

Toward Amodal Instance Segmentation of Overlapping Concrete Aggregates via Occlusion-Conditioned Residual Correction

I Gede Teguh Satya Dharma¹, I Nyoman Gede Arya Astawa^{2,*}, I Nyoman Suardika³,
Made Pradnyana Ambara⁴

^{1,2,4}*Department of Information Technology, Politeknik Negeri Bali, Badung, 80364, Indonesia*

³*Department of Civil Engineering, Politeknik Negeri Bali, Badung, 80364, Indonesia*

(Received: April 15, 2026; Revised: June 15, 2026; Accepted: September 18, 2026; Available online: October 14, 2026)

Abstract

Concrete aggregate characterization is important for quality control, particularly in estimating particle size and shape distributions. Although computer vision provides a faster alternative to conventional sieve analysis, overlapping particles can hide parts of their boundaries, causing visible-only segmentation to underestimate the complete particle geometry. This study proposes an occlusion-conditioned residual correction mechanism for amodal instance segmentation of overlapping concrete aggregates. Rather than predicting the complete mask from scratch, the method preserves the observed visible region and estimates only the missing geometry along its boundary, using the visible mask as an explicit physical constraint. Local boundary information is incorporated through a ring context to guide residual prediction. The proposed mechanism is implemented within a standard Mask R-CNN framework and evaluated on a synthetic aggregate dataset with controlled occlusion levels. Compared with the corresponding direct amodal baseline, the proposed method improves AP from 48.02% to 52.69%, while ablation results show an additional gain from the ring context. Qualitative results further indicate more faithful reconstruction of irregular contours, narrow gaps, and sharp corners with less excessive mask expansion. These findings support constrained residual reconstruction as an effective mechanism for amodal aggregate segmentation under occlusion.

Keywords: Concrete Aggregates, Amodal Instance Segmentation, Occlusion Handling, Residual Correction, Mask R-CNN, Computer Vision

1. Introduction

Concrete aggregates are granular materials, such as sand, gravel, and crushed stone that constitute most of the concrete volume and are essential for the structural integrity of concrete buildings [1], [2]. Aggregate variations in size and shape, also known as aggregate gradation, can significantly impact the workability, strength and durability of concrete structures, thereby controlling the aggregate properties is a crucial aspect in concrete quality control [3], [4], [5]. While traditional sieving remains the dominant approach for aggregate gradation analysis, its time-consuming and labour-intensive nature [6] highlights the need for non-destructive and automated alternatives, particularly computer vision-based techniques. Image-based analysis provides an automated way to estimate grain morphology and volume distribution from digital captures. Consequently, it offers a practical solution to accelerate routine testing in highway and materials engineering.

Building upon these advantages, numerous studies have investigated the application of computer vision techniques for 2D aggregate characterization. Early studies primarily relies traditional image processing such as watershed segmentation and edge detection, to extract particle boundaries and estimate size distributions [7], [8]. While computationally efficient, these methods often require extensive manual parameter tuning. Subsequent works utilized deep learning-based approaches, particularly convolutional neural networks combined with instance segmentation frameworks to learn discriminative features automatically [9], [10], [11], [12]. Although proven to perform well in various scenarios, these models are only able to segment the visible parts of aggregate particles. They are neither designed nor tested in occluded environments typical of real-world conditions, making the predicted aggregate particle

*Corresponding author: I Nyoman Gede Arya Astawa (arya_kmg@pnb.ac.id)

 DOI: <https://doi.org/10.47738/jads.v7i4.1569>

This is an open access article under the CC-BY license (<https://creativecommons.org/licenses/by/4.0/>).

© Authors retain all copyrights

size distribution (PSD) inherently misleading. While segmenting occluded objects (amodal segmentation) has been extensively studied in other domains [13], [14], [15], [16], their implementation in aggregate particles are still largely under-explored. Study by [17] attempted to resolve occlusion by integrating contextual feature modules into the network architecture, whereas the study by [18] leveraged a Diffusion-GAN framework to jointly synthesize data and reconstruct occluded geometry.

However, these prior works fundamentally rely on full mask regeneration, predicting the entire amodal mask from scratch. While effective for general shape reconstruction, this from-scratch approach runs the risks of neglecting the definitive pixel cues already established in the visible region, making the reconstruction prone to over-completion. Instead of predicting the full geometry, the detected region is proposed as an explicit physical constraint to guide hidden region reconstruction. Therefore, this study introduces occlusion-conditioned residual to exploit local context along the visible mask boundary to predict occlusion-specific corrections, rather than estimating the entire amodal mask. The proposed approach is developed on top of a standard Mask R-CNN [19] baseline instead of a complex bilayer architecture such as BCNet [20] enabling the performance gains to be attributed solely to the residual correction component.

The proposed approach was evaluated on the dataset, demonstrating that the occlusion-conditioned residual improves AP from 48.02% to 52.69% compared to the baseline direct amodal, with the largest improvements observed for heavily occluded particles. The primary contributions of this study are threefold. First, we identify asymmetric certainty between visible and invisible regions as a fundamental cause of over-completion in aggregate amodal instance segmentation. Second, we introduce an occlusion-conditioned residual correction mechanism that refines visible masks by predicting constrained residuals conditioned on local boundary context, rather than regenerating the entire amodal mask from scratch. Third, we construct an aggregate dataset with controllable levels of occlusion, enabling systematic evaluation of model performance under varying degrees of occlusion, as described in detail in the Method and Data Availability sections. The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 details the dataset construction and methodology. Section 4 presents the results and discussion. Finally, Section 5 concludes the paper with directions for future work.

2. Literature Review

2.1. Amodal Instance Segmentation

Amodal instance segmentation was first introduced by [21], who defined it as the task of predicting the complete object regions, including both visible (modal) and invisible (amodal) parts, using only modal annotations for model training. A study by [22] then introduced "semantic amodal segmentation" with extensive amodal annotations, which was subsequently adopted by [16] in formulating KINS, one of the most widely studied amodal datasets. Since then, various architectures have been proposed to tackle this task, ranging from early two-stage CNN-based frameworks to modern Generative Adversarial Network (GAN) and Vision Transformer architectures [14], [23], [24], [25], [26]. Apart from their architectural diversity, most of these approaches still predict amodal masks in their entirety, making them prone to over-completion when visual evidence is insufficient to constrain the prediction area.

2.2. Amodal Instance Segmentation for Aggregate Materials

In aggregate characterization domain, the application of amodal instance segmentation remains largely under-explored. A study by [17] adopted BCNet along with Feature Pyramid Network and global context module to segment occluded aggregate particles. On their next study [18], they propose virtual-to-real rock-fill dataset based on DiffusionGAN [27] to reduce annotations cost. While these contributions improve architectural design and data availability, existing approaches still predict the full amodal mask without an explicit mechanism to constrain predictions using visible regions. Furthermore, the datasets used in these approaches are not publicly available, making reproduction and further evaluation difficult.

2.3. Instance Segmentation Backbones

Most amodal segmentation are built upon well-established backbones. Mask-RCNN [19] introduces two-stage approaches that enhance Faster-RCNN [28] by adding mask prediction branch upon region proposal, producing

normalized feature set for each set. Consequently, this paradigm serves as the standard baseline for many amodal frameworks including BCNet [20] and its variants, as disentangled and consistent object representations facilitate the inference of occluded regions. On the other hand, the advancement of Transformer-based detectors like DETR [25] and Mask2Former [26] have shifted the field to fully end-to-end predictions, eliminating hand-crafted heuristics such as anchor generators and non-maximum suppression (NMS), in favor of object queries and self-attention to model global object relations.

