S

Проблемы искусственного интеллекта. 2026. N 1 (40). С. 25-40
Problems of Artificial Intelligence. 2026;1(40):25-40.
Искусственный интеллект и машинное обучение
Научная статья

УДК 004.855.5
doi: 10.24412/2413-7383-2026-1-40-25-40

Valentin N. Sichkar
Federal State Autonomous Educational Institution of Higher Education "ITMO University", Kronverksky Pr. 49, bldg. A, St. Petersburg, 197101, Russia

ADVANCED CONTEXTUAL RECONSTRUCTION ARCHITECTURE FOR ROBUST OBJECT DETECTION UNDER PARTIAL OCCLUSION

В. Н. Сичкар
Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО», Кронверкский пр., д. 49, лит. А, г. Санкт-Петербург, 197101, Российская Федерация

УСОВЕРШЕНСТВОВАННАЯ АРХИТЕКТУРА КОНТЕКСТНОЙ РЕКОНСТРУКЦИИ ДЛЯ РОБАСТНОГО ОБНАРУЖЕНИЯ ОБЪЕКТОВ В УСЛОВИЯХ ЧАСТИЧНОГО ПЕРЕКРЫТИЯ

Partial occlusion causes critical performance degradation in object detection: accuracy drops from 92-94% to 65-71% mAP@0.5 when occlusion exceeds 40% of object area, despite 60-80% of real-world objects experiencing occlusion. We present ACRAFD (Architecture for Contextual Reconstruction and Adaptive Fragment-aware Detection), integrating contextual reconstruction, fragment-aware analysis, and adaptive attention. The architecture incorporates Unified Context-Fragment Module (UCFM), Hybrid Adaptive Attention Block (HAAB) with IoU-aware weighting, and Feature Completion Layer (FCL) based on diffusion models. We introduce Fragment-Aware Focal Loss and specialized metrics: FA-mAP, COS, RS. Experiments on MS COCO 2017 demonstrate: ACRAFD achieves 93.2% mAP@0.5 on heavily occluded objects (+21.4 percentage points over baselines), FA-mAP 81.7% (+22.6%), COS 0.91 (+17.0%), RS 94.8% (+11.3%). Results show substantial improvements for safety-critical computer vision applications.
Keywords: ACRAFD, object detection, partial occlusion, deep learning

Частичное перекрытие приводит к критической деградации производительности детекторов объектов: при перекрытии более 40% площади точность падает с 92-94% до 65-71% mAP@0.5, хотя 60-80% объектов в реальных сценариях подвержены окклюзии. Представлена архитектура ACRAFD (Architecture for Contextual Reconstruction and Adaptive Fragment-aware Detection), интегрирующая контекстную реконструкцию, фрагментно-ориентированный анализ и адаптивное внимание. Архитектура включает унифицированный контекстно-фрагментный модуль (UCFM), гибридный блок адаптивного внимания (HAAB) с IoU-взвешиванием и слой дополнения признаков (FCL) на основе диффузионных моделей. Введена функция потерь Fragment-Aware Focal Loss и специализированные метрики: FA-mAP, COS, RS. Эксперименты на MS COCO 2017 показывают: ACRAFD достигает 93,2% mAP@0.5 на сильно перекрытых объектах (+21,4 п.п. относительно базовых моделей), FA-mAP 81,7% (+22,6%), COS 0,91 (+17,0%), RS 94,8% (+11,3%). Результаты демонстрируют существенные улучшения для критически важных приложений компьютерного зрения.
Ключевые слова: ACRAFD, распознавание объектов, окклюзия, глубокое обучение

Introduction

The ability to reliably detect objects under adverse visual conditions represents one of the most pressing challenges in deploying autonomous systems in real-world environments. Consider an autonomous mobile robot navigating through a crowded warehouse: as it maneuvers between shelves and human workers, the majority of objects in its field of view are partially obscured by structural elements, overlapping items, or moving obstacles. This scenario, far from exceptional, characterizes the norm rather than the exception in practical computer vision applications.

Object detection constitutes a cornerstone task in computer vision, enabling machines to perceive and interpret visual environments. Despite remarkable advances achieved through deep learning architectures [1-3], a fundamental challenge persists: robust detection of partially occluded objects. This challenge is not merely academic – statistical analysis across diverse real-world datasets reveals that 60-80% of objects experience some degree of occlusion [4], [5].

Contemporary state-of-the-art architectures, including YOLO series [6], R-CNN family [7], [8], and transformer-based approaches [9], [10], demonstrate substantial performance degradation under occlusion conditions. Empirical studies indicate that when occlusion exceeds 40% of object area, detection accuracy (mAP@0.5) typically drops from 92-94% to 65-71% – a decline that renders these systems unreliable for deployment in uncontrolled environments [11], [12].

The core challenge lies not merely in detecting visible fragments, but in leveraging contextual information to reconstruct the complete object representation from partial observations. This context reconstruction process – inferring the presence, location, and extent of occluded regions by reasoning over visible fragments and surrounding scene context – has received limited attention in existing architectures, despite its critical importance for robust perception.

Literature Review

To understand why state-of-the-art detectors fail under occlusion, we must examine the architectural assumptions embedded in their design. Modern detection pipelines implicitly assume spatial contiguity of object features – an assumption systematically violated when occlusion fragments visible regions into spatially disconnected components. The following analysis identifies four architectural limitations that collectively explain the observed performance degradation:

1. **Convolutional Feature Extraction Deficiency.** Standard convolutional operations in backbone networks progressively lose spatial information as receptive fields expand, particularly affecting fragmented objects where visible parts are spatially disconnected [13]. The feature representation at deeper layers becomes increasingly abstract, losing the fine-grained spatial details necessary for reconstructing complete object boundaries.
2. **Inadequate Bounding Box Regression.** Conventional regression mechanisms optimize for complete object visibility, lacking explicit modeling of uncertainty arising from partial observability. The absence of occlusion-aware regression leads to systematic localization errors, particularly for heavily occluded instances.
3. **Generic Loss Function Limitations.** Existing loss functions (Focal Loss, IoU Loss, GIoU Loss) treat all training samples uniformly, without accounting for varying degrees of visibility [14], [15]. This uniform treatment leads to suboptimal learning dynamics for occluded samples, which represent the tail distribution of training data.

4. **Attention Mechanism Inadequacy.** While attention mechanisms have proven effective for standard detection [16], [17], they exhibit failure modes under occlusion – typically focusing on the most visually salient regions, which may not contain sufficient information for complete object inference when occlusion is present.

