Volume 9, Issue 3, March – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR1535

Leveraging Mobile Net and Yolo Algorithm for Enhanced Perception in Autonomous Driving

Rithika Grace R[1]
Electronics and Communication Engineering
St. Joseph's College of Engineering Chennai, India

Vaishnavi H[2]
Electronics and Communication Engineering
St. Joseph's College of Engineering Chennai, India

Angelin Ponrani M[3]
Electronics and Communication Engineering
St. Joseph's College of Engineering Chennai, India

Abstract:- This project presents a cutting-edge autonomous navigation system designed to enhance spatial understanding and object recognition for vehicles. Equipped with cameras on both the left and right sides, by capturing comprehensive view of the surroundings, utilizing stereo vision for crucial depth information. The integration of Mobile Net and YOLO (You Only Look Once) algorithms is central to achieving real-time object detection and recognition. Mobile Net is employed for efficient feature extraction, ensuring optimal computational efficiency. Simultaneously, YOLO plays a pivotal role in rapid and accurate identification of objects within the captured images, contributing to the robustness of spatial understanding crucial for autonomous navigation. The result is a comprehensive and reliable autonomous navigation system, showcasing the effectiveness of combining cutting-edge technologies for improved real time decision making.

Keywords:- *YOLO, Mobile Net, Stereo Vision, Autonomous Navigation.*

I. INTRODUCTION

In our quest to enhance autonomous driving, our vehicle utilizes stereo vision technology, mimicking human binocular vision to analyze disparities between images captured from two viewpoints. This depth perception enhances navigation in complex environments. Our system integrates Mobile Net and YOLO algorithms for real-time object detection and recognition. Mobile Net ensures efficient feature extraction, while YOLO's single-pass object detection enhances accuracy. Beyond driving, our project pioneers the fusion of stereo vision with advanced algorithms, setting a precedent for comprehensive self-driving experiences. This integration marks a technological milestone, showcasing the transformative potential of intelligent navigation systems. Stereo vision employs area or feature-based matching, representing a foundational aspect of computer vision research. The paper [1] conveys that the Current methodologies face two critical challenges: 1) inadequate detection accuracy, especially for challenging objects like pedestrians, and 2) significant performance decline with additional noise points. Addressing these issues, this paper introduces TA Net, featuring a Triple Attention (TA) module and a Coarse-to-Fine Regression (CFR) module. The TA module combines channel-wise, point-wise, and voxel-wise attention to enhance target information and suppress unstable points. Operating at approximately 29 frames per second, TA Net improves object detection in 3D point clouds, crucial for applications like autonomous navigation, robotics, and augmented reality, without relying on hand-crafted feature representations. In [2] the authors Zhou, Yin, and Tuzel introduced Voxel Net, a novel approach eliminating manual feature engineering in 3D point clouds. Voxel Net integrates feature extraction and bounding box prediction into a single-stage, end-to-end trainable deep network. It divides point clouds into uniformly spaced 3D voxels. Experimental results on the KITTI car detection benchmark show Voxel Net outperforms state-of-the-art LiDAR-based 3D detection methods. LiDAR-based or RGB-D-based object detection in [3]is highlighted for its applications across domains like autonomous driving and robot vision. Voxel-based 3D convolutional networks are commonly used to enhance information retention in processing LiDAR point cloud data. The paper introduces an enhanced sparse convolution method tailored for such networks, leading to accelerated training and inference speeds. The proposed network achieves state-of-the-art results on the KITTI 3D object detection benchmarks while maintaining fast inference speeds. In their study [4] Shi et al. underscored the increasing significance of 3D object detection across diverse computer vision applications. Traditional methods predominantly depend on LiDAR point cloud data, encountering obstacles related to regional alignment and distribution reliability. In response, they proposed PIF Net, an end-to-end learning framework that seamlessly integrates LiDAR point cloud and camera sensor data. PIF Net capitalizes on smart points and voxel features to mitigate data loss and enhance overall performance. Rigorous evaluation conducted on the KITTI dataset demonstrated PIF Net's superiority over existing state-of-the-art techniques. This innovative approach represents a substantial stride forward in addressing the challenges associated with 3D object detection, promising enhanced accuracy and robustness in real-world scenarios. The paper [5] introduces MV3D, a sensory-fusion framework for high-accuracy 3D object detection in autonomous driving scenarios. MV3D utilizes both LIDAR point clouds and RGB

