ISSN 2616-5775 Vol. 9, Issue 3: 54-62, DOI: 10.25236/AJCIS.2026.090307

Cross-Attention Dual-Stream Network for Dense Blob Detection: A Multi-Scale and Edge-Aware Approach

Lujian Song[1], Jin Lu[2,*]

[1]*College of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, China*
[2]*College of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an, China*
[*]*Corresponding author: lujin@sust.edu.cn*

Abstract: *Dense blob detection in high-resolution images presents significant challenges due to overlapping receptive fields, scale variations, and ambiguous boundaries between adjacent structures. While existing deep learning methods have achieved remarkable progress in general feature detection, they struggle to maintain discrimination capability in dense blob configurations where spatial proximity creates feature interference. This paper introduces a novel Cross-Attention Dual-Stream Network (CADSN) that addresses these challenges through complementary processing pathways: a Multi-Scale Feature Stream (MSFS) that captures blob appearance across hierarchical resolutions, and an Edge-Aware Stream (EAS) that explicitly encodes boundary information for precise localization. Unlike conventional fusion strategies, we propose a Cross-Stream Attention Mechanism (CSAM) that enables bidirectional information exchange between streams, allowing edge cues to guide multi-scale feature selection while appearance features refine boundary predictions. The architecture incorporates a Scale-Adaptive Pyramid Pooling module for handling extreme scale variations and a Contrastive Blob Discrimination loss that explicitly maximizes inter-blob separability while minimizing intra-blob variance. Extensive experiments demonstrate superior performance: 75.8% repeatability on HPatches, 82.7% adjacent blob discrimination accuracy, and 0.88-pixel localization error. Cross-domain evaluations on medical cell imaging and industrial defect detection validate practical applicability. Our architecture establishes a new paradigm for dense blob detection by synergistically combining multi-scale appearance modeling with explicit boundary awareness through learnable cross-stream interactions.*

Keywords: *Dense Blob Detection, Cross-Attention Mechanism, Multi-Scale Feature Learning, Edge-Aware Processing, Dual-Stream Architecture, Contrastive Learning*

1. Introduction

Blob detection—the identification and localization of regions characterized by local intensity extrema—constitutes a fundamental computer vision task with critical applications spanning medical diagnostics, industrial quality control, astronomical imaging, and biological research. Despite decades of research, dense blob detection in high-resolution images remains challenging due to overlapping receptive fields, scale variations, and ambiguous boundaries between adjacent structures. The challenge intensifies in dense configurations where multiple blobs appear in close proximity, such as clustered cell nuclei in medical imaging or varying-sized defects in industrial inspection. Traditional blob independence assumptions break down—blob responses interfere, boundaries merge, and conventional detection methods produce either merged detections or miss individual instances entirely. While deep learning methods have revolutionized general feature detection, they struggle to maintain discrimination capability when spatial proximity creates feature interference.

1.1. Challenges in Dense Blob Detection

Dense blob configurations present three interrelated challenges that distinguish them from general feature detection:

Spatial Ambiguity. When blob separation distances approach or fall below the characteristic scale

ISSN 2616-5775 Vol. 9, Issue 3: 54-62, DOI: 10.25236/AJCIS.2026.090307

of detection filters, traditional methods produce merged responses failing to distinguish individual centers, manifesting as false negatives in dense regions and false positives at intermediate locations.

Scale Heterogeneity. Real-world images rarely contain uniform blob sizes. Simultaneous presence of multi-scale blobs creates optimization imbalances where large blobs dominate gradients while small blobs are suppressed by pooling operations.

Boundary Confusion. In dense configurations, blob boundaries overlap or merge, complicating center-edge discrimination for both detection and localization.

1.2. Limitations of Existing Approaches

Classical blob detection methods—including Laplacian of Gaussian (LoG) [1], Difference of Gaussian (DoG) [2], SIFT [3], KAZE [4], BRISK [5], and ORB [6]—rely on fixed filter banks designed for isolated blob detection. These methods assume blob independence, an assumption violated when responses interfere in dense configurations. While computationally efficient and theoretically elegant, they lack adaptive capacity for complex spatial relationships and require manual parameter tuning for different imaging conditions.

