IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 9, September 2020

DOI 10.17148/IJARCCE.2020.9906

A Comparative Study of MonoLoco with Improvised Loss Functions

N. Pavan Srinivas[1], N. Vamsy Krishna[2], M.V.S. Sanjay[3]

K L University, Vijayawada, India[1]

Bharath Instituite of Engineering and Technology, Mangalpally, Hyderabad, India[2]

GuruNanak Instituitions Technical Campus, Ibrahimpatnam, Hyderabad, India[3]

Abstract: The native concept of 3D human localization, from monocular color images is an ill posed problem. Considering the limitations of neural networks, we compare various loss functions that are based on a variety of distributions. Monoloco is a 3D pedestrian localization architecture which uses a lightweight feed forward neural network and predicts the distance of pedestrians, from the camera and the uncertainty associated with its prediction based on Laplacian loss. We trained two individual models on Kitti dataset with updated unnormalization methods, changed dataset sizes and Losses which are Cauchy and Generalized Extreme Value (Gev) losses. These newly trained models using Monoloco were observed to perform better. We evaluated these trained models on Kitti dataset and found improvised results than existing Monoloco.

Keywords: Deep Learning, Kitti, Nuscenes, MonoLoco, Loss Function,Neural Networks.

I. INTRODUCTION

In the present era, the most commonly used ranging technologies are LIDAR and RADAR. Despite their sparsity of point clouds over long ranges [1,2,3], these technologies are extremely costly to set up and maintain. Adoption of stereo/multiple cameras has been proposed to tackle the drawbacks of the above-mentioned technologies [4,5]. There are many researches going on to push the limits of monocular perceptions by contributing to multi-sensor fusion [6]. While there is a huge progress in the field of estimating 3D positions of the vehicles from monocular data [7,8,9], the pedestrians received low attention due to lack of good performance. In reality, detection of 3D positions of humans from a single 2D image is very uncertain due to the diversity their heights and shapes. The works of Kendall and Gal [10] showed us the real time uncertainty estimation in deep learning for perception tasks, and difference between the aleatoric and epistemic uncertainties [11]. The comparison of the measure of proposed and known uncertainty is shown as task error. Based on the classification of human heights within the adult population, the task error and the upper bound status of the model were set and shown accurate results in the localization without overcoming the limitations [12]. A pose estimator [35] gives the information on the dimensional representation of human 2-D joint. These detected joints and the 3D position (ground truth) of each detected instance are given as inputs to train the light-weight-feed forward network proposed in Monoloco[12].

Fig1: Output of the model with changed sizes of training and validation datasets with Cauchy loss. The color of the dot represents the predicted distance from the camera. To the left is the bird's eye view of the image.

At inference time, a 2D image is given as input to the trained network and it predicts the 3D location of pedestrians along with confidence about its prediction [12]. In this work, we study that uncertainty in locating humans in the scene and overlook learning ability on this data. From the perception task, we aim to provide more data to this problem in uncertainty estimation through deep learning. We develop the existing model with the recently developed loss functions based on Cauchy and Generalized extreme value distributions [13,14]. The new unnormalized spread of data shows better performance in the aspects of uncertainty and predicts the positions better, and the compared results are shown below. The code is publicly available online[1]

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 9, September 2020
DOI 10.17148/IJARCCE.2020.9906

II. RELATED WORK

Monocular 3D Object detection:
Most of the monocular detection methods have been found to be center their focus on vehicles and other objects with fixed shapes. The evaluation of pedestrians from monocular images has been considered insignificant in previous works. The work of Kundegorski and Breckon [36] which combined infrared imagery and real time photogrammetry achieved sensible results. In the method of Mono3D, deep learning was used to create 3D object proposals for car, cyclists and pedestrian categories. Their proposals assumed fixed ground plane and were scored based on scene priors like shape, semantic and instance segmentations. Methods like [8,16,15] used geometrical reasoning to regress 3D pose parameters from 2D detections.

