International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Estimating the Distance of a Moving Object through Optimized DSP-Depthnet

Ms.Shwetambari G.Pundkar [1], Dr. Amit K. Gaikwad [2]

Submitted: 08/12/2023 **Revised**: 15/01/2024 **Accepted**: 29/01/2024

Abstract: Autonomous driving requires understanding the layout of the surroundings, such as the distance to vehicles, pedestrians, and other obstacles. As the demand for enabling high-level autonomous driving has increased in recent years and visual perception is one of the critical features to enable fully autonomous driving, in this research, an efficient approach for simultaneous object detection, and depth estimation is done by the novel Hybridized Driving scene perception network (DSPNet) and Depthnet. The DSPnet is used in this research because the network uses multi-level feature maps and multi-task learning to improve the accuracy and efficiency of object detection, distance estimation, and semantic segmentation tasks from an input image. The Depthnet is the recurrent neural network architecture that is used for depth prediction, which estimates the distance between the object and the relative camera. The complexity of the distance estimation is greatly reduced in this hybridized DSP-Depth net and the performance of the distance estimation also improved. Here the Sparrow Egret Optimization (SEO) will be incorporated for the effective tuning of the hyperparameters in the network by the standard hybridization of the Sparrow Search Optimization (SSO) and Egret Swarm Optimization (ESO), where their characteristics of searching, and hunting will be used for training purposes. The proposed system achieves 99%, 97%, and 95% for accuracy, specificity, and sensitivity respectively.

Keywords: *DSPNet, Depth net, Sparrow Egret Optimization, Distance estimation, object tracking.*

1. Introduction

The Global Status Report on Road Safety published by the World Health Organization ranks road traffic injuries as the eighth most common cause of death. By 2030, it will surpass heart disease as the fifth leading cause of death, according to recent studies [10]. One method to reduce traffic accidents is to develop computer vision systems for use in intelligent transportation systems. Unintentional lane changes continue to be a significant contributor to the high accident frequency of traffic. Due to the significance of the application, a sizable number of driver assistance systems have recently been created, particularly in other developed countries. These researchers' main goal is to help drivers by relieving them of certain responsibilities while driving. The most crucial aspect of these systems is the analysis of the frame sequences captured by the onboard cameras. All of these frames have been examined with the goal of assisting drivers in actual traffic situations. Due to its numerous uses in areas like cybernetics, human-computer interface, and refined video surveillance, visual tracking is a crucial and preeminent topic in computer vision.

Despite the fact that visual tracking has advanced significantly over the past 20 years, the field continues to face difficult problems related to variations in appearance brought on by stance, lighting, and occlusion. Consequently, determining the efficient construction of the object's appearance is crucial to the visual tracker's success. For the precise estimation of variations in the object's appearance, various types of visual features are used till now [9]. Instead of categorizing different images, the researchers must accurately assess the locations as well as concepts of the objects used in the image to gain deep insight into it. This procedure, known as object detection, entails a number of smaller tasks, including the identification of people, structures, pedestrians, and vehicles. Autonomous vehicles are a rapidly developing technology that could revolutionize mobility and transportation systems in recent years. Cameras are frequently used to provide a 2D characterization of the space in order to accurately sense the surroundings of the vehicle [1]. Applications like 3D scene comprehension and restructuring [4] depend on the ability to accurately estimate depth from 2D images.

Monocular depth estimation is a more complex problem than depth estimation from stream images or video sequences, meaning there may be more than one unique solution. Deep convolution neural networks (DCNNs) have shown promise as a solution to this poorly posed problem[4]. For autonomous driving, an understanding of the environment's advanced layout is required. This entails being aware of the distances between vehicles, pedestrians, and other hazards as well as the locations of driving lanes, sidewalks, road markings, and traffic signs. A system with a specific task model that is able to perform object detection, depth prediction, and pixel-level image segmentation concurrently is not accessible because the bounding box-level detection of the drivable road is meaningless and the pixel-level segmentation of a group of vehicles is not intuitive in practice [9]. The performance of object detection and depth estimation were enhanced using vision-based applications. Based on the two components binocular vision cameras are used for car position detection and distance estimation. These two methods can be used with the system as soon as the geometry information

[1] *Research Scholar Computer Science &Engineering Department, G. H. Raisoni University Amravati (Maharashtra),*
ORCID ID: 0009-0005-7533-4077
[2]*Computer Science &Engineering Department, G. H. Raisoni University Amravati (Maharashtra),*
** Corresponding Author Email: shwetapundkar@gmail.com*

for the road, environment objects, and vehicles, including traffic signs, are accurately detected. In intelligent transportation systems, both methods can be very useful. Finding the separation between preceding vehicles can be crucial for emergency braking systems using the distance estimation method. Whether the driver is paying attention or not, these systems function automatically. When drivers temporarily lose attention, this can offer them significant protection [10].

In this research, a deep learning-based object detection, and distance es model is developed in order to measure the dissociation between moving objects and cameras. Pre-processing and RoI extraction are applied to the camera-generated image to improve its quality. The modified hybridized DSP-Depth network is developed in this research to address the need for distance estimation and semantic segmentation. Additionally, the hyper-parameters of the hybridized DSP-Depth network are best tuned using SEO in order to improve system performance. The research's contribution is listed below.

- **Hybridized DSP-Depth network:** The DSP-Depthnet consist of multiple encoder and decoders with multiple convolution layers, LSTM layer, and residual blocks to predict the distinctive characteristics of the image and provide sparse item position information with semantic segmentation.
- **Sparrow Egret Optimization (SEO):** The SEO is a metaheuristic algorithm, where the sparrows expended a greater amount of energy in their hunt for prey, so finding food is not guaranteed. Therefore, the sit-and-wait strategy of ESO will be made available for the sparrows' food foraging in order to conserve their energy. This enabling facilitates the location of the prey without requiring a greater expenditure of energy. The SEO is designed to fine-tune the hyperparameter that improves distance estimation accuracy.

The research article is organized as follows, Section 2 provides motivation, a review of some existing techniques with their drawbacks, and challenges that highlight the need for advanced object detection. The proposed distance estimation of a moving object using a DSP-Depthnet is illustrated in section 3 and a detailed description of the result analysis is illustrated in section 4. Finally, the conclusion is elaborated in section 5.

2. Motivation:

Recognizing, locating, and localizing all recognized objects in a scene are the goal of object detection. Recovering the position of objects in 3D space, preferably, is crucial for automatic control systems. Obstacle avoidance and other interactions with the environment can be done using the information from the object detector. A car needs to be aware of all the objects around it and what they are in order to decide what to do next, such as accelerate, apply the brakes, or turn. The following section explains the literature review, challenges, and problem statement based on the previous papers.

