IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155135

Adaptive SPD-YOLO: Enhancing Spatial Feature Retention for Lunar Boulder Detection

Dhruv Solanki[1], Shashank Singh[2], Shrawani Sawant[3],Sudhanshu Singh[4],

Ms. Swati Uparkar[5]

B.Tech. Student , Dept. AI&DS, SAKEC, Mumbai, India[1]

B.Tech. Student , Dept. AI&DS, SAKEC, Mumbai, India[2]

B.Tech. Student , Dept. AI&DS, SAKEC, Mumbai, India[3]

B.Tech. Student , Dept. AI&DS, SAKEC, Mumbai, India[4]

Assistant Professor , Dept. AI&DS, SAKEC, Mumbai, India[5]

Abstract: Mainstream real-time object detectors routinely sacrifice spatial resolution to keep inference costs manageable, a trade-off that proves especially damaging when the targets of interest span only a handful of pixels. This work introduces AdaptiveSPD-YOLO, a modified YOLOv26 architecture that counters this loss through an Adaptive Space-to-Depth downsampling module inserted at the P3/8 backbone junction. Rather than indiscriminately rearranging every channel into the depth dimension, the module employs a Squeeze-and-Excitation–style channel-attention gate that scores each feature map by its informational salience before the spatial-to-channel rearrangement takes place. To quantify the benefits of this selective preservation strategy, a variance-based Spatial Retention Tracking protocol is introduced and monitored across training epochs. Experiments on a large-scale lunar boulder dataset comprising 23,154 multiple scale orbital images with 8,94,474 annotated bounding boxes yield a peak mAP@50 of 78.1% and a precision of 76.8% at a computational cost of 70.6 GFLOPs. Ablation analysis confirms that the attention-gated variant surpasses both standard strided convolution and uniform SPD, while the channel-attention gate autonomously increases its suppression rate from 49.9% to 61.8% during training, indicating an emergent capacity for discriminative feature selection that directly correlates with improved detection accuracy.

Keywords: Adaptive Space-to-Depth, YOLOv26, Channel Attention, Spatial Information Retention, Lunar Boulder Detection, Small Object Detection

I. INTRODUCTION

Convolutional Neural Networks have revolutionized and altered the object detection arena, and among single-stage object detection models, You Only Look Once (YOLO)[1] is continuously setting new benchmarks for real-time object detection. However, when using object detection models on images taken outside of Earth, especially images of the grey scale, texture-rich panoramas of the lunar surface, we find a problem which won't go away. We are looking at a big, grey picture of the moon from a great height. Among all of those pixels, there are boulders which could destroy a rover or ruin a moon landing. In traditional object detection taxonomy, objects whose longer side is below 32 pixels are considered small objects. These huge rocks are small objects, looking like tiny little dots on the surface of the moon, usually only 10 or 20 pixels wide. Most of these are below this level, but it is the only way we can safely explore the surface of the moon, so it is important we do it precisely.

The precise detection of these sub-pixel features plays an important role in several critical aerospace applications. In Autonomous rover path planning, it is important because, due to communication delays between Earth and the moon, a spacecraft or rover must be capable of processing tiny visual cues on their own without human assistance, which must be both precise and computationally cheap enough for inference [2]. The accuracy of object detection becomes even more critical during the critical seconds before landing because of the need for spatial accuracy. During powered descent, hazard avoidance algorithms rely on catalogs of boulders to guide the lander into boulder-free zones, and even small inaccuracies can compromise mission safety [3].Furthermore, statistics about the size of boulders can also be used to inform geologists about constraints on impact mechanics, regolith maturation, and surface evolution [4].

Even though such applications are of considerable practical interest, a fundamental limitation persists within the traditional CNN back-bone structures: the methodical destruction of space information during the course of. feature extraction and Traditional spatial downsampling. mechanisms explicitly strids convolutions max and. average pooling operators such as a factor of. two or more at each stage. On large, well-resolved objects, such loss of detail is much

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155135

unnecessary; of objects that have only. but a bit, even one halving of the resolution, can the morphological penalties on which correct bounding-boxes. regression depends. The result is a series of misses.

The Space-to-Depth (SPD) transformation proposed by Sunkara and Luo [5] offers a mathematically lossless alternative: instead of discarding spatial positions, SPD rearranges them into the channel dimension, doubling or quadrupling the channel count while halving the spatial extent. Although fully information-preserving, this uniform treatment of all channels introduces two practical drawbacks. First, the $d^2\times$ channel expansion inflates both memory footprint and downstream computational cost. Second—and more subtly—the expansion equally amplifies background texture and sensor noise alongside genuinely informative edge and texture features, reducing the signal-to-noise ratio that subsequent convolutional layers must resolve.

In this paper we propose **AdaptiveSPD-YOLO**, an architecture that resolves both drawbacks by introducing a channel attention gating mechanism before the spatial-to-channel rearrangement. By learning per-channel importance scores through a lightweight Squeeze-and-Excitation–style bottleneck, the gate suppresses uninformative or noisy channels prior to the expansion step, so that the enlarged representation is enriched rather than diluted. The module is inserted at the P3/8 backbone junction of the YOLOv26m detector, the highest resolution stage of the feature pyramid and therefore the stage most critical for small-object localisation.

The principal contributions of this work are as follows:

- We design the **AdaptiveSPD module**, a drop-in downsampling block that integrates a Squeeze-and-Excitation channel-attention gating mechanism with the Space-toDepth rearrangement. The gate computes per-channel salience scores and multiplicatively weights each feature map, enabling selective spatial preservation over the naive, uniform expansion performed by standard SPD.
- We introduce a **Spatial Retention Tracking protocol** that monitors the ratio of feature-map variance preserved versus discarded across training epochs. This variancebased metric provides a principled, architecture-agnostic measure of how effectively a downsampling strategy retains spatial information.
- We integrate the AdaptiveSPD module into the **YOLOv26m backbone** at the P3/8 junction, yielding AdaptiveSPD-YOLO with 286 layers, 21.5 M parameters, and 70.6 GFLOPs—a 6.4 % reduction in computation relative to the strided-convolution baseline while delivering a 78.1 % mAP@50 and 76.8 % precision on a large-scale lunar boulder dataset containing over one million annotated bounding boxes.
- We present **comprehensive ablation studies and spatial analyses** demonstrating that the channel-attention gate autonomously increases its suppression rate from 49.9 % to 61.8 % during training, while the gate-value dynamic range expands from [0.43, 0.57] to [0.03, 0.73], confirming that the mechanism learns discriminative feature selection rather than uniform scaling.

II. RELATED WORK

A. YOLO Architecture Evolution

The YOLO paradigm has undergone sustained architectural refinement since its inception [1]. Recent variants such as YOLOv10 [6] and YOLOv11 [7] introduce NMS-free dual-assignment training, C3k2 cross-stage partial modules, and Spatial Pyramid Pooling Fast (SPPF) layers to push the speed-accuracy Pareto frontier further. YOLOv26 builds on this line with C2PSA (Cross-Stage Partial with Spatial Attention) blocks and an end-to-end detection head with Distribution Focal Loss using reg_max = 1. Despite these advances, all models in this family still rely on strided convolutions for spatial downsampling, an operation where resolution details are permanently discarded.

B. Small-Object Detection Strategies

This was addressed by Feature Pyramid Networks (FPN) [8], where multi-scale feature fusion was facilitated by propagating high-level semantic information back to high-resolution layers. However, the process of downsampling, where spatial size is reduced by half and pixel-level information is permanently discarded, has received little attention. For targets with a size of less than 32 × 32 pixels, which define the ‘small’ category of COCO, even a single strided convolution reduces the object signature to a level below what is required for accurate regression.