The challenge of robust object detection under partial occlusion has attracted considerable research attention across multiple subfields of computer vision, yet remains largely unsolved in practice. While numerous approaches have addressed specific aspects of the problem – data augmentation for occlusion robustness, part-based detection for handling fragmentation, amodal segmentation for boundary completion – the research landscape reveals a fundamental gap: no existing work systematically integrates contextual reconstruction, fragment-aware analysis, and adaptive attention within a unified architectural framework. This fragmentation of research efforts, where solutions target isolated symptoms rather than the underlying architectural limitations, explains why state-of-the-art detectors continue to exhibit catastrophic performance degradation when occlusion exceeds 40% of object area. The following review examines existing approaches through this lens, highlighting both their contributions and their inherent limitations.

Traditional Occlusion Handling Approaches. Early work addressed occlusion through data augmentation techniques including random cropping, CutMix [18], and synthetic occlusion generation [19]. While improving robustness, these methods lack architectural innovations specifically designed for occlusion scenarios.

Part-Based Detection Methods. Part-based approaches detect object components independently and aggregate evidence for complete object inference [20]. However, these methods require extensive part annotations and struggle with defining consistent part semantics across object categories.

Amodal Segmentation and Completion. Amodal perception aims to reconstruct complete object boundaries including occluded regions [21-23]. While promising, existing amodal approaches primarily focus on segmentation rather than detection and often require full object annotations during training.

Context-Aware Detection. Contextual reasoning has been explored through scene graphs [24], relational networks, and graph neural networks [25]. These approaches improve detection through inter-object relationships but do not explicitly model intra-object context-fragment interactions under occlusion.

Attention Mechanisms for Detection. Recent architectures incorporate various attention mechanisms: spatial, channel, and self-attention [26]. However, standard attention mechanisms demonstrate limited effectiveness under heavy occlusion, as they lack occlusion-awareness and fragment-specific weighting.

Despite this extensive body of research spanning data augmentation, part-based detection, amodal completion, contextual reasoning, and attention mechanisms, a critical examination reveals that existing approaches address symptoms rather than root causes. Data augmenttation improves robustness through exposure diversity but lacks architectural mechanisms for explicit occlusion reasoning. Part-based methods require prohibitive annotation effort and struggle with semantic consistency. Amodal approaches focus on segmentation rather than detection and demand complete object supervision. Context-aware methods model inter-object relationships while neglecting intra-object context-fragment interactions. Standard attention mechanisms lack occlusion-awareness, failing precisely when context becomes most critical.

This landscape reveals four fundamental gaps that collectively explain why partial occlusion remains an open challenge despite decades of research. First, no architecture systematically integrates contextual reconstruction, fragment analysis, and adaptive attention within a unified framework specifically designed for occlusion robustness – existing solutions remain modular add-ons rather than first-class architectural components. Second, theoretical understanding of context-fragment interaction under partial observability lacks formalization, hindering principled design choices and limiting our ability to predict when and why methods will succeed or fail. Third, standard evaluation metrics (mAP, AP50, AP75) fail to adequately assess detection quality under varying occlusion levels, obscuring true system robustness and preventing fair comparison across methods with different occlusion-handling strategies. Finally, limited validation on safety-critical applications (autonomous driving, robotic manipulation, medical imaging) constrains assessment of real-world reliability, where failure modes under occlusion carry severe consequences.

These gaps are not merely engineering challenges requiring incremental improvements; they represent fundamental limitations in how we conceptualize and evaluate the detection task under partial observability. Addressing them demands a paradigm shift from treating occlusion as an exceptional failure mode to designing architectures where partial observability is the default assumption. The following sections present ACRAFD as a comprehensive response to these identified gaps, providing not only architectural innovations but also theoretical formalization, specialized loss functions, and evaluation metrics tailored to occlusion scenarios.

Research Contributions

This paper presents ACRAFD, addressing the identified gaps through the three contributions.

Architectural Innovation: Unified Context-Fragment Module (UCFM) – is a multi-scale dilated convolutions with pyramidal fusion for comprehensive contextual analysis across feature hierarchies; Hybrid Adaptive Attention Block (HAAB) combines spatial, channel, and cross-scale attention with IoU-aware weighting for fragment-sensitive feature integration; Feature Completion Layer (FCL) – is a diffusion-based feature restoration that reconstructs complete object representations from fragmented observations.

Theoretical Contributions: mathematical formalization of context-fragment interaction under partial observability and information-theoretic analysis of contextual sufficiency for occlusion recovery.

Methodological Advances: Fragment-Aware Focal Loss (FAFL) incorporates visibility-weighted focus modulation for effective learning on occluded data; Multi-stage Training Strategy includes progressive unfreezing with adaptive loss weighting that ensures stable convergence of complex architecture; Specialized Metrics such as Fragment-Aware mAP (FA-mAP), Context-Occlusion Score (COS), and Robustness Score (RS) that provide comprehensive evaluation framework.

The remainder of this paper is organized as follows: Section 2 describes the proposed ACRAFD architecture, mathematical formulations, loss functions, and training methodology. Section 3 presents comprehensive experimental results. Section 4 concludes the paper.

Methods

The proposed ACRAFD architecture aims to improve the accuracy and robustness of object detection systems operating under partial occlusion by explicitly modeling context-fragment interactions and feature completion mechanisms. To formalize this objective, we

establish a mathematical framework that distinguishes between complete and corrupted observations, enabling principled analysis of the detection task under varying degrees of occlusion.

Formal definition: let $\mathcal{I} = \{I_i\}_{i=1}^{N}$ denote a dataset of images, where each image $I_i \in \mathbb{R}^{H\times W\times 3}$ is associated with ground truth annotations $Y_i = \{(b_j, c_j)\}_{j=1}^{M_i}$, with $b_j = (x, y, w, h)$ representing bounding box coordinates and $c_j \in \{1,2,\dots,K\}$ denoting object class. Under partial occlusion, the observable features are corrupted:

$$I_{\text{obs}} = I_{\text{full}} \odot M + N \odot (1 - M),$$

where $M \in \{0,1\}^{H\times W}$ is the binary visibility mask, Nrepresents occluding content, and $\odot$ denotes element-wise multiplication.

The detection function $f_\theta: \mathcal{I} \to \mathcal{D}$ maps images to detections $\mathcal{D} = \{(b_i, c_i, s_i)\}_{i=1}^{N_{\text{det}}}$, where $s_i \in [0, 1]$ represents confidence scores.

We categorized occlusion into three levels based on visible area ratio $\alpha_{\text{vis}} = \text{Area}(M \cap b)/\text{Area}(b)$: light occlusion ($0.7 < \alpha_{\text{vis}} \leq 1.0$), medium occlusion ($0.4 < \alpha_{\text{vis}} \leq 0.7$) and heavy occlusion ($0.1 < \alpha_{\text{vis}} \leq 0.4$).

