

# Innovative Horizons in Aerial Imagery: LSKNet Meets DiffusionDet for Advanced Object Detection

Ahmed Sharshar\*, Aleksandr Matsun\*

Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)  
AbuDhabi, UAE

{ahmed.sharshar, aleksandr.matsun} @mbzuai.ac.ae

## Abstract

*In the realm of aerial image analysis, object detection plays a pivotal role, with significant implications for areas such as remote sensing, urban planning, and disaster management. This study addresses the inherent challenges in this domain, notably the detection of small objects, managing densely packed elements, and accounting for diverse orientations. We present an in-depth evaluation of an object detection model that integrates the Large Selective Kernel Network (LSKNet) [16] as its backbone with the DiffusionDet [2] head, utilizing the iSAID dataset [38] for empirical analysis. Our approach encompasses the introduction of novel methodologies and extensive ablation studies. These studies critically assess various aspects such as loss functions, box regression techniques, and classification strategies to refine the model's precision in object detection. The paper details the experimental application of the LSKNet backbone in synergy with the DiffusionDet heads, a combination tailored to meet the specific challenges in aerial image object detection. The findings of this research indicate a substantial enhancement in the model's performance, especially in the accuracy-time tradeoff. The proposed model achieves a mean average precision (MAP) of approximately 45.7%, which is a significant improvement, outperforming the RCNN model by 4.7% on the same dataset. This advancement underscores the effectiveness of the proposed modifications and sets a new benchmark in aerial image analysis, paving the way for more accurate and efficient object detection methodologies. The code is publicly available at <https://github.com/SashaMatsun/LSKDiffDet>*

## 1. Introduction

Object detection in aerial imaging has emerged as a dynamic and pivotal area of research, focusing on identifying and localizing objects within high-resolution images captured via airborne platforms, such as satellites, drones, or aircraft [3]. This technology finds application in a diverse range of fields, including but not limited to urban planning [40], precision agriculture [26], disaster management [5], and military surveillance [19]. The integration of cutting-edge machine learning methodologies, notably deep learn-

ing and convolutional neural networks [9], enables these object detection models to process extensive aerial datasets efficiently, identifying specific objects like vehicles, buildings, and vegetation [18]. However, this domain faces several challenges, such as dealing with varying image resolutions [43], managing occlusions [10], the necessity for large and accurately annotated training datasets [22], and the real-time processing of high-resolution images [28]. Addressing these challenges is essential for fully unlocking the potential of aerial imaging in object detection, thereby facilitating more effective, data-informed decision-making across a variety of sectors [3].

In this paper, we present a series of innovative contributions that significantly propel the field of aerial image analysis forward. Our comprehensive and multifaceted approach includes the introduction of a new backbone architecture, the incorporation of diffusion models, and the application of a variety of loss functions. In addition, we explore the impact of different activation functions and the refinement of hyperparameters to achieve optimal performance. These elements collectively represent a leap in advancing aerial image analysis, as detailed below:

1. We employ *large kernel convolutions* and *spatial kernel selection* in tandem with a *feature pyramid network* (FPN) [16], constructing a robust and effective backbone for aerial image analysis. This novel design significantly enhances feature extraction and representation in aerial imagery.
2. The adaptation of the *diffusion model* (DiffusionDet) [2] to aerial imaging, with tailored modifications, leads to marked improvements in object detection accuracy within complex aerial scenes.
3. We introduce an innovative and refined model architecture that substantially elevates the accuracy of aerial image analysis. These modifications result in a more powerful and efficient model tailored to aerial imaging contexts.
4. Extensive experimentation is conducted to evaluate the impact of various *activation functions* on our model's performance. This investigation identifies the most effective activation function, boosting the model's overall performance and robustness.
5. To tackle the prevalent class imbalance issue, we de-

velop a *weighted focal loss* function and examine the adaptation of other loss functions for box regression. This approach effectively addresses class imbalances, simultaneously enhancing the model’s accuracy.

6. We thoroughly analyse the influence of *hyperparameters* and *post-processing methods*, fine-tuning them to optimize results. This meticulous optimization process ensures our proposed model’s peak performance in the complex aerial image analysis domain.

Our work introduces an expansive and innovative methodology for aerial image analysis, achieving superior performance while surpassing existing methods. A key aspect of our approach is its efficiency in resource utilization. Despite using a limited number of GPUs and fewer iterations, our model demonstrates significant advancements in accuracy and robustness. This efficient use of computational resources underscores our model’s effectiveness and highlights its sustainability, making it a more energy-efficient solution in aerial image analysis. By meticulously addressing various facets of model design and optimization, our study sets a new benchmark in delivering high-performance aerial image analysis with a reduced environmental impact.

## 2. Dataset

In this work, we used a patchified version of the Instance Segmentation in Aerial Images Dataset (iSAID) [38], which is a comprehensive dataset designed specifically for aerial image object detection and instance segmentation tasks. It contains diverse high-resolution aerial images collected from various sources, including satellites and unmanned aerial vehicles (UAVs). The dataset encompasses a wide range of scenarios, such as urban, rural, and natural environments, providing a robust foundation for training and evaluating object detection and segmentation models.