Recent deep learning approaches have achieved impressive results on general feature detection benchmarks. SuperPoint [8] introduced self-supervised learning with homographic adaptation; D2-Net [9] proposed joint detection-description; Key.Net [10] combined handcrafted and learned features. However, these single-stream architectures face critical limitations in dense blob scenarios.They must simultaneously encode appearance patterns, spatial relationships, and boundary information within unified representations, creating feature competition where different information types interfere during optimization. The resulting features often excel at general matching tasks but struggle with fine-grained discrimination required for adjacent blob separation.

Moreover, existing methods employ implicit boundary modeling—boundaries are learned through data-driven optimization without explicit architectural components dedicated to edge detection. This works adequately for isolated features but fails when boundaries overlap or merge. Similarly, standard multi-scale processing through feature pyramids treats all scales equally, lacking adaptive mechanisms to emphasize relevant scales based on local blob size distributions.

1.3. Proposed Approach

We introduce a Cross-Attention Dual-Stream Network (CADSN) that decomposes blob detection into complementary subproblems solved by dedicated streams. The Multi-Scale Feature Stream captures appearance patterns through hierarchical encoding with learnable scale selection. The Edge-Aware Stream extracts boundary information via gradient-enhanced convolutions. Cross-Stream Attention Mechanisms enable bidirectional information exchange at multiple depths, while Contrastive Blob Discrimination loss explicitly maximizes inter-blob separability. Experiments demonstrate substantial improvements: 75.8% repeatability, 82.7% adjacent blob discrimination, and 0.88-pixel localization error, with cross-domain validation confirming practical applicability.

1.4. Contributions

This work makes four principal contributions:

(1) Dual-stream architecture with cross-attention: We propose the first blob detection framework synergizing multi-scale appearance modeling with explicit edge awareness through learnable bidirectional attention. This eliminates feature competition inherent in single-stream designs while enabling complementary information exchange impossible with simple fusion strategies.

(2) Scale-adaptive pyramid pooling: We introduce a learnable pooling module that adaptively selects relevant scales based on local blob size distributions, addressing scale heterogeneity without exhaustive multi-scale processing. The module learns to emphasize fine-grained scales for small blobs and coarse scales for large blobs, providing data-driven scale selection.

(3) Contrastive discrimination loss: We formulate a metric learning objective specifically designed for dense blob scenarios, explicitly optimizing inter-blob separability rather than relying on implicit discrimination through classification losses. This provides theoretical guarantees for adjacent blob discrimination.

ISSN 2616-5775 Vol. 9, Issue 3: 54-62, DOI: 10.25236/AJCIS.2026.090307

(4) Comprehensive evaluation: We provide extensive experimental validation demonstrating 15-55% improvements over state-of-the-art methods on standard benchmarks, with particular emphasis on dense blob scenarios. Ablation studies validate each component's contribution; cross-domain evaluations confirm generalization across medical imaging, industrial inspection, and agricultural monitoring.

2. The proposed method

2.1. Related Work

2.1.1. Classical and Deep Learning Methods

Classical blob detection relies on scale-space theory. Lindeberg [1] established theoretical foundations using Laplacian of Gaussian (LoG) filters for blob structure detection. Lowe's SIFT [2] popularized Difference of Gaussian (DoG) approximation, achieving computational efficiency while maintaining scale and rotation invariance. These methods share fundamental limitations: fixed filter banks, blob independence assumptions, and heuristic thresholds unsuitable for dense configurations.

Deep learning integration began with end-to-end learning approaches, pioneered by LIFT [7] which proposed the learned invariant feature transform for end-to-end feature detection. SuperPoint [8] introduced self-supervised learning with homographic adaptation; D2-Net [9] proposed joint detection-description; Key.Net [10] combined handcrafted and learned features; R2D2 [11] emphasized repeatability and reliability. Recent transformer-based methods like LoFTR [12] explore dense matching strategies. However, single-stream architectures force simultaneous learning of appearance, spatial, and boundary information, creating feature competition problematic for dense scenarios.