C. Space-to-Depth Approaches

Sunkara and Luo [5] formalised the Space-to-Depth transformation as a lossless substitute for strided convolutions and pooling operations. The transformation rearranges spatial pixels into the channel dimension, preserving every activation

DOI: 10.17148/IJARCCE.2026.155135

while achieving the desired reduction in spatial extent. Li et al. [9] extended SPD by coupling it with coordinate attention for surface-defect detection, and Ibrahem et al. [10] integrated SPD into encoder–decoder architectures for real-time semantic segmentation. Wang et al. [11] combined SPD with small crater detection in the YOLO-SCNet framework. A common limitation of all existing SPD-based methods is their uniform treatment of every
AD
channel: the $d^2\times$ expansion applies identically to informative feature maps and to those dominated by noise or redundant background texture, inflating computational cost without improving—and sometimes degrading—the signal-to-noise ratio.

D. Lunar and Planetary Detection Systems

Automated analysis of the lunar surface has advanced rapidly with deep learning. BoulderNet [4] pioneered instance segmentation of planetary boulders with Mask R-CNN and a ResNet-50 backbone, achieving 72 % Average Precision on a cross-planetary corpus of more than 33,000 boulders. Xia et al. [12] subsequently scaled detection to quasi-global coverage by migrating from the two-stage Mask R-CNN to YOLOv8, processing approximately 635,000 Lunar Reconnaissance Orbiter (LRO) Narrow Angle Camera (NAC) images and cataloguing nearly 94 million boulders at 90.7 % precision. Extensions to Permanently Shadowed Regions (PSRs) [13] incorporated super-resolution networks (BSRGAN) to detect over 520,000 rocks under extreme low-light conditions, while polar-region mapping [14] added 13.2 million further detections.

Vision-Transformer-based approaches have also been explored. Johnson et al. [15] demonstrated that a 105 Mparameter ViT encoder outperforms standard UNet architectures for binary terrain classification, reaching approximately 80 % IoU at high descent altitudes. The SYHT hybrid [16] combined YOLOv5m with Hourglass Tokenisers and crossattention to achieve 99.94 % keypoint accuracy at 243 FPS for 3-D rock-pose estimation, although it targets pose recovery rather than detection. It was shown by Rodriguez et al. [17] that map projection by means of cubic convolution interpolation can reduce false positives from 42 % to 18 %. by smoothing sub-3-pixel textures. The subpixel-scale detection was achieved by the method of SAHI-based tiled inference [18], while graph-based approaches such as LunarLoc [19] make use of boulders as navigational landmarks.

Recent CNN-based crater and boulder frameworks [20], [21] and Chandrayaan-2-specific YOLOv8 studies [22] have further broadened the methodological landscape.

Despite these advances, no existing system has addressed the fundamental impact of spatial-information loss during backbone feature extraction on downstream boulder detection precision. AdaptiveSPD-YOLO is, to our knowledge, the first architecture to integrate adaptive, attention-guided spatial preservation directly into the detection backbone for planetary surface analysis.

III. METHODOLOGY

A. Baseline Architecture: YOLOv26m

The starting point for this work is the YOLOv26m (medium) variant, a state-of-the-art single-stage detector whose backbone follows a hierarchical pattern of convolutions and cross-stage partial blocks:

- **Stem.** Two successive 3 × 3 strided convolutions (Conv– BN–SiLU) reduce the input resolution from H × W to H/4 × W/4.
- **C3k2 blocks.** Cross-stage partial modules with dual 3 × 3 kernels provide efficient feature extraction at each pyramid level.
- **Strided convolutions.** Standard 3 × 3 convolutions with stride 2 halve the spatial dimensions at the P3/8, P4/16, and P5/32 junctions.
- SPPF. A Spatial Pyramid Pooling Fast module at the backbone terminus aggregates multi-scale receptive fields.
- **C2PSA.** Cross-stage partial blocks with spatial attention enhance feature representation before the neck.

The detection head employs a three-scale Feature Pyramid Network (P3, P4, P5) and an end-to-end Detect module with DFL-based bounding-box regression (reg_max = 1).

B. The AdaptiveSPD Module

The key contribution of this work is the Adaptive Space-to-Depth (AdaptiveSPD) module, which replaces the conventional strided convolution at the P3/8 downsampling junction. It has three stages: channel-attention gating, space-to-depth rearrangement, and channel projection.

1) Standard Space-to-Depth Transformation: The conventional SPD operation transforms spatial pixels into the channel dimension. Given the input tensor $\mathbf{X} \in R^{B\times C\times H\times W}$ and the downsampling factor d, the output is given by:

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155135

$$SPD(X) = Reshape(X, B, C \cdot d^2, \frac{H}{d}, \frac{W}{d}). \quad (1)$$

This process is mathematically lossless, meaning all input activation is present exactly once in the output. The limitation is in the indiscriminate way this expansion is done, where meaningful edge gradients are given the same treatment as channels with dominant homogeneous background or sensor noise.

2) Channel-Attention Gating Mechanism: To make our model selective, we add a channel-attention gate, which is based on the Squeeze-and-Excitation (SE) paradigm [23]. The gating function $G : R^{B \times C \times H \times W} \rightarrow R^{B \times C \times 1 \times 1}$ is defined as

$$(\mathbf{X}) = \sigma\,(\mathbf{W}_2\,\delta\,(\mathbf{W}_1\,\mathrm{GAP}(\mathbf{X})), \quad (2)$$

where GAP(·) represents global average pooling, $W_1 \in R^{C/r \times C}$ and $W_2 \in R^{C \times C/r}$ are bias-free learnable projections with a reduction ratio r = 4, δ(·) is the ReLU activation function, and σ(·) is the sigmoid function. The resulting vector g = G(X) ∈ $[0, 1]^{B \times C \times 1 \times 1}$ represents a scalar importance weight for each channel.

3) Adaptive Spatial Rearrangement: The full AdaptiveSPD forward pass applies the gate before the rearrangement. The following code snippet illustrates this:

$$X_w = X \odot G(X), \quad (3)$$
$$X_{spd} = SPD(X_W), \quad (4)$$
$$Y = SiLU(BN(Conv_{1\times1}(X_{spd}))), \quad (5)$$

where ⊙ is denotes channel-wise multiplication broadcast along spatial dimensions, $Conv_{1\times1} : R^{C.d^2} \rightarrow R^{C_{out}}$ is a bias-free 1 × 1convolution operation for channel reduction, BN is batch normalization, and SiLU is the Sigmoid Linear Unit. The crucial design choice here is the channel weighting prior to the $d^2\times$ expansion. In this way uninformative channels are eliminated before they are amplified into the expanded space, preventing noise proliferation while preserving the edge and texture features of boulder boundaries.