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR1535

images to predict oriented 3D bounding boxes. It encodes the sparse 3D point cloud into a concise multi-view representation. The network consists of two subnetworks: one for generating 3D object proposals and another for multi-view feature fusion. The proposal network efficiently generates 3D candidate boxes from the bird's eye view representation of the point cloud. A deep fusion scheme combines region-wise features from multiple views, facilitating interactions between intermediate layers of different paths. The bird's eye view map is encoded with (M +2)-channel features, where M represents equally divided slices of point clouds. The coordinates of a 3D point in the front view map can be computed using $\Delta\theta$ and $\Delta\varphi$, representing the horizontal and vertical resolution of laser beams.

In their paper [6] the authors delve into the pursuit of optimal speed and accuracy in object detection through a comprehensive exploration of various features aimed at enhancing Convolutional Neural Networks (CNNs). This investigation involves empirical testing across extensive datasets, coupled with theoretical justifications to elucidate the efficacy of each feature. Notably, certain features demonstrate efficacy tailored to specific models or problems, while others such as batch normalization and residual connections offer broad applicability across various contexts. Key findings from this study highlight universally beneficial features, including Weighted-Residual-Connections (WRC), Cross-Stage-Partial-connections (CSP), Cross mini-Batch Normalization (Cm BN), Self-adversarial-training (SAT), and Mish-activation. These identified features contribute significantly to advancing the state-of-the-art in object detection, promising improved performance and efficiency across a range of applications in computer vision. In paper [7] the authors Zhang, K., Wang, Y., Qi, X., & Qi, X. Discusses the applications of LiDAR-based or RGB-D-based object detection in fields like autonomous driving and robot vision. Voxel-based 3D convolutional networks are utilized to improve information retention in processing LiDAR point cloud data. However, challenges remain, including slow inference speeds and suboptimal orientation estimation performance. The study explores an enhanced sparse convolution method for these networks, significantly improving both training and inference speeds. The paper [8] aims to perform 3D object detection for autonomous driving, focusing on generating high-quality 3D object proposals using stereo imagery. The approach minimizes an energy function incorporating object size priors, ground plane object placement, and depth-informed features. A Convolutional Neural Network (CNN) is then applied to these proposals for object detection. Experiments show significant performance improvements over existing RGB and RGB-D methods on the KITTI benchmark. When combined with CNN, the approach outperforms existing results in object detection and orientation estimation across all KITTI object classes. It also explores scenarios where LIDAR information is available, demonstrating the best results when utilizing both LIDAR and stereo inputs. In [9] they propose an unsupervised framework for training deep Convolutional Neural Networks (CNNs) for single view depth prediction, eliminating the need for vast amounts of labeled data. Our method employs an auto encoder-like approach, training the network to predict depth maps from image pairs with known camera motion. By reconstructing the source image from the predicted depth and inter-view displacement, we minimize photometric error as the reconstruction loss. This approach simplifies data acquisition, requiring no manual annotation or depth sensor calibration. With training on less than half of the KITTI dataset, our network achieves performance comparable to state-of-the-art supervised methods for single view depth estimation. In their research, V. Prasad and B. Bhowmick introduce a method for monocular Visual Odometry (VO) in [10] that relies on leveraging Epi polar constraints to ensure robust geometric learning. By imposing meaningful constraints using the Essential matrix obtained from Nister's Five Point Algorithm, the accuracy of depth and relative pose estimations sees notable enhancement. Despite its simplicity and fewer parameters, their approach achieves comparable performance to state-of-the-art techniques, highlighting the effectiveness of incorporating Epi polar constraints. This geometrically constrained learning strategy demonstrates its efficacy particularly in scenarios where relying solely on minimizing photometric error may prove inadequate. The proposed method represents a significant advancement in monocular VO methodologies, promising improved accuracy and reliability in various real-world applications. The paper [11] presents a novel approach for depth estimation in single images using binocular stereo footage instead of explicit depth data. By leveraging epi polar geometry constraints, our convolutional neural network learns to predict depth without ground truth data. We introduce a training objective that enforces consistency between disparities relative to both left and right images, yielding superior performance on the KITTI dataset. Our method outperforms supervised approaches, demonstrating state-of-the-art results in monocular depth estimation. [12] Tackles the integration of depth and ego motion networks in self-supervised structure-from-motion (SfM) learning, proposing a novel tightly-coupled approach that leverages iterative view synthesis for contextual information exchange without weight sharing. Through comprehensive experiments, our method enhances consistency between depth and ego motion predictions, improves generalization on new data, and achieves state-of-the-art accuracy on indoor and outdoor depth and ego motion benchmarks. Deep SLAM, developed by R. Li, D. Gu and S. Wang, in [13] disrupts visual SLAM through its use of unsupervised deep learning trained solely on stereo imagery, seamlessly transitioning to monocular sequences during testing. Its components include MappingNet for 3D structure depiction, Tracking-Net for camera motion tracking, and Loop-Net for loop closure identification, providing pose estimations, depth maps, and outlier rejection masks simultaneously. Renowned for its remarkable accuracy and robustness, Deep SLAM showcases exceptional performance across diverse datasets, particularly excelling in challenging environments. This groundbreaking approach solidifies Deep SLAM as a pioneering paradigm in monocular SLAM methodologies, promising significant advancements in the field of simultaneous localization and mapping. In their work,[14] the authors present advancements in self-supervised monocular depth estimation, aiming to overcome challenges associated with obtaining per-pixel ground-truth