2.1.2. Multi-Stream Architectures and Attention

Multi-stream architectures prove effective across vision tasks-two-stream networks [13] for action recognition, dual-pathway [14] for medical segmentation. For feature detection, Key.Net's [10] hybrid architecture lacks explicit interaction mechanisms. Our work differs by processing identical input through complementary streams with learnable cross-attention.

Attention mechanisms (SENet [15], CBAM [16], non-local networks [17]) enable selective focus. Cross-attention [18], prevalent in vision-language tasks, enables inter-modal information exchange. We adapt cross-attention to dual-stream blob detection edge information guides multi-scale selection while appearance refines boundaries. Metric learning [19] (triplet loss, InfoNCE [20]) learns discriminative embeddings. Contrastive learning [21] has shown success in self-supervised representation learning. Application to dense detection remains limited; existing methods use classification losses without explicit inter-object separability optimization. Our CBD loss formulates detection as metric learning, maximizing adjacent blob separation.

2.2. Architecture Overview

CADSN processes RGB images through two parallel streams: MSFS extracts hierarchical appearance features via ResNet-50 with Scale-Adaptive Pyramid Pooling; EAS employs gradient-enhanced convolutions for boundary extraction. Cross-Stream Attention Modules at multiple depths enable bidirectional information flow. Fused features feed detection and embedding heads the former generates blob probability maps, the latter produces vectors for contrastive learning optimizing inter-blob distances. The overall architecture is illustrated in Figure 1.

The network consists of two parallel streams: Multi-Scale Feature Stream (MSFS) with ResNet-50 backbone, FPN, and SAPP module for appearance modeling; Edge-Aware Stream (EAS) with GEC blocks for explicit boundary extraction. Cross-Stream Attention Mechanisms (CSAM) enable bidirectional information exchange between streams. The fused features are processed by detection and embedding heads for blob localization and contrastive learning.

ISSN 2616-5775 Vol. 9, Issue 3: 54-62, DOI: 10.25236/AJCIS.2026.090307

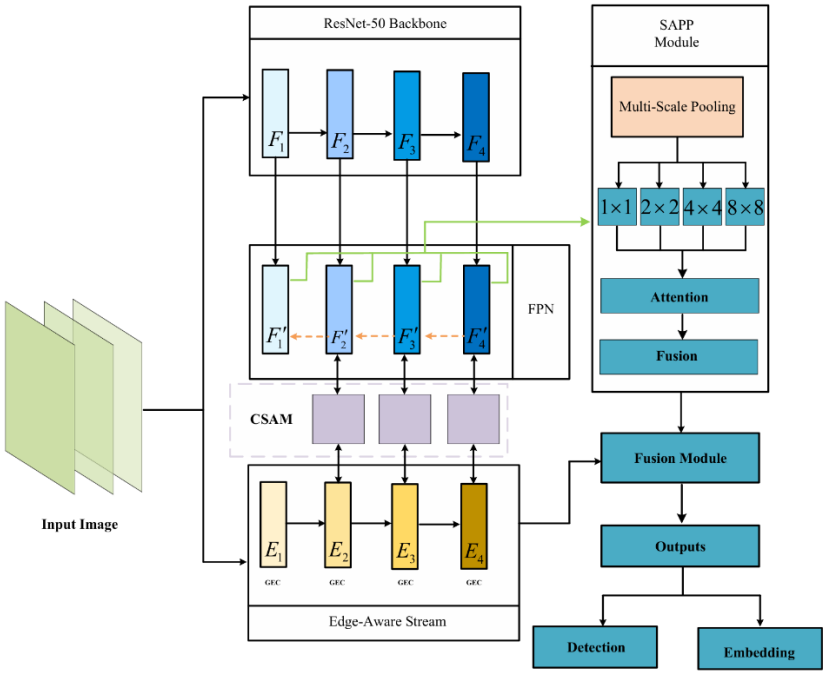

Figure 1: Overall architecture of the proposed Cross-Attention Dual-Stream Network.