2.1 Literature Review:

To complete the depth of a LiDAR point cloud, Lin Bai *et al*. [1] created a simple network. This neural network achieves comparable error performance while significantly reducing the number of parameters. The state-of-the-art (SOTA) network's parameter count is further optimized using a depthwise separable technique. The LiDAR depth completion network is designed with an effective hardware architecture. In particular, all additional multiplication with zeros is avoided when implementing deconvolution operations. The proposed depth completion neural network can be executed by the FPGA implementation by cautiously matching on-chip memory and multipliers. Depthnet uses only 3.8% of the parameters compared to state-of-the-art networks while still achieving a comparable error performance. Optimize the network using a depthwise separable technique, which reduces the number of parameters by an additional factor of 7.3 at the expense of a minor drop in error performance and is intended for low-power embedded platforms like autonomous vehicles. Using sparse consecutive measurements from a non-repetitive circular scanning (NRCS) Lidar, Orkeny Zovathi *et al*. [2] propose an innovative depth image completion method, showcasing the potential of new, portable, and affordable sensor technology for dense range mapping of highly dynamic scenes. A new deep learning-based approach called STDepthnet adds a spatiotemporal downscaling branch to the traditional U-Net architecture to make use of a series of sparse measurements obtained by NRCS Lidars. This model uses a spatial upscaling branch after efficient temporal pooling steps to generate spatially precise high-density depth data. Utilized is a brand-new artificial urban dataset called Livox-Carla, which combines dense depth Ground Truth (GT) data with simulated NRCS Lidar data. However, this approach needs to have better spatial resolution. By utilizing pose information, Jinyoung Jun *et al*. [3] propose a novel monocular depth estimator that increases the prediction accuracy of human regions. The two networks PoseNet and Depthnet used in the suggested algorithm are used to estimate keypoint heatmaps and a depth map, respectively. The network architecture must be modified to account for the presence of humans in images in order to create adaptive schemes for feature transfer from PoseNet to Depthnet. A highly compact self-normalizing network for monocular depth estimation, called Depthnet Nano, is introduced by Linda Wang *et al*. [4] using a human-machine collaborative design strategy that combines machine-driven design exploration with principled network design prototyping based on encoder-decoder design principles. However, it must investigate methods for integrating temporal data into the Depthnet Nano infrastructure in a way that enhances performance while preserving minimal computational complexity. In order to predict depth from a monocular video sequence, Arun CS Kumar *et al*. [5] propose a novel convolutional LSTM (ConvLSTM)-based network architecture. In the proposed ConvLSTM network architecture, we take advantage of the long short-term memory (LSTM)-based RNNs' capacity for sequential reasoning to predict the depth map for an image frame as a function of the appearances of scene objects in the image frame and image frames in its temporal neighborhood. It is necessary to model each independently moving object in the scene individually in order for explainability masks to automatically learn them. To get around these issues, Shi Zhou *et al*. [6] proposed a brand-new self-supervised monocular depth estimation technique based on occlusion mask and edge awareness. Edge information gathered using conventional methods and used to design edge awareness loss has a strong

robustness to varying lighting conditions and is useful for depth estimation. To limit the training of TrajNet, Chaoqiang Zhao *et al*. [7] propose a novel pose-to-trajectory constraint. By using the TrajNet outcome as an original amount in the image alignment algorithm, the initialization and tracking are enhanced. However, there are still issues that must be resolved in the future. For instance, the proposed DDSO still suffers from scale drift, so we intend to train TrajNet on stereo image sequences to assist it learn the absolute scale information, thereby reducing scale ambiguity and improving scale drift. To optimize the depth net with high-level data, Kunhong Li*et al*. [8] initiate an adversarial loss for self-supervised depth estimation. The discriminator guides the data distribution of synthetic images that are close to the real ones by learning the data distribution of real images and synthetic images. An adaptive loss balance strategy is used to achieve balance among various losses.

2.2. Challenges

The challenges in the existing models are mentioned below:

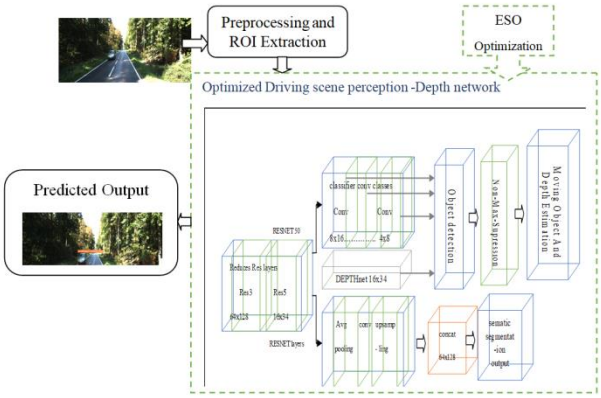
- Object classification and location determination present the biggest obstacles in object detection. It is challenging for detection algorithms to distinguish between different objects at various scales and views because objects always vary in size and ratio. [9].
- Some objects are only partially visible, making it challenging to find them. Objects that take up more space are simpler to grasp than objects that take up less space. [3].
- Despite belonging to the same class, some objects can have various sizes. All of these objects should be able to be classified as belonging to the same class by a good detector, which should also be sensitive to variations between classes. [6].
- Another issue that object detection frequently encounters is the dearth of training data. [2].
- Due to numerous challenging factors like occlusion, blur, and other issues, generic visual object tracking is challenging. Each of these elements could lead to significant issues for a tracking unit [10].

2.3 Problem Statement:

To comprehend and analyze the scenes in videos and images, object detection works flawlessly with other related computer vision techniques, such as image segmentation and recognition. But this identification comes with a number of difficulties, and automation systems especially those used in autonomous vehicles need to estimate distance for safety reasons. Because object detection algorithms must work quickly in order to accurately classify and localize significant moving objects in order to support real-time video processing, object detection in videos can also be challenging. The small amount of training data is also a significant issue for object detection.

3. Proposed distance estimation of a moving object using a modified DSP-Depth net Network.

The main intention of the research is to estimate the distance of the moving object using the Hybridized modified Driving scene perception network (DSPNet) and Depthnet. Initially, the images will be collected from the Kitti dataset [14], and then the preprocessing will be performed for the removal of noisy and irrelevant data. After preprocessing, the extraction of Region of Interest (RoI) takes place, which will help in focusing on the specific regions that contain important information, rather than considering the entire image. The RoI will be then fed forward to the hybridized DSP-Depth network for object detection. The DSP net is used in this research because the network uses multi-level feature maps and multi-task learning to improve the accuracy and efficiency of object detection, distance estimation, and semantic segmentation tasks from an input image. The Depthnet is the recurrent neural network architecture that is used for depth prediction, which estimates the distance between the images and the relative camera. The complexity of the distance estimation is greatly reduced in this hybridized DSP-Depth net and the performance of the distance estimation also improved. In addition, a significant enhancement is made in the hybridized DSP-Depth net using the optimization process that will help in improving the robustness and efficiency of the training. Here the Egret Search Optimization will be incorporated for the effective tuning of the hyper parameters in the network that helps in improvising and stabilizing the process. A Sparrow Egret Optimization (SEO) is developed by the standard hybridization of the Sparrow Search Optimization (SSO)[16] and Egret Swarm Optimization (ESO)[15], where their characteristics of searching, hunting, and escaping will be used for training purposes, thus improving the efficiency of the DSP-depth net performance.