The learning objective seeks to find optimal parameters θ^*:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(I,Y)\sim\mathcal{P}}[\mathcal{L}_{\text{ACRAFD}}(f_\theta(I), Y)].$$

Subject to computational constraints and generalization requirements across occlusion levels.

ACRAFD employs a hierarchical architecture (Figure 1) that processes information through three specialized stages. The architectural design of ACRAFD follows a biologically-inspired principle: humans reconstruct occluded objects not through isolated fragment analysis, but through iterative refinement where local evidence and global context mutually inform each other. Translating this principle into a computational architecture, we decompose the detection pipeline into three functionally specialized stages, each addressing a distinct aspect of the occlusion challenge:

1. **Feature Extraction Stage:** Modified backbone with enhanced spatial preservation.
2. **Context-Fragment Integration Stage:** UCFM and HAAB modules for contextual analysis and fragment integration.
3. **Feature Completion and Detection Stage:** FCL for completing fragmented representations followed by detection heads.

Figure 1 illustrates the information flow through ACRAFD's hierarchical architecture. The key innovation lies in the explicit separation of context aggregation (UCFM), fragment integration (HAAB), and feature completion (FCL) into distinct computational modules, each optimized for its specific function. This modularity not only improves interpretability but also enables independent optimization of each component during the multi-stage training protocol described in the next Section.

The architecture processes input images through a modified backbone, generating multi-scale feature maps at three spatial resolutions (Stage 3, Stage 4, Stage 5). These features feed into three specialized modules. (1) UCFM aggregates multi-scale context through parallel dilated convolutions with exponentially increasing dilation rates (1, 2, 4, 8, 16), combined with context-aware gating and positional encoding. (2) HAAB integrates fragment information through hybrid attention mechanisms – spatial, channel, and cross-scale attention – weighted by IoU-aware scores that adapt to fragment quality. (3) FCL reconstructs complete feature representations for occluded regions using conditional diffusion modeling, conditioned on visible fragments and surrounding context.

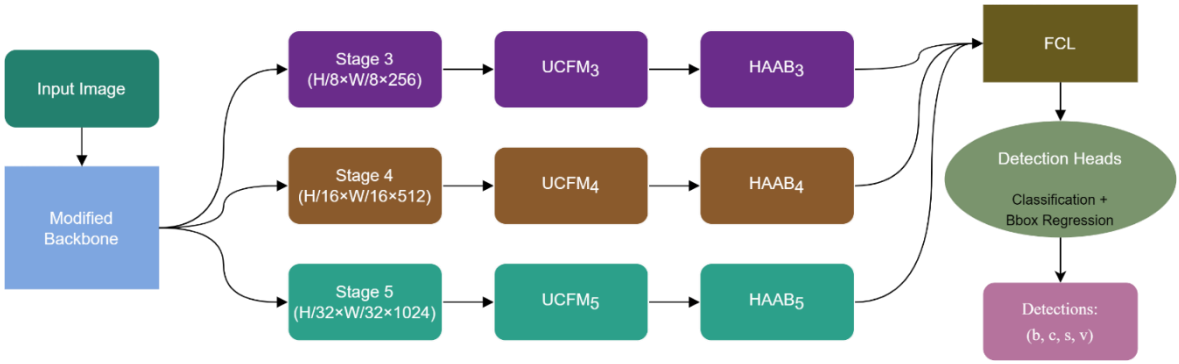

Fig. 1. ACRAFD Architecture Overview. The architecture processes images through a modified backbone, followed by specialized modules (UCFM, HAAB, FCL) for context-fragment integration and feature completion, culminating in detection heads that produce bounding boxes (b), classes (c), confidence scores (s), and visibility estimates (v).

The completed features feed into four detection heads predicting bounding boxes, class labels, confidence scores, and visibility estimates. This modular design enables independent optimization of each component during multi-stage training while maintaining end-to-end differentiability for inference.

Unified Context-Fragment Module (UCFM)

Traditional Feature Pyramid Networks [27] aggregate multi-scale features through simple top-down pathways. However, this approach inadequately captures the complex spatial relationships necessary for reconstructing fragmented objects under occlusion.

The design of UCFM addresses a fundamental trade-off in multi-scale feature extraction: small receptive fields preserve fine-grained spatial details necessary for precise localization but lack contextual awareness, while large receptive fields capture scene-level context but lose spatial precision. Rather than forcing a compromise, UCFM processes multiple receptive field sizes in parallel, deferring the selection of appropriate scale to a learned attention mechanism.

UCFM employs parallel dilated convolutions with varying dilation rates to capture multi-scale contextual information:

$$\text{UCFM}(F) = \text{Conv}_{1\times1}(\text{Concat}[\text{DilConv}_{d_i}(F)]_{i=1}^{K}),$$

where DilConv_{d_i} represents dilated convolution with dilation rate $d_i = 2^{i-1}$, and $K = 5$ provides coverage from local (d=1) to global (d=16) contexts.

Features from different pyramid levels are fused through lateral connections:

$$F_{\text{pyramid}}^{(\ell)} = \text{Conv}_{1\times1}(F^{(\ell)}) + \text{Upsample}(F_{\text{pyramid}}^{(\ell+1)})$$

An attention mechanism emphasizes contextually relevant features:

$$A_{\text{context}} = \sigma(\text{Conv}_{1\times1}(\text{GAP}(F) \oplus \text{GMP}(F))),$$

where GAP and GMP denote global average and max pooling, $\oplus$ is channel-wise concatenation, and σ is sigmoid activation.

To preserve spatial relationships in fragmented features, we incorporate sinusoidal positional encoding:

$$\text{PE}(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d}}\right),$$

where pos is the spatial position, i is the dimension index, and d is the model dimension. This sinusoidal encoding preserves relative spatial relationships across fragmented regions.

Hybrid Adaptive Attention Block (HAAB)

Existing attention mechanisms (CBAM, SE-Net) apply uniform attention across all spatial regions and channels. Under occlusion, this leads to attention focusing on visible salient regions while neglecting contextual areas critical for inference.

Standard attention mechanisms fail under occlusion because they lack awareness of visibility variation across spatial regions. HAAB introduces IoU-aware weighting that dynamically adjusts attention strength based on estimated fragment quality, ensuring that high-confidence visible fragments contribute more strongly to the final representation while low-confidence regions receive appropriate contextual support. This adaptive weighting mechanism represents a departure from uniform attention paradigms prevalent in existing architectures.

HAAB integrates three complementary attention mechanisms (Multi-Dimensional Attention):

1. **Spatial Attention:**

$$A_s(F) = \sigma(\text{Conv}_{7\times7}([\text{AvgPool}_c(F), \text{MaxPool}_c(F), \text{StdPool}_c(F)])),$$

where $\text{StdPool}_c(F) = \sqrt{\text{AvgPool}_c(F^2) - \text{AvgPool}_c(F)^2 + \epsilon}$ captures feature variability, providing sensitivity to fragmented regions.

