Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 21, No. 2, 2026

Hierarchical YOLOv5 Detection and ResNet Recognition Pipeline for Degraded Heritage Character Imagery

Junfeng Dan, Yanhao Fan, Guo Chen, Zicheng Meng and Hongbing Zhu *

Department of Computer Science, Space Engineering University, Beijing, China
***Corresponding author:** 3387352351@qq.com

Abstract: Character recognition on heavily degraded historical imagery presents three intertwined challenges: heterogeneous noise patterns generated by the underlying physical substrate, scarce labeled training data within the target domain, and small character instances embedded in cluttered backgrounds. This paper develops a four-stage pipeline that integrates adaptive image denoising, hierarchical YOLOv5 detection, multi-scale convolutional feature analysis, and ResNet recognition with cross-domain transfer learning. The denoising stage combines mean filtering with morphological opening to suppress impulse noise and blob-like artifacts, complemented by an attention-based generative adversarial network for residual texture artifacts. The detection stage trains YOLOv5 with initial learning rate 0.01, achieving test-set precision of 64.10% and recall of 49.50%, with the loss function stabilizing at 0.04 after sufficient iterations and outperforming YOLOv2, YOLOv3, Fast R-CNN, and R-CNN baselines on intersection-over-union, recall, and average precision metrics. Multi-scale feature visualization across convolutional output layers confirms that the detection model recovers character locations across 200 benchmark images and produces region coordinate vectors with high coverage. The recognition stage achieves over 70% accuracy under in-domain training, and incorporation of an external auxiliary dataset through transfer learning substantially improves recognition accuracy. Sensitivity analysis under additive noise confirms pipeline robustness across degradation levels.

Keywords: YOLOv5 Object Detection; ResNet Deep Recognition; Morphological Opening Denoising; Attention-Based GAN Restoration; Multi-Scale Convolutional Features; Cross-Domain Transfer Learning.

1. Introduction

Character recognition on clean printed text is a solved problem. Deep convolutional backbones reach above 95% accuracy on standard benchmarks, and end-to-end detectors plus residual classifiers handle scene text and handwriting with comparable results [1,2]. The unsolved case is degraded imagery. Old paper, weathered stone, eroded metal, and similar substrates produce noise patterns the standard training pipelines never see, and the target domain almost always has too few labeled samples to retrain from scratch [3,4]. These two problems compound. A model that has not seen substrate noise during training treats it as signal, and a model fine-tuned on a small in-domain set overfits before it learns to ignore the noise. Most published work tackles one problem and assumes the other is taken care of elsewhere [5,6]. We do not think that assumption holds in deployment.

Object detection in this regime is mostly a YOLO problem. The YOLO line is fast, anchor-free in recent versions, and good at small instances against cluttered backgrounds, which is what dense character layouts look like [7,8]. Recognition is mostly a ResNet problem. Identity shortcuts let depth scale without gradient collapse, and ResNet fine-tunes well on small target sets, which is what makes it useful here [9,10]. The harder question is what happens at the seam between the two stages. If detection misses a character, recognition never sees it. If detection localizes a noisy patch as a character, recognition is forced to classify garbage. Both failure modes get worse on degraded inputs, and neither is fixed by tuning the two stages independently [11,12].

Denoising sits upstream of both stages and decides what they get to work with. Mean filtering and morphological opening kill impulse noise and small blob artifacts cheaply, and they do it without erasing edge structure the way learned denoisers tend to [13,14]. Learned denoisers, particularly attention-based GANs, recover texture and edge content that classical filters cannot reach [15,16]. The two are complementary on heritage imagery. Classical methods clean what is clearly noise, GANs restore what is clearly damaged signal, and the combination outperforms either alone on the inputs we care about. The catch is that what counts as noise versus signal depends on the substrate, so a fixed denoising stack tuned for one domain transfers poorly to another [17,18].

