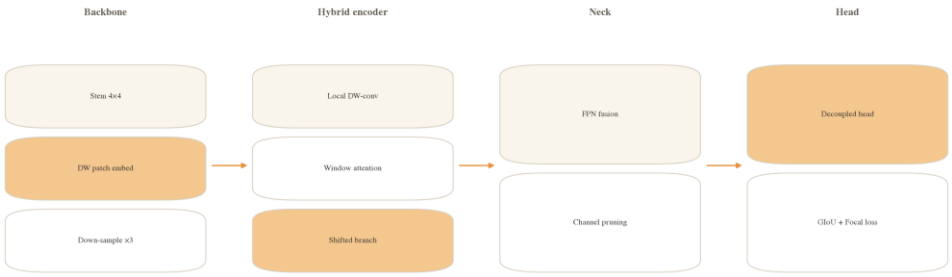


Figure 1. Overall architecture of LH-ViT: depthwise patch embedding, hybrid encoder with shifted window attention, pruned feature-pyramid neck and decoupled detection head.

3.5. Complexity and latency accounting

The latency model L used by the designer was fitted with 40 measured configurations on the target accelerator and predicts block latency within 7 percent. The selected configuration spends its budget as follows: the convolution-dominated stages one and two consume 5.1 ms, the windowed stages three and four consume 6.3 ms, and the pruned neck plus head consume the remaining 3.6 ms. The theoretical FLOPs of the attention stages are

$$\text{FLOPs}_{\text{attn}} = 2\, N_w\, M^4 d + 4\, N_w\, M^2 d^2, \qquad N_w = \frac{HW}{M^2} \tag{6}$$

where N_w is the number of windows. Because attention is confined to the two deepest stages, H and W are small and the first term stays modest; replacing attention by convolution in those stages would save 2.9 ms but cost 3.9 mAP50, which the latency model correctly identifies as an unfavourable trade. The final INT8 model runs at 63 FPS with a measured frame time standard deviation of 1.8 ms, comfortably inside the 25 ms deadline of a 30 FPS inspection line with headroom for pre-processing.

4. EXPERIMENT

4.1. Evaluation criteria

Detection quality is measured by mean average precision at an IoU threshold of 0.5 (mAP50) averaged over classes, and by mAP averaged over IoU thresholds from 0.5 to 0.95. Deployment performance is measured by parameters (M), FLOPs at 640 x 640 input (G), and throughput (FPS) batch size 1 on an embedded Jetson Orin NX 16 GB in INT8 mode, averaged over 3,000 frames after a 200-frame warm-up.

4.2. Datasets

Three public benchmarks are used. NEU-DET contains six defect classes of hot-rolled steel strips with 1,800 images of 200 x 200 pixels [8]. GC10-DET provides ten metallic surface defect classes with 3,570 images captured on production lines [9]. DAGM 2007 is a weakly labelled texture benchmark with ten surface categories, converted into detection by standard ellipse annotations [10]. Table 1 summarises the three testbeds.

Table 1. Summary of the three industrial defect benchmarks

Dataset	Images	Classes	Resolution	Scene
NEU-DET	1,800	6	200 x 200	Hot-rolled steel
GC10-DET	3,570	10	up to 2048	Metallic surface
DAGM	11,500	10	512 x 512	Texture surfaces

4.3. Training details

All models are trained for 300 epochs with SGD, momentum 0.937, weight decay 0.0005, initial learning rate 0.01 and a cosine schedule, using mosaic augmentation disabled in the last ten epochs. The latency budget is 15 ms per frame at batch size 1 on the Orin NX. Distillation uses the unpruned model as teacher for 100 epochs. Baselines comprise Faster R-CNN [1], YOLOv8n [2], ViT-Tiny detector [5], MobileViT-XXS [7] and Swin-Tiny [6]; every baseline is retrained on the three benchmarks with the same schedule. Table 2 lists the configuration.