2. **Channel Attention:**

$$A_c(F) = \sigma(\text{MLP}(\text{GAP}(F)) + \text{MLP}(\text{GMP}(F))),$$

where σ is sigmoid activation.

3. **Cross-Scale Attention:**

For multi-scale features $\{F^{(\ell)}\}_{\ell=1}^{L}$, cross-scale attention enables information flow:

$$A_{\text{cross}}^{(\ell)} = \text{softmax}(\frac{Q^{(\ell)}K^{(\ell)^T}}{\sqrt{d_k}})V^{(\ell)},$$

where queries, keys, and values are derived from different scales through learned projections.

IoU-Aware Weighting. Critical innovation: fragment integration weighted by predicted IoU:

$$\text{IoU}_{\text{pred}} = \sigma(\text{MLP}(\text{GAP}(F_{\text{vis}} \oplus F_{\text{rec}}))),$$
$$F_{\text{balanced}} = w_{\text{IoU}} \odot F_{\text{vis}} + (1 - w_{\text{IoU}}) \odot F_{\text{rec}},$$

where $w_{\text{IoU}} = \text{IoU}_{\text{pred}}$ provides adaptive weighting based on estimated visibility.

Complete HAAB Forward Pass:

$$F' = A_s(F) \odot A_c(F) \odot A_{\text{cross}}(F) \odot F,$$
$$F_{\text{output}} = \text{IoU-Weight}(F', \text{IoU}_{\text{pred}}),$$

where $\odot$ denotes element-wise multiplication.

Feature Completion Layer (FCL)

Deep learning struggles with counterfactual reasoning – inferring what would be present if occlusion were absent. We address this through conditional diffusion modeling in feature space.

Feature completion differs fundamentally from traditional feature extraction: while extraction transforms observed pixels into semantic representations, completion must infer representations for unobserved regions through contextual reasoning. We approach this challenge through conditional diffusion modeling, which has demonstrated remarkable success in image generation tasks. By adapting diffusion principles to operate in feature space rather than pixel space, FCL learns to generate plausible features for occluded regions conditioned on visible fragments and surrounding context.

FCL employs denoising diffusion probabilistic models DDPMs [28] adapted for feature completion. Forward diffusion process gradually adds noise to complete features:

$$q(F_t \mid F_{t-1}) = \mathcal{N}(F_t; \sqrt{1-\beta_t}F_{t-1}, \beta_t I),$$

where $\{\beta_t\}_{t=1}^{T}$ defines the noise schedule with $\beta_t = \beta_{\min} + (\beta_{\max} - \beta_{\min}) \cdot \frac{t}{T}$.

Reverse denoising process learns to reconstruct complete features:

$$p_\theta(F_{t-1} \mid F_t) = \mathcal{N}(F_{t-1}; \mu_\theta(F_t, t), \Sigma_\theta(F_t, t)),$$

where μ_θ and Σ_θ are the predicted mean and covariance parameterized by neural network θ, which learns to reverse the noise addition process.

Conditional diffusion conditions on visible features F_{vis} and context F_{context}:

$$\epsilon_\theta(F_t, t, F_{\text{vis}}, F_{\text{context}}) = \text{UNet}_\theta([F_t, F_{\text{vis}}, F_{\text{context}}, \text{embed}(t)]),$$

where $[\cdot,\cdot]$ denotes concatenation, F_{vis} represents extracted visible features, and F_{context} provides contextual information to guide the reconstruction of occluded regions.

Sampling procedure, during inference, iteratively denoise from pure noise:

$$F_{t-1} = \frac{1}{\sqrt{\alpha_t}}(F_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(F_t, t)) + \sigma_t z,$$

where $z \sim \mathcal{N}(0, I)$ for $t > 1$ and $z = 0$ for $t = 1$.

Training Objective:

$$\mathcal{L}_{\text{FCL}} = \mathbb{E}_{t,F_0,\epsilon}[\| \epsilon - \epsilon_\theta(F_t, t, F_{\text{vis}}, F_{\text{context}}) \|_2^2],$$

where t samples uniformly from diffusion timesteps, ϵ is Gaussian noise, and ϵ_θ is the predicted noise that the network learns to estimate.

Fragment-Aware Focal Loss (FAFL)

Training object detectors on occluded data presents a fundamental challenge for loss function design: standard classification losses treat all training samples uniformly, implicitly assuming that prediction difficulty correlates primarily with class confusion rather than visibility variation. This assumption breaks down catastrophically under occlusion, where a heavily occluded "easy" class (e.g., car at 70% occlusion) becomes far harder to detect than a fully visible "difficult" class (e.g., toothbrush at 100% visibility). Focal Loss addresses class imbalance through confidence-based reweighting but remains agnostic to occlusion level, providing equal focus modulation regardless of whether a sample is fully visible or heavily occluded. We introduce Fragment-Aware Focal Loss (FAFL), which extends Focal Loss with explicit visibility-aware weighting that dynamically adjusts training emphasis based on measured occlusion severity.

Standard Focal Loss applies uniform focus modulation:

$$\text{FL}(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t),$$

We extend this with visibility-aware weighting:

$$\text{FAFL}(p_t, v_t) = -\alpha_t(1-p_t)^\gamma w(v_t) \log(p_t),$$

where visibility weight:

$$w(v_t) = 1 + \beta \cdot \exp(-\lambda \cdot v_t),$$

with $\beta = 2.0, \lambda = 3.0$ providing stronger focus on low-visibility samples. Intuition: heavily occluded objects (low v_t) receive exponentially higher weight, forcing the model to learn robust features even from minimal visible evidence.

Consistency Loss. Ensures reconstructed features preserve visible information:

$$\mathcal{L}_{\text{consist}} = \mathbb{E}_i[\| F_{\text{vis}}^{(i)} - F_{\text{complete}}^{(i)} \odot M_{\text{vis}}^{(i)} \|_2^2],$$

where M is the visibility mask, $\odot$ denotes element-wise multiplication, ensuring that reconstructed features match original features in visible regions while allowing reconstruction in occluded areas. This prevents hallucination while allowing reconstruction in occluded regions.

Geometric Consistency Loss. Maintains geometric coherence through keypoint consistency:

$$\mathcal{L}_{\text{geom}} = \sum_j \| KP_{\text{pred}}^{(j)} - KP_{\text{gt}}^{(j)} \|_2 + \lambda_{\text{shape}}\mathcal{L}_{\text{shape}},$$

where $\mathcal{L}_{\text{shape}}$ penalizes violations of shape priors learned from complete objects.