Daigram 1

3.1 Input dataset

A collection of vision tasks created using an autonomous driving platform is included in the Kitti dataset [14]. Numerous tasks, including stereo, optical flow, visual odometry, etc., are included in the complete benchmark. The object detection dataset is included in this dataset, along with bounding boxes and monocular images. The dataset includes 7481 training images with 3D bounding boxes annotated. The input video is first processed before being turned into images, which are shown as follows:

$$V = \sum_{i=1}^{n} V_i \tag{1}$$

where, V is the input video, n is the number of frames in the video, and i is the frame number. The input image is fed forward to the preprocessing stage.

3.2 Preprocessing and RoI extraction

Noise degrades the quality of the actual footage produced by the rearview cameras, which has a negative impact on the accuracy of the outcome. Pre-processing is, therefore, necessary to remove those noises and improve image quality. Unwanted distortions are suppressed and some features are improved during the pre-processing stage to improve accuracy and reliability. The filter that reduces high-frequency components and blurring regions is used to preprocess the input image. The processed image is represented as,

$$V = \sum_{i=1}^{n} V_i^* \tag{2}$$

where, V_i^* is the pre-processed image, which is given as the input to the ROI extraction.

The important and pertinent regions of the input image are extracted during the RoI extraction to streamline the processing. As it doesn't process the image's irrelevant data, it increases computational efficiency and cuts down on computational time. The ROI extraction process is mathematically referred to as follows;

$$R_i = \left\{ V_{i+1}^{roi} \quad for\ i = 1.....n \right\} \tag{3}$$

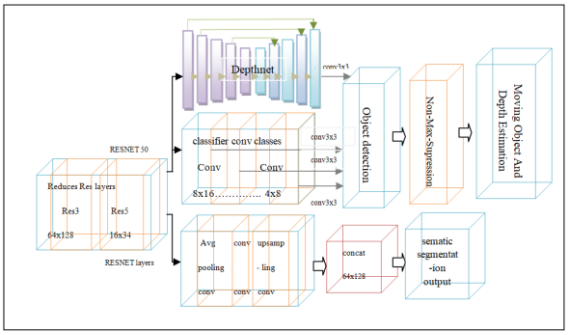
where, V_{i+1} is the frame continuity representation, V^{roi} is the ROI in the selected frame.

3.3 Optimized Driving scene perception-Depth network:

This paper presents an effective method for concurrent object detection, depth estimation, and pixel-level semantic segmentation using a shared convolution architecture. High-level autonomous driving has become more popular in recent years, and visual perception is one of the essential features to enable fully autonomous driving. The proposed network model, DSP-Depth net, enhances object detection, depth estimation, and image segmentation tasks from a single input image using multi-level feature maps and multi-task learning. With the aid of numerous resizing skip connections and up-convolution layers, which further improve its depth estimation abilities, the DSPNet is merged with a Depthnet framework to anticipate the relative depth of objects placed in the picture of an RGB image regarding one another. Deep learning techniques based on training have been used to create depth maps from RGB images from a single camera, reducing the dependence on stereo vision systems.

3.3.1 DSP-Depthnet:

The proposed DSP-Depthnet output will obtain an accurate depth estimation of the moving object by using a hybrid network model, which is shown in Figure 3. The performance of the suggested system will improve with the addition of Depthnet. The DSP network consists of two decoders for object detection, semantic segmentation, and distance estimation, as well as encoders for feature extraction. The decoders take advantage of multiple convolution layers and the multi-level feature maps that were obtained from the ResNet-based encoders to decode the features. By altering the network's training, the pixel-level semantic and intense-level features are preserved. For the street scene images that are taken into consideration as the input, this architecture is assured efficient feature extraction. The ResNet is used to carry out the feature extraction task, as was already mentioned. Image recognition tasks in ResNet architecture, which includes residual connections, perform noticeably better than earlier state-of-the-art models. The convolution typically includes the first four residual blocks, which have been pre-trained for image classification. The block is scaled down to half its actual size in both width and height. In addition, there are now twice as many channels as in earlier blocks. The weights of the model have previously been trained to forecast the image category, and the residual blocks gather the low-level features and store the position data. The leftover blocks closer to the output layer attempt to predict the distinctive characteristics of the image and provide sparse item position information.



i) Semantic Segmentation:

The process of giving a semantic label to each pixel of an image is known as semantic segmentation, also known as scene labeling. Understanding the environment is a crucial data processing step for robots and other automated systems. In order to perform further image analysis and visual comprehension, a semantic segmentation divides a given image into several visually interesting or meaningful regions. In a wide range of applications, including understanding a scene, image analysis in the medical field, robot view, and image segmentation, semantic segmentation is crucial. To adaptively decrease the number of outcomes in the global prior and local feature maps, semantic segmentation is used. Specifically, to apply a deconvolution operator to the encoded low-level features in order to extract low-level feature maps and use learnable up sampling filters. Convolution layers are used for semantic segmentation after ReLU non-linearity, which slightly speeds up convergence at the beginning of training and reduces memory usage during inference. This aims to categorize every pixel in relation to the represented image. As it predicts each pixel of the image, this

process is commonly referred to as a dense prediction. Both the local and global feature maps' outputs are effectively reduced by segmentation. In situations where higher-level feature maps provide the global prior, it specifically decreases the number of outputs in both the global prior at higher levels and the local prior at lower levels. In terms of width and height, the output size is reduced to a size that is four times smaller. The pixel-wise softmax must be implemented with a large computational scale and memory footprint. By halving the size of the segmentation output, the softmax computation is significantly decreased. The architecture of the semantic segmentation in the DSP network is shown in Figure 3.

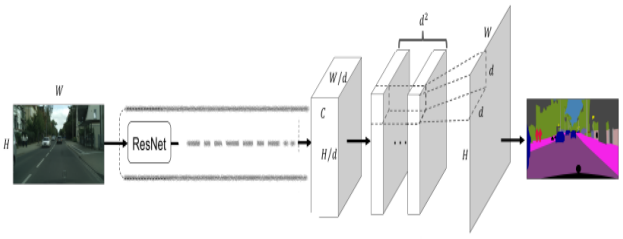


Figure 2: Semantic segmentation

In Figure 3, it is clearly shown that each object in the input image is separated with different colors for easy understanding of the surroundings. Semantic segmentation produces more than just a class label or a set of bounding box parameters. In actuality, the output is a complete, high-resolution image with pixel classification.

ii) Distance Estimation as well as Object Detection using Hybrid DSP-Depthnet:

In various street scenes, an image that is given as input can efficiently extract features thanks to the shared convolutional architecture design in DSPnet. Tasks related to feature extraction are carried out using the residual network. More specifically, the shared convolutional architecture for image feature encoding uses a 3x3 convolution along with the first four residual blocks that were pre-trained to complete an image classification task on the Kitti dataset [14]. There are 3,4,6,3 residual units in the 1,2,3,4th residual blocks for resent-50, which was used in the suggested network model. The resolution of the image features within each residual block is cut in half, both in terms of height and width, compared to the previous block size. There are now twice as many channels as in the previous block. The residual blocks near the input layers often extract low-level features and maintain the location information of objects in the input image because the weights in the network model were previously trained to predict image categories. The residual blocks near the output layers typically predict categorical image features and have fewer details about the locations of objects.