While it is effective to infer relations between instances, this approach does not align with the present scenario, where amodal prediction errors arise when the reconstruction of the occluded region is not strictly constrained by the boundary of the visible object mask. Moreover, transformer-based methods demand substantial computational resources and large-scale training data. Therefore, the proposed method is based on a region-based detector, as it yields independent per-instance representations that allow for strictly controlled feature attribution.

3. Methodology

3.1. Problem Formulation

Given an RGB image I and M region proposal features $X \in \mathbb{R}^{M \times C \times h \times w}$ extracted from Feature Pyramid Network (FPN) backbone, the goal of amodal instance segmentation is to predict visible mask $\hat{V}$ and amodal mask $\hat{A}$. The hidden region is defined by:

$$\hat{H} = \hat{A} \odot (1 - \hat{V}) \approx H^* = \max(A^* - V^*, 0) \quad (1)$$

On granular objects, direct amodal prediction approach $\hat{A} = \sigma(A_0)$ yields a strong AP50, yet degrades sharply at AP75 as occlusion increases. This suggests the main bottleneck is not object localization, but rather the precise reconstruction of amodal shapes and boundaries where direct visual cues are absent. To tackle this problem, as shown in figure 1, amodal prediction is formulated as a base prediction refined through visible-conditioned residuals:

$$\hat{A}_{\text{logits}} = A_0 + \alpha \cdot \mathcal{R}(z) \quad (2)$$

Here, A_0 denotes the initial amodal prediction, z is a conditioning representation summarizing object characteristics and visual context, while $\mathcal{R}(z)$ represents a structured residual that refines occluded geometry. Through this formulation, amodal inference is framed as residual shape completion, learning the necessary shape adjustments to complete unobserved regions while remaining grounded in visible evidence.

3.2. Feature Extraction

The proposed architecture extract multi-scale features through ResNet-50 and Feature Pyramid Network (FPN). Given an input image $I \in \mathbb{R}^{B \times C_{in} \times H \times W}$ ($C_{in} = 3$), with ResNet-50 extract hierarchical features $\{C_2, C_3, C_4, C_5\}$ using stride factor $\{4, 8, 16, 32\}$. FPN is then merge the channels to $C = 256$ through top-down connection, producing feature pyramids $\mathcal{P} = \{P_2, P_3, P_4, P_5, P_6\}$ with $P_k \in \mathbb{R}^{B \times C \times \frac{H}{2^k} \times \frac{W}{2^k}}$.

Region Proposal Network (RPN) then proposes object bounding box candidates from $\mathcal{P}$. To infer the context around the occluded area, each bounding box is expanded by a scale factor γ . The expanded feature map is subsequently cropped and spatially aligned based according to their pyramids level to form a Region of Interest (RoI) feature map $X \in \mathbb{R}^{M \times C \times H \times W}$, where M denotes the number of region proposals.

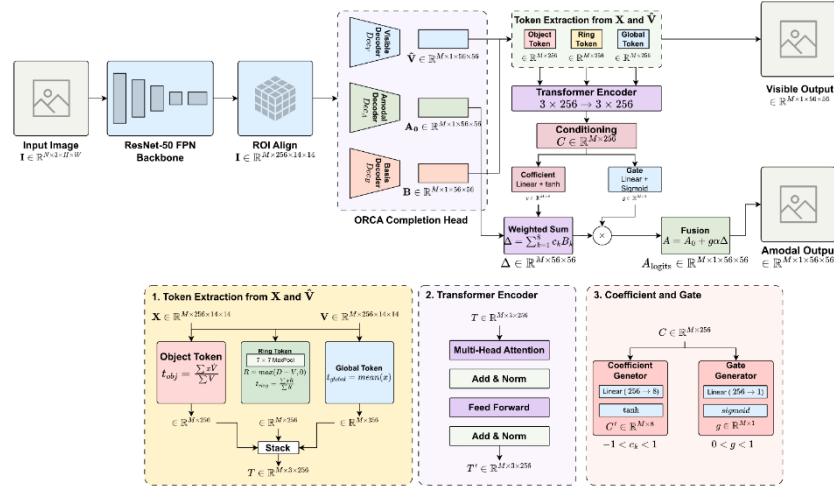


Figure 1. Proposed network architecture. Please zoom in for details.

3.3. Visible Decoder

In amodal instance segmentation, the visible region is the only reliable cue for estimating the complete object shape, whereas the occluded region is inferred from incomplete evidence [29], [30]. Based on this observation, visible geometry processing and residual reconstruction processing are explicitly separated to avoid redundant computation on already observed regions. This study realize this idea through a visible decoder branch Dec_{vis} applied to the obtained ROI features X .

The decoder consists of two transposed convolution layers that upsample the feature map from $H \times W$ to the target resolution $H_m \times W_m$, with $H_m = W_m = 56$, and reduce the channel dimension to one to form a raw spatial logit $Z_{vis} \in \mathbb{R}^{M \times 1 \times H_m \times W_m}$. The visible mask probability map $\hat{V}$ is then obtained through sigmoid function:

$$Z_{vis} = Dec_{vis}(X) \in \mathbb{R}^{M \times 1 \times H_m \times W_m} \quad (3)$$

$$\hat{V} = \sigma(Z_{vis}) \in \mathbb{R}^{M \times 1 \times H_m \times W_m} \quad (4)$$

This branch is trained with independently supervised ground-truth visible mask aligned with the corresponding ROI region to ensure that $\hat{V}$ can accurately capture visible spatial details and pixel level boundaries from the actual visual evidence in the image.

3.4. Amodal Decoder

The visible decoder captures the observed object regions, whereas the base amodal decoder Dec_{base} predicts an initial amodal logit map $A_0 \in \mathbb{R}^{M \times 1 \times H_m \times W_m}$ that provides a coarse estimate of the complete object shape. Following prior two-stage amodal segmentation and region-based detection frameworks [16], [19], [28], direct feature-to-mask prediction is treated as a reliable geometric prior, particularly for un-occluded or lightly occluded regions:

$$A_0 = Dec_{amodal}(X). \quad (5)$$

The decoder contains a 3×3 convolution with 128 channels and ReLU activation, followed by two 2×2 transposed convolutions with stride 2. This architecture progressively upsamples ROI features from 14×14 to 56×56 while producing a single-channel output map. Importantly, A_0 remains in logit space without an initial Sigmoid activation, which preserves linear shape information and avoids early gradient saturation.

Rather than using A_0 as a fixed prediction, A_0 is preserved as a geometric anchor and refine it through a residual correction:

$$Z_{amodal} = A_0 + \alpha g \Delta \quad (6)$$

Z_{amodal} denotes the final amodal logit, Δ is a learned spatial residual, α is a fixed scaling factor, and $g \in [0,1]$ is a learned scalar gate. The gate is initialized near zero by setting its bias to -4 , so training begins from the baseline prediction and introduces residual corrections only when necessary. This design stabilizes optimization and helps maintain strong performance on easy, lightly occluded objects.

This residual formulation differs from approaches that predict invisible regions independently [17], [18], which can lead to excessive mask growth and boundary leakage under severe occlusion. In the proposed model, Δ acts only as a localized additive correction around occlusion boundaries. While sharing the motivation of prior-guided methods [31], [32] in using structural shape constraints, the proposed approach avoids external shape codebooks A_0 within the network (see figure 2). Consequently, the residual branch focuses on completing the missing regions rather than reconstructing the visible object structure from scratch.