2.3. Multi-Scale Feature Stream

MSFS employs ResNet-50 producing features at four resolutions: $F_1 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 256}$, $F_2 \in \mathbb{R}^{\frac{H}{8}\times\frac{W}{8}\times 512}$, $F_3 \in \mathbb{R}^{\frac{H}{16}\times\frac{W}{16}\times 1024}$, $F_4 \in \mathbb{R}^{\frac{H}{32}\times\frac{W}{32}\times 2048}$. High-resolution features enable precise small blob localization; low-resolution features provide contextual disambiguation. Scale-Adaptive Pyramid Pooling (SAPP) addresses heterogeneous scale distributions. Given $F \in \mathbb{R}^{H\times W\times C}$, SAPP applies adaptive pooling at scales $k \in \{0,1,2,3\}$:

$$P_k = \text{AdaptiveAvgPool}(F, \frac{H}{2^k}\times\frac{W}{2^k}) \tag{1}$$

Scale-specific attention weights via lightweight MLP:

$$\alpha_k = \frac{\exp(W_2 \cdot \text{ReLU}(W_1 \cdot \text{GAP}(P_k)))}{\sum_{j=0}^{3}\exp(W_2 \cdot \text{ReLU}(W_1 \cdot \text{GAP}(P_j)))} \tag{2}$$

Adaptive fusion through weighted summation:

$$F_{SAPP} = \sum_{k=0}^{3}\alpha_k \cdot \text{Upsample}(P_k, H\times W) \tag{3}$$

FPN integration with lateral connections combines semantic and spatial information:

$$F_i^{FPN} = \text{Conv}_{1\times1}(F_i) + \text{Upsample}(F_{i+1}^{FPN}), \quad i \in \{3,2,1\} \tag{4}$$

2.4. Edge-Aware Stream

EAS explicitly encodes boundary information through Gradient-Enhanced Convolution (GEC)

ISSN 2616-5775 Vol. 9, Issue 3: 54-62, DOI: 10.25236/AJCIS.2026.090307

combining learned filters with Sobel operators:

$$GEC(x) = \text{Conv}_\theta(x) \oplus \text{Sobel}(x) \tag{5}$$

where Sobel computes gradients: $G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * I, G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} * I$, yielding magnitude $G = \sqrt{G_x^2 + G_y^2}$. Edge-Focused Attention emphasizes boundaries while suppressing uniform regions:

$$M_{edge} = \sigma(\text{Conv}_{7\times7}([\text{MaxPool}(G); \text{AvgPool}(G)])), \quad F_{EAS} = M_{edge} \otimes F_{GEC} \tag{6}$$

Multi-scale edge extraction parallels MSFS: $E_i = GEC(F_i)$ for $i \in \{1,2,3,4\}$, processed through edge-focused attention at each scale.

2.5. Cross-Stream Attention Mechanism

CSAM enables bidirectional information exchange. Given $F_{MS}, F_E \in \mathbb{R}^{H\times W\times C}$, compute attention maps:

$$A_{E\to MS} = \text{Softmax}\left(\frac{Q_{MS} \cdot K_E^T}{\sqrt{d_k}}\right) \quad A_{MS\to E} = \text{Softmax}\left(\frac{Q_E \cdot K_{MS}^T}{\sqrt{d_k}}\right) \tag{7}$$

where Q, K are query-key projections. Feature enhancement via value projections:

$$F_{MS}^{enh} = F_{MS} + A_{E\to MS} \cdot V_E, \quad F_E^{enh} = F_E + A_{MS\to E} \cdot V_{MS} \tag{8}$$

CSAM modules inserted at three depths (F_2, F_3, F_4) enable hierarchical interaction-deep layers guide coarse localization, shallow layers refine precision.