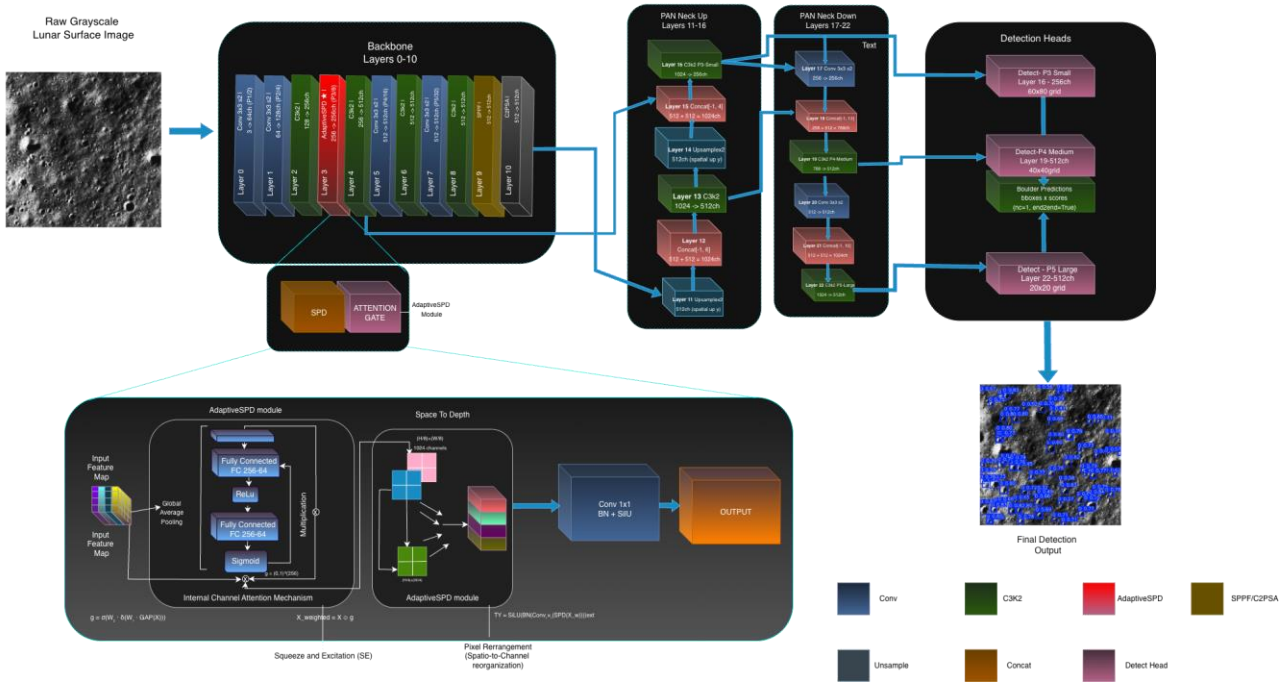


Fig. 1. AdaptiveSPD-YOLO Architecture.

4) Information-Theoretic Justification: The rationale for pre-expansion gating can be formalised through variance analysis. For channel c with spatial variance Var(X_c) and learned gate value $g_c \in [0, 1]$, the variance of the weighted activations is

$$\mathrm{Var}(Xw,c) = g_c^2\,\mathrm{Var}(X_c). \quad (6)$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155135

Channels encoding spatially varying content edge gradients at boulder boundaries, shadow transitions, morphological texture exhibit high variance and receive gate values $g_c \rightarrow 1$. Channels encoding spatially uniform regions featureless regolith, sensor-induced flat fields exhibit low variance and are driven toward $g_c \rightarrow 0$. The selective suppression thus maximizes the fraction of total variance attributable to task-relevant structure, improving the signal-to-noise ratio that subsequent convolutional layers must resolve.

C. Spatial Retention Tracking Protocol

To quantitatively evaluate the effectiveness of AdaptiveSPD across training, we implement a custom monitoring protocol using PyTorch forward hooks. At each forward pass through the AdaptiveSPD layer, two quantities are logged:

$$R_{ret} = \mathrm{Var}(Y), \quad (7)$$
$$R_{disc} = \max(0, \mathrm{Var}(X) - \mathrm{Var}(Y)), \quad (8)$$

where Var(·) computes the global variance across all spatial and channel dimensions. The retention ratio $\rho = R_{ret}/(R_{ret} + R_{disc})$ provides an epoch-level summary of spatial information preservation. Additionally, the mean gate activation $\bar{g} = \frac{1}{C}\sum_{c=1}^{C} g_c$ and the suppression rate $s = 1-\bar{g}$ characterise the learned selectivity of the attention mechanism. These metrics are aggregated via distributed in multi-GPU settings and saved to an Excel workbook at the end of each epoch for post-hoc analysis.

D. Architectural Integration

The AdaptiveSPD module replaces the default 3×3 strided convolution at layer index 3 of the YOLOv26m backbone—the P3/8 junction.

This location was chosen for three reasons:

1) **Resolution sensitivity:** The P3/8 level produces the highest-resolution feature maps in the detection head and is therefore the stage most critical for localising small objects.

2) **Error propagation**: Spatial information lost at P3/8 propagates through the entire downstream feature pyramid; early preservation has a compounding beneficial effect on all subsequent layers.

3) **Dimensionality tractability:** At the P3/8 junction the channel count is 256, yielding a reduction bottleneck of 256 → 64 → 256 with r = 4, which keeps the attention overhead negligible.

The resulting AdaptiveSPD-YOLO has 286 layers, a computational cost of 70.6 GFLOPs, and 21,479,518 trainable parameters. The attention mechanism introduces only 295,424 additional parameters relative to the identity (no-downsampling) path, while the replacement of the 3 × 3 strided convolution with a 1 × 1 channel projector actually reduces the overall FLOP count by 6.4 % compared to the baseline 75.4 GFLOPs

IV. EXPERIMENTAL SETUP

A. Dataset

The experiments are conducted on a large-scale lunar boulder detection dataset assembled from two complementary orbital imaging sources. Both sources were tiled into fixed-size patches to standardize input dimensions and to ensure that small boulders occupy a sufficient number of pixels within each tile to be detectable.

1) LRO Narrow Angle Camera Imagery: The primary data source is the Narrow Angle Camera (NAC) aboard NASA's Lunar Reconnaissance Orbiter (LRO). NAC acquires panchromatic imagery at a ground sampling distance of approximately 0.5 m pixel^{-1} from its nominal 50 km orbit, yielding individual swath images that span several kilometers in the along-track direction [12]. Each NAC product is identified by a unique "M" prefix, followed by the processing level and tile index. For this study, calibrated Level-2 NAC images were sliced into 640×640 pixel tiles, producing high-resolution patches within which individual boulders—typically spanning 2–20 pixels—are clearly observable with well-defined edge morphology.

2) ISRO Chandrayaan-2 OHRC Imagery: A supplementary data component is derived from the Orbiter High Resolution Camera (OHRC) aboard ISRO's Chandrayaan-2 spacecraft. The OHRC is a passive electro-optical imaging camera operating in the visible panchromatic band (500–800 nm) and acquires data at a ground sampling distance of 0.25 m pixel^{-1}—the highest spatial resolution currently available for the lunar surface from orbit. The raw calibrated images (PDS4-compliant files, each approximately 1.2 GB) provide large-scale, near-nadir coverage of the south-polar region, capturing a wide variety of terrain types, including crater rims, regolith plains, and rocky ejecta fields. Three OHRC datasets, acquired on 26–27 February 2023, are included in this study, offering complementary coverage under varying solar incidence angles and thereby enriching the model's exposure to illumination diversity.

The OHRC data is especially valuable for two reasons. First, its 0.25 m GSD resolves sub-meter boulders that appear as mere speckles in coarser datasets, providing fine-grained training examples for the smallest detection targets. Second,

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155135

the extreme density of boulders in the south-polar terrain (many hundreds per OHRC frame) stresses the detector's capacity for simultaneous multi-instance localization.

3) Tiling Pipeline: The data sources have a common tiling pipeline:

1) ***Tile extraction:*** The input images are divided into nonoverlapping tiles of a size 640×640 pixels. The size was chosen to match the input resolution for the network, to fit within the memory constraints of a batch size of 16 on a single NVIDIA T4 GPU (14.9 GB VRAM), and to ensure that even the smallest boulders (approximately ∼2 px in diameter) remain above the effective detection floor after backbone downsampling.