The original version of iSAID contains 655,451 object instances belonging to 15 categories within 2,806 high-resolution images. The images within the iSAID dataset are the same as those in the DOTA-v1.0 dataset [36], which are primarily collected from Google Earth. Some of the images are captured by the JL-1 satellite, while others are taken by the GF-2 satellite, both operated by the China Centre for Resources Satellite Data and Application.

Our study employed a patchified version of the iSAID dataset, which includes 28029 images. These were derived by segmenting the original dataset into patches of size  $800 \times 800$ . This segmentation led to a substantial increase in the number of instances in the training subset, amounting to 704428. The rise in instance count can be largely attributed to overlaps in the dataset patches. Using this batchified approach provided two key advantages: first, it significantly enriched the training data’s diversity and complexity, thereby enhancing our model’s robustness in various aerial imaging scenarios. Second, it allowed for a more comprehensive training experience, exposing the model to various instances and contextual environments. It is crucial

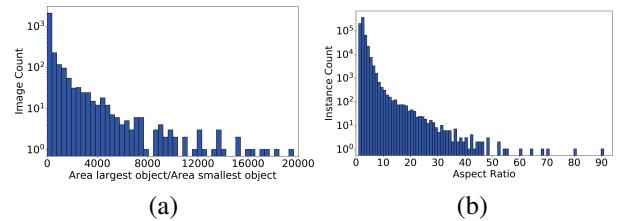


Figure 1. (a) Ratio between the largest and smallest objects’ areas shows scale variation. (b) Aspect ratio variations between instances.

for improving its generalization capabilities across diverse aerial images. Crucially, despite the segmentation and augmentation process, we meticulously maintained the separation of the train, validate, and test sets, ensuring no data leakage between them. This careful partitioning guarantees the integrity and reliability of our evaluation process.

Despite its richness, the iSAID dataset presents several challenges to researchers and practitioners in the field of aerial object detection. Addressing these challenges requires the development of more advanced and robust machine-learning models capable of handling the unique complexities of aerial imagery. These challenges include:

1. *Varied Object Sizes*: Aerial images feature objects of diverse sizes, from large buildings to small vehicles, making it difficult for models to accurately identify and segment instances.
2. *Occlusion*: Objects in aerial images often overlap or are partially hidden by other objects or natural features, making it challenging to segment and detect them accurately.
3. *Scale Variation*: The dataset contains images with varying scales and aspect ratios, which can impact the performance of models trained on this data, as shown in Figure 1.
4. *Complex Backgrounds*: Aerial images often feature intricate and cluttered backgrounds, making it difficult for models to distinguish between objects and their surroundings.
5. *Illumination and Weather Conditions*: Changes in lighting and atmospheric conditions can impact the visibility and appearance of objects in aerial images, posing additional challenges for object detection and segmentation models.

## 3. Related Works

Aerial imaging and object detection have evolved into indispensable tools across various applications, offering profound insights into terrestrial phenomena and human activities. The advent of remote sensing technologies, encompassing satellite and drone imagery, has revolutionized observational and analytical methods in large-scale environmental and geographical studies [37]. Aerial imaging facilitates the acquisition of high-resolution data over extensive areas, proving invaluable in diverse tasks like land use

classification [42], disaster response [4], and environmental monitoring [31]. In this context, object detection in aerial images is pivotal for distilling actionable information from these vast datasets, aiding in identifying and localizing various objects and features.

Recent advancements in deep learning-based object detection algorithms have markedly enhanced the accuracy and efficiency of analyzing aerial images [14, 36]. The synergy of aerial imaging and object detection has profoundly influenced numerous fields, contributing significantly to our understanding and capabilities in addressing global challenges.

One-stage object detection models have gained prominence due to their operational efficiency and efficacy. These models amalgamate object localization and classification tasks into a single streamlined network, thereby reducing inference time. A notable example of this model type is the You Only Look Once (YOLO) framework [28], which segments the input image into a grid system, predicting bounding boxes and class probabilities for each grid cell. Similarly, the Single Shot MultiBox Detector (SSD), proposed by Liu et al. [24], employs multi-scale feature maps to handle objects of various sizes effectively. These one-stage models have shown commendable performance in a range of real-world scenarios.

Conversely, two-stage object detection models typically consist of a region proposal network followed by a classification network. The initial stage generates candidate regions for objects, while the subsequent stage classifies these regions into specific object categories. The Region-based Convolutional Networks (R-CNN) family [9] exemplifies two-stage models. The R-CNN model utilizes a selective search for region proposal generation, followed by classification via a CNN. Successors like Fast R-CNN [8] and Faster R-CNN [29] have enhanced the efficiency and accuracy of the original R-CNN through innovations like ROI pooling and region proposal networks.

State-of-the-art in object detection within aerial imagery has progressed rapidly, spurred by the availability of large-scale aerial image datasets like DOTA [36] and xView [14]. These datasets have enabled focused training and evaluation of object detection models tailored for aerial contexts. Aerial images often pose unique challenges, including significant object scale variations, cluttered backgrounds, and diverse viewpoints. To address these challenges, specialized models have been developed, such as the Oriented Region Proposal Network (ORPN) [39], designed to detect objects irrespective of their orientations. The High-Resolution Network (HRNet) [35] effectively manages objects of varying sizes by maintaining high-resolution feature maps throughout the model. Furthermore, the Adaptively-Sized Object Detector (ASOD) [34] adapts to varying object sizes by adjusting receptive fields and anchor scales. These advancements have substantially improved object detection in aerial images, enhancing the accuracy and efficiency of data analysis.