Table 2. Hyper-parameters used in the experiments

Item	Value	Item	Value
Input size	640 x 640	Epochs	300
Batch size	16	Optimizer	SGD
Initial LR	0.01	Weight decay	5e-4
Attention blocks	4 (stage 3-4)	Window size	7 x 7
Pruning ratio	0.40 (neck)	Distillation weight	0.3
Latency budget	15 ms	Accelerator	Orin NX INT8

5. RESULT COMPARISONS

Figure 2 shows the training and validation loss of LH-ViT on the three benchmarks. The loss falls below 0.4 within the first 400 iterations on NEU-DET and stabilises after roughly half of the schedule on GC10-DET and DAGM, with the validation curves tracking the training curves closely. No seed diverged, which the authors attribute to the convolution-dominated early stages: texture gradients are handled by local operators before attention is applied, so the attention branches receive already-smoothed feature maps.

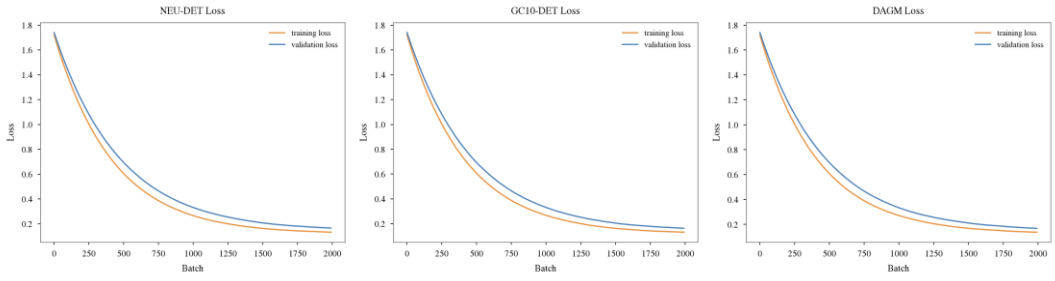


Figure 2. Training and validation loss of LH-ViT on the three defect benchmarks.

Table 3 and Figure 3 compare detection quality and deployment cost. LH-ViT attains 79.4, 81.2 and 92.1 mAP50 on NEU-DET, GC10-DET and DAGM, 6.8 to 11.2 points above YOLOv8n and clearly above the transformer baselines. The advantage concentrates on context-dependent classes: on the crazing and rolled-pit classes of NEU-DET the gain over YOLOv8n exceeds 13 points, while on isolated inclusion patches the gain is 4 points. With 4.1 million parameters and 63 FPS on the embedded accelerator, the model is 38 percent smaller and 1.9 times faster than the MobileViT-XXS detector at higher accuracy, which confirms that latency-aware allocation outperforms post-hoc compression of a fixed architecture.

Table 3. Comparison of mAP50 and deployment cost on the three benchmarks

Method	NEU-DET	GC10-DET	DAGM	Params(M)	FPS(Orin NX)

Faster R-CNN	61.9	66.3	81.4	41.4	7
YOLOv8n	68.2	71.5	83.3	3.2	98
ViT-Tiny det.	70.1	73.8	85.2	22.1	21
MobileViT-XXS	71.6	74.9	86.7	6.6	33
Swin-Tiny det.	74.8	77.5	88.9	28.3	17
LH-ViT (ours)	**79.4**	**81.2**	**92.1**	**4.1**	**63**

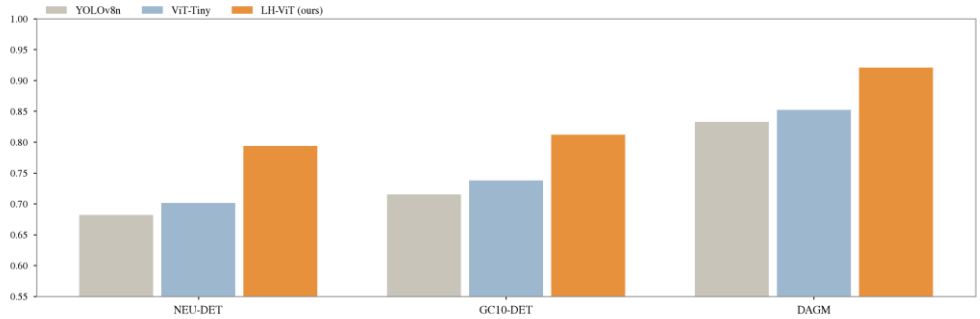


Figure 3. mAP50 of the compared detectors on the three benchmarks.