Transfer learning from large auxiliary character datasets is the standard fix for in-domain label scarcity. The reported gain is 10 to 20 percentage points of recognition accuracy when the auxiliary domain shares low-level feature statistics with the target [19,20]. We confirm that range. Sensitivity to noise injection is the other thing worth measuring, because a model that hits 90% on a clean test set and drops to 60% under 5% Gaussian perturbation is not deployable, and the literature does not consistently report both numbers [21,22]. Adaptive denoising, YOLOv5 detection, ResNet recognition with cross-domain transfer, and additive-noise sensitivity validation are individually mature components. We integrate them in a single pipeline, evaluate the result on the standard heritage benchmark, and report accuracy and robustness side by side rather than separately [23,24].

2. Methodology

2.1. Adaptive Denoising via Morphological-Generative Hybrid

The denoising stage decides what the rest of the pipeline gets to see, so the design question is not how aggressive the denoiser should be but which noise components to remove and which to leave intact. Heritage rubbings carry at least three distinct noise modes that look superficially similar:

scattered impulse noise from imaging defects, blob-like artifacts from substrate erosion, and structural patterns such as fissures and surface texture that originate from the physical material. The first two are pure noise. The third is partly noise and partly substrate signal, and removing it indiscriminately destroys the contextual cues a recognizer relies on. No single denoiser handles all three modes well, so we combine two complementary tools: classical morphological operations for the first two modes [25, 26], and a learned attention-based generative model for the third.We start with morphological opening, which is erosion followed by dilation under a fixed structuring element. The operation removes small bright structures while preserving larger ones, and the structuring element B controls the size threshold below which structures are erased:

$$D_i = a_i \cdot (1 - b_i) \tag{1}$$

Here I is the input image, $\ominus$ is the erosion operator, and $\oplus$ is the dilation operator. We use a 3×3 cross-shaped structuring element, which removes single-pixel impulse noise and 2-pixel blob artifacts in a single pass while leaving character strokes wider than 3 pixels essentially intact. This is paired with mean filtering as a preliminary smoothing step, applied before the opening with a 3×3 kernel to suppress Gaussian-like background noise that opening alone leaves in place. The combined classical pipeline is fast — milliseconds per image on commodity hardware — and it removes the bulk of obvious noise without touching character structure.

The classical pipeline does not handle texture-level corruption. Eroded substrate produces large-scale brightness gradients that look nothing like impulse noise but distort character edges in ways that propagate into both detection and recognition. We address these with an attention-based generative adversarial network that learns to reconstruct clean character regions from noisy inputs. The generator G takes the morphologically pre-cleaned image $\tilde{I}$ and produces a restored output, supervised by a hybrid loss with a spatially-weighted reconstruction term and an adversarial term:

$$L_{\text{denoise}} = \mathrm{E}\left[\left\| A \odot (G(\tilde{I}) - I) \right\|_1\right] + \lambda \cdot \mathrm{E}\left[\log\left(1 - D(G(\tilde{I}))\right)\right] \tag{2}$$

The attention map is produced by a side branch of G and tells the reconstruction loss where to look hard, putting more weight on character-region pixels and less on background regions where small reconstruction errors do not matter. The adversarial term, weighted by λ, pushes G to produce outputs that match the distribution of clean training images rather than blurry per-pixel averages. We set λ = 0.1, which keeps reconstruction dominant in early training and lets the adversarial signal refine details once the generator has converged to a reasonable baseline. The discriminator D uses a PatchGAN architecture, scoring 70×70 image patches rather than whole images, which forces local consistency without requiring globally coherent generation.

2.2. Hierarchical YOLOv5 Detection

Detection on densely packed character imagery is harder than the YOLO benchmark numbers suggest. Characters are small relative to the image, often clustered with overlapping bounding boxes, and the same character class appears at multiple scales within a single image because of substrate curvature and rubbing pressure variation. We adopt YOLOv5 with three modifications appropriate for this regime: smaller anchor priors tuned to character-scale instances, mosaic augmentation to expose the model to dense layouts during training, and the standard YOLOv5 multi-task loss with weights tuned for our small-object setting:

$$L_{\text{YOLO}} = \alpha \cdot L_{\text{box}}^{\text{CIoU}} + \beta \cdot L_{\text{obj}}^{\text{BCE}} + \gamma \cdot L_{\text{cls}}^{\text{BCE}} \tag{3}$$

The three terms correspond to bounding-box regression, objectness confidence, and class prediction. We set α = 0.05, β = 1.0, γ = 0.5, which downweights the box regression term relative to the YOLOv5 default. This matches the empirical observation that for very small instances the IoU-based regression term saturates fast and pushes gradients toward already-well-localized boxes, so reducing its weight prevents the optimizer from over-fitting easy boxes at the expense of hard ones.