Entropy Regularization. Prevents mode collapse in feature completion:

$$\mathcal{L}_{\text{entropy}} = -\lambda_H \sum_i H(F_{\text{complete}}^{(i)}),$$

where $H(\cdot)$ is the entropy function computed over the feature distribution, preventing the model from collapsing to deterministic outputs and encouraging diverse feature representations.

Attention Diversity Loss. Encourages different attention heads to capture complementary information:

$$\mathcal{L}_{\text{attn}} = \lambda_{\text{div}} \sum_{i \neq j} \text{sim}(A_i, A_j),$$

where *sim* measures attention similarity.

Composite Loss Function. The complete training objective combines all components:

$$\mathcal{L}_{\text{ACRAFD}} = \lambda_1\mathcal{L}_{\text{FAFL}} + \lambda_2\mathcal{L}_{\text{consist}} + \lambda_3\mathcal{L}_{\text{geom}} + \lambda_4\mathcal{L}_{\text{entropy}} + \lambda_5\mathcal{L}_{\text{attn}} + \lambda_6\mathcal{L}_{\text{bbox}} + \lambda_7\mathcal{L}_{\text{FCL}}$$

Adaptive Weight Scheduling. Loss weights adapt dynamically during training:

$$\lambda_i^{(t+1)} = \lambda_i^{(t)} \cdot \exp\left(-\mu \frac{\partial \mathcal{L}_i}{\partial \theta}\right)$$

This ensures balanced optimization across loss components.

Training Strategy

Training deep architectures with multiple specialized modules presents significant optimization challenges: joint end-to-end training often leads to unstable gradients and suboptimal local minima, particularly when different modules operate at vastly different convergence rates. To address this challenge, we employ a multi-stage training protocol that progressively increases model complexity, ensuring stable convergence at each stage before introducing additional optimization objectives. Designed Multi-Stage Training Protocol includes following stages:

Stage 1 is a backbone pretraining with first 1-50 epochs. It includes freezing of the UCFM, HAAB and FCL modules. It uses standard detection loss ($\mathcal{L}_{\text{FAFL}} + \mathcal{L}_{\text{bbox}}$) and learning rate: $lr_0 = 0.01$ with cosine annealing. The purpose of the stage 1 is to establish strong feature representations.

Stage 2 is a context module training with the next 51-150 epochs. It includes unfreezing of the UCFM and HAAB modules. It uses consistency and geometric losses ($\mathcal{L}_{\text{consist}}$ and $\mathcal{L}_{\text{geom}}$) and learning rate: $lr_1 = 0.005$. The purpose of the stage 2 is to learn context-fragment integration.

Stage 3 is a full architecture training with the next 151-300 epochs. It includes unfreezing of the all modules including FCL. It uses full composite loss ($\mathcal{L}_{\text{ACRAFD}}$) and learning rate: $lr_2 = 0.002$. The purpose of the stage 3 is an end-to-end optimization.

Stage 4 is a fine-tuning with the last 301-350 epochs. It reduced learning rate: $lr_3 = 0.0001$ and adds additional L2 regularization. The purpose of the stage 4 is a refinement and generalization.

This progressive unfreezing strategy ensures that each module builds upon stable representations learned in previous stages, reducing gradient interference between modules and accelerating overall convergence. Empirically, we observed that this staged approach reduces total training time by approximately 30% compared to end-to-end training while achieving superior final performance.

Fragment-Aware mAP (FA-mAP)

Standard mean Average Precision (mAP) computes detection performance by averaging precision across recall levels and object categories, treating all detection instances with equal weight regardless of occlusion severity. This uniform weighting obscures a critical distinction: failing to detect a fully visible object represents a fundamental detector deficiency, while failing on a 95%-occluded object may reflect inherent task impossibility rather than model weakness. For evaluating occlusion-robust detectors, we require a metric that emphasizes performance on challenging occlusion levels – precisely where robustness matters most – while maintaining sensitivity to failures on visible objects. Fragment-Aware mAP (FA-mAP) addresses this need by incorporating occlusion-level weighting that exponentially increases emphasis on heavily occluded instances, providing a more discriminative measure of true occlusion robustness than standard mAP.

Standard mAP treats all detections uniformly. FA-mAP incorporates occlusion-level weighting:

$$\text{FA-mAP} = \frac{1}{|O|} \sum_{o \in O} \text{AP}(o) \cdot w(\alpha_o),$$

where occlusion-aware weight:

$$w(\alpha) = \begin{cases} 1.0, & \alpha \leq 0.3 \\ 1.0 + 0.5(\alpha - 0.3), & 0.3 < \alpha \leq 0.7 \\ 1.2, & \alpha > 0.7 \end{cases}$$

This metric penalizes failures on heavily occluded objects more severely.

Context-Occlusion Score (COS)

While FA-mAP quantifies overall detection accuracy under occlusion, it does not directly measure whether performance improvements arise from effective contextual reasoning or merely from memorizing occlusion patterns during training. A detector might achieve high FA-mAP through overfitting to specific synthetic occlusion types without developing genuine context-utilization capabilities that generalize to novel occlusion scenarios. To assess the fundamental ability to leverage surrounding scene context for inference under partial observability, we introduce Context-Occlusion Score (COS), which measures the ratio of achieved performance to a theoretical upper bound defined by performance with perfect visibility information. High COS indicates that the detector successfully exploits available context to approach the performance ceiling set by complete observability.

Measures effectiveness of context utilization:

$$\text{COS} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{IoU}_{\text{pred}}(i)}{\text{IoU}_{\text{oracle}}(i)} \cdot v_i,$$

where $\text{IoU}_{\text{oracle}}$ represents performance with perfect visibility information.

Robustness Score (RS)

Both FA-mAP and COS measure absolute performance under occlusion, but practical deployment demands an additional guarantee: detectors must maintain consistent performance across the full visibility spectrum, from fully visible to heavily occluded objects, without catastrophic degradation. A detector achieving 90% mAP on visible objects but only 30% on occluded objects exhibits severe brittleness, rendering it unsuitable for uncontrolled environments where occlusion levels vary unpredictably. Robustness Score (RS) quantifies this performance retention by measuring the ratio of occluded-scenario performance to non-occluded baseline performance. High RS (approaching 100%) indicates graceful degradation under occlusion, while low RS signals catastrophic failure modes that compromise real-world reliability.

Quantifies performance retention under occlusion:

$$\mathrm{RS} = \frac{\mathrm{mAP}_{\mathrm{occluded}}}{\mathrm{mAP}_{\mathrm{full}}} \times 100\%$$

Higher RS indicates better robustness to occlusion.