Each ConvLSTM layer in the Depthnet's encoder layer contains N states, where N denotes the number of timestamps in total. The decoder layer reconstructs the depth maps that were individually learned for each state. The decoder layer skips connections between the encoder and decoder layers and has N distinct deconvolutional layers. The depth map can be recreated more precisely thanks to this decoder architecture, which has been demonstrated to perform well for several reconstruction tasks. In the Depthnet, while the decoder layer learns to reconstruct the N individual depth maps, the encoder layer learns the spatiotemporal relationships between N image frames in order to predict the depth maps.

Instead of using the conventional fully-connected LSTM, we can use the ConvLSTM, which preserves all spatial information, to jointly exploit the ability of the various convolutional layers to capture appearance cues at various spatial scales and the ability of the LSTM to make sense of the temporal variations in the input data. A set of 3x3 ConvLSTM layers with various filters make up the encoding phase. Each convolutional step of the ConvLSTM layer uses the relu activation function, but the recurrent step uses the hard sigmoid activation function. All layers have the padding set to the same value. The first ConvLSTM layer accepts an input R_i of the form $[B*T*H*\omega*\subset]$, which B represents the batch size, T shows the total number of time steps, and H , ω are the height and width of the image, and $\subset$ stands for the number of image channels. For use in the decoding stage, each ConvLSTM layer is built to return the entire sequence rather than just the result. A convolutional layer with sigmoid activation toward the end of the expansion phase to obtain the depth map is used. In the object detection phase, the DSPnet output is combined with the Depthnet output. The non-maximum suppression combines both networks' output and provides the depth estimation, and bounding box with distance. The output obtained from the hybrid DSP-Depthnet is fed forward to the novel ESO optimization method.

3.4 Sparrow Egret Optimization:

A maximum and minimum function evaluation is produced by the optimization process, which involves iteratively training the model. One of the most significant trends in machine learning (ML) is the desire for better outcomes. In this research, the SSO has the benefits of quick convergence and high accuracy but has a poor ability to leave the local optimum. ESO and SSO are combined to achieve the global optimum value, which puts forth a parallel structure that balances exploitation and exploration while providing outstanding performance and stability. To obtain the global optimal value, the SSO's position update strategy and the ESO's hunting strategy are combined in this research.

3.4.1 Motivation:

Around the world, farms and forests are home to a variety of small birds known as sparrows. Sparrows can be found in a wide range of environments and climes, but they typically stay closer to populated areas and stay away from thick forests, grasslands, and terrains. This species has two different kinds of groups such as producers and scroungers, while the producers look for sources of food, the scroungers get their food by pestering the producers. Additionally, it can be said that sparrows typically combine their producer and scrounger strategies to find food. Each sparrow in the group keeps an eye on how the others behave, according to the researchers. In the meantime, the flock's predatory birds compete for the food sources of their friends who eat a lot, increasing their predation rate. The birds in the middle of the flock move closer to their closest neighbors to reduce their likelihood of danger, while the birds at the edges of the population are more likely to be attacked by predatory animals and are constantly looking for a better position. Group

intelligence, foraging, and anti-predator behaviors of sparrows are SSO's main advantages. The sit-and-wait strategy of the Snowy Egret and the aggressive strategy of the Great Egret served as inspiration for ESO, which combined the benefits of both strategies and built a correlating computational formula to quantify the behaviors. Three key elements make up the parallel algorithm known as ESO: the sit-and-wait strategy, the aggressive strategy, and the discriminant condition. One Egret squad consists of three Egrets; Egret A uses a guiding forward mechanism while Egret B and Egret C, respectively, use random walking and encircling mechanisms. In terms of exploitation and exploration, ESO will be more balanced and able to conduct quick searches for workable solutions. ESO is less likely to reach the saddle point of the optimization problem than gradient descent because it uses stochasticity and historical data in the gradient estimation. By estimating the optimization problem's tangent plane, ESOA distinguishes itself from other meta-heuristic algorithms and enables a quick descent to the current optimal point. In this research, the optimization characteristics like sparrows searching for prey, and the ESO sit-and-wait strategy makes it easier to find prey without requiring more energy to do so is utilized to create the best optimal result. The SEO's primary function is to optimize the hyperparameter, which increases the precision of distance estimation.

3.4.2 Mathematical formulation:

The SSA is a metaheuristic algorithm that can be applied to various fields of optimization. The SSA is a competitive algorithm in that individuals have a better chance of obtaining food in the solution space if their cost values are higher. The locations of the sparrows are randomly distributed throughout the solution space. The following positions are attained by the sparrows when they first become aware of the threat:

$$X_{i,j}^{t+1} = \begin{cases} X_B^t + \delta \times \left| X_{i,j}^t - X_B^{t+1} \right| , & if\ f_i \rangle\ f_g \\ X_{i,j}^t + R \times \left(\dfrac{\left| X_{i,j}^t - X_W^{t+1} \right|}{(f_i - f_w) + \sigma} \right), & if\ f_i = f_g \end{cases} \tag{4}$$

where, δ is the random value with 0 mean and 1 variance value. σ is the smallest constant to avoid zero division error, X_B is the global optimal position. R is the random value between -1 and +1. f_i is the individual cost value, f_g is the current global best, and f_w is the worst fitness value. $X_{i,j}^t$ is the value of the i^{th} individual at j^{th} dimension at t^{th} iteration. X_W is the global worst position.

Instead of the scroungers, the producers search in a variety of search spaces for food. In the absence of a predator, the producer's position is mathematically modeled as follows:

$$X_{i,j}^t = \left\{ X_{i,j}^t \times EXP\left(\frac{i}{\phi \times Itr_{\max}} \right) \right\} if\ A_{v2} < S_t \tag{5}$$

where $Itr_{\max}$ is the constant with a higher iteration number, t show the present iteration, the value of j varying from 1 to d, ϕ is the random variable with the value of 0 or 1. A_{v2} is the alarm value varying from 0 or 1. S_t is the safest threshold value of 0.5 and 1.

The sparrows expended more energy in their hunt for prey, and there is no guarantee that they will find food. Therefore, the sit-and-wait strategy will be made possible for the sparrows' food foraging of ESO, in order to conserve their energy. With less energy expended, the position of the prey can be determined because of this enabling.

$$X_{i,j}^t = \left\{ X_{i,j}^t \times EXP\left(\frac{i}{S_E(x_i) \times Itr_{\max}} \right) \right\} \ if\ A_{v2} < S_t \tag{6}$$

where, $S_E(x_i)$ is the snowy egrets' position in n dimension for varying i.

The position of the producer in the case of predator is mathematically modeled as follows:

$$X_{i,j}^t = \left\{ X_{i,j}^t + N \times D \right\} \ if\ A_{v2} \geq S_t \tag{7}$$

where, N is the normal distributed random number, and D is the d dimension vector where every element inside hold the value of 1.