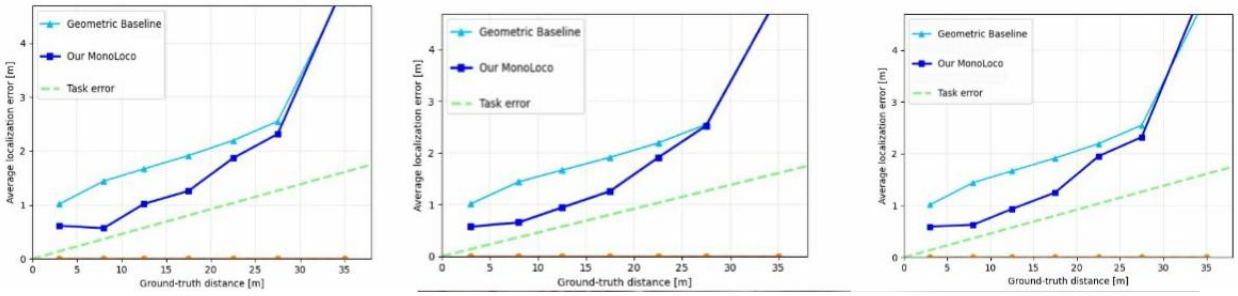

Fig 2: The above images show the comparison between ALE (Average Localization Error) vs GT (Ground Truth) for Monoloco model with changed training and validation dataset sizes with different loss functions. These graphs are obtained when the loss functions used are is Laplace, Cauchy and Gev losses respectively (from left to right)

[17] and [18] used a module for depth estimation. [9] used a bird's eye view to map image-based features instead of image domain using integral images. Most of these methods only achieved reasonable performances for vehicles. Monocular method to address categories like pedestrians and cyclists in the case of autonomous driving has been addressed by Monoloco.

Uncertainty in Computer Vision:
In Autonomous driving framework, where safety is very crucial, the DNN's used should not only provide correct outputs but also a measure of uncertainty. Bayesian Neural Networks (BNNs) [19,20] were usually used to model epistemic uncertainty through probability distribution over the weights and biases of the models. But these distributions were found hard to be dealt with. Some intriguing solutions to perform Bayesian inference to measure uncertainty were provided by researchers which include [21,22,23]. [24,25] showed that applying dropout during inference will result in a form of variational inference where parameters of the network are modelled as a mixture of multivariate Gaussian distributions with small variances. This technique, which is quite popular, is called Monte Carlo (MC) dropout, because of its flexibility to non-probabilistic deep learning frameworks. This has been used in [10,26,27].

Human Pose Estimation:
Recognising humans from images and evaluate their skeletons is a broadly examined issue. There are two groups of state-of-the-art methods namely Top-down [28,29,30] and bottom-up [31], which are based on CNN's. [32] showed the efficacy of concealed information in 2D joints stimuli. They used a fully connected network with light weights to predict 3D joints form 2D poses and achieved state of the art results. Similar work was done by [33] without providing the 3D location of the person in the image.

MonoLoco:
Monoloco is a 3D pedestrian localization method which outputs the predicted distance of a person from the camera, in 2D images along with some uncertainty. It uses PifPaf as a standard decoder for pose estimation and calculates the distance based on the output of the pose estimator. [12]

III. PROPOSED MODEL

Convolutional Neural Networks are widely used in the field of deep learning and pattern recognition. The key parameters of neural network are weights, biases, and activation functions. The pattern in the data is to be recognised as it can be used to predict the output for a random input. The weights and the biases are the variables which are responsible to recognise the inherent pattern in the data. These parameters result in the linear polynomial which is the predicted relationship between input and output data. While training the neural network, a special function called loss function or

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 9, September 2020
DOI 10.17148/IJARCCE.2020.9906

cost function is used to evaluate the prediction and quantify its error in prediction (if any). Loss function plays a crucial role in calculating the weights and biases. Loss function inputs predicted and true value and outputs the error in prediction. Different losses can be calculated from several probability distributions such that the distribution fits the data.

MonoLoco:
The input to Monoloco is a set of 2D joints extracted from a raw image and the output is the 3D location of a pedestrian μ and the spread b which represents the associated aleatoric uncertainty. μ±b is the confidence interval for the prediction. Epistemic is obtained through stochastic forward passes applying MC dropout is applied to calculate the Epistemic Uncertainty in a stochastic forward pass [24].

The dashed ellipse represents the two combined uncertainties are represented using dashed ellipse. Every fully connected layer output 256 features. It is followed by a Batch Normalization layer [34] and a ReLU activation function. MonoLoco v4.0 uses a Laplacian loss to calculate and alter the parameters of the network so that loss is minimized. There are several other distributions that can be used to derive a loss function for the model. The Monoloco model outputs the predicted distance (μ) and the spread (b). These two values are to be used along with the ground true distance (x) and the constructed loss function should be able to change the network's parameters so that it matches the inherent pattern in the data. L1 loss is the absolute difference between the predicted and true distances. Using this difference as the error in prediction, the weights and biases are modified. Laplacian loss is one of the few losses which are based on L1 loss. Firstly, we changed the sizes of training and validation datasets. The number of images used for training were increased by a factor of 1.2. Two models are generated using (i) Cauchy and (ii) Generalized extreme value losses based on Cauchy and Generalized extreme value distributions respectively by using Monoloco to enhance the performance.