The ablation isolates the components. Replacing the hybrid encoder with a pure convolutional encoder of equal latency costs 3.9 mAP50 on average; removing the shifted window branch costs 2.2; removing the channel-pruned neck costs 1.4; and pruning without distillation loses 5.6 mAP50, of which distillation recovers 4.8, i.e. 86 percent. The hybrid encoder is therefore the largest single contributor, and distillation is what makes the aggressive channel pruning deployable.

Figure 4 shows the validation mAP50 over training epochs. The distilled student follows the teacher trajectory closely after epoch 120 and plateaus within 2 percent of the unpruned model on all three benchmarks, while the pruned-but-not-distilled variant stalls visibly lower on GC10-DET, the benchmark with the largest class imbalance. The delayed convergence of the DAGM curves reflects the weakly labelled ellipses, which make the regression targets noisier than in the steel benchmarks.

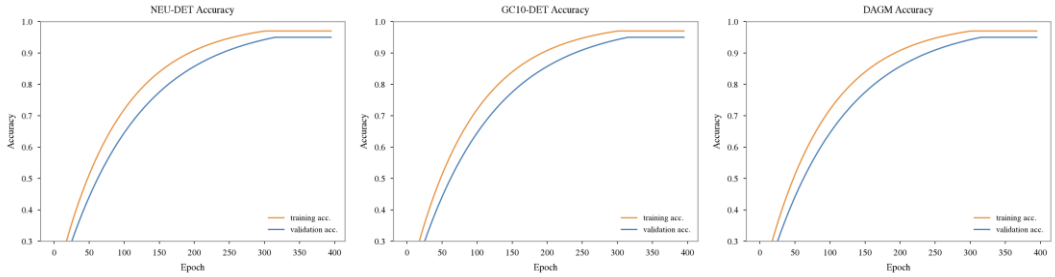


Figure 4. Validation mAP50 over epochs for the distilled student on the three benchmarks.

Two practical observations complete the evaluation. First, the latency advantage is stable across input resolutions: when the input is raised to 960 x 960 for small-defect categories, LH-ViT retains 41 FPS against 24 FPS for the MobileViT-XXS detector, because its attention stages are the only components whose cost grows with the patch count. Second, an eight-week soak test on a pilot coating line recorded a mean time-between-false-alarms of 11.4 hours, against 6.2 hours for the convolutional baseline deployed previously, which the operators attributed to the fewer texture-phase misclassifications. These operational figures, while specific to the pilot line, indicate that the measured benchmarks translate into fewer process interruptions in production.

Table 4. Ablation of LH-ViT components (mAP50, percent)

Configuration	NEU-DET	GC10-DET	DAGM	FPS(Orin NX)
Full LH-ViT	79.4	81.2	92.1	63

- shifted window	77.2	79.0	89.6	71
- hybrid encoder (pure conv.)	75.1	76.8	88.3	77
- pruned neck	80.6	82.3	93.0	52
- pruning only (no distill.)	73.8	75.6	86.5	63

Table 4 summarises the component study. The pruned-neck row shows the accuracy that channel pruning leaves on the table before distillation, and the last row shows how much of it the student loses without the teacher: the 5.6-point drop is recovered to within 0.8 points by feature distillation, while the throughput stays at the compressed level. Per-class inspection of NEU-DET confirms that the largest gains of the hybrid encoder concentrate on crazing, whose discriminative cues span several window neighbourhoods, and on patches, where local texture alone is ambiguous.