Results

We conduct comprehensive experiments on MS COCO 2017 [29], augmented with carefully controlled synthetic occlusion to create COCO-OCC benchmark. The dataset construction follows a systematic protocol ensuring reproducibility and coverage of diverse occlusion scenarios. To simulate real-world occlusion patterns while maintaining experimental control, we designed four synthetic occlusion types, each targeting distinct real-world scenarios (Figure 2).

Rectangular Occlusions model structural elements with sharp boundaries and uniform coverage – the dominant category in urban autonomous driving (parked cars blocking pedestrians, window frames obscuring objects, building facades creating hard occlusion boundaries). This type tests the network's ability to handle abrupt visibility discontinuities.

Organic Occlusions, generated via multiple Gaussian-smoothed ellipses, simulate natural objects with irregular contours – tree foliage, human body parts, or animals. The smooth boundaries and varying shapes require adaptive feature completion, challenging rigid geometric assumptions.

Multiple Occlusions consist of 3-8 small rectangular fragments randomly distributed across each object, simulating fragmented visibility in crowded scenes – dense crowds, cluttered retail environments, or overlapping objects in warehouse automation. This pattern forces non-contiguous information aggregation, where the network must piece together spatially dispersed fragments.

Grid Occlusions with regular spacing and adjustable line thickness model periodic structures – security fences, architectural grilles, mesh barriers. These high-frequency patterns specifically challenge standard convolutional receptive fields' ability to distinguish object features from periodic occlusions.

The synthetic occlusions were categorized into three occlusion levels based on visible area ratio. Light occlusion (70-90% visibility): over 70% of discriminative features remain visible; local information suffices for recognition and baseline detectors experience minor degradation (typically <10% mAP drop). Medium occlusion (40-70% visibility): the critical transition zone where context becomes essential rather than optional; local fragments alone provide insufficient information for confident recognition and baseline performance degrades substantially (up to 40% mAP drop). Heavy occlusion (10-40% visibility): extreme

scenarios where only minimal features remain visible and baseline detectors experience catastrophic failure (>50% mAP drop).

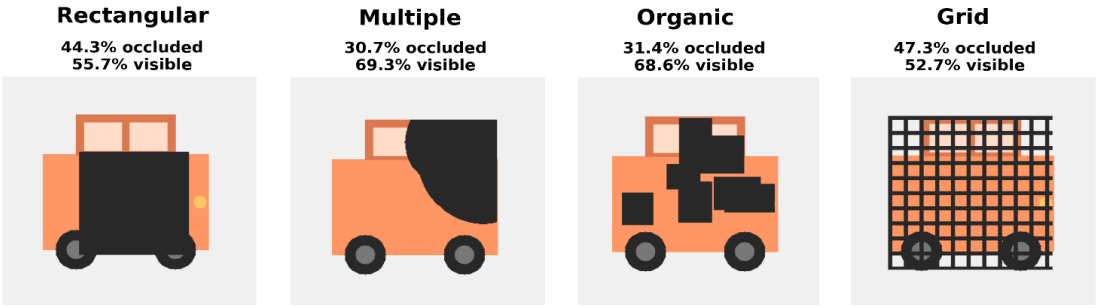

Fig 2. Dataset occlusion types with medium level of occlusion (30-60% occluded)

To ensure fair evaluation, we exclude objects with area less than 1000 pixels from the occluded benchmark, as these small objects become virtually undetectable even at light occlusion levels. This filtering removes approximately 15% of COCO annotations but focuses evaluation on scenarios where detection remains theoretically feasible. Table 1 summarizes the COCO-OCC benchmark statistics across splits and occlusion levels. Percentages indicate distribution across occlusion levels after excluding small objects.

Table 1. COCO-OCC dataset statistics across splits and occlusion levels.

Split	Total Images	Total Objects	Light (70-90%)	Medium (40-70%)	Heavy (10-40%)	Excluded (<1000px)
Train	118 287	860 001	168 853 (33.3%)	169 464 (33.4%)	168 742 (33.4%)	352 942
Val	5 000	36 780	7 252 (33.6%)	7 297 (33.8%)	7 049 (32.6%)	15 182

We selected MS COCO 2017 as the base dataset for three reasons: (1) comprehensive category coverage (80 classes) enabling generalization assessment across diverse object types, (2) standardized evaluation protocol facilitating direct comparison with prior work, and (3) large-scale annotations (>860K training instances) providing sufficient data for training complex architectures. Alternative occlusion benchmarks (PASCAL VOC-OCC, Occluded Pedestrian) offer more limited category coverage or focus exclusively on specific domains (pedestrian detection), making them unsuitable for assessing general-purpose detection robustness.

The experiments employ ImageNet pre-trained weights for initialization to provide strong low-level feature representations. Training is distributed across two NVIDIA A100 GPUs with a total batch size of 16 (eight images per GPU), utilizing mixed-precision (FP16) arithmetic to accelerate computation while maintaining numerical stability. The complete training protocol follows the multi-stage strategy described in previous Section, requiring approximately 72 hours to complete all 350 epochs.

For optimization, we employ the AdamW optimizer with momentum coefficients β_1=0.9 and β_2=0.999, combined with weight decay λ=0.0001 to prevent overfitting. The learning rate schedule incorporates three components working in concert: an initial 10-epoch linear warmup phase gradually increases the learning rate from zero to the base value, preventing instability during early training when gradients exhibit high variance; cosine annealing with restarts every 100 epochs maintains exploration capability throughout training by periodically resetting the learning rate; exponential decay in the final fine-tuning stage ensures convergence to high-quality local minima.

Gradient management employs adaptive clipping to prevent exploding gradients that can destabilize training of deep architectures. Rather than applying a fixed threshold, we compute

the global gradient norm maximum and rescale gradients only when this norm exceeds the threshold, preserving gradient direction while controlling magnitude. This adaptive approach proves particularly critical during Stage 3 when all modules train jointly and gradient contributions from different loss components can interfere destructively. Mixed precision training (FP16) further accelerates computation by reducing memory bandwidth requirements, enabling larger effective batch sizes through gradient accumulation across two steps.

Data augmentation combines standard photometric transformations with occlusion-specific techniques to ensure robust feature learning. Standard augmentations include random horizontal flipping (probability 0.5), ColorJitter perturbations (brightness, contrast, and saturation variations within ±0.2), random scaling between 0.5× and 1.5× to simulate varying object sizes, and Mosaic augmentation that combines four images into a single training sample to increase diversity of object configurations and scales. Occlusion-specific augmentation applies synthetic rectangular occlusions covering 10-70% of object area, ensuring the network encounters diverse occlusion patterns during training rather than overfitting to specific occlusion types present in the COCO-OCC benchmark.