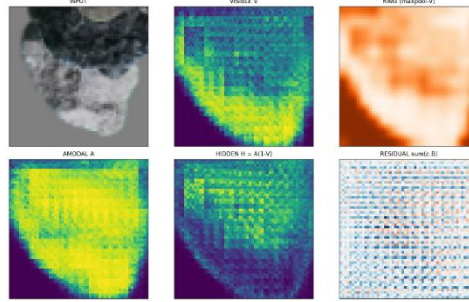


Figure 2. Illustration of the proposed ring context for capturing local boundary information around the visible mask.

3.5. Shape-based Decoder

Amodal region reconstruction is vulnerable to boundary leakage when predicted solely from raw visual features. To constraint the solution space, the basis decoder Dec_{basis} extracts geometric dictionary consisting of N_b spatial basis maps $B \in \mathbb{R}^{M \times N_b \times H_m \times W_m}$. Following the shape dictionary paradigm [33], [34], this approach represents shape variations via adaptively combinable components, while leveraging structured shape representations to restrict the feasible shape space:

$$B = Dec_{basis}(X) \in \mathbb{R}^{M \times K \times H_m \times W_m} \quad (7)$$

Two transposed convolution layers upsample the RoI features X from 14×14 to 56×56 , with a final projection resulting $K = 8$ basis maps B_k . Each B_k captures local spatial pattern and is dynamically combined to form a residual correction Δ . Setting $K = 8$ provides a balance between compactness and expressiveness, an aspect further evaluated in Section 4.5.2.

3.6. Conditioning Mechanism

To evaluate how the basis dictionary B is utilized, this study do not apply B directly as a residual update to the amodal logits A_0 . Instead, it is assumed that individual basis $b \in B$ should contribute unequally across instances. Therefore, this study derives instance-specific vectors and use them to predict 8 adaptive coefficients that modulate the contribution of each basis to the residual. These specific vectors are: visible object information $\hat{V}$ (Object Token), surrounding boundary context (Ring Token), and global object features (Global Token). The first representation, the Object Token, is obtained by using $\hat{V}$ as a spatial weighting map over X . Specifically, the feature responses are weighted by their corresponding visible probabilities and aggregated over the spatial dimensions:

$$t_{obj} = \frac{\sum_{h,w} X_{:,h,w} V_{:,h,w}}{\sum_{h,w} V_{:,h,w}} \quad (8)$$

This operation yields $t_{obj} \in \mathbb{R}^{M \times 256}$ where each vector describes the feature content of the predicted visible region. The Object Token provides object-centric information describing what is directly observable from the target instance. The second representation, the Ring Token, is introduced to capture local contextual information surrounding the visible object boundary. Starting from the same visible probability map V , a spatially dilated representation is first constructed using a 7×7 max-pooling operation:

$$D = \text{MaxPool}_{7 \times 7}(V) \quad (9)$$

$$R = \max(D - V, 0) \quad (10)$$

$$t_{ring} = \frac{\sum_{h,w} X_{:,h,w} R_{:,h,w}}{\sum_{h,w} R_{:,h,w}} \quad (11)$$

The surrounding region is then obtained by removing the original visible response, as formally written in Eq. (10). Thus, R acts as a soft spatial weighting map emphasizing the neighborhood around the predicted visible region. This is distinct from the visible mask itself: the Object Token focuses on evidence inside the visible region, whereas the Ring Token focuses on evidence surrounding that region. Using R as the spatial weighting function, the Ring Token is computed analogously to the Object Token as annotated in Eq. (11), yielding $t_{\text{ring}} \in \mathbb{R}^{M \times 256}$.

This representation provides local contextual evidence around the visible boundary of the target instance. In an occluded setting, such surrounding information can complement the appearance information obtained from the visible object itself. The source implementation explicitly constructs the ring through dilation followed by subtraction of the visible response and uses it as a weighting function for feature aggregation.

$$t_{\text{obj}} = \frac{\sum_{h,w} X_{:,h,w} V_{:,h,w}}{\sum_{h,w} V_{:,h,w}} \quad (12)$$

The third representation is the Global Token. Unlike the Object and Ring Tokens, the Global Token does not use a spatial weighting mask. Instead, it summarizes the complete ROI feature map through global spatial averaging:

$$t_{\text{global}} = \frac{1}{HW} \sum_{h,w} X_{:,h,w} \quad (13)$$

Each instance is therefore represented by three complementary token embeddings, corresponding respectively to visible object evidence, local surrounding context, and global ROI information. The resulting token sequence is processed using a two-layer Transformer encoder with an embedding dimension of 256 and four attention heads. The purpose of this stage is to allow information from the three conditioning sources to interact before producing the final instance representation. Rather than concatenating the three tokens directly, self-attention provides a mechanism through which each token can incorporate information from the other conditioning sources.

3.7. Dynamic Residual Completion

After the three decoder branches yield the geometric anchor A_0 and the basis dictionary B , the final stage synthesizes the spatial residual refinement Δ to complete the amodal mask composition. The Transformer conditioning engine processes the ROI features X to extract object context, occlusion boundary, and global spatial tokens ($t_{\text{obj}}, t_{\text{ring}}, t_{\text{glob}}$). Through lightweight linear projection heads, these tokens generate a dynamic coefficient vector $w \in [-1,1]^K$ via a tanh activation alongside a scalar gate $g \in [0,1]$. The spatial residual map is synthesized through a weighted linear combination of the basis primitives:

$$\hat{H} = \sigma(\sum_{k=1}^K w_k \cdot B_k) \odot (1 - \hat{V}) \quad (14)$$

$\sigma(\cdot)$ is the Sigmoid function, B_k and w_k are the spatial bases and their weights, $\hat{V}$ is the visible mask, and $\odot$ denotes element-wise multiplication. This formulation deterministically derives $\hat{H}$, ensuring strict topological consistency ($\hat{V} \subseteq \hat{A}$) with zero parameter overhead.

As demonstrated in figure 3, this residual mechanism dynamically adapts to occlusion severity. In scenarios featuring minor occlusion or unambiguous object boundaries, the network autonomously dampens the scalar gate. This behaviour validates the core architectural design: the model relies predominantly on the baseline geometric anchor A_0 to maintain structural stability. Consequently, this approach suppresses over-expansion artifacts while retaining high boundary fidelity under complex occlusion patterns.

3.8. Dataset Construction

To the best of current knowledge, existing aggregate instance segmentation datasets remain predominantly closed source. To bridge this gap, a comprehensive public dataset is released alongside an extensible codebase to facilitate customization and reproducibility. The coarse aggregate dataset was collected from the Materials Laboratory, Department of Civil Engineering, Politeknik Negeri Bali (8°48'02.55"S, 115°09'41.31"E; WGS84).

Large aggregate samples were photographed in non-contact arrangements of 30-40 randomly distributed particles on a green backdrop, with a 2003 500 Coin IDR serving as a scale reference, as shown in figure 3. Each object is then

Figure 1 illustrates the experimental setup and results. The setup includes a Sony ILCE 7603 camera mounted on a K&F 210 Tripod, positioned 45 cm above the sample table. The camera is connected to a horizontal window extension of 45 cm. The sample table is covered with a green background mat and contains aggregate samples. A horizontal axis of imaging is shown, with a distance of 1.15 m from the camera to the sample table. A scale reference (IDR 500 Cells) is placed on the sample table. The setup is illuminated by natural window light (diffuse). The results are shown in a 4x4 grid of images, comparing the input image, the modal + amodal result, the modal only result, and the amodal only result, under low and high occlusion conditions.