## 4. Methodology

In this section, we elucidate the technical intricacies of the methods employed in constructing our model.

### 4.1. Model Architecture

#### 4.1.1 LSKNET

The backbone of our model mirrors the general architecture of prevalent models such as those discussed in [25], consisting of repeated blocks with a similar structure. A key innovation in our approach is integrating the Large Selective Kernel (LSK) mechanism [17] into each backbone block. This integration is pivotal in enhancing the feature extraction capability of the model by providing a broader contextual area.

*Large Kernel Convolutions:* Implemented as a sequence of depthwise convolution layers [6], these convolutions utilize increasing kernel sizes and dilation rates. This configuration rapidly expands the receptive field, as detailed in [7, 23]. The structure's primary benefits are twofold: it facilitates extracting multiple features encompassing various contextual areas. It offers superior efficiency to a single large kernel with an equivalent receptive field. For instance, for an input with 64 channels, a sequential mechanism with a structure of  $(3, 1) \rightarrow (5, 2) \rightarrow (7, 3)$  requires only 11.3K parameters. In contrast, a single convolution layer of size 29 would necessitate 60.4K parameters.

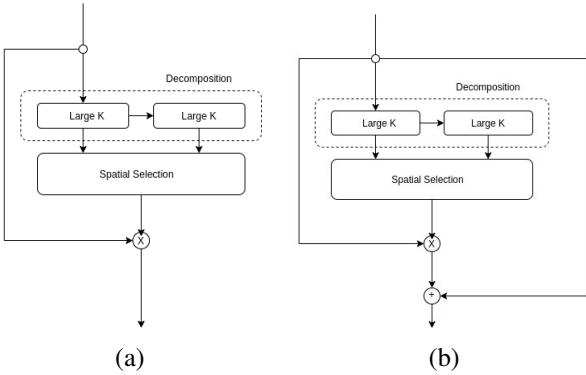
*Spatial Kernel Selection:* As per [15], this process dynamically selects kernels suited to different objects based on the extracted features. Initially, features from various-sized kernels are concatenated into a feature map of size  $\mathbb{R}^{h \times w \times N}$ . Subsequently, the channel-wise average and maximum are computed and integrated into a feature descriptor of size  $\mathbb{R}^{h \times w \times 2}$ . A convolution layer followed by a sigmoid activation function is then applied to these feature descriptors, resulting in a spatial attention map of size  $\mathbb{R}^{h \times w \times N}$ . The final output of this module is the element-wise product of the concatenated input feature maps and the spatial attention map.

Finally, a *feature pyramid network* [20] is constructed using a series of downsampling blocks, each comprising a sequence of Large Selective Kernel blocks. This configuration ensures that the final output of the backbone comprises multiple feature maps of varying resolutions obtained by passing the input through a differing number of blocks. As an innovative modification, we have also incorporated residual connections parallel to the spatial filtering operation. This addition allows the preservation and passage of features potentially filtered out by the preceding LSK block. Figure 2 shows our modification to the LSK block compared to the original block.

#### 4.1.2 DiffusionDet

We chose DiffusionDet head for our model [2]. DiffusionDet is a novel framework that approaches object detection as a denoising diffusion process, transitioning from noisy boxes to actual object boxes. During the training phase, object boxes diffuse from ground-truth boxes to a random dis-

Figure 2. (a) The original LSK Module (b) Our modified LSK Module with a residual connection.



tribution, and the model learns to reverse this process. During inference, the model progressively refines a set of randomly generated boxes to produce the final output. Comprehensive evaluations on standard benchmarks, including MS-COCO and LVIS, demonstrate that DiffusionDet outperforms many well-established detectors. This work reveals two key insights in object detection: First, random boxes, despite being significantly different from predefined anchors or learned queries, can still serve as effective object candidates. Second, object detection, as a representative perception task, can be addressed using a generative approach.

The diffusion model generates data samples iteratively, requiring multiple runs of the model  $f_\theta$  during the inference stage. To address the computational intractability of applying  $f_\theta$  on the raw image at every step, the model is divided into an image encoder and a detection decoder. The image encoder extracts high-level features from the raw input image, while the detection decoder refines box predictions from noisy boxes using these features. Inspired by Sparse R-CNN [33], the detection decoder takes in proposal boxes, crops ROI-features from the feature map, and sends them to the detection head for box regression and classification. The main differences between DiffusionDet and Sparse R-CNN include the use of random boxes instead of learned ones, the input requirements, and the re-use of the detector head in iterative sampling steps with shared parameters across different steps.

We prefer this model over others in this coherence because of its ability to handle noisy images and pay attention to small objects. Those are two main challenges in aerial images, so we think it may be suitable for that task. We replaced the default Swin transformer backbone with our modified version of LSKNet and initialized the model using pre-trained weights from COCO dataset.