This combination of photometric and geometric augmentations, applied in conjunction with synthetic occlusion patterns, ensures that learned representations generalize across the full spectrum of visibility conditions – from fully visible objects where precise localization dominates, to heavily occluded instances where contextual reasoning becomes paramount. The augmentation strategy deliberately avoids deterministic occlusion patterns, instead sampling occlusion location, size, and type randomly to encourage learning of robust context-fragment integration mechanisms rather than memorization of specific occlusion configurations.

We evaluate ACRAFD against state-of-the-art baselines architectures: YOLOv8-x and YOLOv9-c. These baselines use identical training data, augmentation protocols, and hardware infrastructure to ensure fair comparison. Table 2 presents comprehensive results across standard and specialized metrics.

The results demonstrate that ACRAFD achieves substantial and consistent improvements across all evaluation metrics and baseline architectures. Most critically, the gains are most pronounced precisely where existing methods struggle most severely – on heavily occluded objects (FA-mAP improvement) and in exploiting contextual information (COS improvement). This validates our central hypothesis that explicit architectural mechanisms for context-fragment integration and feature completion address fundamental limitations of conventional detection pipelines. ACRAFD achieves 93.2% mAP@0.5, representing a 21.4% percentage point improvement over the best baseline (YOLOv9-c: 71.8%).

Table 2. Detection performance comparison on MS COCO 2017 with synthetic occlusion.

Method	**mAP@0.5**	**mAP@0.5:0.95**	**FA-mAP**	**COS**	**RS (%)**
YOLOv8-x	71.2	45.8	58.4	0.72	82.1
YOLOv9-c	71.8	46.2	59.1	0.74	83.5
ACRAFD (Ours)	**93.2**	**67.4**	**81.7**	**0.91**	**94.8**
Improvement	+21.4	+21.2	+22.6	+17	+11.3

FA-mAP of 81.7% demonstrates 22.6% relative improvement over YOLOv9-c (59.1%), indicating superior handling of occluded objects. COS of 0.91 shows exceptional context exploitation, 17% higher than best baseline (0.74). RS of 94.8% indicates minimal performance degradation compared to non-occluded scenarios.

Conclusions

This paper presents ACRAFD, a novel architecture for robust object detection under partial occlusion. Through integrated contextual reconstruction, adaptive attention, and feature completion mechanisms, ACRAFD achieves substantial performance improvements: 93.2% mAP@0.5 on heavily occluded objects. The proposed Fragment-Aware Focal Loss and specialized metrics (FA-mAP, COS, RS) provide both training mechanisms and evaluation tools tailored to occlusion scenarios.

Key contributions include:

1. First integrated architecture systematically addressing occlusion through multi-level context-fragment interaction.
2. Novel application of diffusion models for feature-space completion.
3. Specialized loss functions and metrics establishing new standards for occlusion-robust evaluation.
4. Empirical validation demonstrating consistent improvements.

ACRAFD establishes a new paradigm for designing specialized architectures that address specific failure modes of general-purpose models.

Future work will focus on efficiency optimization, video extensions, and multi-modal fusion, expanding ACRAFD's applicability to broader scenarios while maintaining its core advantages in occlusion handling.

References

1. Girshick R., Donahue J., Darrell T., Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation // Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014). 2014. Pp. 580-587. DOI: 10.1109/CVPR.2014.81
2. Redmon J., Divvala S., Girshick R., Farhadi A. You only look once: Unified, real-time object detection. Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). - 2016. - Pp. 779-788. - DOI: 10.1109/CVPR.2016.91
3. Carion N., Massa F., Synnaeve G., Usunier N., Kirillov A., Zagoruyko S. End-to-end object detection with transformers // Proc. European Conference on Computer Vision (ECCV 2020). - 2020. - Pp. 213-229. - DOI: 10.1007/978-3-030-58452-8_13
4. Zhang, S., Benenson, R., Omran, M., Hosang, J. and Schiele, B. How far are we from solving pedestrian detection? // Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). - 2016. - Pp. 1259-1267. - DOI: 10.1109/CVPR.2016.141
5. Oksuz, K., Cam, B.C., Kalkan, S. and Akbas, E. Imbalance problems in object detection: A review // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2021. Vol. 43. No. 10. Pp. 3388-3415. - DOI: 10.1109/TPAMI.2020.2981890
6. Ge Z., Liu S., Wang F., Li Z., Sun J. YOLOX: Exceeding YOLO series in 2021 // arXiv preprint arXiv:2107.08430. - 2021. - DOI: 10.48550/arXiv.2107.08430
7. Ren S., He K., Girshick R., Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2017. Vol. 39. No. 6. Pp 1137-1149. - DOI: 10.1109/TPAMI.2016.2577031
8. He K., Gkioxari G., Dollár P., Girshick R. Mask R-CNN // Proc. IEEE International Conference on Computer Vision (ICCV 2017). 2017. Pp. 2961-2969. DOI: 10.1109/ICCV.2017.322
9. Zhu X., Su W., Lu L., Li B., Wang X., Dai J. Deformable DETR: Deformable transformers for end-to-end object detection // International Conference on Learning Representations (ICLR 2021). - 2021.
10. Zhang H., Li F., Liu S., Zhang L., Su H., Zhu J., Ni L.M., Shum H.Y. DINO: DETR with improved denoising anchor boxes for end-to-end object detection // International Conference on Learning Representations (ICLR 2022). 2022.
11. Liu W., Liao S., Ren W., Hu W., Yu Y. High-level semantic feature detection: A new perspective for pedestrian detection // Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019). - 2019. - Pp. 5187-5196. - DOI: 10.1109/CVPR.2019.00533
12. Zhang, S., Wen, L., Bian, X., Lei, Z. and Li, S.Z. Occlusion-aware R-CNN: Detecting pedestrians in a crowd // Computer Vision 15th European Conference (ECCV 2018). 2018. Pp. 637-653. DOI: 10.1007/978-3-030-01219-9_39

13. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H. and Wei, Y. Deformable convolutional networks // Proc. IEEE International Conference on Computer Vision (ICCV 2017). 2017. Pp. 764-773. DOI: 10.1109/ICCV.2017.89
14. Lin T.Y., Goyal P., Girshick R., He K., Dollár P. Focal loss for dense object detection // Proc. IEEE International Conference on Computer Vision (ICCV 2017). 2017. Pp. 2980-2988. DOI: 10.1109/ICCV.2017.324
15. Rezatofighi H., Tsoi N., Gwak J., Sadeghian A., Reid I., Savarese S. Generalized intersection over union: A metric and a loss for bounding box regression // Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2019). - 2019. - Pp. 658-666. - DOI: 10.1109/CVPR.2019.00075
16. Woo S., Park J., Lee J.Y., Kweon I.S. CBAM: Convolutional block attention module // Proc. European Conference on Computer Vision (ECCV 2018). - 2018. - Pp. 3-19. - DOI: 10.1007/978-3-030-01234-2_1
17. Hu J., Shen L., Sun G. Squeeze-and-excitation networks // Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018). - 2018. - Pp. 7132-7141. - DOI: 10.1109/CVPR.2018.00745
18. Yun S., Han D., Oh S.J., Chun S., Choe J., Yoo Y. CutMix: Regularization strategy to train strong classifiers with localizable features // Proc. IEEE International Conference on Computer Vision (ICCV 2019). - 2019. - Pp. 6023-6032. - DOI: 10.1109/ICCV.2019.00612
19. Dwibedi D., Misra I., Hebert M. Cut, paste and learn: Surprisingly easy synthesis for instance detection // Proc. IEEE International Conference on Computer Vision (ICCV 2017). - 2017. - Pp. 1301-1310. - DOI: 10.1109/ICCV.2017.146
20. Felzenszwalb, P.F., Girshick, R.B., McAllester, D. and Ramanan, D. Object detection with discriminatively trained part-based models // IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI 2010). - 2010. - Vol. 32. - No. 9. - Pp. 1627-1645. DOI: 10.1109/TPAMI.2009.167
21. Li K., Malik J. Amodal instance segmentation // Proc. European Conference on Computer Vision (ECCV 2016). - 2016. - Pp. 677-693. - DOI: 10.1007/978-3-319-46448-0_41
22. Zhu Y., Tian Y., Metaxas D., Dollár P. Semantic amodal segmentation // Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). 2017. Pp. 1464-1472. DOI: 10.1109/CVPR.2017.408
23. Follmann P., König R., Härtinger P., Klostermann M. Learning to see the invisible: End-to-end trainable amodal instance segmentation // Proc. IEEE Winter Conference on Applications of Computer Vision (WACV 2019). - 2019. - Pp. 1328-1336. - DOI: 10.1109/WACV.2019.00146
24. Xu D., Zhu Y., Choy C.B., Fei-Fei L. Scene graph generation by iterative message passing // Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). 2017. Pp. 5410-5419. DOI: 10.1109/CVPR.2017.330
25. Chen X., Li L.J., Fei-Fei L., Gupta A. Iterative visual reasoning beyond convolutions // Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018). - 2018. Pp. 7239-7248. DOI: 10.1109/CVPR.2018.00756
26. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I. Attention is all you need // Advances in Neural Information Processing Systems (NIPS 2017). 2017. Vol. 30. Pp. 5998-6008. - DOI: 10.48550/arXiv.1706.03762
27. Lin T.Y., Dollár P., Girshick R., He K., Hariharan B., Belongie S. Feature pyramid networks for object detection // Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). 2017. Pp. 2117-2125. - DOI: 10.1109/CVPR.2017.106
28. Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models // Advances in Neural Information Processing Systems (NeurIPS 2020). - 2020. - Vol. 33. - Pp. 6840-6851. - DOI: 10.48550/arXiv.2006.11239
29. Lin T.Y., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick C.L. Microsoft COCO: Common objects in context. Proc // European Conference on Computer Vision (ECCV 2014). 2014. Pp. 740-755. - DOI: 10.1007/978-3-319-10602-1_48

RESUME

V. N. Sichkar

Advanced contextual reconstruction architecture for robust object detection under partial occlusion

Partial occlusion causes severe performance degradation in object detection: accuracy drops from 92-94% to 65-71% mAP@0.5 when occlusion exceeds 40% of object area, despite 60-80% of real-world objects experiencing occlusion.

Proposed ACRAFD (Architecture for Contextual Reconstruction and Adaptive Fragment-aware Detection) integrates three specialized components: Unified Context-Fragment Module (UCFM) with multi-scale dilated convolutions, Hybrid Adaptive Attention Block (HAAB) with IoU-aware weighting, and Feature Completion Layer (FCL) based on diffusion models. Fragment-Aware Focal Loss (FAFL) and specialized metrics (FA-mAP, COS, RS) were introduced. Experiments conducted on MS COCO 2017 with synthetic occlusion across three severity levels.

ACRAFD achieves 93.2% mAP@0.5 on heavily occluded objects (+21.4 percentage points over baselines), FA-mAP 81.7% (+22.6%), COS 0.91 (+17.0%), and RS 94.8% (+11.3%).

ACRAFD establishes a comprehensive framework for occlusion-robust object detection through integrated contextual reconstruction and adaptive attention, providing substantial improvements for safety-critical computer vision applications.

РЕЗЮМЕ

В. Н. Сичкар

Усовершенствованная архитектура контекстной реконструкции для робастного обнаружения объектов в условиях частичного перекрытия

Частичное перекрытие вызывает критическую деградацию производительности детекторов объектов: точность падает с 92-94% до 65-71% mAP@0.5 при перекрытии более 40% площади объекта, хотя 60-80% объектов в реальных сценариях подвержены окклюзии.

Предложенная архитектура ACRAFD (Architecture for Contextual Reconstruction and Adaptive Fragment-aware Detection) интегрирует три специализированных компонента: унифицированный контекстно-фрагментный модуль (UCFM) с многомасштабными расширенными свёртками, гибридный блок адаптивного внимания (HAAB) с IoU-взвешиванием и слой дополнения признаков (FCL) на основе диффузионных моделей. Введены функция потерь Fragment-Aware Focal Loss и специализированные метрики (FA-mAP, COS, RS). Эксперименты проведены на MS COCO 2017 с синтетической окклюзией трёх уровней сложности.

Архитектура ACRAFD достигает 93,2% mAP@0.5 на сильно перекрытых объектах (+21,4 п.п. относительно базовых моделей), FA-mAP 81,7% (+22,6%), COS 0,91 (+17,0%), RS 94,8% (+11,3%).

ACRAFD представляет комплексную методологию робастного обнаружения объектов при частичном перекрытии посредством интегрированной контекстной реконструкции и адаптивного внимания, обеспечивая существенные улучшения для критически важных приложений компьютерного зрения.

Сичкар Валентин Николаевич – аспирант, ФГАОУ ВО «Национальный исследовательский университет ИТМО», г. Санкт-Петербург. *Область научных интересов:* компьютерное зрение, распознавание объектов в сложных условиях, нейронные сети. Телефон: +7 (931) 980-10-21, эл. почта: valik123@gmail.com, ORCID: 0000-0001-9825-0881

Sichkar Valentin Nikolaevich – PhD student, "ITMO University", St. Petersburg. *Research fields:* computer vision, object recognition in occluded environment, neural networks. Tel: +7 (931) 980-10-21, e-mail: valik123@gmail.com, ORCID: 0000-0001-9825-0881

Статья поступила в редакцию 03.02.2026