In constructing the BIMA dataset, scenes are synthesized at 512×512 pixels (16 px padding) with identity-based asset partitioning across train/val/test splits (70: 15: 15) to eliminate visual leakage. Coins are always rendered intact on the top layer, whereas aggregates are distributed via z-ordered rejection sampling (128 candidates on an 8×8 grid) guided by a multi-objective cost function $\mathcal{C}(p)$ targeting occlusion, density, spatial spread, and boundary compliance. A pixel assignment map $\mathbf{M} \in \{0, \dots, N\}^{H \times W}$ deterministically yields the visible mask $V_i = [\mathbf{M} = i]$ and amodal mask $A_i = (\text{Alpha}_i > 0)$, enabling exact occlusion ratio computation $\eta_i = 1 - \frac{|V_i|}{|A_i|}$ without manual labeling. Annotations are exported as extended COCO JSON alongside a YAML-configurable generator script.

4.1. Experimental Setup

Table 1. Dataset split ratios for 512×512 generated images, each featuring 12 non-fully overlapping aggregates and one reference coin.

Each image contains 12 aggregates of randomized shapes and sizes alongside a single reference coin. To maintain physical plausibility, aggregates exhibit partial occlusion but no complete (100%) overlapping, as fully occluded objects cannot be visually identified or reconstructed from a 2D image. To analyse the model's performance under varying density conditions, the test set is further divided into five occlusion intervals based on the average visible area of aggregates, as summarized in [table 2](#).

Occlusion Interval	Images	Num. of Aggregates	Mean Occlusion	Mean Visible Fraction
[0, 10)	60	720	5%	95%
[10, 20)	60	720	15%	85%
[20, 30)	60	720	25%	75%
[30, 50)	60	720	40%	60%
[50, 70)	60	720	60%	40%

All model variants were trained on a single NVIDIA A100 GPU using SGD with a batch size of 4, a learning rate of 2.5×10^{-4} , momentum of 0.9, weight decay of 1×10^{-4} , and the training schedule specified in [table 3](#).

Table 3. Hyperparameter settings and training specifications for the proposed model.

Training Hyperparameter	Setting
Optimizer	SGD
Batch Size	4
Learning rate	2.5×10^{-4}
Momentum	0.9
Weight decay	1×10^{-4}
Warm-up	500 iterations
Training Iterations	6000
Learning Rate (LR) Scheduler	Step decay

To optimize the proposed framework, an unweighted joint objective function was formulated that simultaneously supervises the reconstruction of visible and amodal spatial representations:

$$\mathcal{L}_{total} = \mathcal{L}_{visible} + \mathcal{L}_{amodal} \quad (15)$$

Both $\mathcal{L}_{visible}$ and $\mathcal{L}_{amodal}$ are instantiated using Binary Cross-Entropy (BCE) with Logits over all N pixels, defined as:

$$\mathcal{LBCE}(z, y) = -\frac{1}{N} \sum_i \log \sigma(z_i) = -\frac{1}{N} \sum_i [y_i \log \sigma(z_i) + (1 - y_i) \log(1 - \sigma(z_i))] \quad (16)$$

here z_i denotes the predicted feature logit, $y_i \in \{0,1\}$ represents the corresponding ground-truth binary label, and $\sigma(\cdot)$ signifies the sigmoid activation function. Specifically, $\mathcal{L}_{visible} = \mathcal{L}_{BCE}(V_{logit}, V_{gt})$ enforces feature learning on unoccluded regions, whereas $\mathcal{L}_{amodal} = \mathcal{L}_{BCE}(A_{logit}, A_{gt})$ guides complete object reconstruction. Notably, no explicit auxiliary loss is applied to the hidden region during training; the hidden probability map H is instead implicitly derived at inference time via $H = A(1 - V)$:

To evaluate segmentation quality and amodal completion fidelity across varying occlusion states, standard COCO Average Precision metrics (AP, AP₅₀, and AP₇₅) are computed across visible, amodal, and hidden target masks. Furthermore, to quantify the geometric accuracy of the reconstructed boundaries independently of detection confidence scoring, Matched Amodal IoU and Matched Hidden IoU are evaluated specifically over correctly localized aggregate instances. To compute the Matched Amodal IoU and Matched Hidden IoU, predicted aggregate instances and their corresponding ground-truth targets are first paired using a greedy one-to-one matching procedure based on amodal bounding box IoU with a threshold of 0.5. Consequently, only prediction-ground-truth pairs satisfying:

$$IoU_{box}(B_{pred}, B_{gt}) \geq 0.5 \quad (17)$$

are retained for subsequent evaluation. For each successfully matched pair, the Matched Amodal IoU is computed from the spatial overlap between the predicted amodal mask A_{pred} and the ground-truth amodal mask A_{gt} :

$$IoU_A = \frac{|A_{pred} \cap A_{gt}|}{|A_{pred} \cup A_{gt}|} \quad (18)$$

The hidden mask is not explicitly predicted as an independent output branch; instead, it is derived from the relationship between the amodal and visible masks via $H = A \odot (1 - V)$. Accordingly, the Matched Hidden IoU for each paired instance is formulated as:

$$IoU_H = \frac{|H_{pred} \cap H_{gt}|}{|H_{pred} \cup H_{gt}|} \quad (19)$$

The reported Matched Amodal IoU and Matched Hidden IoU values represent the mean IoU across all N_{match} successfully matched instance pairs:

$$\text{Matched IoU} = \frac{1}{N_{match}} \sum_{i=1}^{N_{match}} IoU_i \quad (20)$$

Because IoU is naturally normalized by the union area, its value remains on a uniform scale across individual instances. However, no explicit normalization is applied with respect to the occlusion fraction. Therefore, cross-interval comparisons of Hidden IoU should be interpreted with the understanding that the absolute size and proportion of the hidden region inherently shift as occlusion severity scales. To assess boundary fidelity beyond region-level mask IoU,

this study additionally report Boundary IoU following [35]. Boundary IoU evaluates the overlap between boundary bands extracted from the predicted and ground-truth masks, providing a more boundary-sensitive measure than conventional Mask IoU. For each matched amodal instance, Boundary IoU is computed as:

$$BIOU = \frac{B_{pred} \cap B_{gt}}{B_{pred} \cup B_{gt}} \quad (21)$$

4.2. Overall Performance Evaluation

To evaluate the overall occlusion-conditioned residual correction mechanism, the proposed network is compared with a plain/direct Mask R-CNN under an identical evaluation protocol. The vanilla Mask R-CNN act as a baseline model that predict amodal instance segmentation without any residual correction, and the proposed model (denoted as ORCA) preserves Mask R-CNN settings with additional Occlusion-Residual Completion Head (ORCA). All evaluations are conducted within occlusion range of 0 –70%. Furthermore, the findings are reported using amodal Average Precision (AP, AP₅₀, AP₇₅) and matched amodal Intersection over Union (IoU). As shown in table 4, ORCA consistently gains greater performance compared to Direct Mask R-CNN. Mean amodal AP increased from 48.02% to 52.69% which shows 4.57 percentage points (pp) gains. Performance gains persist across different IoU thresholds, where AP₅₀ increases from 89.93% to 92.53% and AP₇₅ improves from 48.92% to 56.66%.

Table 4. Overall Performance Comparison.