## 4.2. Augmentations

Data augmentation stands as a cornerstone technique in enhancing machine learning models' performance, particularly in aerial image analysis. This process involves generating new training samples by applying a range of transformations to existing images. Such transformations, which

include rotation, scaling, flipping, and colour modifications, substantially increase the training dataset's diversity. This, in turn, equips models with better generalization capabilities for new, unseen data.

Our study focused on two prominent data augmentation techniques: flipping and Albumentations. Flipping, a simple yet effective method, creates new training instances by mirroring the original images horizontally or vertically. This approach aids in diversifying the dataset and plays a crucial role in mitigating the risk of overfitting.

Albumentations, a specialized data augmentation library tailored for computer vision tasks [1], was another key component of our methodology. This library offers a comprehensive suite of image transformations designed to enhance the model's generalization ability. These transformations encompass geometric operations, such as rotations, translations, scaling, flipping, and photometric adjustments, like altering brightness, contrast, and colour.

Applying these data augmentation techniques is particularly beneficial in addressing the unique challenges posed by aerial images. These challenges include variations in scale and resolution, geometric distortions, diverse environmental conditions, and seasonal changes. By employing flipping and Albumentations, our machine learning models are better equipped to adapt to and accurately interpret the complex characteristics of aerial imagery.

## 4.3. Loss Functions

Loss function selection is a pivotal aspect of object detection, fundamentally guiding the learning process. Broadly, loss functions in object detection can be categorized into two types: bounding box regression loss and classification loss. The bounding box regression loss measures the similarity between predicted and ground truth bounding boxes, considering attributes such as shape, orientation, aspect ratio, and centre distance. Various loss functions, or their linear combinations, are utilized for this purpose:

1. *Intersection Over Union (IOU)*: A widely-used metric for evaluating the accuracy of object detection models. The IOU loss is defined as:

$$IOU_{loss} = 1 - \frac{A_{pred} \cap A_{gt}}{A_{pred} \cup A_{gt}} \quad (1)$$

2. *Generalized Intersection Over Union (GIOU)*: An extension of IOU, GIOU considers the smallest convex hull that encloses both the ground truth and predicted boxes. It is more robust than IOU, as it accounts for the shape and orientation of the boxes, thereby reducing the impact of misaligned boxes on the final loss value [30].

$$GIOU_{loss} = 1 - (IOU - \frac{A_C - A_U}{A_C}) \quad (2)$$

3. *Complete Intersection Over Union (CIOU)*: CIOU further enhances GIOU by incorporating the aspect ratio and centre distance between the ground truth and predicted boxes. This integration improves convergence

and localization accuracy, particularly beneficial for objects with varying aspect ratios [41].

$$CIOU = 1 - (GIOU - \alpha \cdot \frac{d_{center}^2}{A_C} - \beta \cdot \frac{v_{aspect.ratio}^2}{1 - IOU}) \quad (3)$$

4. *SmoothL1 Loss*: A variant of L1 loss, SmoothL1 Loss is less sensitive to outliers. It applies a smooth approximation to the absolute function, transitioning from L1 to L2 loss near the origin. This approach results in a more stable learning process and mitigates the impact of noisy samples.

$$L_{\text{SmoothL1}}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \quad (4)$$

Various loss functions are applicable for classification tasks, with Focal Loss being of particular interest. Focal Loss is designed to address the class imbalance in object detection tasks by introducing a modulating factor that down-weights the contribution of easy examples and concentrates on harder, misclassified examples [21].

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \quad (5)$$

As Figure 3 illustrates, significant class imbalance might limit Focal Loss's efficacy. To address this, we implement Weighted Focal Loss, combining the principles of Focal Loss with class weighting. This method assigns varying weights to each class, enabling the model to prioritize minor classes or those with higher misclassification costs. This approach can enhance overall performance, especially in pronounced class imbalances. Equation 6 depicts the Weighted Focal Loss, where  $\alpha_t$  is the weighting factor for the target class, calculated for simplicity as the inverse ratio of the number of samples of each class to the total number of samples.

$$WFL(p_t) = -\alpha_t w_c(1 - p_t)^\gamma \log(p_t) \quad (6)$$

#### 4.4. Activation Function

Activation functions are crucial in neural networks, introducing non-linearity that enables models to learn complex patterns from input data. This paper discusses three advanced activation functions: Mish, Hardswish, and Gaussian Error Linear Units (GELU), each bringing unique benefits to the model's learning capability.

1. *Mish*: Mish is a self-regularized novel activation function that has demonstrated superior performance to traditional functions like ReLU, Leaky ReLU, and Swish [27]. It introduces attributes of smoothness and non-monotonicity, fostering enhanced gradient flow and expedited convergence. The Mish function is defined as:

$$\text{Mish}(x) = x \cdot \tanh(\text{softplus}(x)) \quad (7)$$

This function has effectively promoted deeper feature extraction with improved learning dynamics.