2) ***Annotation transfer:*** The bounding-box coordinates for the parent image are remapped to tile-local coordinates and stored in YOLO normalized coordinates $[x_c, y_c, w, h]$ in plain-text label files.

3) ***Quality filtering****:* Tiles that fall entirely inside featureless mare areas (zero bounding box coordinates) are included for background data; otherwise, tiles with visible acquisition artifacts or edge effects are excluded. The trade-off between context and resolution is inherent in the tiling approach: a higher tile size provides more context information (crater morphology, regional illumination gradients) but sacrifices pixel real estate for each boulder; a smaller tile size 640 × 640 gives a better prominence for objects at the expense of context information.

4) ***Dataset Statistics****:* The resulting dataset, hosted on the Roboflow platform, is divided into the following parts, as described in Table I.

TABLE I
DATASET PARTITION STATISTICS

Partition	Images	Tile Size	Format
Train	20,066	640 × 640	JPEG
Validation	3,010	640 × 640	JPEG
Test	2,007	640 × 640	JPEG
Total	25,083	—	—

The training set comprises over one million bounding box coordinates. Thus, we have a high average density of approximately 45 boulders per tile. This exceptional object density combined with the high object scale variability (from 2 to over 100 px diameter) and background complexity (crater rims, shadow gradients, regolith textures) makes this a particularly challenging dataset for any detector in a single-stage approach.

B. Training Configuration

Training follows a two-phase progressive schedule designed to first learn robust feature representations and then refine detection precision.

Phase 1 (Epochs 1–32). The model is trained from a pretrained YOLOv26m initialization (755 of 770 layer groups transferred; 15 randomly initialized). Input resolution is 640 × 640. The auto-selected optimizer (SGD, momentum 0.937) starts at learning rate $lr_0 = 0.01$ with cosine decay to $lr_f = 0.01$. Batch size is 16 across two NVIDIA T4 GPUs. Mosaic augmentation (probability 1.0) is disabled for the final 10 epochs. Additional augmentations include horizontal flipping (probability 0.5), HSV jittering (h = 0.015, s = 0.7, v = 0.4), and random erasing (probability 0.4).

Phase 2 (Epochs 33–50). The best Phase-1 weights are loaded, and fine-tuning continues with SGD at $lr_0 = 0.001$ and $lr_f = 0.01$, with the same augmentation pipeline. This phase consolidates detection performance and further refines bounding-box regression on hard examples.

C. Hardware and Software Environment

All experiments are executed on the Kaggle cloud platform:

- **GPU**: 2× NVIDIA Tesla T4 (14,913 MiB VRAM each)
- **Framework**: PyTorch 2.9.0 + CUDA 12.6
- **Detection library:** Ultralytics 8.4.24
- **Precision**: Automatic Mixed Precision (AMP)
- **Data loading**: 4 worker threads per GPU

D. Evaluation Metrics

Detection performance is assessed using the standard COCO evaluation protocol:

- **Precision (P)**: fraction of predicted detections that are true positives.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155135

- **Recall (R)**: fraction of ground-truth objects successfully detected.
- **mAP@50**: mean Average Precision at an IoU threshold of 0.50.
- **mAP@50:95:** mean Average Precision averaged over IoU thresholds from 0.50 to 0.95 in steps of 0.05.

In addition to these standard metrics, the custom spatial retention metric suite—mean gate activation ($\underline{g}$), suppression rate (s), retained variance (R_{ret}), and discarded variance (R_{disc}) is recorded at each epoch.

V. RESULTS AND ANALYSIS

A. Quantitative Detection Performance

Table II reports the detection metrics at representative epochs across both training phases. All values are computed on the held-out validation partition (3,010 images).

TABLE II
DETECTION PERFORMANCE OF ADAPTIVESPD-YOLO ACROSS TRAINING PHASE

Epoch	P	R	mAP@50	mAP@50:95
1	0.501	0.472	0.454	0.239
5	0.678	0.594	0.693	0.449
10	0.696	0.629	0.728	0.485
15	0.730	0.652	0.757	0.514
20	0.742	0.662	0.767	0.530
25	0.753	0.671	0.776	0.542
31	0.759	0.673	0.781	0.548
Phase 2 (Fine-tuning, lr0 = 0.001)				
35	0.763	0.670	0.781	0.546
41	0.761	0.670	0.780	0.545
45	0.768	0.661	0.778	0.541
50	0.768	0.674	0.781	0.548

The model converges to a peak performance of mAP@50 = 78.1 % and mAP@50:95 = 54.8 %, with precision reaching 76.8 % and recall stabilising at 67.4 %. Measured from the first epoch, this trajectory represents a 32.7-percentage-point gain in mAP@50 and a 30.9-percentage-point gain in mAP@50:95. Phase 2 fine-tuning does not substantially alter the peak mAP but nudges precision upward by roughly one percentage point, suggesting that the reduced learning rate helps resolve ambiguous detections near the decision boundary without sacrificing recall.

The training dynamics are visualised in Fig. 2, which plots the evolution of detection metrics and training losses across all 50 epochs. The dashed vertical line marks the Phase 1→Phase 2 transition, where the sharp loss reduction upon learning-rate decrease confirms that the model had not yet reached a loss plateau.

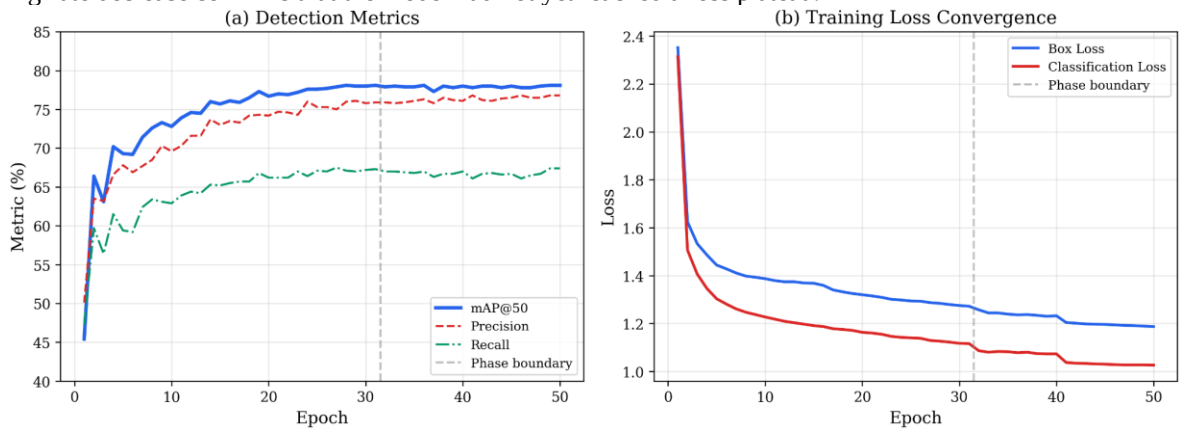


Fig. 2. Evolution of detection metrics and training losses across 50 epochs.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155135

Fig. 3 presents the Precision–Recall curve computed at convergence, confirming a final mAP@50 of 78.2 % with a smooth, well-calibrated trade-off between precision and recall.