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR1535

depth data. Their approach introduces several key enhancements, including the integration of a minimum reprojection loss to effectively handle occlusions, utilization of full-resolution multi-scale sampling to mitigate artifacts, and incorporation of an auto-masking loss mechanism to disregard training pixels that violate camera motion assumptions. These enhancements collectively contribute to superior depth map predictions compared to existing self-supervised methods. Notably, the proposed approach achieves state-of-the-art results on the KITTI benchmark, showcasing its effectiveness and reliability in real-world scenarios. The findings underscore the significance of these advancements in advancing the accuracy and robustness of self-supervised monocular depth estimation techniques, with implications for various applications in computer vision and autonomous systems. The paper [15] proposes a self-supervised monocular depth estimation method, Pack Net, leveraging geometry and novel symmetrical packing and unpacking blocks. Without labeled data, Pack Net surpasses other self, semi, and fully supervised methods on the KITTI benchmark. Its 3D inductive bias enables scalability with input resolution and parameters, generalizing well on Nu Scenes data. Pack Net operates in real-time without large-scale supervised pretraining and benefits from DDAD, a new urban driving dataset with accurate depth evaluation from high-density Li DARs on self-driving cars.

II. METHODOLOGY

In our proposed system the vehicle is equipped with cameras positioned on both the left and right sides, providing a comprehensive view of its surroundings. The utilization of stereo vision for depth information is integral to achieving robust spatial understanding. Automation is realized through the integration of Mobile Net and the YOLO (You Only Look Once) algorithms, crucial for real-time object detection and recognition.