If the alarm value is below the safety threshold, there is no predator nearby, and the producer has switched to the wide search mode. However, if the alarm value is greater than or equal to the safety threshold, sparrows have identified the predator, and everyone needs to fly quickly to other secure areas. Considering previous explanations, some of the Scroungers frequently follow the producers as soon as they locate the producer of high-quality food, they rush to contend for it. Once they triumph over the rivals, they can purchase food from the producer, otherwise, they keep following the rules. The revised formulation of the best scrounger's position is as follows:

$$X_{i,j}^{t+1} = X_P^{t+1} + \left| X_{i,j}^t - X_P^{t+1} \right| \times A^+ \times D \tag{9}$$

where, A describes the 1xd vector such that the elements are randomly assigned from -1 or 1.

$$X_{i,j}^{t+1} = N \times EXP(X_W^t - X_P^{t+1}) \ \ if\ i > n/2 \tag{8}$$

where, X_W^t is the present global worst, and X_P optimal position found by the producer. Here the worst scrounger is considered and to improve their characteristics the following behavior of the scroungers is integrated with the egrets' following behavior so that the efficiency of the worst scenario also gets improves.

$$X_{i,j}^{t+1} = 0.5\left\{ N \times EXP(X_W^t - X_P^{t+1}) + \left(\frac{X_{iB} - X_i}{|X_{iB} - X_i|} \cdot \frac{f_{iB} - f_i}{|X_{iB} - X_i|} \times d_{iB} \right) \right\} \quad (9)$$

where d_{iB} is the directional correction of the best location, and f_{iB} is the best fitness value in the i^{th} location in n dimension. Here the egrets following behavior and scroungers following behavior to obtain food are combined to produce the best optimal equation, which will tune the weight and bias of the classifier.

The pseudo-code for the optimization is given in Table 1

s.n	Pseudo code of the proposed SEO-Hybridized DSP-Depth network
1	Start
2	Initialize the position of sparrows
	$X_{i,j}^{t+1} = \begin{cases} X_B^t + \delta \times \left| X_{i,j}^t - X_B^{t+1} \right| & if\ f_i > f_g \\ X_{i,j}^t + R \times \left(\frac{\left| X_{i,j}^t - X_W^{t+1} \right|}{(f_i - f_w) + \sigma} \right) & f_i = f_g \end{cases}$
3	Determine the fitness function
4	*if*
	$A_{v2} < S_t$
5	$X_{i,j}^t = \{ X_{i,j}^t + N \times D \}$
6	*if*
	$A_{v2} \geq S_t$
7	$X_{i,j}^t = \{ X_{i,j}^t + N \times D \}$
8	For, Best Scrounger
9	$X_{i,j}^{t+1} = \{ X_P^{t+1} + \left| X_{i,j}^t - X_P^{t+1} \right| \times A^+ \times D \}$
10	For, the worst Scrounger
11	$X_{i,j}^{t+1} = 0.5\left\{ N \times EXP(X_W^t - X_P^{t+1}) + \left(\frac{X_{iB} - X_i}{\lvert X_{iB} - X_i \rvert} \cdot \frac{f_{iB} - f_i}{\lvert X_{iB} - X_i \rvert} \times d_{iB} \right) \right\}$
12	Worst Scrounger change to Best Scrounger
13	end

4. Result and discussions:

Results from segmentation and object detection are included in this section. By changing the training percentage, the performance is assessed, and existing models are used for comparative analysis.

4.1 Experimental setup

The proposed method is written in Python, and the research system configuration uses Python 2020, which runs on Windows 10 and has 8 GB of internal memory.

4.1.1 Dataset Description

Kitti dataset [14]: The vision tasks in the Kitti dataset were created using an autonomous driving platform. Numerous tasks, including stereo, optical flow, and visual odometry, are included in the complete benchmark. The object detection dataset is included in this dataset, along with bounding boxes and monocular images.

4.1.2. Performance metrics

a) Mean Absolute Error (MAE): MAE is the error measure between the predicted values and actual values that express the same aspects.

$$\mu_{abs} = \frac{\sum_{i=1}^{S_{size}} |\rho_i - \tau_i|}{S_{size}} \quad (10)$$

where ρ represents the predicted value, τ represent the actual value, and S_{size} represents the sample size.

b) Mean Squared Error (MSE): MSE estimates the average square difference between the predicted and the real values is represented as;

$$\mu_{MSE} = \frac{\sum_{i=1}^{S_{size}} |\rho_i - \tau_i|^2}{S_{size}} \quad (11)$$

c) Root Mean Square Error (RMSE): The RMSE is estimated by taking the root of the MSE and it is mathematically illustrated as follows;

$$R\mu_{MSE} = \sqrt{\frac{\sum_{i=1}^{S_{size}} |\rho_i - \tau_i|^2}{S_{size}}} \quad (12)$$

The performance metrics are used for the estimation of the object detection model. The elaboration of the metrics is provided as follows.

d) Accuracy: It is a statistical measure, which demonstrates the effectiveness of the classifier in identifying the objects. The mathematical description of the accuracy is given as,

$$\alpha c = \frac{\Gamma_{pos} + \Gamma_{neg}}{\Gamma_{pos} + \Gamma_{neg} + fl_{pos} + fl_{neg}} \quad (13)$$

where Γ_{pos} represents the true positive, Γ_{neg} denotes the true negative, fl_{pos} denotes the false positive, and fl_{neg} denotes the false negative

e) Sensitivity: It is the percentage of positive class appropriately identified from the image

$$sn = \frac{\Gamma_{pos}}{\Gamma_{pos} + fl_{neg}} \quad (14)$$

f) Specificity: It is the percentage of the negative class that was correctly identified from the image

$$sp = \frac{\Gamma_{neg}}{\Gamma_{neg} + fl_{pos}} \quad (15)$$

4.2 Performance evaluation:

This section illustrates the performance of the proposed model in segmentation, distance estimation, and object detection.

4.2.1 Performance evaluation for segmentation:

The performance achieved by the proposed DSPDepthnet with SEO is illustrated in Figures 3 a), b), and c). At 40% of training and the 30th epoch, the proposed model provides an MAE of 5.04, which is reduced with a rise in training percentage. Hence, the model attains an error of 3.16 at 90% of training. In the case of MSE, the proposed DSPDepthnet with SEO gains the MSE of 6.26 at 40% of training and in the 30th epoch, which gradually decreased to 3.89 at 60% of training and attains the lowest MSE of 2.3 at 90% of training. While considering RMSE, the error attained by DSPDepthnet with SEO is found to be 4.16 at 40% training and the 30th epoch. The RMSE is minimized to 2.25 at

90% of training. From the above observation, it is demonstrated that the proposed approach attains the lowest errors, in terms of MAE, MSE, and RMSE. This lowest error is due to the tuning of parameters of the DSPDepthnet network.