Performance Metrics	Mask RCNN (Direct)	Mask-RCNN + Occlusion Residual Head (Proposed)	Delta
Mean amodal AP (%)	48.02 ± 0.42	52.69 ± 0.36	+4.67 pp
Mean amodal AP ₅₀ (%)	89.93 ± 0.31	92.53 ± 0.27	+3.20 pp
Mean amodal AP ₇₅ (%)	48.92 ± 0.68	56.66 ± 0.61	+6.75 pp
Mean matched amodal IoU	0.773 ± 0.004	0.785 ± 0.003	+0.0123 pp
Mean matched hidden IoU	0.162 ± 0.04	0.165 ± 0.014	+0.0027pp
Boundary IoU	0.675 ± 0.015	0.718 ± 0.005	+0.022 pp
Mean aggregate reall (%)	95.20 ± 0.27	95.80 ± 0.22	+0.60 pp
FPS ↑	18.63 ± 0.3	16.9 ± 0.3	-1.7

Specifically, the higher margin of improvement in AP₇₅ over AP₅₀ highlights that ORCA's superiority is especially evident under stricter spatial alignment demands rather than just coarse overlap thresholds. Furthermore, ORCA introduces a moderate increase in model complexity compared with the Direct Mask R-CNN baseline, from 44.5M to 47.1M parameters and from 180.2G to 191.6G FLOPs. This additional complexity results in a moderate reduction in inference speed, from 18.6 to 16.9 FPS under the same evaluation setting. To assess the robustness of the main comparison, both models were evaluated using three independent random seeds (seeds 42, 43, and 44), and the results are reported as mean ± standard deviation in table 4. Despite the additional computational cost and slightly lower inference speed, ORCA provides a 4.67 percentage-point improvement in amodal AP, suggesting that the observed performance gain is not solely attributable to increased model capacity. Due to the substantially higher computational cost of repeating the extended ablation experiments, those experiments were conducted using a fixed seed (42). Particle Size Distribution analysis was also performed, following [17], as shown in figure 4.

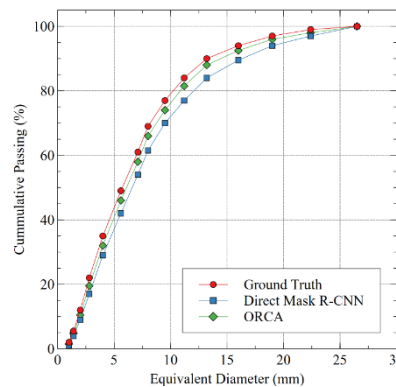


Figure 4. PSD Analysis.

The PSD curve obtained from ORCA follows the reference distribution more closely than the Direct Mask R-CNN prediction, particularly across the dominant particle-size range. This indicates that the proposed residual correction not only improves segmentation overlap metrics but also better preserves the overall size distribution of the reconstructed aggregates. The consistent increase of AP, AP_{50} and AP_{70} , indicates that ORCA's improvement is not limited to only a certain level of occlusion level and proves that residual correction can enhance amodal prediction quality against direct prediction. In later sections, an extensive evaluation across multiple occlusion intervals and ablation study are presented, highlighting the dominant factors contributing to the observed performance gains.

4.3. Overall Performance Evaluation

To analyse the robustness of the proposed model across varying degrees of aggregate occlusion, the performance is evaluated across five occlusion intervals: [0,10)%, [10,20)%, [20,30)%, [30,50)%, and [50,70)%. This stratification enables a detailed comparative analysis between Direct Mask R-CNN and ORCA as the proportion of occluded regions increases. Beyond the mean amodal AP, the evaluation within each interval incorporates AP_{50} , AP_{75} , matched amodal Intersection over Union (IoU), matched hidden IoU, and aggregate recall to comprehensively capture degradation dynamics as occlusion severity scales.

As detailed in table 5, ORCA consistently outperforms Direct Mask R-CNN across the entire occlusion intervals. At the lowest occlusion level ([0,10)), mean AP increases from 65.17% to 69.22%, yielding a gain of 4.05 percentage points. The most substantial improvements are observed in the [10,20)% and [20,30)%, rising by 6.26 and 5.86 percentage points, respectively (56.19% vs. 62.45% and 53.06% vs. 58.92%). In the [30,50)% range, ORCA sustains its performance lead over the baseline, advancing from 42.17% to 47.13%. Finally, under severe occlusion ([50,70)%), the performance gap reduces to 2.02 percentage points, with ORCA reaching 25.54% vs. 23.52% for the baseline.

Table 5. Performance Comparison under Different Occlusion Levels.

Occlusion Range (%)	Average Precision (AP)		Proposed Model Metrics				
	Mask RCNN	Proposed Model	AP_{50}	AP_{75}	Amodal IoU	Hidden IoU	Recall (%)
0-10	65.17	69.22	99.98	82.02	0.8542	0.0619	99.8
10-20	56.19	62.45	97.81	74.81	0.8263	0.1244	98.6
20-30	53.06	58.92	98.83	69.02	0.8035	0.1677	99.4
30-50	42.17	47.13	96.13	40.25	0.7542	0.2102	98.3
50-70	23.52	25.54	69.91	12.21	0.6866	0.2607	82.8

Aside of mean AP changes, the supporting metrics exhibit a consistent degradation pattern as occlusion becomes more severe. In ORCA, AP_{75} decrease from 82.02% to 74.81% on [0,10)% as well as 69.02%, 40.25%, and 12.21 on [10,20)%, [20,30)%, [30,50)% and [50,70)% respectively. The matched amodal IoU exhibits a corresponding decline, dropping from 0.8542 to 0.6866 under severe occlusion. Similarly, aggregate recall remains high through [30,50)% but falls to 82.8% at [50,70)%. Collectively, these trends demonstrate that escalating occlusion severity adversely impacts mask overlap accuracy and model localization performance.

Interestingly, the matched hidden IoU exhibits an opposite trend, increasing monotonically with occlusion severity from 0.0619 in the [0,10)% interval to 0.2607 in the [50,70)% range. Overall, this occlusion-level analysis demonstrates that ORCA's performance advantage varies across occlusion regimes. Although gains are observed in all intervals, the largest improvements occur at moderate occlusion levels, with more modest gains at [50,70)% occlusion. Simultaneously, both amodal mask fidelity and instance detection rate decrease under higher occlusion severity. This behavior highlights that residual correction relies on the residual visual signal available, an observation further analysed in later sections.

4.4. Qualitative Performance Evaluation

Figure 5 shows qualitative performance comparison between Direct Mask R-CNN against ORCA on different occlusion level. In low occlusion settings, while both approaches produce predictions near the ground truth, ORCA exhibits contours that better capture the aggregate geometry, especially in regions with irregular boundaries. The discrepancy becomes more evident at moderate occlusion levels ([20,30)%), where parts of the aggregate structure are hidden. Under these conditions, ORCA generates amodal shapes that remain more consistent with the ground truth, whereas Direct Mask R-CNN tends to produce simplified contours that fail to capture the underlying object structure

in occluded regions. These qualitative observations confirm that residual correction helps preserve accurate extent estimates as available visual cues decrease.

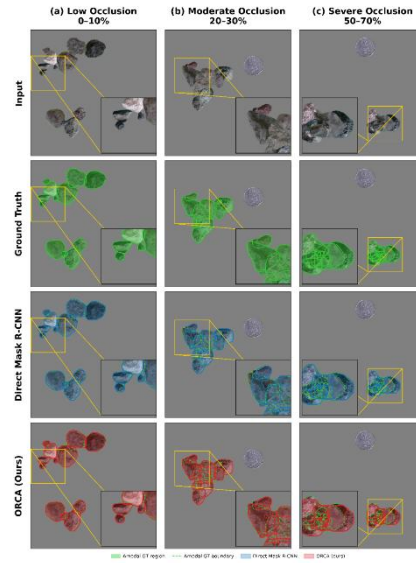


Figure 5. Qualitative Evaluation. The proposed model perform relatively better in amodal completion.