Mobile Net is chosen for its efficiency in feature extraction, ensuring optimal computational efficiency, while YOLO facilitates swift and accurate identification of objects within captured images. This project not only advances autonomous driving but also highlights the potential of combining stereo depth perception with cutting-edge algorithms for a more comprehensive and reliable autonomous navigation system. The primary research objective is to create and assess a perception system for autonomous driving, leveraging the Mobile Net lightweight Convolutional Neural Network (CNN) and the You Only Look Once (YOLO) algorithm. The approach involves integrating Mobile Net CNN for efficient feature extraction with the YOLO algorithm for real-time object detection, optimizing the combined model to achieve a balance between computational efficiency and high accuracy in object recognition.

Fig 1 Left and Right Stereo Images

The parameters describing the stereo depth erroneous values are Non Occulation, all and disc. This method is proposed with the keen intention of reducing the number of bad pixels present in the fed image and reducing the Non Occulative errors. To bring out this enhancement the model images are trained in prior with the help of host post processing elements.

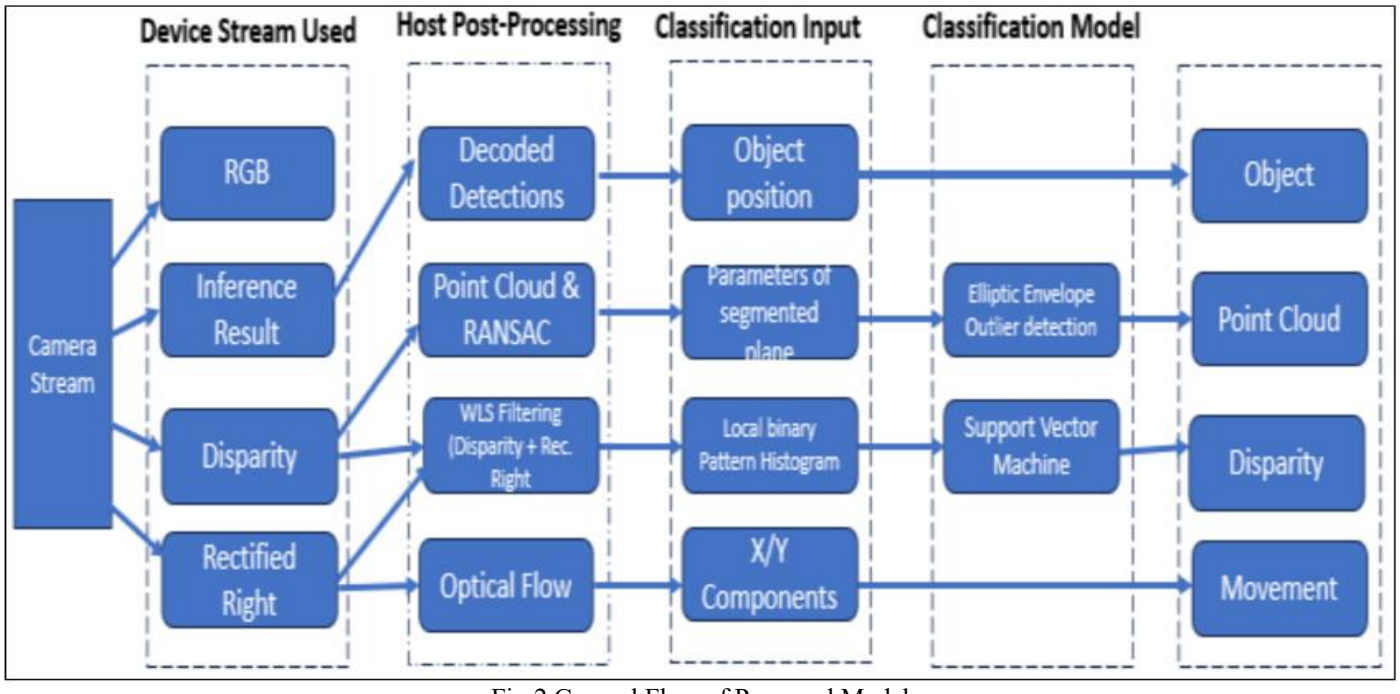

Fig 2 General Flow of Proposed Model

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR1535

In the motive of object detection and feature extraction we have proposed an enhanced version of Mobile Net algorithm for feature extraction and YOLO algorithm for object detection in this paper. Integrating stereo depth perception with sophisticated computer vision algorithms like Mobile Net and YOLO can indeed enhance the vehicle's understanding of its environment in real-time.