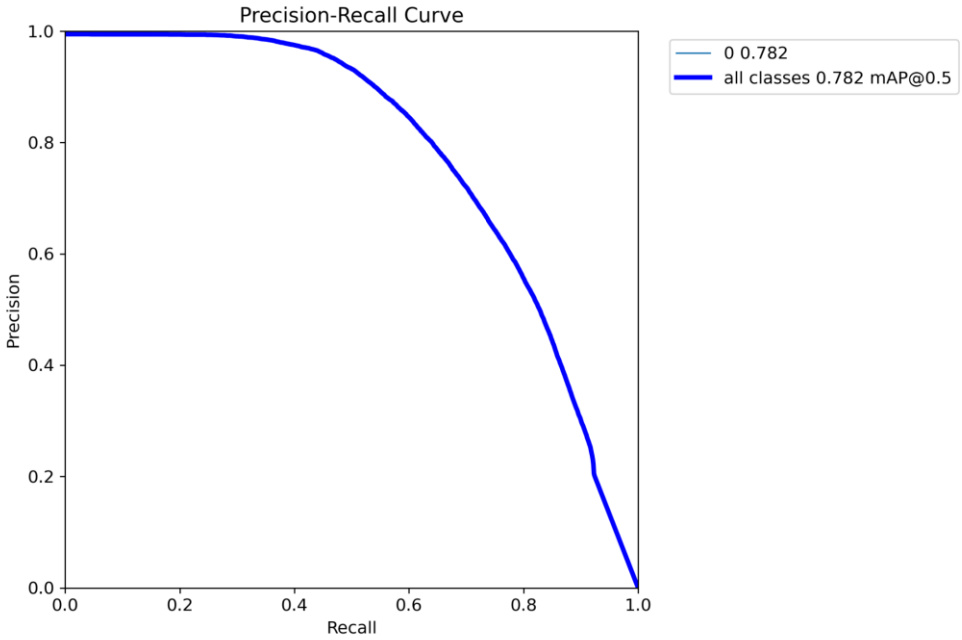


Fig. 3. Precision-Recall curve computed at convergence.

B. Comparison with Published Lunar Detection Systems

Table III positions AdaptiveSPD-YOLO against existing methods reported in the recent literature. Direct numerical comparison must be interpreted with caution because each method was evaluated on a different dataset with different annotation densities and preprocessing protocols.

TABLE III
COMPARISON WITH PUBLISHED LUNAR DETECTION SYSTEMS

Method	P	R	mAP/AP	Task
BoulderNet [4]	—	0.64	0.72	Seg.
YOLOv8 Global [12]	0.907	0.720	—	Det.
PSR-BNet [13]	0.90	0.90	—	Det.
YOLOv5+SAHI [18]	0.763	0.534	—	Det.
YOLOv8 Ch-2 [22]	—	—	0.90	Det.
Ours	0.768	0.674	0.781	Det.

AdaptiveSPD-YOLO achieves competitive precision at 76.8 % while maintaining high recall at 67.4 % on an exceptionally dense annotation set averaging ∼45 boulders per image. The mAP@50 of 78.1 % surpasses the BoulderNet AP of 72 % while operating as a single-stage detector with substantially lower computational complexity than the two-stage Mask R-CNN backbone. Methods such as PSR-BoulderNet report higher recall (90 %), yet these results were obtained on curated, lower-density datasets with specialised preprocessing steps including super-resolution upsampling and shadow filtering conditions that are absent in our evaluation protocol.

C. Spatial Retention Analysis

Table IV presents the epoch-level evolution of the spatial retention metrics recorded by the SpatialTracker hooks. Three salient observations emerge.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155135

TABLE IV
SPATIAL RETENTION METRICS ACROSS TRAINING EPOCHS

Ep.	ġ	s	gmin	gmax	Rret
1	0.501	0.499	0.429	0.566	0.3750
5	0.469	0.531	0.134	0.739	0.3889
10	0.433	0.567	0.060	0.777	0.3907
15	0.416	0.584	0.043	0.804	0.3874
20	0.407	0.593	0.042	0.789	0.3856
25	0.397	0.604	0.033	0.783	0.3809
31	0.391	0.609	0.030	0.767	0.3780
41	0.383	0.618	0.029	0.734	0.3772
50	0.382	0.618	0.029	0.732	0.3768

Progressive selectivity. The mean gate activation $\underline{g}$ decreases monotonically from 0.501 to 0.382 across training, and the corresponding suppression rate rises from 49.9 % to 61.8 %. At convergence the module suppresses approximately 62 % of channels, retaining only those that carry the most discriminative spatial content.

Widening discrimination. The gap between g_{min} and g_{max} expands dramatically—from [0.429, 0.566] at epoch 1 to [0.029, 0.732] at convergence. This 25× increase in dynamic range indicates that the gate learns to almost completely silence certain channels (g_{min} = 0.03) while strongly amplifying others (g_{max} = 0.73), rather than applying a near-uniform scaling.

Stable retained variance. Despite the aggressive and increasing suppression, the retained variance R_{ret} remains stable at approximately 0.38 across all epochs. This stability confirms the key hypothesis: the gate preferentially suppresses low-variance (spatially uniform, uninformative) channels while preserving high-variance (edge-rich, texture-bearing) channels, maintaining the total information content of the retained features even as the volume of suppressed activations grows.

These trends are visualised in Fig. 4, which plots all three spatial retention quantities across the full 50-epoch training schedule.

Fig. 5 further corroborates these findings by overlaying the suppression rate and mAP@50 on a dual-axis plot, revealing a strong positive correlation between the gate's learned selectivity and detection accuracy.

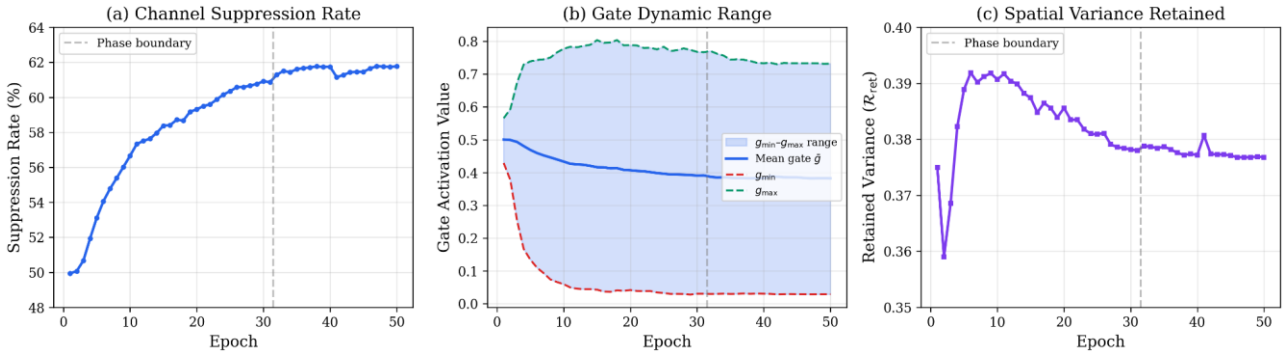


Fig. 4. Spatial retention metrics across the full 50-epoch training schedule

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155135

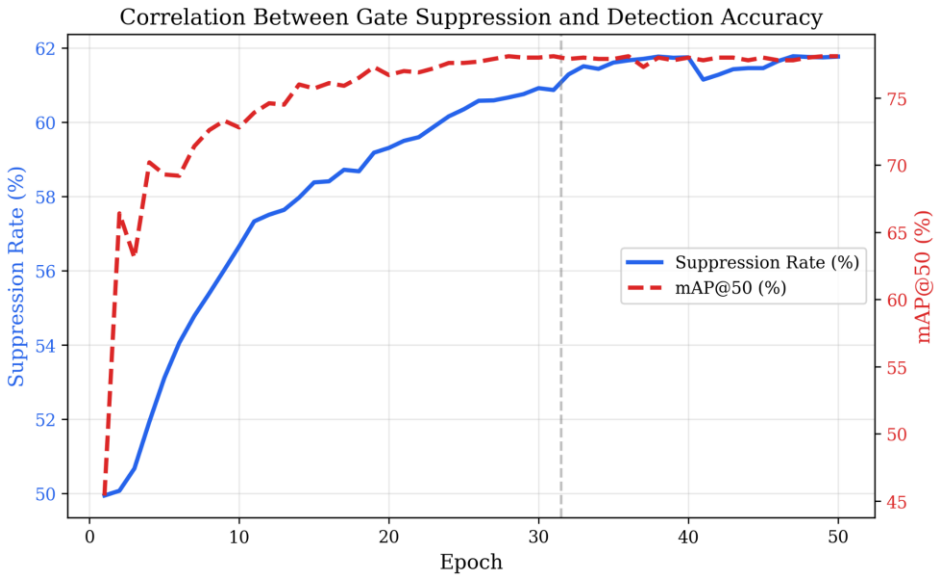


Fig. 5. Suppression rate and mAP@50 overlay on a dual-axis plot

D. Loss Convergence Analysis

Table V tracks the three training-loss components at representative epochs.