Under severe occlusion ($[50,70)\%$), both approaches show elevated errors when reconstructing hidden aggregate parts. While ORCA occasionally yields predictions closer to the ground truth, the qualitative gap narrows relative to moderate occlusion levels. This observation corroborates the quantitative trend across occlusion tiers, where ORCA achieves substantial AP improvements of 6.26 and 5.86 percentage points at $[10,20)\%$ and $[20,30)\%$, but only 2.02 percentage points at $[50,70)\%$. Thus, qualitative evidence supports that ORCA provides maximal benefit when visual information is sufficient for residual refinement, whereas severe occlusion heavily constrains both models.

4.5. Ablation Studies

4.5.1. Component Contribution Analysis

To evaluate the contribution of the main components of ORCA, an ablation study was conducted by comparing the Direct Mask R-CNN baseline, ORCA without Ring Context, and the full ORCA model. All other experimental configurations were kept consistent across the three variants. The results are presented in [table 6](#).

Table 6. Component ablation results.

Model	AP	AP ₅₀	AP ₇₅	Amodal IoU	Hidden IoU	Recall
Direct Mask R-CNN	48.02	89.10	58.30	0.7730	0.1620	95.10
ORCA w/o Shape-based Decoder	51.12	92.30	61.72	0.7805	0.1631	95.30
ORCA w/o Dynamic Basis Combination	50.74	91.95	60.84	0.7792	0.1628	95.20
ORCA w/o Transformer Conditioning	51.36	92.35	62.01	0.7810	0.635	95.40
ORCA w/o Ring Context	50.85	92.40	62.15	0.7795	0.1638	95.40
ORCA w/o Learned Scalar Gate	51.84	92.18	55.72	0.7839	0.1640	95.60
ORCA w/o Residual Scaling	51.47	92.10	61.34	0.7816	0.1634	95.50
Full ORCA	52.69	94.55	65.18	0.7850	0.1650	95.80

As shown in [table 6](#), the Direct Mask R-CNN baseline achieves an AP of 48.02%. Introducing the residual correction without Ring Context increases the AP to 50.85%, corresponding to a gain of 2.83 percentage points. Adding Ring Context further improves the AP to 52.69%, providing an additional gain of 1.84 percentage points. Similar improvements are observed in AP₅₀, AP₇₀, amodal IoU, hidden IoU, and recall. These results demonstrate that both residual correction and Ring Context contribute positively to the overall performance of ORCA. The residual correction provides the larger contribution, while Ring Context provides an additional improvement by incorporating local contextual information.

The improvement obtained from the residual correction is consistent across the evaluated metrics, rather than being limited to average precision alone. Compared with the Direct Mask R-CNN baseline, the residual variant improves AP50 from 89.10% to 92.40% and AP70 from 58.30% to 62.15%, while amodal IoU increases from 0.7730 to 0.7795. The corresponding increase in hidden IoU remains smaller, from 0.1620 to 0.1638, suggesting that the residual branch primarily improves the overall amodal mask quality rather than producing a large change in the isolated hidden-region overlap. Aggregate recall also increases from 95.10% to 95.40%, indicating that the improvement is not accompanied by a reduction in the number of successfully retrieved aggregate instances.

The addition of Ring Context provides a further improvement across the same evaluation criteria. Relative to the residual-only variant, the full model increases AP by 1.84 percentage points, from 50.85% to 52.69%, while AP50 increases from 92.40% to 94.55% and AP70 from 62.15% to 65.18%. The increase in amodal IoU from 0.7795 to 0.7850 is accompanied by a smaller improvement in hidden IoU from 0.1638 to 0.1650. This pattern suggests that the additional contextual information is beneficial primarily for improving the consistency of the complete amodal prediction, while its effect on the isolated hidden region is more limited.

Altogether, the ablation results show a clear incremental contribution from the proposed components: the residual correction establishes the main performance gain over direct amodal prediction, while the Ring Context provides an additional improvement when incorporated into the residual-based formulation. The fact that the full model achieves the highest values across the reported metrics supports the use of visible-object information together with local contextual information for conditioning the residual correction. Further ablations on the internal design of the residual mechanism, including the basis representation, dynamic basis combination, conditioning engine, gating, and residual scaling, are presented in the following sections.

4.5.2. Basis Dictionary Capacity Analysis

The contribution of basis dictionary capacity to the residual correction mechanism is evaluated using a varying number of basis maps, denoted by K . Specifically, $K \in \{2, 4, 8, 16\}$ is tested throughout the evaluation, and the results are presented in table 7. The results show that increasing the basis size from $K = 2$ to $K = 8$ consistently improves the overall segmentation performance, with AP increasing from 50.90% to 52.69%. A similar trend is observed in AP50, AP75, amodal IoU, hidden IoU, and recall. However, increasing the basis size to $K = 16$ does not provide further improvement, with AP slightly decreasing to 52.50%. This suggests that $K = 8$ provides sufficient capacity to represent the residual shape variations while avoiding unnecessary increases in model capacity. Therefore, $K = 8$ is selected as the basis dictionary size for the final model.

Table 7. Effects of basis dictionary capacity on ORCA performance.

Basis Size K	AP	AP ₅₀	AP ₇₅	Amodal IoU	Hidden IoU	Recall
2	50.90	91.70	58.80	0.779	0.163	95.40
4	51.35	92.20	55.10	0.782	0.164	95.60
8	52.69	94.55	65.18	0.7850	0.1650	95.80
16	52.50	92.50	56.30	0.784	0.165	95.70

4.5.3. Residual Gating Analysis

The contribution of the residual gating mechanism is evaluated by comparing the full ORCA model with a variant without the adaptive gate under the same evaluation settings. The results are presented in table 8. The full model achieves an AP of 52.69%, compared with 51.84% for the model without gating. Similar improvements are observed in AP50, AP75, amodal IoU, hidden IoU, and recall. These results indicate that the gating mechanism helps regulate the magnitude of the residual correction based on the available contextual information, leading to more stable and accurate amodal predictions. Therefore, the adaptive residual gate is retained in the final ORCA configuration.

Table 8. Effects of basis dictionary capacity on ORCA performance

Model Variant	AP	AP ₅₀	AP ₇₅	Amodal IoU	Hidden IoU	Recall
ORCA w/o Gating Mechanism	51.84	92.18	55.72	0.7839	0.164	95.60
Full ORCA	52.69	94.55	65.18	0.7850	0.1650	95.80

4.6. Discussion

4.6.1. Effectiveness of Residual Correction

The ablation results provide direct evidence that the residual correction mechanism is the main contributor to the performance improvement of ORCA. As reported in Table 4, the Direct Mask R-CNN baseline achieves an AP of 48.02%, which increases to 50.85% when the residual branch is introduced without the Ring Context, yielding a gain of 2.83 percentage points. Adding the Ring Context further increases AP to 52.69%, giving an additional gain of 1.84 percentage points. The improvement from the Direct baseline to the residual-only variant is important because the two models differ primarily in how the amodal mask is generated. The direct baseline relies on a single amodal prediction, whereas ORCA starts from a base amodal prediction and applies a bounded residual correction conditioned on the visible object. This formulation therefore provides a mechanism to refine the initial shape prediction using occlusion-related contextual information, rather than requiring the model to recover the complete object extent solely from the direct amodal prediction. The implementation explicitly formulates the final amodal logits as a base prediction plus a gated residual term.