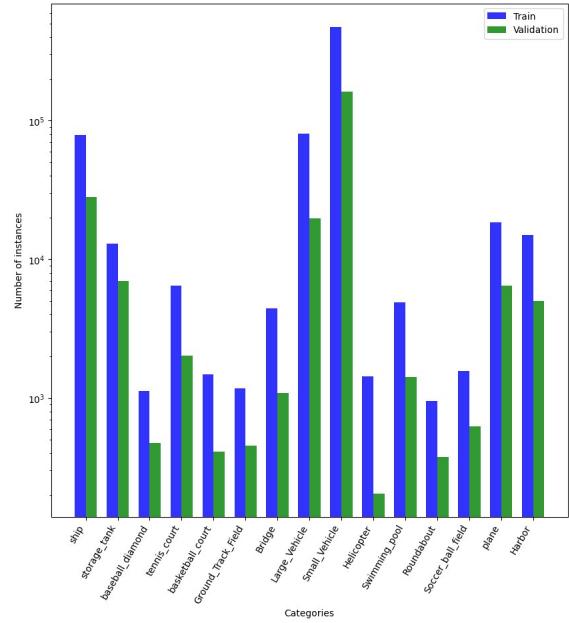


Figure 3. number of instances in each class for train and validation

2. *Hardswish*: As a computationally efficient alternative to the Swish activation function, Hardswish offers comparable performance while reducing computational overhead [12]. It has found utility in lightweight models such as MobileNetV3 and EfficientNet, which aim to balance high accuracy with low computational complexity. The following equation defines Hardswish:

$$\text{Hardswish}(x) = x \cdot \frac{\text{ReLU6}(x + 3)}{6} \quad (8)$$

The primary advantage of Hardswish lies in its efficiency, particularly in resource-constrained environments.

3. *Gaussian Error Linear Unit (GELU)*: Inspired by the Gaussian error function, GELU has gained popularity in various models, including BERT and GPT, especially in natural language processing tasks [11]. It is characterized by:

$$\text{GELU}(x) = x \cdot \frac{1}{2} \left( 1 + \text{erf} \left( \frac{x}{\sqrt{2}} \right) \right) \quad (9)$$

GELU is renowned for its ability to facilitate more nuanced and probabilistic feature transformations, contributing to the model's overall expressiveness and performance.

#### 4.5. Hyper-parameters

There are many hyper-parameters that can be tuned to increase performance. From these, we made an ablation study on the effect of each, which are:

1. *The learning rate* determines the step size taken during gradient descent optimization. A crucial hyper-parameter influences the convergence rate and model

performance [32]. A too-large learning rate can lead to divergence, while a too-small learning rate can result in slow convergence.

2. *Number of proposals* refers to the number of candidate bounding boxes generated by a region proposal network (RPN) in object detection models [29]. This hyperparameter affects the trade-off between recall and computational complexity.
3. *Aspect ratios* are the different proportions of anchor boxes used in object detection models [28]. They help the model detect objects with varying shapes and sizes.
4. *The number of epochs* is the number of times the entire training dataset is processed during training. A higher number of epochs can result in better model performance, but it may also increase the risk of overfitting if the model is trained for too long. Choosing the optimal number of epochs depends on the specific problem and dataset.
5. *Batch size* is the number of training samples used to compute the gradient during a single optimization step [13]. Larger batch sizes can lead to more stable gradient estimates but may require more memory and computational resources.
6. *Images per batch* is the number of images used in each batch during training. This hyperparameter is related to the batch size and influences the memory requirements and the stability of the gradient estimates.

## 5. Experiment Setup

To comprehensively understand the impact of each modification in our model, we adopted a systematic approach in our experiments, altering only one variable at a time. This systematic process allowed us to isolate and examine the specific effect of each change.

Our initial focus was on the model architecture. We explored the combination of LSKNet as the backbone with DiffusionDet heads. We conducted five distinct experiments to assess the impact of varying the backbone and the head. In each, we altered either the backbone or the head only. Specifically, we utilized ResNet and LSKNet as backbones with the Faster RCNN and replicated this configuration for DiffusionDet. Additionally, Swin Transformer was employed as the backbone in conjunction with DiffusionDet in our fifth experiment.

For all subsequent experiments, we standardized the architecture, employing LSKNet as the backbone and DiffusionDet for the heads. This model configuration included GeLU as the activation function and utilized focal loss and GIOU as loss functions. By default, the model also employed Non-Maximum Suppression (NMS) and default aspect ratios of [0.5, 2, 4].

The hyper-parameters were set as follows: batch size = 512, number of images per batch = 3, learning rate = 0.00005, number of proposals = 300, and a maximum of

100000 iterations. Given our limited hardware resources, specifically the constraint of utilizing only a single GPU with 24 GB of memory, we were mindful of the complexities added to our model. This hardware limitation influenced decisions regarding the number of images per batch, iterations, and other aspects that could potentially increase performance. Each modification was underpinned by a rationale aimed at performance improvement:

- *Activation Functions*: Mish was selected to address the vanishing gradient problem, known to enhance model accuracy, especially in deep networks. Hard Swish, recognized for its computational efficiency, has been shown to yield similar or slightly better accuracy than ReLU in certain tasks.
- *Architectural Modifications*: We expanded the model by adding a block with a depth of 32 at the start of the model’s sequence. This expansion aimed to enhance the detection of smaller objects, leveraging high-resolution feature maps. Furthermore, to prevent the potential loss of important features by spatial selection, we experimented with direct residual connections from the unfiltered feature map to the output of the LSK block.
- *Loss Functions*: CIOU was chosen for considering overlaps, aspect ratios, and centre distances between predicted and ground-truth boxes. Weighted Focal Loss targeted class imbalance by focusing on harder examples. Smooth L1 Loss was introduced to mitigate the effect of outliers, potentially improving regression performance.