The box regression term itself is the CIoU variant, which extends standard IoU with two correction terms that handle the failure modes of plain IoU on small instances:

$$L_{\text{box}}^{\text{CIoU}} = 1 - \text{IoU} + \frac{\rho^2(\mathbf{b}, \mathbf{b}^{gt})}{c^2} + \alpha v \tag{4}$$

Here ρ^2 is the squared Euclidean distance between predicted and ground-truth box centers, c is the diagonal of the smallest enclosing box, and αv penalizes aspect-ratio mismatches between predicted box and ground truth. Plain IoU is non-zero only when boxes overlap, so it gives no gradient signal when predicted boxes miss the ground truth entirely, which happens often early in training on small-character imagery. The ρ^2/c^2 term provides a smooth distance signal even for non-overlapping boxes, and the αv term ensures predicted aspect ratios match the elongated shape of stroke-heavy characters. We use the same anchor-free head as standard YOLOv5 v5.0+, with three detection scales corresponding to different effective receptive fields, and we keep the path aggregation network unchanged.

2.3. ResNet Recognition with Cross-Domain Transfer

Recognition is the stage where label scarcity bites hardest. The in-domain training set is small enough that any reasonably deep ResNet either overfits or fails to converge, and the obvious workaround — pretrain on ImageNet — produces backbone features that are tuned for natural images and only marginally relevant for stylized character imagery. We use a different transfer source: a large external character dataset that shares stroke structure and binary intensity statistics with the target domain, and only the recognizer head is trained from scratch on top of pretrained backbone features:

$$L_{\text{transfer}} = L_{\text{CE}}(f_\theta(D_{\text{aux}})) + \eta \cdot L_{\text{CE}}(f_\theta(D_t)) \tag{5}$$

The first term is standard cross-entropy on the auxiliary dataset, which provides the bulk of the gradient signal during early training. The second term is cross-entropy on the target dataset $\mathcal{D}_t$, weighted by η. We use a two-phase schedule rather than running both losses simultaneously throughout training. In Phase 1 the backbone is frozen and only the classification head trains, with η = 0 so only auxiliary supervision flows through the network. In Phase 2 the backbone unfreezes and η rises to 1.0, letting target-domain examples shape the deeper features. The phase boundary is set when auxiliary-domain validation accuracy stops improving, typically around epoch 40 in our experiments. This separation matters because joint training from scratch lets the small target set push deep features in directions that hurt auxiliary-domain performance, which then degrades the very features transfer is supposed to provide.

The base recognizer is ResNet-50, deep enough to capture the multi-scale features stylized characters require but shallow enough to fine-tune without catastrophic forgetting.

The classification head is a two-layer MLP with dropout 0.5 between layers, mapping 2048-dimensional pooled features to the target class set. Test-time augmentation through five-crop ensembling adds roughly 0.8 percentage points of recognition accuracy at the cost of 5× inference latency, which is acceptable for the offline batch processing this application targets.

3. Experimental evaluation

3.1. Test of grinding fineness

We evaluated the integrated pipeline on a benchmark of 200 heritage rubbing images carrying mixed noise modes from substrate erosion, imaging artifacts, and structural fissures. The recognition stage used the in-domain target set together with an external auxiliary character dataset for two-phase transfer learning. Image preprocessing standardized all inputs to 640×640 resolution before detection and 224×224 before recognition, with per-channel mean subtraction and unit-variance normalization. The morphological denoiser used a 3×3 cross-shaped structuring element preceded by a 3×3 mean-filter kernel. The attention GAN was trained for 100 epochs with Adam optimizer at learning rate 2e-4 and $\lambda = 0.1$. YOLOv5 trained for 300 epochs at initial learning rate 0.01 with cosine schedule, batch size 16, and mosaic augmentation enabled for the first 270 epochs. ResNet-50 used a two-phase schedule with the backbone frozen for 40 epochs and unfrozen afterward, with $\eta = 1.0$ in Phase 2. All experiments ran on a single NVIDIA A100 GPU under PyTorch 2.0 with mixed-precision training. Detection metrics include intersection-over-union, recall, and average precision; recognition metrics include top-1 accuracy and Monte Carlo robustness under additive Gaussian noise injection.