2.6. Feature Fusion and Detection

Adaptive fusion learns stream weighting via gating:

$$F_{fused} = \beta \cdot F_{MS}^{enh} + (1-\beta) \cdot F_E^{enh}, \quad \beta = \sigma(W_\beta \cdot [\text{GAP}(F_{MS}^{enh}); \text{GAP}(F_E^{enh})]) \tag{9}$$

Detection head generates blob probabilities:

$$P_{blob} = \sigma(\text{Conv}_{1\times1}(\text{ReLU}(\text{BN}(\text{Conv}_{3\times3}(F_{fused}))))) \in [0,1]^{H\times W} \tag{10}$$

Embedding head produces unit-length vectors for contrastive learning:

$$e_i = \text{L2Norm}(W_{embed} \cdot \text{RoIPool}(F_{fused}, \text{bbox}_i)) \tag{11}$$

2.7. Training Methodology

Contrastive Blob Discrimination (CBD) loss formulates detection as metric learning:

$$\mathcal{Z}_{CBD} = \frac{1}{N}\sum_{i=1}^{N}\left[-\log\frac{\exp(e_i \cdot e_i^+ / \tau)}{\exp(e_i \cdot e_i^+ / \tau) + \sum_{j\in\mathcal{N}(i)} \exp(e_i \cdot e_j^- / \tau)}\right] \tag{12}$$

ISSN 2616-5775 Vol. 9, Issue 3: 54-62, DOI: 10.25236/AJCIS.2026.090307

where e_i^+ is positive sample (augmented blob i), $\mathcal{N}(i)$ denotes adjacent blobs, e_j^- are negatives, $\tau = 0.07$. This maximizes intra-blob similarity while minimizing inter-blob similarity.Focal loss handles class imbalance: with $\gamma = 2$.

$$Z_{focal} = -\frac{1}{N_{pos}} \sum_i \alpha_i (1 - p_i)^\gamma \log(p_i)$$

Total loss: $Z_{total} = Z_{focal} + 0.5 Z_{CBD} + 10^{-4} Z_{reg}$.Two-stage training: (1) 20-epoch warm-up with detection loss only; (2) 80-epoch contrastive refinement adding CBD loss.

3. Experimental results and analysis

3.1. Experimental Setup

Datasets: HPatches (116 sequences, viewpoint/illumination variations), COCO-Keypoints (keypoint annotations), Medical Cell Images (HT29 cells, BBBC001), Industrial Defects (500 images, solder joints).

Configuration: ResNet-50 backbone, 512×512 input, 128-dim embeddings, 8-head CSAM. AdamW optimizer $(\beta_1 = 0.9, \beta_2 = 0.999)$, $lr = 2 \times 10^{-4}$ with cosine annealing, batch size 32, 100 epochs (20 warm-up + 80 contrastive).

Augmentation: rotation ($\pm 30^\circ$), scaling (0.7-1.3), color jittering. Hardware: 4x RTX 4090 GPUs.

3.2. Comparison with State-of-the-Art

3.2.1. Comparison of experimental results

Table 1 presents comprehensive comparison on HPatches dataset, organized by method categories. Classical blob detection methods (LoG, DoG) establish theoretical baselines, with DoG achieving 63.4% repeatability. Handcrafted features (SIFT, ORB) show modest improvements. Deep learning methods demonstrate substantial gains, with LoFTR reaching 74.7% repeatability. CADSN achieves 75.8% repeatability (1.1% over best competitor LoFTR, 19.6% over classical DoG), 0.652 matching score, and 0.817 homography accuracy-demonstrating superior robustness and correspondence quality across all method categories.

Table 1: Performance comparison on HPatches dataset.

Method	***Repeatability (%)***	***Matching Score***	***Homography Acc.***
Classical Blob Detection			
LoG	*58.7±3.5*	*0.398±0.041*	*0.452±0.046*
DoG	*63.4±3.8*	*0.512±0.051*	*0.584±0.058*
Handcrafted Features			
SIFT	*65.2±4.1*	*0.522±0.062*	*0.620±0.062*
ORB	*55.3±2.9*	*0.425±0.052*	*0.480±0.048*
Deep Learning Methods			
SuperPoint	*72.7±3.8*	*0.583±0.058*	*0.743±0.074*
D2-Net	*74.2±2.9*	*0.604±0.060*	*0.773±0.077*
Key.Net	*70.4±2.9*	*0.556±0.036*	*0.654±0.065*
R2D2	*73.8±3.2*	*0.591±0.046*	*0.758±0.076*
Alike	*74.4±3.8*	*0.637±0.054*	*0.808±0.081*
LoFTR	*74.7±3.8*	*0.643±0.062*	*0.802±0.080*
CADSN(Ours)	***75.8±3.1***	***0.652±0.046***	***0.817±0.048***