The additional gain from Ring Context indicates that the residual branch benefits from information concentrated around the visible object boundary. In the proposed architecture, the Ring Context is obtained from a dilated region surrounding the visible mask and is used together with the object and global tokens to construct the conditioning representation. Thus, the ablation suggests that the effectiveness of ORCA does not come solely from adding a residual branch, but from conditioning that correction on spatial information related to the occlusion boundary.

4.6.2. Performance under Different Occlusion Levels

The effectiveness of ORCA varies with occlusion severity. As shown in table 5, ORCA consistently outperforms the Direct Mask R-CNN baseline across all evaluated ranges, with the largest AP gains observed at 10–20% and 20–30% occlusion, reaching +6.26 and +5.86 percentage points, respectively. The gain remains substantial at 30–50% (+4.96 pp) but decreases to +2.02 pp at 50–70% occlusion.

The observed trend suggests that ORCA is most effective under moderate occlusion, where sufficient visible boundary information remains to guide the residual correction while the hidden region is large enough to require meaningful shape completion. Under severe occlusion, the reduced visible evidence limits the reliability of the inferred hidden shape, resulting in a smaller performance gap. This interpretation is consistent with the overall results: although mean amodal AP increases from 48.02% to 52.69%, mean matched hidden IoU improves only slightly from 0.162 to 0.165. Therefore, the main benefit of ORCA is improved overall amodal mask quality, while precise hidden-region reconstruction remains constrained by the amount of visible shape information available. Figure 6 shows the comparative performance of ORCA and the Direct Mask R-CNN baseline across different occlusion severity levels.

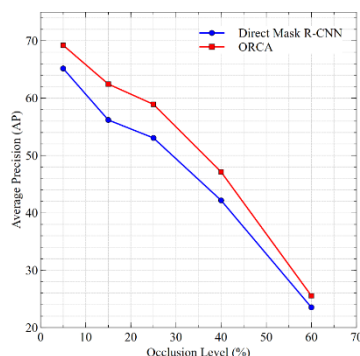


Figure 6. AP performance across different occlusion levels for Direct Mask R-CNN and the proposed model.

4.6.3. Limitations and Generalizability

A primary limitation of this study is the lack of fully reproducible external baselines. Several existing approaches use closed-source implementations, proprietary training configurations, or dataset formats that differ from the proposed experimental setting. As a result, full replication of all competing methods under fair conditions was not possible. The

comparisons should therefore be interpreted as evaluations against the implementations and protocols available to the user, rather than as a complete reproduction of every previously reported system.

Another limitation concerns the evaluation of real-world aggregate images. Reliable evaluation of amodal segmentation requires annotations of the complete shape of each aggregate, including portions that are fully hidden behind other particles. For densely overlapping concrete aggregates, obtaining such amodal annotations directly from real images is difficult because the boundaries of the occluded regions are not observable. Consequently, the present study relies on synthetic images with controlled occlusion, where the complete ground-truth shape is available by construction. Although this enables quantitative evaluation across predefined occlusion levels, it does not fully represent the visual appearance and complex occlusion geometry of real aggregate scenes.

These limitations constrain the generalizability of the reported findings. The improvement of ORCA should therefore be interpreted as empirical evidence under the evaluated synthetic occlusion protocol rather than as a general claim of superiority on real-world aggregate imagery. Future work will focus on validating ORCA using more realistic data generation or reliable real-world amodal annotation protocols, and on adapting the ORCA correction mechanism to multiple backbone and segmentation architectures. Evaluating ORCA across different architectural backends would help determine whether the observed improvement arises from the proposed residual formulation itself and whether the mechanism can be transferred beyond the current Mask R-CNN-based implementation.

5. Conclusion

This study introduced an occlusion-conditioned residual correction mechanism for amodal instance segmentation of overlapping concrete aggregates. Compared with the direct Mask R-CNN baseline, the proposed ORCA model improves mean amodal AP from 48.02% to 52.69%, corresponding to a gain of 4.67 percentage points under the evaluated synthetic occlusion protocol. The occlusion-level analysis further shows that the largest improvements occur at moderate occlusion levels, while the performance gain becomes more limited under severe occlusion. Ablation results also indicate that both the residual correction mechanism and local Ring Context contribute to the overall improvement. These findings demonstrate that occlusion-conditioned residual refinement can improve amodal segmentation performance over direct prediction under the evaluated experimental conditions.

The scope of these findings is limited by the use of a synthetic dataset with controlled occlusion, which enables systematic evaluation but does not fully represent the appearance and occlusion patterns of real aggregate images. In addition, the proposed mechanism is currently implemented within a Mask R-CNN-based segmentation framework, and its effectiveness on other architectures has not yet been established. Future work will focus on validation using realistic or real-world data with reliable amodal annotations and on adapting the ORCA residual correction mechanism to different segmentation backbones and architectures, including transformer-based models.

6. Declarations

6.1. Author Contributions

Conceptualization: I.G.T.S.D., I.N.G.A.A.; Methodology: I.G.T.S.D., I.N.G.A.A., I.N.S.; Software: I.G.T.S.D.; Validation: I.G.T.S.D., I.N.G.A.A., I.N.S.; Formal Analysis: I.G.T.S.D., I.N.S.; Investigation: I.G.T.S.D.; Data Curation: I.G.T.S.D.; Visualization: I.G.T.S.D.; Writing – Original Draft Preparation: I.G.T.S.D.; Writing – Review and Editing: I.N.G.A.A., I.N.S.; Supervision: I.N.G.A.A., I.N.S.; Project Administration: I.N.G.A.A.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The dataset and source code is available at <https://github.com/hippocampa/orca-rcnn>.

6.3. Funding

This work was supported by the Ministry of Higher Education, Science, and Technology of the Republic of Indonesia through the BIMA Research Funding Program (Grant No. 09249/DST/PL8/AL.04/2026).