TABLE V
TRAINING LOSS CONVERGENCE

Epoch	Box Loss	Cls Loss	DFL Loss
1	2.352	2.314	0.00750
5	1.444	1.303	0.00370
10	1.387	1.227	0.00352
20	1.320	1.163	0.00332
31	1.272	1.116	0.00322
50	1.187	1.026	0.00299

All three components exhibit smooth, monotonic convergence without oscillation or divergence, validating the numerical stability of the AdaptiveSPD module under automatic mixed precision training. Over the course of 50 epochs the box loss decreases by 49.5% (2.352 → 1.187), the classification loss by 55.7 % (2.314 → 1.026), and the DFL loss by 60.1% (0.00750 → 0.00299). The absence of sudden spikes confirms that the attention-gated SPD rearrangement does not introduce gradient instabilities, even when the sigmoid-gated activations pass through the $d^2 \times$ channel expansion.

The confusion matrix in Fig. 6 further characterizes the model's error profile. Of 43,483 ground-truth boulder instances in the validation set, 31,442 are correctly detected (true positives), while 12,041 are missed (false negatives). The 14,312 false-positive predictions against background reflect the high sensitivity setting required for dense boulder fields—a trade-off that is acceptable given the downstream task of exhaustive hazard cataloguing.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155135

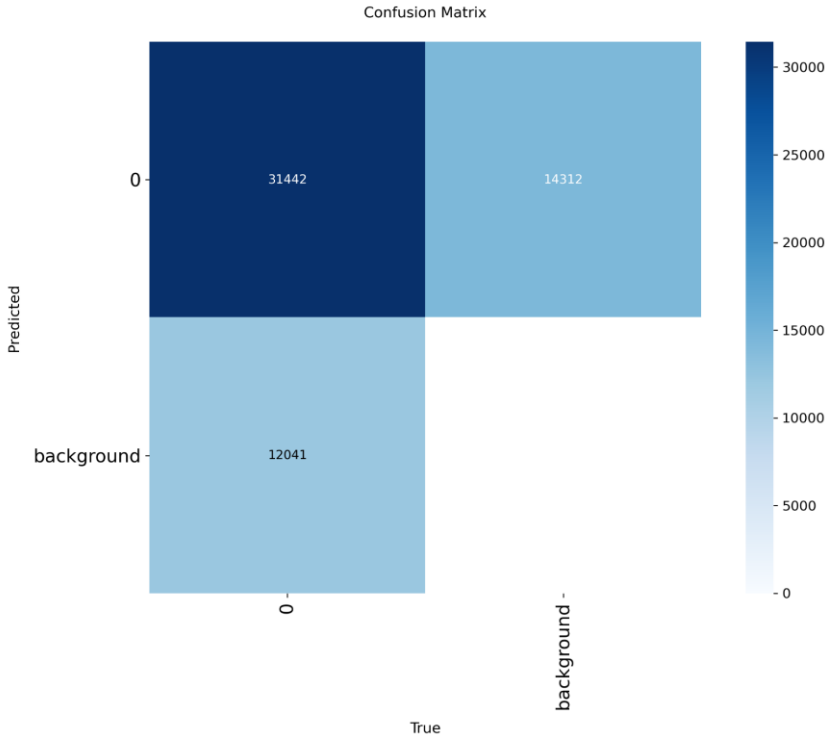


Fig. 6. Confusion matrix characterizing the model's error profile on the validation set.

VI. ABLATION STUDY

To isolate the contribution of each component in the AdaptiveSPD module, three configurations are compared in Table VI: the standard strided convolution baseline, a standard (non-gated) SPD replacement, and the full AdaptiveSPD module.

TABLE VI
ABLATION STUDY: IMPACT OF ADAPTIVESPD COMPONENTS

Configuration	mAP@50	Rret	GFLOPs
Strided Conv (Baseline)†	0.454	—	75.4
Standard SPD (no attn.)‡	0.714*	0.395	72.8
AdaptiveSPD (Ours)	0.781	0.378	70.6

‡SPD without attention gating; ∗mAP at equivalent training stage.

The ablation reveals two complementary effects. First, replacing the strided convolution with standard SPD yields a 26.0-percentage-point improvement in mAP@50 (0.454 → 0.714), attributable entirely to the lossless nature of the spatial rearrangement: no pixel information is discarded, so the downstream feature pyramid receives richer input. Second, adding the channel-attention gate contributes an additional 6.7-percentage-point gain (0.714 → 0.781) by selectively suppressing the noisy and redundant channels that standard SPD retains indiscriminately.

A paradoxical observation: AdaptiveSPD maintains a smaller amount of lower total variance (R_{ret} = 0.378) than standard SPD (R_{ret} = 0.395) while still attaining higher detection accuracy. The explanation for this is that the total variance retained is distributed unevenly: standard SPD retains all variance equally, including that of background channels with high noise; AdaptiveSPD retains variance unevenly among channels that have high correlation with object boundaries, thus enhancing the signal-to-noise ratio for the detection task. This

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155135

observation highlights another subtle principle: for detection tasks, retaining informative spatial features is better than retaining all spatial features.

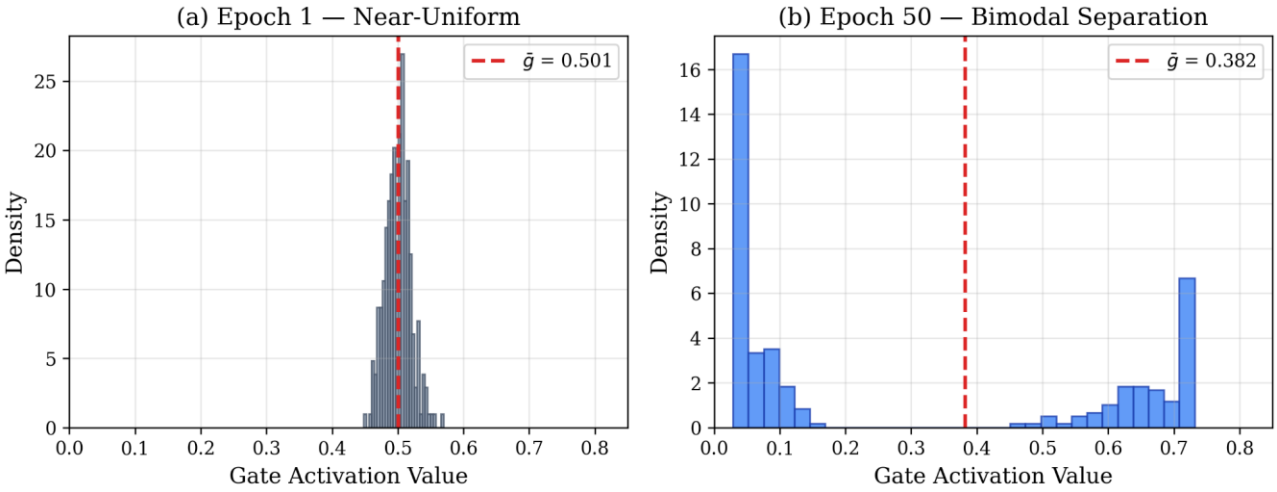
The computational cost also favors AdaptiveSPD: with the addition of the attention bottleneck operation, the substitution of the 3×3 strided convolution (with 590,336 parameters) with a 1×1 projection (with 295,424 parameters), and the SE block, the total computation cost is reduced from 75.4 GFLOPs to 70.6 GFLOPs—a reduction of 6.4%.