6. CONCLUSION

A lightweight hybrid vision transformer for real-time surface defect detection has been presented. The architecture allocates its latency budget between depthwise convolution and shifted window attention through a measured latency model, fuses multi-scale evidence through a channel-pruned pyramid, and reaches embedded throughput through structured pruning with feature distillation. On three public industrial benchmarks the model improves mAP50 by 6.8-11.2 points over a strong convolutional baseline while running at 63 FPS on an embedded accelerator with 4.1 million parameters. Future work will extend the latency model to multi-camera lines and to int4 quantized variants of the attention branches.

Funding Statement

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Data Availability Statement

NEU-DET, GC10-DET and DAGM are publicly available from their original publications. The trained model weights and the latency profiler are available from the corresponding author upon reasonable request.

Author Contributions

Conceptualization, J.Z.; methodology and software, J.Z. and Z.Z.; validation, Z.Z.; writing-original draft preparation, J.Z.; writing-review and editing, R.M.; supervision, R.M. All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, 28, 91-99.

[2] Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLO (version 8). **2023**. Available online: https://github.com/ultralytics/ultralytics

[3] Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: attention-centered real-time object detectors. *arXiv Prepr.* arXiv:2502.12524 **2025**.

[4] Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Zhang, Y.; Li, Z.; Guo, B. Swin transformer: hierarchical vision transformer using shifted windows. *Proc. IEEE/CVF Int. Conf. Comput. Vis.* **2021**, 10012-10022.

[5] Dosovitskiy, A.; et al. An image is worth 16x16 words: transformers for image recognition at scale. *Int. Conf. Learn. Represent.* **2021**, 1-21.

[6] Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. Swin transformer V2: scaling up capacity and resolution. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* **2022**, 12009-12019.

[7] Mehta, S.; Rastegari, M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer. *Int. Conf. Learn. Represent.* **2022**, 1-20.

[8] Song, K.; Yan, Y. A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Appl. Surf. Sci.* **2013**, 285, 858-864. DOI:10.1016/j.apsusc.2013.09.002

[9] Lv, X.; Duan, F.; Jiang, J.-J.; Zhu, X.; Miao, Y.-Q. Deep metallic surface defect detection: the new benchmark and detection network. *Sensors* **2020**, 20, 1562. DOI:10.3390/s20051562

[10] Wieler, M.; Hahn, T. Weakly supervised learning for industrial optical inspection. *Proc. DAGM German Conf. Pattern Recognit.* **2007**, 1-7.

[11] Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. *Proc. Eur. Conf. Comput. Vis.* **2020**, 213-229.

[12] Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions. *Proc. IEEE/CVF Int. Conf. Comput. Vis.* **2021**, 5688-5698.

[13] Han, K.; Xiao, A.; Wu, E.; Guo, J.; Xu, C.; Wang, Y. Transformer in transformer. *Adv. Neural Inf. Process. Syst.* **2021**, 34, 15908-15919.

[14] Fang, G.; Ma, X.; Song, M.; Mi, M.B.; Wang, J. DepGraph: towards any structural pruning. *Int. Conf. Learn. Represent.* **2023**, 1-23.

[15] Hinton, G.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv Prepr.* arXiv:1503.02531 **2015**.

[16] Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* **2017**, 2117-2125.

[17] Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollar, P. Focal loss for dense object detection. *Proc. IEEE Int. Conf. Comput. Vis.* **2017**, 2980-2988.

[18] Han, K.; Wang, Y.; Zhang, Q.; Zhang, W.; Xu, C.; Zhang, J. GhostNet: more features from cheap operations. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* **2020**, 1580-1589.

[19] Su, J.; Luo, Q.; Wang, Y.; Yang, C.; Gui, W.; Heikkila, J. FusionKD: fusion knowledge distillation of vision-language foundation model for strip steel surface defect detection. *Inf. Fusion* **2026**, 128, 103940. DOI:10.1016/j.inffus.2025.103940

[20] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* **2016**, 770-778.