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

- [1] B. Wang, L. Yan, Q. Fu, and B. Kasal, "A comprehensive review on recycled aggregate and recycled aggregate concrete," *Resour. Conserv. Recycl.*, vol. 171, no. 1, pp. 1–15, Aug. 2021, doi: 10.1016/j.resconrec.2021.105565.
- [2] Z. Abedin Khan, U. Balunaini, S. Costa, and N. H. T. Nguyen, "A review on sustainable use of recycled construction and demolition waste aggregates in pavement base and subbase layers," *Cleaner Mater.*, vol. 13, no. 1, pp. 1–12, 2024, doi: 10.1016/j.clema.2024.100266.
- [3] M. Coenen, D. Beyer, C. Heipke, and M. Haist, "Learning to sieve: Prediction of grading curves from images of concrete aggregate," *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. V-2-2022, no. 1, pp. 227–235, 2022, doi: 10.5194/isprs-annals-V-2-2022-227-2022.
- [4] G. Cao et al., "Influence of aggregate shape on the flow properties of fresh concrete," *Powder Technol.*, vol. 415, no. 1, pp. 1–12, Feb. 2023, doi: 10.1016/j.powtec.2022.118186.
- [5] V. M. Nguyen, T. T. Nguyen, H. M. Gioi, T. Van Phan, and R. Siddique, "Optimizing aggregate grading and residual mortar coefficient in low-cement concrete by using crushed coal mine waste rock as base layers for low-traffic roads," *Constr. Build. Mater.*, vol. 449, no. 1, pp. 1–15, 2024, doi: 10.1016/j.conbuildmat.2024.138468.
- [6] L. Wang, Y. Yao, J. Li, Y. Tao, and K. Liu, "Review of visualization technique and its application of road aggregates based on morphological features," *Appl. Sci.*, vol. 12, no. 20, pp. 1–20, Oct. 2022, doi: 10.3390/app122010571.
- [7] X. Hu, H. Fang, J. Yang, L. Fan, W. Lin, and J. Li, "Online measurement and segmentation algorithm of coarse aggregate based on deep learning and experimental comparison," *Constr. Build. Mater.*, vol. 327, no. 1, pp. 1–12, Apr. 2022, doi: 10.1016/j.conbuildmat.2022.127033.
- [8] M. Moaveni, S. Wang, J. M. Hart, E. Tutumluer, and N. Ahuja, "Evaluation of aggregate size and shape by means of segmentation techniques and aggregate image processing algorithms," *Transp. Res. Rec.*, vol. 2335, no. 1, pp. 50–59, 2013, doi: 10.3141/2335-06.
- [9] W. Wang, C. Su, and H. Zhang, "Automatic segmentation of concrete aggregate using convolutional neural network," *Autom. Constr.*, vol. 134, no. 1, pp. 1–12, Feb. 2022, doi: 10.1016/j.autcon.2021.104106.
- [10] Y. Xiao et al., "A novel method for predicting coarse aggregate particle size distribution based on segment anything model and machine learning," *Constr. Build. Mater.*, vol. 429, no. 1, pp. 1–15, 2024, doi: 10.1016/j.conbuildmat.2024.136429.
- [11] H. Fan et al., "Rockfill material segmentation and gradation calculation based on deep learning," *Case Stud. Constr. Mater.*, vol. 17, no. 1, pp. 1–12, Dec. 2022, doi: 10.1016/j.cscm.2022.e01216.
- [12] J. Zhao, D. Li, and Y. Yu, "Identification of rock fragments after blasting by using deep learning-based segment anything model," *Minerals*, vol. 14, no. 7, pp. 1–15, Jul. 2024, doi: 10.3390/min14070654.
- [13] R. Kanke and M. Takahashi, "Amodal instance segmentation of thin objects with large overlaps by seed-to-mask extending," *IEICE Trans. Inf. Syst.*, vol. E107-D, no. 7, pp. 908–911, Jul. 2024, doi: 10.1587/transinf.2023EDL8068.
- [14] M. Tran, W. Bounsavy, K. Vo, A. Nguyen, T. Nguyen, and N. Le, "ShapeFormer: Shape prior visible-to-amodal transformer-based amodal instance segmentation," *Proc. Int. Joint Conf. Neural Netw.*, vol. 2024, no. Jun., pp. 1–8, Jun. 2024, doi: 10.1109/IJCNN60899.2024.10650837.
- [15] R. Mohan and A. Valada, "Perceiving the invisible: Proposal-free amodal panoptic segmentation," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 9302–9309, Oct. 2022, doi: 10.1109/LRA.2022.3189425.
- [16] L. Qi, L. Jiang, S. Liu, X. Shen, and J. Jia, "Amodal instance segmentation with KINS dataset," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 2019, no. Jun., pp. 3009–3018, Jun. 2019, doi: 10.1109/CVPR.2019.00313.

-
- [17] Y. Hu, J. Wang, X. Wang, and T. Guan, "Occlusion-aware particle size distribution detection of gravel material based on the improved bilayer convolutional network," *Constr. Build. Mater.*, vol. 404, no. 1, pp. 1–12, Nov. 2023, doi: 10.1016/j.conbuildmat.2023.133263.
- [18] Y. Hu, J. Wang, X. Wang, J. Yu, and J. Zhang, "Efficient virtual-to-real dataset synthesis for amodal instance segmentation of occlusion-aware rockfill material gradation detection," *Expert Syst. Appl.*, vol. 238, no. 1, pp. 1–15, Apr. 2024, doi: 10.1016/j.eswa.2023.122046.
- [19] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017, no. Oct., pp. 2980–2988, Oct. 2017, doi: 10.1109/ICCV.2017.322.
- [20] L. Ke, Y.-W. Tai, and C.-K. Tang, "Deep occlusion-aware instance segmentation with overlapping bilayers," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 2021, no. Jun., pp. 4018–4027, Jun. 2021, doi: 10.1109/CVPR46437.2021.00401.
- [21] K. Li and J. Malik, "Amodal instance segmentation," *Computer Vision—ECCV 2016*, vol. 9906, no. 1, pp. 677–693, Oct. 2016, doi: 10.1007/978-3-319-46475-6_42.
- [22] Y. Zhu, Y. Tian, D. Metaxas, and P. Dollár, "Semantic amodal segmentation," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2017, no. Jul., pp. 3001–3009, Jul. 2017, doi: 10.1109/CVPR.2017.320.
- [23] K. Ehsani, R. Mottaghi, and A. Farhadi, "SeGAN: Segmenting and generating the invisible," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 2018, no. Jun., pp. 6144–6153, Jun. 2018, doi: 10.1109/CVPR.2018.00643.
- [24] E. Ozguroglu et al., "pix2gestalt: Amodal segmentation by synthesizing wholes," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 2024, no. Jun., pp. 3931–3940, Jun. 2024, doi: 10.1109/CVPR52733.2024.00377.
- [25] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *Computer Vision—ECCV 2020*, vol. 12346, no. 1, pp. 213–229, Aug. 2020, doi: 10.1007/978-3-030-58452-8_13.
- [26] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 2022, no. Jun., pp. 1280–1289, Jun. 2022, doi: 10.1109/CVPR52688.2022.00135.
- [27] Z. Wang, H. Zheng, P. He, W. Chen, and M. Zhou, "Diffusion-GAN: Training GANs with diffusion," *arXiv*, vol. 2022, no. Jun., pp. 1–10, Jun. 2022, doi: 10.48550/arXiv.2206.02262.
- [28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.
- [29] Z. Li, W. Ye, T. Jiang, and T. Huang, "2D amodal instance segmentation guided by 3D shape prior," *Computer Vision—ECCV 2022*, vol. 13689, no. 1, pp. 165–181, Oct. 2022, doi: 10.1007/978-3-031-19818-2_10.
- [30] G. Zhan, C. Zheng, W. Xie, and A. Zisserman, "Amodal ground truth and completion in the wild," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 2024, no. Jun., pp. 28003–28013, Jun. 2024, doi: 10.1109/CVPR52733.2024.02645.
- [31] S. Raghavendra, D. Rao, S. K. Abhilash, V. M. Nookala, and P. G. Bharathi, "Elevating amodal segmentation using ASH-Net architecture for accurate object boundary estimation," *IEEE Access*, vol. 11, no. 1, pp. 83377–83389, 2023, doi: 10.1109/ACCESS.2023.3301724.
- [32] J. Gao et al., "Coarse-to-fine amodal segmentation with shape prior," *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, vol. 2023, no. Oct., pp. 1262–1271, Oct. 2023, doi: 10.1109/ICCV51070.2023.00122.
- [33] J. Wei et al., "Mixture-of-Shape-Experts (MoSE): End-to-end shape dictionary framework to prompt SAM for generalizable medical segmentation," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, vol. 2025, no. Jun., pp. 6450–6460, Jun. 2025, doi: 10.1109/CVPRW67362.2025.00642.
- [34] X. Zhao, J. Xie, S. Chen, and J. Liu, "Convex combination star shape prior for data-driven image semantic segmentation," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 2025, no. Jun., pp. 14068–14077, Jun. 2025, doi: 10.1109/CVPR52734.2025.01313.
- [35] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, "Boundary IoU: Improving object-centric image segmentation evaluation," *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 2021, no. Jun., pp. 15329–15337, Jun. 2021, doi: 10.1109/CVPR46437.2021.01508.