Additional empirical adjustments to the hyperparameters were made based on specific characteristics of the dataset and expected outcomes:

- *Aspect Ratios*: Adjusted to [0.25, 0.75, 2, 4] to better suit the dataset’s large variance in aspect ratios.
- *More Proposals*: Increased to 700, aiming to enhance accuracy by offering a broader range of regions for evaluation.
- *More Images Per Batch*: Increased from 3 to 4 to provide a richer variety of data for the algorithm, enhancing learning and generalization.
- *Soft NMS*: Implemented as an alternative to traditional NMS, this technique lowers scores of overlapping boxes rather than discarding them, potentially retaining more accurate predictions in scenarios with closely packed or partially occluded objects.

## 6. Results & Discussion

This section delineates the outcomes of each experiment and provides insights into the implications of these results. It is important to note that all experiments involved singular modifications to the base model, which consists of LSKNet and DiffusionDet, making it the reference point for performance comparisons.

Table 5. Results of Best Model on Validation & Test Sets: Notations: ST: Storage tank, TC: Tennis court, BD: Baseball Diamond, BC: Basketball court, GTF: Ground Track Field, LV: Large Vehicle, SV: Small Vehicle, HC: Helicopter, SP: Swimming pool, RA: Roundabout, SBF: Soccer ball field.

Data	AP	AP75	AP50	Ship	ST	TC	BD	BC	GTF	Bridge	LV	SV	HC	SP	RA	SBF	Plane	Harbor
Test	45.7	50.6	66.8	55.4	33.4	50.9	25.0	66.1	34.1	45.3	58.5	37.2	46.6	22.4	25.0	45.3	78.8	45.7
Validation	44.8	48.6	67.0	57.8	37.6	80.9	55.1	44.5	49.1	25.2	47.3	20.8	21.3	38.2	34.3	49.6	70.8	49.8



Figure 4. Some Qualitative Analysis on Our Model

Table 1. Results of Different Model Architectures

Experiment (Model)	AP	$AP_s$	$AP_m$	$AP_l$
Resnet + Faster RCNN	41.09	26.52	48.15	53.13
LSKNet + Faster RCNN	41.53	27.46	48.78	53.40
Resnet + DiffusionDet	40.72	27.17	46.67	52.70
Swin + DiffusionDet	36.53	22.74	44.56	50.37
<b>LSKNet + DiffusionDet*</b>	<b>42.48</b>	<b>27.83</b>	<b>49.08</b>	<b>55.23</b>
<b>*Residual Connection</b>	<b>42.94</b>	<b>28.11</b>	<b>49.64</b>	<b>55.48</b>
*Added Block	40.21	25.93	46.48	51.77

## 6.1. Model Architectures and Their Impact

Table 1 presents the results from varying model architectures. The integration of LSKNet as the backbone with DiffusionDet heads yielded the most favourable results, enhancing the mAP by approximately 1.8% compared to the base Diffusion model with a Resnet backbone. Incorporating the Residual Connection improved performance by nearly 0.5%, likely due to its ability to preserve initial image features before deeper feature extraction. However, adding a new block with randomly initialized parameters, as opposed to pre-trained weights like the rest of the blocks, resulted in reduced performance, indicating challenges in adapting these new weights within the overall backbone.

## 6.2. Effects of Loss and Activation Functions

As shown in Table 2, experimenting with different loss and activation functions generally led to improvements in mAP, except for Weighted Focal Loss. Replacing GIOU with CIOU marginally increased accuracy but required more time to converge. Weighted Focal Loss underperformed, possibly due to imbalanced or excessive class weighting. Among the activation functions, Hardswish notably excelled in detecting smaller objects, a critical area for accuracy enhancement.

Table 2. Results of Different Losses & Activation Functions

Experiment (Model)	AP	$AP_s$	$AP_m$	$AP_l$
<b>Smooth L1</b>	<b>43.40</b>	<b>28.57</b>	<b>49.87</b>	<b>55.74</b>
CIOU	43.11	28.46	49.84	54.83
Weighted Focal Loss	41.51	27.29	49.13	53.99
MISH	42.76	27.67	50.17	54.40
<b>Hardswish</b>	<b>43.52</b>	<b>32.72</b>	<b>49.84</b>	<b>55.75</b>

## 6.3. Hyperparameter Tuning and Its Effectiveness

Table 3 reflects the outcomes of fine-tuning hyperparameters, demonstrating performance boosts from more customized settings. Altering aspect ratios appeared to benefit medium and large objects significantly. The most impactful individual modification was the increase in the number of proposals, elevating the model’s mAP to 44.71%. More