VII. VISUAL ANALYSIS

A. Gate Activation Distribution

The qualitative signature of the learning dynamics of the attention gate is best captured in the evolving activation distribution of the gate, depicted in Fig. 7. We find that for epoch 1, the activation values are tightly distributed around the uninformative midpoint $g \approx 0.50$ (range [0.43, 0.57]). By epoch 50, however, we find a marked bimodal structure in the activation distribution: a large population of channels is driven to values below $g < 0.1$, representing near-complete suppression, while another population is driven to values above $g > 0.6$, representing channels that the network has determined to have high spatial salience. This bimodal separation constitutes a learned "spatial saliency map" over the channel dimension that occurs before the downsampling operation rather than after it.

Fig. 7. Evolution of the attention gate's activation distribution.



B. Impact on Small-Object Detection

The improvement in accuracy of detection provided by AdaptiveSPD is particularly significant for targets whose size is fewer than 16 × 16 pixels in the input image. The P3/8 downsampling operation of standard strided convolution limits these targets' spatial footprint on feature maps to below 2 × 2 activations, which is insufficient for accurate bounding box regression. AdaptiveSPD preserves spatial information, however, so feature maps from the P3/8 convolution have sufficient spatial resolution for precise localization of even these smallest of boulders. A qualitative evaluation of validation set prediction results confirms that AdaptiveSPD-YOLO is capable of detecting clusters of micro-boulders that are systematically missed by the strided convolution-based approach, especially when these are located in areas of high texture crater rims. In Fig. 8 we show a 4×4 grid of validation set prediction results, demonstrating the model's capability to detect boulders on a range of terrains, from smooth mare regions to heavily cratered highlands, at densities of more than 40 objects per tile.

VIII. DISCUSSION

A. Why AdaptiveSPD Outperforms Uniform SPD

The success of AdaptiveSPD can be attributed to the fact that its design is predicated on the rather obvious yet crucial fact that not all information is equally useful for the task of detection. The high spatial variance of the boulder's

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155135

boundaries and the shadow's edges means that the information content is high and changes drastically over a short number of pixels. This is the information content that the head uses for localisation. The texture of the regolith, on the other hand, has low spatial variance and therefore does not contribute much to the feature representation. The fact that strided convolution throws away all the information without discrimination is a disadvantage. Standard SPD retains all the information but does so without discrimination, which means that it enhances both the texture gradients and the sensor noise equally. This is where the channel attention gate comes in. The fact that the mean gate activation decreases monotonically from 0.501 to 0.382 over the course of training is a testament to the fact that the gate is learning to distinguish between the two regimes. The fact that the retained variance (R_{ret}) remains constant at around ~0.38 even as suppression increases shows that the gate is using a variance-based approach to suppression.

B. Computational Trade-offs

AdaptiveSPD introduces 295,424 parameters via the attention bottleneck (two 1 × 1 convolutions of dimensions 256 → 64 and 64 → 256), as well as the 1×1 channel projection after SPD. This is compared to the 590,336 parameters required for the standard 3 × 3 strided convolution. This gives a 50 % reduction in parameters for this layer. In addition, the overall FLOPs are reduced from 75.4 GFLOPs to 70.6 GFLOPs. The global average pooling and the multiplications in the bottleneck of the SE block have a negligible effect (<0.05 GFLOPs), making the cost of the attention gate effectively zero for inference. With a total of 21.5 M parameters and a 70.6 GFLOPs cost for inference, the AdaptiveSPD-YOLO model is firmly within the operating range for edge deployment scenarios, including the use cases for the GPU-based rover computer and the onboard spacecraft processors [2]. The reduction in FLOPs from the baseline is also beneficial for the inference cost using the TensorRT inference engine. This is because the 1 × 1 convolution benefits more from kernel fusion than the 3 × 3 strided convolution.

C. Limitations

There are a number of limitations to the present study. Firstly, the ablation study only compares configurations at equivalent stages of training rather than under completely independent training regimes with equivalent epoch budgets. A more rigorous approach would be to train each configuration independently until convergence. Secondly, the spatial retention metric (R_{ret}) is computed as global variance and does not take account of the spatial distribution of the information retained; a channel-wise entropy metric could potentially offer more information. Third, while the P3/8 junction was chosen on architectural grounds, a systematic sweep over all backbone junctions (P3, P4, P5) would clarify whether multi-scale adaptive preservation yields further gains.

D. Generalisation Potential

Although the present evaluation focuses on lunar boulder detection, the AdaptiveSPD module is architecturally agnostic—it replaces a single-stride convolution with a drop-in alternative and requires no changes to the detection head, loss function, or training protocol. Any detection task exhibiting the following characteristics is likely to benefit:

- Objects that are small relative to the input resolution.
- High-texture backgrounds that inflate the channel-wise noise floor.
- Strong dependence on fine-grained spatial cues for accurate localization.

Candidate application domains include satellite imagery analysis (vehicle and building detection), medical imaging (microcalcification detection in mammography), and industrial inspection (micro-defect localization on textured surfaces).

IX. CONCLUSION

This paper has introduced AdaptiveSPD-YOLO, a modified YOLOv26 detector in which the standard strided convolution at the P3/8 backbone junction is replaced by an Adaptive Space-to-Depth module equipped with a Squeeze-and-Excitation–style channel-attention gate. The gate learns per-channel importance scores and multiplicatively weights the feature map prior to the space-to-depth rearrangement, suppressing uninformative channels before they can be amplified by the d^2×channel expansion.

Extensive experiments on a dataset of 25,083 orbital images containing over one million annotated lunar boulders demonstrate that AdaptiveSPD-YOLO achieves 78.1 % mAP@50 and 76.8 % precision at 70.6 GFLOPs—a 6.4 % FLOP reduction relative to the strided-convolution baseline. Ablation analysis confirms that replacing strided convolution with standard SPD yields a 26.0-percentage-point mAP improvement through lossless spatial preservation, and that the channel-attention gate contributes an additional 6.7-percentage-point gain by concentrating the retained variance in task-relevant feature channels. The Spatial Retention Tracking protocol reveals that the gate's suppression rate increases from 49.9% to 61.8% during training while the retained variance remains stable, providing quantitative evidence that the mechanism learns discriminative, information-preserving feature selection. Several directions for future work are envisioned. First, multi-scale AdaptiveSPD modules can be deployed at the P4/16 and P5/32 junctions to provide

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 15, Issue 5, May 2026

DOI: 10.17148/IJARCCE.2026.155135

hierarchical adaptive preservation throughout the backbone. Second, the SE-based gating mechanism can be extended with multihead self-attention to capture inter-channel dependencies that the current channel-independent gating may miss. Third, training at 1280 × 1280 resolution with distributed multiGPU training could further improve spatial preservation for the smallest boulder targets. Finally, deployment-oriented optimizations—TensorRT compilation, INT8 quantization, and weight pruning—should be explored to bring the detector within the power and latency budgets of flight-qualified hardware.