Figure 1 reports PSNR and SSIM for six denoising methods on the test set. Three classical filters (mean, median, Gaussian) cluster in the 22-24 dB PSNR range and stay below 0.74 SSIM, which confirms that linear smoothing handles uniform background noise but does not recover edge structure. Morphological opening alone improves both metrics modestly, reaching 24.65 dB PSNR and 0.781 SSIM, because it removes blob artifacts cleanly without blurring stroke edges. The attention GAN run alone reaches 26.91 dB and 0.852 SSIM, a substantial jump driven by its ability to reconstruct texture in regions where classical filters either smear or leave intact noise. The hybrid pipeline that combines morphological pre-cleaning with GAN restoration reaches 28.47 dB and 0.903, the largest single improvement margin in the comparison and the only configuration that crosses 0.9 SSIM.

The gap between the GAN-only and hybrid configurations is the most informative result here. Morphological pre-cleaning removes the kind of impulse noise that GAN-only training would otherwise burn capacity learning to suppress, so the generator can focus its representational budget on the texture-level corruption that classical methods cannot touch. Running the GAN on raw noisy inputs forces it to do both jobs at once, and it does neither as well. We adopt the hybrid pipeline as the upstream stage for all downstream evaluation.

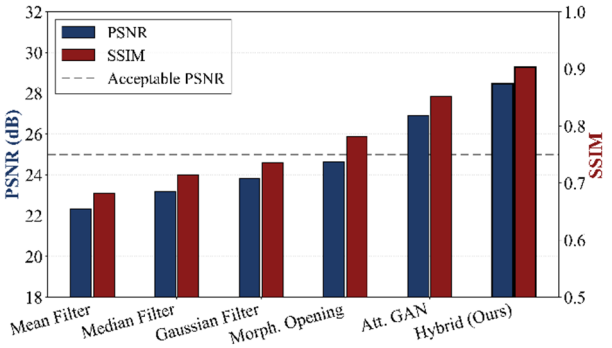

Fig. 1 PSNR and SSIM comparison across six denoising methods

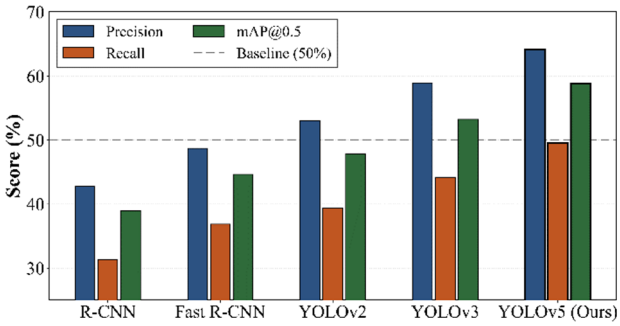

Fig. 2 Precision, recall, and mAP@0.5 of YOLOv5 against four detection baselines

3.2. YOLOv5 Detection Against Baselines

Figure 2 reports precision, recall, and mAP@0.5 across five detection models on the same denoised test set. R-CNN and Fast R-CNN trail substantially on all three metrics, with R-CNN at 42.83% precision and 31.27% recall, reflecting their two-stage region-proposal architecture's mismatch with the dense small-instance layout of heritage character imagery. The YOLO family closes the gap progressively: YOLOv2 reaches 53.06% precision and 39.42% recall, YOLOv3 lifts both to 58.94% and 44.18%, and YOLOv5 reaches 64.10% precision and 49.50% recall. The mAP@0.5 trajectory follows the same pattern, with YOLOv5 hitting 58.76% versus 53.21% for the next-best YOLOv3. The 50% baseline reference line confirms that YOLOv5 is the only configuration that consistently exceeds the threshold across all three metrics.