By leveraging stereo cameras to mimic human binocular vision, the system can capture depth information crucial for understanding the 3D layout of the scene. This depth perception enables the vehicle to better assess distances to objects and obstacles, improving navigation and decision-making.

➢ *Mobile Net:*

Mobile Net is chosen for its efficiency in feature extraction, ensuring optimal computational efficiency. Its efficiency in feature extraction is advantageous for real-time processing, allowing the system to analyze stereo images swiftly while maintaining accuracy. Its reduced computational complexity is particularly beneficial for onboard implementation in resource-constrained environments like autonomous vehicles.

Fig 3 Feature Extraction Using Mobile Net Algorithm

➢ *YOLO (You Only Look Once):*

YOLO 'S ability to rapidly detect and classify objects within images is crucial for identifying and tracking various elements in the vehicle's surroundings. By utilizing YOLO, the system can efficiently recognize pedestrians, vehicles, traffic signs, and other relevant objects, enabling proactive decision-making to ensure safe navigation.

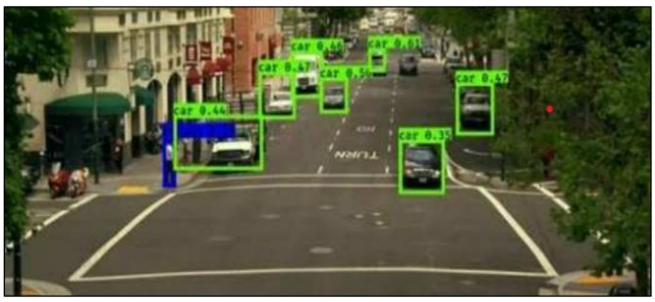

Fig 4 Object detection Using YOLO Algorithm.

Overall, the combination of stereo depth perception, Mobile Net, and YOLO holds significant potential for revolutionizing autonomous driving systems, making them more efficient, reliable, and capable of navigating complex environments with heightened awareness.

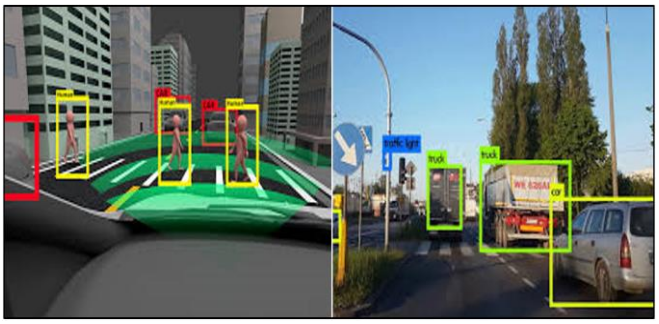

Fig 5 Simulation Results for Proposed Method

Table 1 Erroneous Disparity Values

Matching method	Non Occ	All	Disc	No of bad pixels
Proposed Mobile Net and YOLO Algorithm	0.3413	1.207	4.769	117
Triangulation Method	0.391	1.31	3.297	810
CNN (Convolutional Neural Network)	0.4267	1.45	2.503	1157
Bayesian Regression	0.4981	1.62	1.386	1376