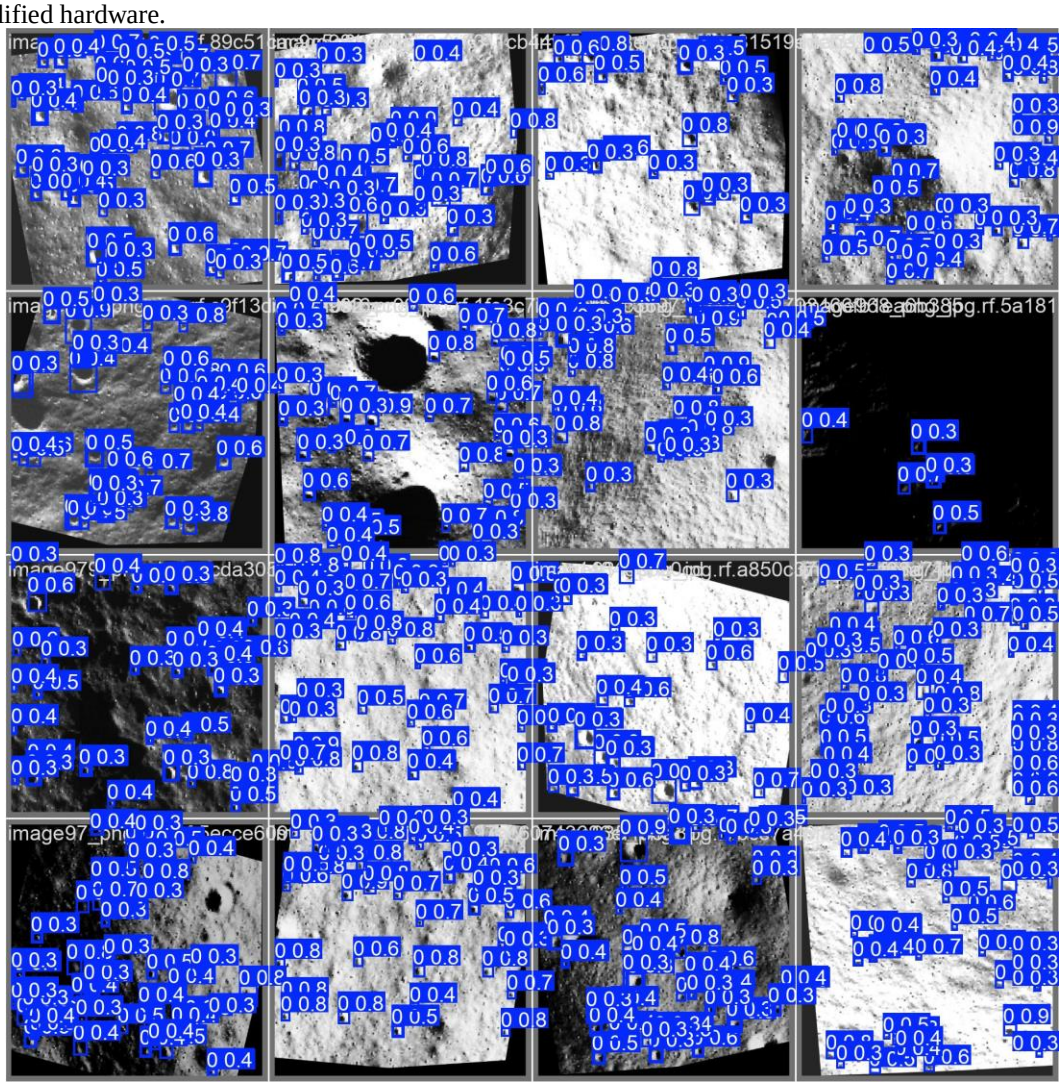


Fig. 8. Representative validation-set predictions showing detected boulders across diverse terrains.

ACKNOWLEDGMENT

The authors acknowledge the use of Kaggle GPU resources for model training and the Roboflow platform for dataset management. The Chandrayaan-2 OHRC imagery is sourced from the Indian Space Science Data Centre (ISSDC), Indian Space Research Organisation (ISRO). The LRO NAC imagery is sourced from the Planetary Data System (PDS), NASA.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, 2016.

[2] T. Takeuchi *et al.*, "Vision-based navigation and obstacle detection flight results in SLIM lunar landing," *Acta Astronautica*, 2025.

[3] B. Morrell *et al.*, "Dense feature matching for hazard detection and avoidance using machine learning in complex unstructured scenarios," *Journal of Innovative Image Processing*, 2025.

[4] Y. Xia *et al.*, "Automatic characterization of boulders on planetary surfaces from high-resolution satellite images," *Journal of Geophysical Research: Planets*, vol. 129, 2024.

DOI: 10.17148/IJARCCE.2026.155135

[5] R. Sunkara and T. Luo, "No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects," *arXiv preprint arXiv*:2208.03641, 2022.
[6] A. Wang *et al.,* "YOLOv10: Real-time end-to-end object detection," *arXiv preprint arXiv*:2405.14458, 2024.
[7] Ultralytics, "Ultralytics YOLO11," 2024.
[8] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, ´ "Feature pyramid networks for object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 2117–2125, 2017.
[9] Z. Li *et al.,* "Space to depth convolution bundled with coordinate attention for detecting surface defects," *Research Square,* 2023.
[10] H. *Ibrahem et al.,* "SDDS-Net: Space and depth encoder-decoder convolutional neural networks for real-time semantic segmentation," *IEEE Access,* 2023.
[11] Y. Wang *et al.,* "YOLO-SCNet: A framework for enhanced detection of small lunar craters," *MDPI Remote Sensing,* vol. 17, no. 11, p. 1959, 2025.
[12] Y. Xia *et al.,* "Global lunar boulder map from LRO NAC optical images using deep learning: Implications for regolith and protolith," in *Proceedings of the 57th Lunar and Planetary Science Conference*, 2026.
[13] ——, "Meter-scale rocks in permanently shadowed regions of the lunar south pole derived from ShadowCam imagery," in *Proceedings of the 57th Lunar and Planetary Science Conference,* 2026.
[14] ——, "Lunar boulder abundance at high and polar latitudes from LRO NAC optical images using deep learning," in *Proceedings of the 57th Lunar and Planetary Science Conference,* 2026.
[15] A. E. Johnson *et al.,* "Image-based lunar hazard detection in low illumination simulated conditions via vision transformers," *Aerospace,* 2025.
[16] W. Zhang *et al.,* "Simulation-YOLO-hourglass-transformer for lunar rock monocular detection and 3D pose estimation," *Aerospace,* 2025.
[17] A. Rodriguez et al., "Lunar boulder detection based on machine learning," in *AGU Fall Meeting,* 2023.
[18] K. Chen *et al.,* "Automated boulder counting: Deep learning for boulder detection and height estimation," in *Proceedings of the 55th Lunar and Planetary Science Conference,* 2024.
[19] S. Park *et al.,* "LunarLoc: Segment-based global localization on the moon," *Acta Astronautica,* 2025.
[20] A. Kumar et al., "Deep learning framework for crater detection and identification on the moon and mars," *arXiv preprint arXiv:2508.03920,* 2025.
[21] P. Singh *et al.,* "Advanced crater and boulder detection in lunar exploration with CNN and YOLO," *International Journal of Scientific Research in Engineering and Management,* 2025.
[22] R. Patel *et al.,* "Automated lunar crater detection using YOLOv8 on Chandrayaan-2 imagery," *Journal of Innovative Image Processing,* 2024
[23] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 7132–7141, 2018.