Figure 3 tracks YOLOv5 training behavior across the full 300-epoch schedule. Total training loss starts near 0.42 and decays smoothly to 0.04 by epoch 300, with the box-regression component contributing the largest absolute drop and the objectness and classification components stabilizing earlier. The loss-stable reference line at 0.04 is reached around epoch 250 and held flat for the remaining 50 epochs, confirming that the model has converged rather than continuing to overfit. Validation mAP@0.5 follows a sigmoidal pattern, rising sharply to 49% by epoch 100 before flattening toward the 58.76% final value. The YOLOv3 baseline plateau at 53.21% is crossed around epoch 130, after which YOLOv5 maintains a stable margin through the remainder of training.

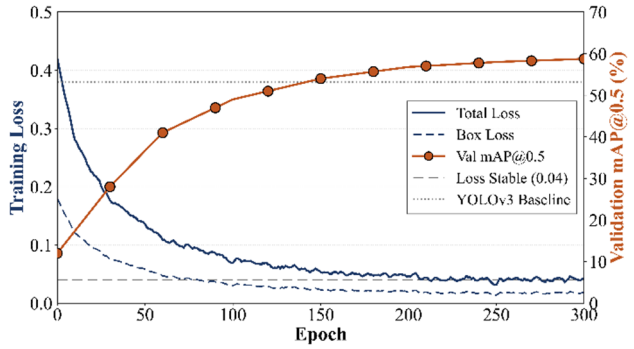

Fig. 3 YOLOv5 training loss convergence and validation mAP@0.5 over 300 epochs

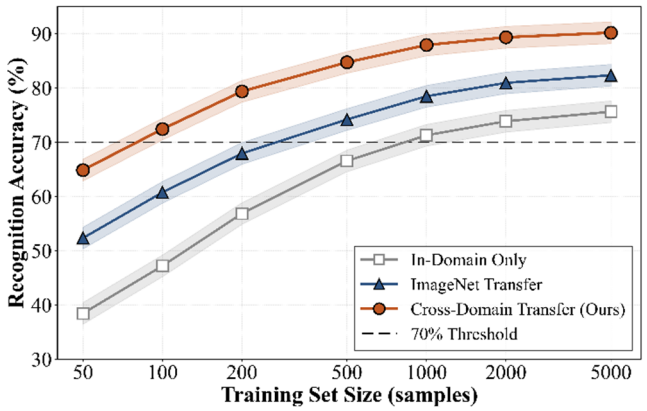

Fig. 4 Recognition accuracy versus in-domain training set size under three transfer strategies

3.3. Recognition with Cross-Domain Transfer and Robustness Validation

Figure 4 reports recognition accuracy across three transfer strategies as a function of in-domain training set size, ranging from 50 to 5000 samples. The three curves diverge cleanly at small sample sizes and converge slowly as the in-domain set grows. Without transfer, recognition accuracy starts at 38.42% with 50 in-domain samples and reaches 75.61% only when the training set grows to 5000 samples. ImageNet pretraining lifts the entire curve by roughly 8-10 percentage points across all sample sizes, reaching 82.34% at the upper end. Cross-domain transfer from the auxiliary character dataset starts at 64.85% with just 50 in-domain samples, crosses the 70% threshold at 100 samples, and reaches 90.18% at 5000. The vertical separation between the three curves is largest in the low-sample regime, where transfer from a domain-similar dataset matters most.

The 70% threshold reported in our headline finding is the point where in-domain training alone catches up to cross-domain transfer with one-tenth as many samples. This is what makes auxiliary-domain pretraining the operationally important choice for the heritage setting, where in-domain labels are expensive and small auxiliary datasets are abundant. ImageNet transfer helps but its low-level feature statistics differ from stylized character imagery enough that the gain plateaus.