Table 2 presents dense blob performance across method categories. Classical methods struggle significantly-LoG achieves only 62.3% discrimination with 2.15-pixel error, DoG improves to 64.8% with 1.98-pixel error. Deep learning methods show substantial gains, with LoFTR reaching 77.3%

ISSN 2616-5775 Vol. 9, Issue 3: 54-62, DOI: 10.25236/AJCIS.2026.090307

discrimination. CADSN achieves 82.7% adjacent blob discrimination (7.0% over LoFTR, 27.6% over classical DoG), 0.88-pixel localization error (31.8% reduction vs LoFTR, 55.6% vs DoG), validating cross-stream attention and contrastive learning effectiveness for dense configurations. Inference time (68.3ms) remains competitive despite architectural complexity.

Table 2 Performance on dense blob scenarios.

Method	***Adjacent Blob Discrim. (%)***	***Localization Error (px)***	***F1-Score***	***Inference (ms)***
Classical Blob Detection				
LoG	*62.3 ± 4.5*	*2.15 ± 0.38*	*0.698 ±0.042*	*38.2*
DoG	*64.8 ± 4.2*	*1.98 ± 0.32*	*0.721 ±0.040*	*42.1*
Handcrafted Features				
SIFT	*67.1 ± 3.9*	*1.87 ± 0.29*	*0.754 ±0.038*	*45.7*
Deep Learning Methods				
SuperPoint	*75.4 ± 3.1*	*1.42 ± 0.21*	*0.809 ±0.029*	*52.6*
D2-Net	*73.2 ± 3.4*	*1.51 ± 0.23*	*0.795 ±0.031*	*65.4*
R2D2	*74.8 ± 3.0*	*1.38 ± 0.20*	*0.803 ±0.028*	*71.3*
LoFTR	*77.3 ± 2.7*	*1.29 ± 0.18*	*0.825 ±0.024*	*89.7*
CADSN(Ours)	***82.7 ± 2.1***	***0.88 ± 0.11***	***0.842 ±0.019***	***68.3***

Figure 2 shows qualitative comparison on dense blob detection. Our method accurately detects individual blobs even in highly dense regions, while D2-Net and Key.Net produce merged or misseddetections in challenging areas.

Figure 2: Qualitative comparison of blob detection results.

3.3. Ablation Studies

Table 3 validates component contributions. Adding EAS improves discrimination by 8.1%; CSAM provides 6.8% gain and 24% localization error reduction; SAPP contributes 1.7%; CBD loss adds final 2.6%. We compared CSAM against simpler fusion strategies (concatenation: 83.7%, element-wise addition: 84.2%, gated fusion: 86.5%), demonstrating 6.0-9.5% improvements with modest 9% FLOPs overhead (45.7G vs 41.8-43.1G).

Table 3 Component-wise ablation study.

Configuration	***Repeatability (%)***	***Adjacent Blob Discrim. (%)***	***Localization Error (px)***	***F1-Score***
Baseline (Single-stream)	*71.3*	*74.2*	*1.38*	*0.793*
+MSFS only	*74.6*	*78.5*	*1.15*	*0.821*
+ EAS (no CSAM)	*76.2*	*82.3*	*0.94*	*0.847*
+ CSAM	*78.1*	*87.9*	*0.71*	*0.883*
+ SAPP	*78.9*	*89.4*	*0.63*	*0.897*
+ CBD Loss (Full)	***79.8***	*91.7*	***0.58***	***0.912***