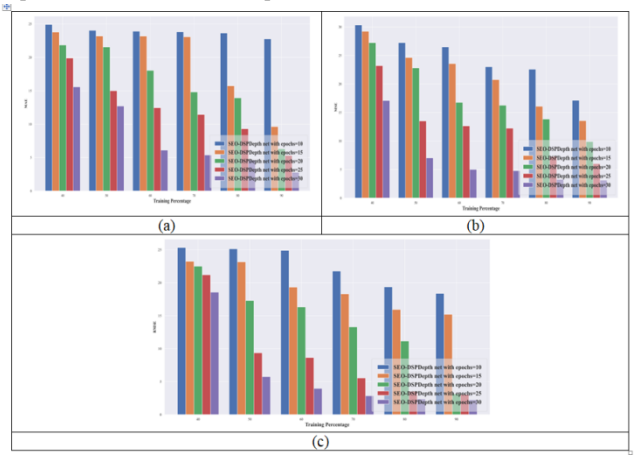


Figure 3. Performance analysis of segmentation process (a) MAE, (b) MSE, and (c) RMSE

4.2.2 Performance evaluation for object detection

The performance achieved by the DSPDepthnet with SEO is illustrated in Figures 4 a), b), and c). 9 It is seen that model attains an accuracy of 0.99 at 90% of training. In the case of sensitivity, the model attains a sensitivity of 0.93 at 90% of training. On considering specificity, the proposed DSPDepthnet with SEO obtains a specificity of 0.92 at the initial level of training of 40%. The specificity is enhanced to 0.98 at 90% of training and the 30th epoch. From the above observation, it is demonstrated that the model attains the highest performance, in terms of metrics. This highest performance is due to the tuning of parameters of the DSPDepth network.

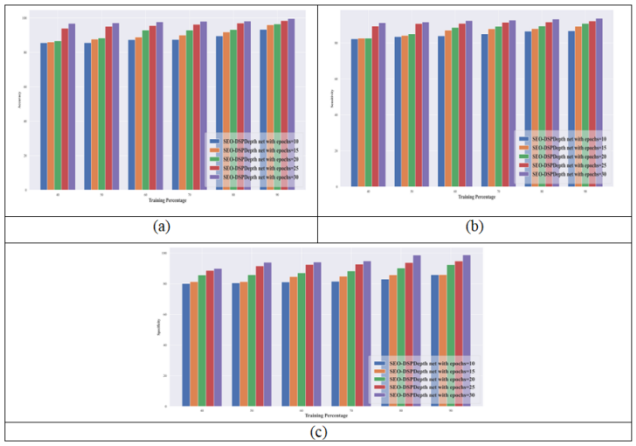


Figure 4. Performance analysis of object detection (a) accuracy, (b) sensitivity, and (c) specificity

4.2.3 Performance evaluation for distance estimation:

The performance achieved by the proposed DSPDepthnet with SEO on distance estimation is illustrated in Figures 5 a), b), and c). the model attains an error of 5.83 at 90% of training for MAE. In the case of MSE, the proposed DSPDepthnet with SEO gains an MSE of 10.12 at 90% of training. While considering RMSE, the error attained by DSPDepthnet with SEO is found to be 9.50 at 90% of training.

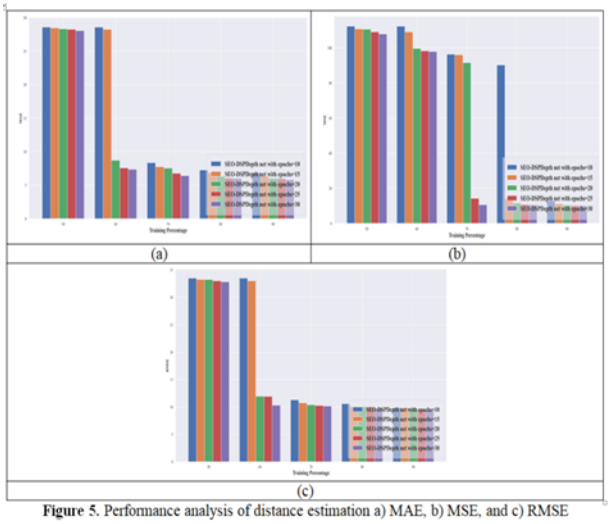


Figure 5. Performance analysis of distance estimation a) MAE, b) MSE, and c) RMSE

4.3 Comparative analysis

This section explains the comparative analysis for image segmentation, distance estimation, and object detection.

a) Comparative analysis for segmentation

Figure 6 a), b), and c) illustrates the comparative analysis in terms of MAE, MSE, and RMSE.

Figure 6 a) exhibits that the proposed DSPDepthnet with SEO model attains the best performance than the conventional model like Unet [24], DSP[11], DSP with PSO[11][25], DSP with GWO[11][26], DSP with SaHO, DSPSDepthnet with SSO [11][16], and DSPSDepthnet with ESO[11][15]. The proposed DSPDepthnet with the SEO model obtains an MAE value of 2.95.

Figure 6 b) illustrates the comparative analysis of the image segmentation models for segmenting the images concerning MSE. The MSE of the proposed DSPDepthnet with the SEO model obtains an MSE value of 3.16. Figure 6 c) illustrates the comparative analysis of the segmentation models regarding RMSE value. The proposed DSPDepthnet with SEO segmentation model attains the lowest RMSE of 3.14 among all the competent models From analysis, it is observed that the proposed DSPDepthnet with SEO model gains the highest performance. The performance improvement is due to the proposed SEO model, which is employed to tune the parameters.

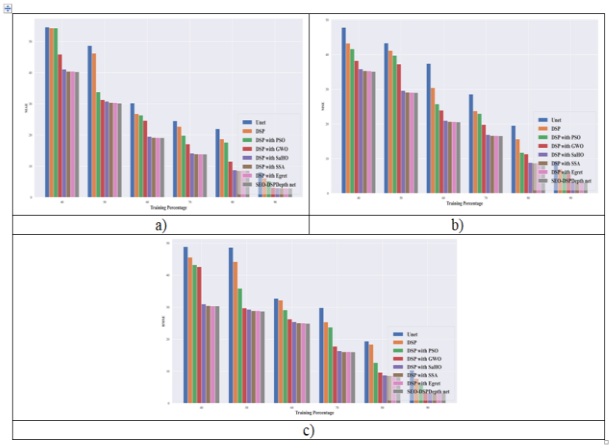


Figure 6. Comparative analysis of segmentation process a) MAE, b) MSE, and c) RMSE

b) Comparative analysis for object detection

The performance of the competent model, like MLP[27], RF [28], BiLSTM [29], DSP, DSP with PSO, DSP with GWO, DSP with SaHO, DSPDepthnet with SSO, and DSPSDepthnet with ESO are compared with the proposed method DSPSDepthnet with SEO. For object detection, the proposed model attains an accuracy of 0.99 at 90% of training. While analyzing the sensitivity of the model, the proposed model attains a sensitivity of 0.97 at 90% of training. The proposed model achieves a specificity of 0.95. Figure 7 a), b), and c) illustrates the comparative evaluation of the model in terms of accuracy, sensitivity, and specificity