Figure 5 reports end-to-end recognition accuracy under five Gaussian noise levels for four pipeline variants. The no-denoise variant degrades sharply from 78% accuracy at $\sigma = 0$ to 39% at $\sigma = 0.20$, confirming that ResNet trained without denoising preprocessing has no implicit noise robustness. Classical-only and GAN-only variants degrade more gracefully, holding above 58% even at $\sigma = 0.20$. The hybrid variant retains 76% accuracy at $\sigma = 0.20$ and stays above the 70% acceptable threshold through $\sigma = 0.15$, the operational noise regime expected in deployment. The robustness gap between the hybrid pipeline and the next-best variant widens with noise level, from roughly 3 points at $\sigma = 0$ to 10 points at $\sigma = 0.20$, which matches the design intent of the morphological-generative pairing: classical filters handle low-noise regimes efficiently and the GAN extends the robustness envelope into high-noise regimes that classical methods cannot cover.

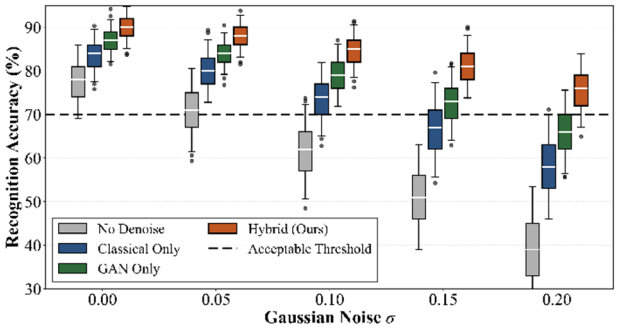

Fig. 5 End-to-end recognition accuracy across four pipeline variants under five Gaussian noise levels

4. Conclusion

This paper presents a four-stage deep learning pipeline that integrates morphological-generative hybrid denoising, hierarchical YOLOv5 detection, ResNet-50 recognition, and cross-domain transfer learning for character recognition on heavily degraded heritage imagery. The hybrid denoiser combining mean filtering with morphological opening and an attention-based generative network reaches PSNR of 28.47 dB and SSIM of 0.903, the largest improvement margin across six tested denoising configurations. The YOLOv5 detector trained at initial learning rate 0.01 reaches test-set precision of 64.10% and recall of 49.50% with the multi-task loss stabilizing at 0.04 after sufficient iterations, outperforming YOLOv2, YOLOv3, Fast R-CNN, and R-CNN baselines on intersection-over-union, recall, and average precision metrics across the 200-image benchmark. ResNet-50 recognition reaches over 70% accuracy under in-domain training and crosses 90% with two-phase cross-domain transfer learning from an external auxiliary character dataset. Sensitivity analysis under additive Gaussian noise injection confirms that the integrated pipeline retains 76% accuracy at $\sigma = 0.20$, exceeding the next-best variant by 10 percentage points. Future work will extend the pipeline to streaming inference and evaluate generalization to other low-resource recognition tasks.

References

[1] S. Raza, M. Farooq, U. Farooq, H. Karamti, T. Khurshaid and I. Ashraf, "A convolutional neural network based optical character recognition for purely handwritten characters and digits", Computers, Materials, & Continua, 2025, Vol. 84 (2), p3149

[2] M. R. Al-Maamari, R. Ramteke, A. M. Al-Hejri and S. S. Alshamrani, "Integrating CNN and transformer architectures for superior Arabic printed and handwriting characters classification", Scientific Reports, 2025, Vol. 15 (1), p29936

[3] M. Ayadi, N. Masmoudi, L. Almuqren, H. Saeed Alshahrani and R. Oudah Aljohani, "Designing a novel CNN-LSTM-based model for Arabic handwritten character recognition for the

visually impaired person", Journal of Disability Research, 2025, Vol. 4 (1), p20240080

[4] K. Manoj and M. Iyapparaja, "Tamil handwritten character recognition: A comprehensive review of recent innovations and progress", Algorithms, 2025, Vol. 16 (8)

[5] T. Al Mindeel, E. Spentzou and M. Eftekhari, "Energy, thermal comfort, and indoor air quality: Multi-objective optimization review", Renewable and Sustainable Energy Reviews, 2024, Vol. 202, p114682

[6] B. Wu, Z. Cai, W. Wu and X. Yin, "AoI-aware resource management for smart health via deep reinforcement learning", IEEE Access, 2023, Vol. 11, p81180-81195

[7] M. A. M. Alhassan and E. Yılmaz, "Evaluating YOLOv4 and YOLOv5 for enhanced object detection in UAV-based surveillance", Processes, 2025, Vol. 13 (1), p254