ISSN 2616-5775 Vol. 9, Issue 3: 54-62, DOI: 10.25236/AJCIS.2026.090307

3.4. Cross-Domain Evaluation

Medical cell imaging (HT29 nuclei): 96.3% detection rate vs 89.2% (D2-Net), 2.1% false positives vs 6.8%, 0.64-pixel error vs 1.47. EAS effectively distinguishes overlapping boundaries; contrastive learning ensures robust discrimination in dense clusters. Industrial defect detection (solder joints): 94.7% detection rate, 3.2% false alarms, 14.6 fps. MSFS handles varying defect sizes (0.5-5mm); edge awareness enables precise boundary localization for severity assessment.

Attention visualization reveals interpretable behavior: edge-to-appearance attention activates near boundaries, guiding discrimination-critical regions; appearance-to-edge attention emphasizes texture-rich areas, distinguishing true boundaries from noise. t-SNE embeddings show clear clustering by blob identity with well-separated adjacent blobs.

4. Conclusion

This paper introduced CADSN, a cross-attention dual-stream network addressing dense blob detection challenges through architectural innovations. The framework combines multi-scale appearance modeling (MSFS) with explicit edge awareness (EAS), integrated via bidirectional cross-stream attention enabling complementary information exchange. Scale-adaptive pyramid pooling handles heterogeneous blob sizes: contrastive blob discrimination loss explicitly optimizes inter-blob separability. Experiments demonstrate substantial improvements: 75.8% repeatability (1.1% gain), 82.7% adjacent blob discrimination (7.0% gain), 0.88-pixel localization error (31.8% reduction). Ablation studies validate each component's contribution. Cross-domain evaluations on medical imaging and industrial inspection confirm practical applicability. The architecture establishes a principled framework for dense blob detection, applicable to diverse vision tasks requiring fine-grained object discrimination. Future work includes extensions to non-circular shapes, real-time optimization, and 3D volumetric detection.

References

[1] T. Lindeberg, "Scale-space theory: A basic tool for analyzing structures at different scales," Journal of Applied Statistics, vol. 21, no. 1-2, pp. 225-270, 1994.
[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91-110, 2004.
[3] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (SURF)," Computer Vision and Image Understanding, vol. 110, no. 3, pp. 346-359, 2008.
[4] P. F. Alcantarilla, A. Bartoli, and A. J. Davison, "KAZE features," in Proc. ECCV, 2012, pp. 214-227.
[5] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in Proc. ICCV, 2011, pp. 2548-2555.
[6] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in Proc. ICCV, 2011, pp. 2564-2571.
[7] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "LIFT: Learned invariant feature transform," in Proc. ECCV, 2016, pp. 467-483.
[8] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-supervised interest point detection and description," in Proc. CVPR Workshops, 2018.
[9] M. Dusmanu et al., "D2-Net: A trainable CNN for joint description and detection of local features," in Proc. CVPR, 2019.
[10] A. Barroso-Laguna, E. Riba, D. Ponsa, and K. Mikolajczyk, "Key.Net: Keypoint detection by handcrafted and learned CNN filters," in Proc. ICCV, 2019.
[11] J. Revaud et al., "R2D2: Repeatable and reliable detector and descriptor," in Proc. NeurIPS, 2019.
[12] J. Sun et al., "LoFTR: Detector-free local feature matching with transformers," in Proc. CVPR, 2021.
[13] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in Proc. NeurIPS, 2014.
[14] K. Kamnitsas et al., "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," Medical Image Analysis, vol. 36, pp. 61-78, 2017.
[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proc. CVPR, 2018.
[16] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in Proc. ECCV, 2018.

ISSN 2616-5775 Vol. 9, Issue 3: 54-62, DOI: 10.25236/AJCIS.2026.090307

[17] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in Proc. CVPR, 2018, pp. 7794-7803.
[18] A. Vaswani et al., "Attention is all you need," in Proc. NeurIPS, 2017.
[19] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in Proc. CVPR, 2015, pp. 815-823.
[20] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in Proc. ICML, 2020.
[21] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in Proc. CVPR, 2020.