Table 1. Comparative discussion for a segmentation model

Methods	90% of training		
	MAE	MSE	RMSE
Unet	8.10	9.27	10.32
DSP	6.16	7.01	7.73
DSP with PSO	5.29	5.93	5.78
DSP with GWO	3.14	5.50	3.48
DSP with SaHO	3.01	3.25	3.21
DSPDepthnet with SSA	2.96	3.18	3.14
DSPDepthnet with ESO	2.97	3.17	3.16
DSPDepthnet with SEO	2.95	3.16	3.14

Table 2. Comparative discussion for object detection model

Methods	90% of training		
	Accuracy	sensitivity	specificity
MLP	75.00	67.00	75.00
RF	81.00	73.00	78.00
BiLSTM	85.00	85.00	82.00
DSP	85.00	87.00	85.00
DSP with PSO	90.00	89.00	88.00
DSP with GWO	96.00	95.00	89.00
DSP with SaHO	98.00	96.00	94.00
DSPDepthnet with SSA	99.16	97.14	95.14
DSPDepthnet with ESO	99.23	97.23	95.29
DSPDepthnet with SEO	99.25	97.23	95.62

Table 3. Comparative discussion for distance estimation model

Methods	90% of training		
	MAE	MSE	RMSE
DCNN	27.09	102.45	32.01
BiLSTM	27.09	102.45	32.01
DSP based Depthnet	14.13	42.81	17.12
DSPDepthnet with SSA	7.95	29.31	10.92
DSPDepthnet with ESO	6.03	11.94	9.80
DSPDepthnet with SEO	3.28	10.12	5.59

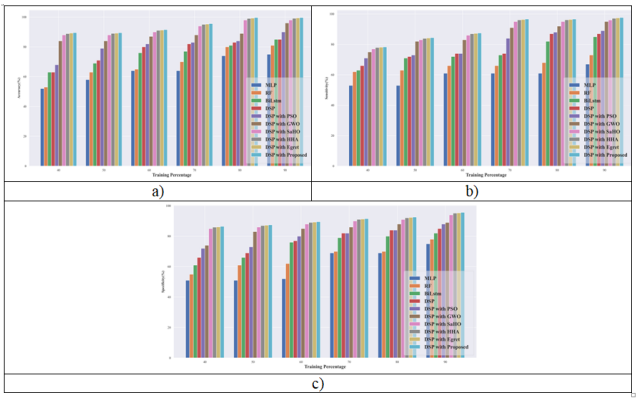


Figure 7. Comparative analysis of object detection process a) accuracy, b) sensitivity, and c) specificity

c) Comparative analysis for distance estimation:

For distance estimation, DCNN [30], BiLSTM, DSP-based Depthnet, DSPDepthnet with SSA, and DSPDepthnet with ESO, are taken as the competent model. Figure 8, illustrates the proposed DSPDepthnet with the SEO model obtains an MAE value of 3.27 at 90% of training. The MSE of the proposed DSPDepthnet with the SEO model obtains an MSE value of 10.12 at 90% of training. The proposed DSPDepthnet with SEO distance estimation model attains the lowest RMSE of 5.58 among all the competent models at 90% of training. From analysis, it is observed that the proposed DSPDepthnet with SEO model gains the highest performance in terms of MSE, MAE, and RMSE.

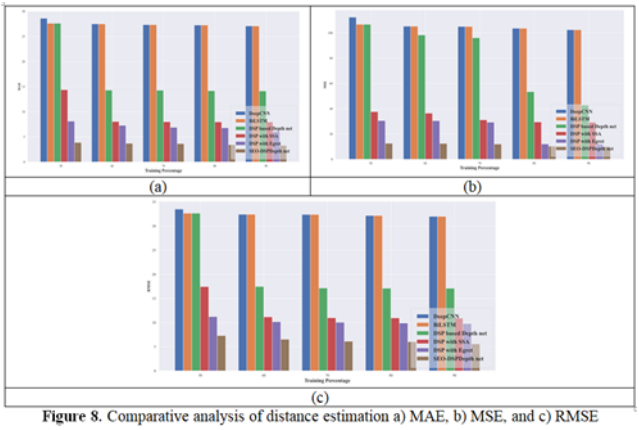


Figure 8. Comparative analysis of distance estimation a) MAE, b) MSE, and c) RMSE

4.4 Comparative discussion

This section illustrates the comparative discussion of the SEO - DSPDepthnet with the existing model for segmentation and object detection. The comparative analysis is accomplished by modifying the training percentage from 40 to 90%. For segmentation, the classifiers such as Unet, DSP, DSP with PSO, and DSP with GWO are used as the competent techniques, for object detection, the classifiers like MLP, RF, BiLSTM, DSP, DSP with PSO, DSP with GWO, DSP with SaHO, DSPDepthnet with SSO, and DSPSDepthnet with ESO are taken as the competent model, and for distance estimation, DCNN, BiLSTM, DSP based Depthnet, DSPDepthnet with SSA, and DSPDepthnet with ESO, are taken as the competent model. The performance metrics are used for the analysis of segmentation, while the metrics are used for the analysis of object detection. From the analysis, it is observed that all the classifier attains maximum performance at 90% of training. The performance attained by the segmentation, object detection, and distance estimation model at 90% of training is listed in Tables 1, 2, and 3 respectively. The improvement in the performance of the classifiers is due to the hyperparameter tuning of the DSPDepthnet classifiers. Further, the errors are significantly reduced by the optimal hyperparameter tuning by the SEO.

Table 3. Comparative discussion for distance estimation model

Methods	90% of training		
	MAE	MSE	RMSE
DCNN	27.09	102.45	32.01
BiLSTM	27.09	102.45	32.01
DSP based Depthnet	14.13	42.81	17.12
DSPDepthnet with SSA	7.95	29.31	10.92
DSPDepthnet with ESO	6.03	11.94	9.80
DSPDepthnet with SEO	3.28	10.12	5.59

Conclusion:

The novel Hybridized DSPDepthnet provides an effective method for concurrent object detection and depth estimation is proposed in this research because the demand for high-level autonomous driving has increased recently and visual perception is one of the essential features to enable fully autonomous driving. The proposed DSP-Depth net demonstrates the capability of LSTM-based RNNs to sequentially predict the depth map for an image frame as a function of the appearances of scene objects in the image frame in its temporal

neighborhood. The proposed method obtained convincing state-of-the-art results on the KITTI dataset when compared to existing depth-supervised approaches, demonstrating that the proposed method is capable of outperforming conventional CNN models at depth prediction. Although the proposed method is trained to predict depth for image sequences, it can also accurately predict depth maps on images. The comparative discussion show that the proposed model attains value for MAE, MSE, and RMSE are 2.95, 3.16, and 3.14 respectively for segmentation, the proposed method attains a value for accuracy, specificity, and sensitivity are 0.99, 0.97, and 0.95 respectively for object detection, and the proposed model attains value for MAE, MSE, and RMSE are 3.28, 10.12, and 5.59 respectively for distance estimation. In the future video-based estimation of the proposed method can be done to analyze the real-time implementation of this method.

Author contributions

Shwetambari G.Pundkar : Conceptualization, Methodology, Software, Field study, Writing-Reviewing and Editing. Amit K. Gaikwad: Field study, Visualization

Conflicts of interest

The authors declare no conflicts of interest.