Table 3. Results of Different Configurations

Experiment (Model)	AP	$AP_s$	$AP_m$	$AP_l$
<b>Customized Aspect Ratio</b>	<b>44.58</b>	<b>30.15</b>	<b>51.66</b>	<b>56.75</b>
<b>More Proposals</b>	<b>44.71</b>	<b>29.82</b>	<b>50.15</b>	<b>57.22</b>
More Images Per Batch	44.42	29.24	51.36	57.97
<b>Soft NMS</b>	<b>43.00</b>	<b>27.93</b>	<b>50.29</b>	<b>54.11</b>
<b>Augmentations</b>	<b>42.55</b>	<b>27.94</b>	<b>48.72</b>	<b>54.40</b>

images per batch also showed improvement, although it was sometimes limited by GPU memory constraints, occasionally leading to crashes. Both Soft NMS and augmentations exhibited potential for enhancing model performance.

#### 6.4. The Best Model and Its Superior Performance

In pursuing the optimal model configuration, we combined various modifications individually, enhancing performance. The best model, highlighted in Table 4, culminates these alterations. It integrates DiffusionDet with an LSKNet backbone, augmented with the extra residual connection, Hardswish activation function, Smooth L1, focal loss, and GIOU. Additionally, it incorporates the customized aspect ratio, increased proposals, and augmentations, with Soft NMS as the post-processing technique.

This best model achieved impressive results, notably a mean Average Precision (mAP) of 45.7% on the test set, a significant accomplishment considering past literature and reports. The performance metrics for each class and across different object sizes in both test and validation sets are detailed in Table 4.

Table 4. Performance Metrics of the Best Model

Model	AP	$AP_s$	$AP_m$	$AP_l$
<b>Best Model (Validation)</b>	<b>46.23</b>	<b>30.12</b>	<b>51.45</b>	<b>57.98</b>
<b>Best Model (Test)</b>	<b>45.70</b>	<b>29.87</b>	<b>51.05</b>	<b>57.31</b>

Qualitatively, Figure 4 showcases the efficacy of our best model on random images from the dataset. The results demonstrate the model’s capability to accurately detect objects, particularly smaller ones that were challenging for the base model.

#### 6.5. Observations

Our rigorous experimentation has resulted in a model that notably enhances object detection in aerial imagery, establishing new performance benchmarks. Remarkably, our model, constrained to a single GPU and limited iterations, demonstrates competitive efficiency compared to state-of-the-art models [38] that utilized 8 GPUs and 180,000 iterations. This achievement highlights our model’s energy efficiency and potential for scalability. The focused modifications in model architecture, loss functions, activation functions, and hyperparameters, even within hardware limitations, have significantly advanced aerial image analysis. These results indicate that our model could achieve further substantial improvements in the field with access to more robust computational resources.

## 7. Conclusion

In this research, we have introduced a suite of innovative enhancements that substantially elevate the good results in aerial image analysis. Our methodology entailed the development of a robust and sophisticated backbone, integrating large kernel convolutions, spatial kernel selection, and feature pyramid networks. This backbone was further augmented with adapted diffusion models tailored specifically for the complexities of aerial imaging, thereby improving object detection and classification.

The architectural advancements in our proposed model yielded a more powerful and efficient tool for aerial image analysis. We conducted extensive investigations into various activation functions, ultimately identifying the most effective option for our specific application. To address the prevalent class imbalance issue, we devised a weighted focal loss function and explored the adaptation of additional loss functions for box regression.

Our exhaustive examination and fine-tuning of hyperparameters and post-processing methods culminated in an optimized model with notable improvements. These efforts culminated in a significant increase in mean Average Precision (mAP), achieving a mAP of 45.7% on the test dataset. This comprehensive approach marks a significant stride forward in enhancing the accuracy and robustness of aerial image analysis.

## 8. Limitations & Future Work

This study encountered certain limitations that impacted our ability to achieve even higher performance metrics. A primary constraint was the GPU memory capacity, which restricted our ability to increase the number of images per batch. This limitation was particularly notable, as our best model demonstrated potential for accuracy improvement with larger batch size, but we often faced memory overflow or system crashes.

Another challenge was the unavailability of LSKNet pre-trained weights for the COCO dataset [22]. Our experiments revealed that COCO pre-trained weights generally outperform those from ImageNet for object detection tasks. We partially mitigated this by using COCO pre-trained weights for the DiffusionDet heads, but this led to inconsistent weight distributions between the backbone and the head. Despite efforts to fine-tune the model on COCO, time constraints and frequent GPU crashes limited our progress.

Looking forward, our future work aims to address these limitations. Access to higher-capacity GPUs and extended training time on the COCO dataset are among our primary objectives. We anticipate that enabling a larger number of images per batch could significantly enhance model performance, further advancing the field of aerial image analysis.