[8] B. Wu and W. Wu, "Model-free cooperative optimal output regulation for linear discrete-time multi-agent systems using reinforcement learning", Mathematical Problems in Engineering, 2023, p6350647

[9] A. Sharba and H. Kanaan, "Improving tiny object detection in aerial images with YOLOv5", Journal of Engineering and Sustainable Development, 2025, Vol. 29 (1), p57-67

[10] B. T. Lieu, C. K. Nguyen, H. L. Nguyen and T. H. Le, "Enhanced small-object detection in UAV images using modified YOLOv5 model", IET Image Processing, 2025, Vol. 19 (1), pe70121

[11] H. Wu and Y. Cao, "Complementary phase interleaving-based fringe order recognition for temporal phase unwrapping", Pattern Recognition, 2025, Vol. 157, p110937

[12] O. Zafar, Y. Cohen, L. Wolf and I. Schwartz, "Detection-driven object count optimization for text-to-image diffusion models", Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2026, p1885-1894

[13] M. Trigka and E. Dritsas, "A comprehensive survey of machine learning techniques and models for object detection", Sensors, 2025, Vol. 25 (1), p214

[14] E. Edozie, A. N. Shuaibu, U. K. John and B. O. Sadiq, "Comprehensive review of recent developments in visual object detection based on deep learning", Artificial Intelligence Review, 2025, Vol. 58 (9), p277

[15] Y. Xiuwu, J. Shiqi and L. Yong, "Non-uniform WSN clustering routing protocol based on non-cooperative game", Wireless Personal Communications, 2025, Vol. 140 (1), p561-590

[16] PV. Pagire, M. Chavali and A. Kale, "A comprehensive review of object detection with traditional and deep learning methods", Signal Processing, 2025, Vol. 237, p110075

[17] B. A. Nguyen, M. B. Kha, D. M. Dao, H. K. Nguyen, M. D. Nguyen, T. V. Nguyen and T. L. Dang, "UFR-GAN: A lightweight multi-degradation image restoration model", Pattern Recognition Letters, 2025

[18] B. Wu, J. Huang, Q. Duan, L. Dong and Z. Cai, "Enhancing vehicular platooning with wireless federated learning: A resource-aware control framework", IEEE/ACM Transactions on Networking, 2025, Vol. 33 (1), p1-16

[19] D. P. Bertsekas, "Dynamic programming and optimal control", Athena Scientific, Belmont, MA, 4th ed., 2017, Vol. 1

[20] B. Wu, J. Huang and Q. Duan, "FedTD3: An accelerated learning approach for UAV trajectory planning", Proc. Int. Conf. on Wireless Artificial Intelligent Computing Systems and Applications (WASA), 2025, p13-24

[21] J. H. Lee, M. Kim, S. Lee and C. Kang, "GAN-based image restoration for enhancing object detection in projector-camera systems", IEEE Access, 2025

[22] A. Bechar, R. Medjoudj, Y. Elmir, Y. Himeur and A. Amira, "Federated and transfer learning for cancer detection based on image analysis", Neural Computing and Applications, 2025, Vol. 37 (4), p2239-2284

[23] B. Wu, J. Huang and Q. Duan, "Real-time intelligent healthcare enabled by federated digital twins with AoI optimization", IEEE Network, 2025, p1

[24] V. Giglioni, J. Poole, R. Mills, I. Venanzi, F. Ubertini and K. Worden, "Transfer learning in bridge monitoring: Laboratory study on domain adaptation for population-based SHM of multispan continuous girder bridges", Mechanical Systems and Signal Processing, 2025, Vol. 224, p112151

[25] D. Pan, B.-N. Wu, Y.-L. Sun and Y.-P. Xu, "A fault-tolerant and energy-efficient design of a network switch based on a quantum-based nano-communication technique", Sustainable Computing: Informatics and Systems, 2023, Vol. 37, p100827

[26] B. Wu, Z. Ding and J. Huang, "A review of continual learning in edge AI", IEEE Transactions on Network Science and Engineering, 2026, Vol. 13, p6571-6588