Reference:

[1] Bai, Lin, Yiming Zhao, Mahdi Elhousni, and Xinming Huang. "DepthNet: Real-time LiDAR point cloud depth completion for autonomous vehicles." *IEEE Access* 8 (2020): 227825-227833.

[2] Zováthi, Örkény, Balázs Pálffy, Zsolt Jankó, and Csaba Benedek. "ST-DepthNet: A spatio-temporal deep network for depth completion using a single non-repetitive circular scanning Lidar." *IEEE Robotics and Automation Letters* (2023).

[3] Jun, Jinyoung, Jae-Han Lee, Chul Lee, and Chang-Su Kim. "Monocular human depth estimation via pose estimation." *IEEE Access* 9 (2021): 151444-151457.

[4] Wang, Linda, Mahmoud Famouri, and Alexander Wong. "DepthNet nano: A highly compact self-normalizing neural network for monocular depth estimation." *arXiv preprint arXiv:2004.08008* (2020).

[5] CS Kumar, Arun, Suchendra M. Bhandarkar, and Mukta Prasad. "Depthnet: A recurrent neural network architecture for monocular depth prediction." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 283-291. 2018.

[6] Zhou, Shi, Miaomiao Zhu, Zhen Li, He Li, Mitsunori Mizumachi, and Lifeng Zhang. "Self-supervised monocular depth estimation with occlusion mask and edge awareness." *Artificial Life and Robotics* 26, no. 3 (2021): 354-359.

[7] Zhao, Chaoqiang, Yang Tang, Qiyu Sun, and Athanasios V. Vasilakos. "Deep direct visual odometry." *IEEE transactions on intelligent transportation systems* 23, no. 7 (2021): 7733-7742.

[8] Li, Kunhong, Zhiheng Fu, Hanyun Wang, Zonghao Chen, and Yulan Guo. "Adv-depth: Self-supervised monocular depth estimation with an adversarial loss." *IEEE Signal Processing Letters* 28 (2021): 638-642.

[9] Chen, Liangfu, Zeng Yang, Jianjun Ma, and Zheng Luo. "Driving scene perception network: Real-time joint detection, depth estimation and semantic segmentation." In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1283-1291. IEEE, 2018.

[10] Ali, Abduladhem Abdulkareem, and Hussein Alaa Hussein. "Distance estimation and vehicle position detection based on monocular camera." In *2016 Al-Sadeq International Conference on Multidisciplinary in IT and Communication Science and Applications (AIC-MITCSA)*, pp. 1-4. IEEE, 2016.

[11] Zeng, Xin, Yunpeng Wu, Shizhe Hu, Ruobin Wang, and Yangdong Ye. "DSPNet: Deep scale purifier network for dense crowd counting." *Expert Systems with Applications* 141 (2020): 112977.

[12] Zhu, Hu, Yiming Qiao, Guoxia Xu, Lizhen Deng, and Yu-Feng Yu. "Dspnet: A lightweight dilated convolution neural networks for spectral deconvolution with self-paced learning." *IEEE Transactions on Industrial Informatics* 16, no. 12 (2019): 7392-7401.

[13] Yu, Hyeonwoo, and Jean Oh. "Anchor distance for 3d multi-object distance estimation from 2d single shot." *IEEE Robotics and Automation Letters* 6, no. 2 (2021): 3405-3412.

[14] Kitti-dataset https://www.kaggle.com/datasets/klemenko/kitti-dataset - (Accessed on July 2023)

[15] Chen, Zuyan, Adam Francis, Shuai Li, Bolin Liao, Dunhui Xiao, Tran Thu Ha, Jianfeng Li, Lei Ding, and Xinwei Cao. "Egret swarm optimization algorithm: an evolutionary computation approach for model free optimization." *Biomimetics* 7, no. 4 (2022): 144.

[16] Zhu, Yanlong, and Nasser Yousefi. "Optimal parameter identification of PEMFC stacks using adaptive sparrow search algorithm." *International journal of hydrogen energy* 46, no. 14 (2021): 9541-9552.

[17] Eigen, David, Christian Puhrsch, and Rob Fergus. "Depth map prediction from a single image using a multi-scale deep network." *Advances in neural information processing systems* 27 (2014).

[18] Li, Tianyi, Yanmei Liu, and Zhen Chen. "Application of Sine Cosine Egret Swarm Optimization Algorithm in Gas Turbine Cooling System." *Systems* 10, no. 6 (2022): 201.

[19] Wang, Yukun, Aiying Zhao, Xiaoxue Wei, and Ranran Li. "A Novel Ensemble Model Based on an Advanced Optimization Algorithm for Wind Speed Forecasting." *Energies* 16, no. 14 (2023): 5281.

[20] Wang, Peng, Yu Zhang, and Hongwan Yang. "Research on economic optimization of microgrid cluster based on chaos sparrow search algorithm." *Computational Intelligence and Neuroscience* 2021 (2021): 1-18.

[21] Dong, Jun, Zhenhai Dou, Shuqian Si, Zichen Wang, and Lianxin Liu. "Optimization of capacity configuration of wind–solar–diesel–storage using improved sparrow search algorithm." *Journal of Electrical Engineering & Technology* 17 (2022): 1-14.

[22] Fan, Yanyan, Yu Zhang, Baosu Guo, Xiaoyuan Luo, Qingjin Peng, and Zhenlin Jin. "A hybrid sparrow search algorithm of the hyperparameter optimization in deep learning." *Mathematics* 10, no. 16 (2022): 3019.

[23] Wang, Rui, Stephen M. Pizer, and Jan-Michael Frahm.

"Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5555-5564. 2019.

[24] Sun, Y., Bi, F., Gao, Y., Chen, L. and Feng, S., "A Multi-Attention UNet for Semantic Segmentation in Remote Sensing Images", Symmetry, vol.14, no.5, pp.906, 2022.

[25] Poli, R., Kennedy, J. and Blackwell, T., "Particle swarm optimization", Swarm intelligence, vol.1, no.1, pp.33-57, 2007.

[26] Bansal, Jagdish Chand, and Shitu Singh, "A better exploration strategy in Grey Wolf Optimizer," Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 1, pp. 1099-1118, 2021.

[27] Qin, L., Yang, Y., Huang, D., Zhu, N., Yang, H. and Xu, Z., "Visual Tracking With Siamese Network Based on Fast Attention Network", IEEE Access, vol.10, pp.35632-35642, 2022.

[28] Zhang, L., Varadarajan, J., Nagaratnam Suganthan, P., Ahuja, N. and Moulin, P., "Robust visual tracking using oblique random forests", In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5589-5598, 2017.

[29] Pan, C., Shi, D., Guan, N., Zhang, Y., Wang, L. and Jin, S., "Learning to Track by Bi-Directional Long Short-Term Memory Networks", In proceedings of 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pp. 783-790, 2019.

[30] Ma, Benteng, Xiang Li, Yong Xia, and Yanning Zhang. "Autonomous deep learning: A genetic DCNN designer for image classification." *Neurocomputing* 379 (2020): 152-161.