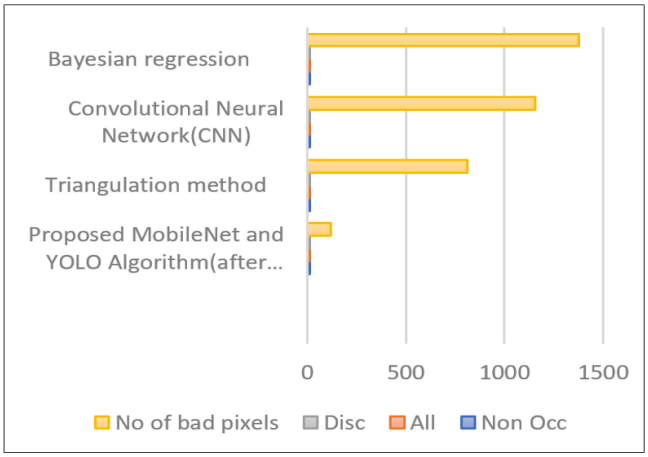

Fig 6 Graph Representing the Erroneous Disparity

The depth information obtained through stereo vision and the automation achieved through the integration of Mobile Net and YOLO Algorithms is shown in Figure 6 and the performance metrics of the proposed approach as a percentage of erroneous disparity is tabulated in Table 1.

From the above results it can be inferred that the proposed algorithm shows remarkable improvement in all three scores on par with minimum mismatches.

III. CONCLUSION

In conclusion, our proposed autonomous navigation system represents a significant advancement in the field, harnessing the power of stereo vision, MobileNet, and YOLO algorithms to create a robust and efficient solution. By

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR1535

strategically positioning cameras on both sides of the vehicle, we ensure a comprehensive view of the surroundings, crucial for safe and reliable autonomous driving. The integration of stereo vision enables accurate depth perception, enhancing spatial understanding. Moving forward, continued research and refinement of these techniques will be instrumental in further improving the performance and adaptability of autonomous vehicles, ultimately contributing to safer and more efficient transportation systems. Moreover, the selection of MobileNet for feature extraction and YOLO for real-time object detection and recognition demonstrates a thoughtful approach to balancing computational efficiency with accuracy.

REFERENCES

[1]. Liu, Z.; Zhao, X.; Huang, T.; Hu, R.; Zhou, Y.; Bai, X. Tanet: Robust 3D object detection from point clouds with triple attention. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 11677–11684.

[2]. Zhou, Y.; Tuzel, O. Voxelnet: End-to-end learning for point cloud based 3D object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4490–4499.

[3]. Yan, Y.; Mao, Y.; Li, B. Second: Sparsely embedded convolutional detection. *Sensors* 2018, *18*, 3337.

[4]. Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; Li, H. Pv-rcnn: Point-voxel feature set abstraction for 3D object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10529–10538.

[5]. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-View 3D Object Detection Network for Autonomous Driving. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 June 2017; pp. 1907–1915.

[6]. Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). YOLOv4: Optimal Speed and Accuracy of Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 960-961).

[7]. Zhang, K., Wang, Y., Qi, X., & Qi, X. (2018). Real-time object detection in autonomous vehicles using deep learning. In Proceedings of the IEEE Intelligent Vehicles Symposium (IV) (pp. 1279-1284).

[8]. X. Chen, K. Kundu, Y. Zhu, A. Berneshawi, H. Ma, S. Fidler, and R. Urtasun. 3d object proposals for accurate object class detection. In NIPS, 2015

[9]. 9.R. Garg, V. K. BG, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in Proc. Eur. Conf. Comput. Vis., 2016, pp. 740–756.

[10]. V. Prasad and B. Bhowmick, "SfMLearner: Learning monocular depth & ego-motion using meaningful geometric constraints," in Proc. IEEE Winter Conf. Appl. Comput. Vis., 2019, pp. 2087–2096.

[11]. C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 270–279.

[12]. B. Wagstaff, V. Peretroukhin, and J. Kelly, "Self-supervised deep pose corrections for robust visual odometry," in Proc. IEEE Int. Conf. Robot. Automat., 2020, pp. 2331–2337.

[13]. R. Li, D. Gu, and S. Wang, "DeepSLAM: A robust monocular SLAM system with unsupervised deep learning," IEEE Trans. Ind. Electron., vol. 68, no. 4, pp. 3577–3587, Apr. 2021.

[14]. C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 3828–3838.

[15]. V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3D packing for self-supervised monocular depth estimation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 2485–2494