## References

- [1] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. Albumentations: Fast and flexible image augmen-

- tations. <https://github.com/albumnetations-team/albumnetations>, 2018. 4
- [2] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection, 2022. 1, 3
  - [3] Guoqing Cheng, Junwei Han, and Xin Lu. A survey and comparative study of satellite remote sensing image classification techniques. *International Journal of Remote Sensing*, 37(12):2978–3015, 2016. 1
  - [4] G. Cheng, J. Han, and P. Zhou. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS Journal of Photogrammetry and Remote Sensing*, 139:30–45, 2018. 3
  - [5] Guoqing Cheng, Xin Lu, and Junwei Han. Automatic disaster monitoring and assessment for urban areas from satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 56(7):3832–3845, 2018. 1
  - [6] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 3
  - [7] Xiaohan Ding, Xiangyu Zhang, Jungong Han, and Guiguang Ding. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11963–11975, 2022. 3
  - [8] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015. 3
  - [9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 1, 3
  - [10] Fuxian Han, Yiqun Wang, Haichang Zhang, and Hui Hu. Optimized object detection in remote sensing images using improved faster r-cnn. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 8459–8462. IEEE, 2018. 1
  - [11] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 5
  - [12] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1314–1324, 2019. 5
  - [13] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016. 6
  - [14] Samuel Lamm, Vasileios P. Kemerlis, and Ken Birman. xview: Objects in context in overhead imagery. In *International Symposium on Visual Computing*, pages 334–346. Springer, 2018. 3
  - [15] Xiang Li, Wenhui Wang, Xiaolin Hu, and Jian Yang. Selective kernel networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 510–519, 2019. 3
  - [16] Yuxuan Li, Qibin Hou, Zhaohui Zheng, Ming-Ming Cheng, Jian Yang, and Xiang Li. Large selective kernel network for remote sensing object detection, 2023. 1
  - [17] Yuxuan Li, Qibin Hou, Zhaohui Zheng, Ming-Ming Cheng, Jian Yang, and Xiang Li. Large selective kernel net-
  - work for remote sensing object detection. *arXiv preprint arXiv:2303.09030*, 2023. 3
  - [18] Yinjie Li, Haifeng Qi, Payam Khorrami, Chen-Han Kuo, Xiaodong Wang, Yejun Liu, Siyuan Chen, Wenyuan Li, and Neng Sun. Vehicle detection from 3d lidar using fully convolutional network. *Robotics: Science and Systems*, 2017. 1
  - [19] Zechun Li, Changan Peng, Gang Yu, Xuan Zhang, Yangdong Deng, and Jian Sun. Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, 128(2):261–318, 2018. 1
  - [20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3
  - [21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2017. 5
  - [22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*, 2014. 1, 8
  - [23] Shiwei Liu, Tianlong Chen, Xiaohan Chen, Xuxi Chen, Qiao Xiao, Boqian Wu, Mykola Pechenizkiy, Decebal Mocanu, and Zhangyang Wang. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. *arXiv preprint arXiv:2207.03620*, 2022. 3
  - [24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016. 3
  - [25] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 3
  - [26] Surya S Mali, Bhabani S Nayak, and Deepak Mishra. High-precision aerial image segmentation with convolutional networks. *IEEE Geoscience and Remote Sensing Letters*, 14(11):2067–2071, 2017. 1
  - [27] Diganta Misra. Mish: A self regularized non-monotonic neural activation function. *arXiv preprint arXiv:1908.08681*, 2019. 5
  - [28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1, 3, 6
  - [29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99. 2015. 3, 6
  - [30] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression, 2019. 4
  - [31] N. Singh and P. Kumar. Urban land use and land cover classification using high-resolution ikonos satellite data and multi-layer perceptron neural network. *Environment, Development and Sustainability*, 19(6):2447–2466, 2017. 3
  - [32] Leslie N Smith. Cyclical learning rates for training neural networks. *arXiv preprint arXiv:1506.01186*, 2017. 6

- [33] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. Sparse r-cnn, 2020. 4
- [34] Hao Wang, Qilong Sun, Meng Zhou, Feng Yang, and Peipei Li. Asod: An adaptive-sized object detector based on structure information. *IEEE Transactions on Geoscience and Remote Sensing*, 58(8):5661–5671, 2020. 3
- [35] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, and Wenyu Liu. Deep high-resolution representation learning for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):1074–1080, 2021. 3
- [36] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Beßlongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018. 2, 3
- [37] C. Yang, F. Rottensteiner, and M. Y. Yang. Object semantic representation and recognition in aerial images. *International Journal of Remote Sensing*, 39(15-16):5332–5358, 2018. 2
- [38] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images, 2019. 1, 2, 8
- [39] Jian Zhang, Yuchao Ding, Shuai Li, Gui-Song Xia, and Waqar Z Qureshi. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Transactions on Multimedia*, 20(11):3111–3122, 2018. 3
- [40] Yanguo Zhang and Hongpu Liu. Building detection in very high resolution satellite imagery using otsu’s multiple-threshold segmentation and aggregation. *International Journal of Remote Sensing*, 37(11):2395–2420, 2016. 1
- [41] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression, 2020. 5
- [42] Y. Zhong, L. Zhang, C. Kang, and J. Yang. Learning to divide and conquer for online multi-target tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5669–5678, 2017. 3
- [43] Shanshan Zhou, Jiang Wang, and Zhenxing Wang. Cad: Scale invariant framework for real-time object detection. *arXiv preprint arXiv:1607.05475*, 2016. 1