Cauchy Loss:
The original distribution:

$$f(x|x_0,\gamma) = \frac{1}{\pi\gamma[1+(\frac{x-x_0}{\gamma})^2]}$$

where x_0 is location parameter and γ is spread [13].

The Cauchy loss is derived using negative log likelihood of the Cauchy distribution. The loss function is as follows:

$$\log(\pi\gamma) + \log(1 + (\frac{x - x_0}{\gamma})^2)$$

The error is taken relative to ground truth distance. The light tails of Cauchy distribution results in the finer performance in certain distance intervals.

Gev Loss:
The original distribution:

$$f(x|\mu,\sigma,\xi) = \frac{1}{\sigma}t(x)^{\xi+1}e^{-t(x)} \;;\; \xi \in R$$

where μ is location parameter, σ is the spread and ξ is the shape parameter [14].

The shape parameter is considered zero and distribution is as follows:

$$t(x) = e^{-(x-\mu)/\sigma} \;;\; \xi = 0$$

The Gev loss is derived using negative log likelihood of the Gev distribution. The loss function is as follows:

$$\log(\sigma) - \frac{(x-\mu)}{\sigma} - e^{-\frac{(x-\mu)}{\sigma}}$$

The error is taken relative to ground truth distance. The outliers in the errors are not amplified here which results in same error in outlier conditions.

Unnormalization:
The unnormalization is the process which sets the spread (b) in the valid range of the predicted distances. The method for unnormalization is changed from the exponential function to the square root function.

$$(\mu, b) \rightarrow (\mu, \mu e^b) \quad Unnormalization\ in\ Monoloco\ V4.0$$

$$(\mu, b) \rightarrow \left(\mu, \mu\sqrt{b}\right)\ Unnormalization\ in\ modified\ Monoloco$$

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 9, September 2020
DOI 10.17148/IJARCCE.2020.9906

IV. EXPERIMENTAL RESULTS

After changing the sizes of training and validation datasets and the loss functions, the following results were obtained.

Pose based Accuracy:
The accuracy of the estimation ultimately depends upon the pose of the person. High uncertainty has been associated with the person who is not in general human pose as it is considered as an outlier case.

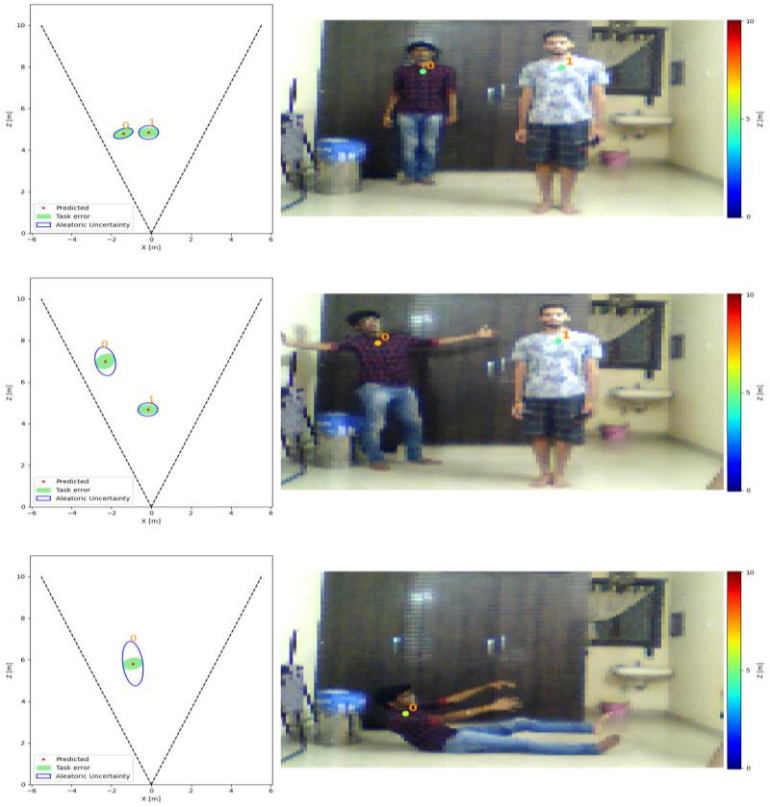

Fig 3: Outlier cases are shown in the image. In the top image, the predicted locations are almost similar to the real locations and the associated uncertainty is small. In the middle and the bottom images, the changes in the prediction & the increase in the uncertainty can be observed with different human poses which can be considered as outlier cases.

Model output with Cauchy Loss:

Fig 4: The top image shows the predicted location of the pedestrians and the bird's eye view with associated aleatoric and total uncertainty. The bottom image shows the average aleatoric uncertainty in different clusters with certain confidence intervals.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 9, September 2020
DOI 10.17148/IJARCCE.2020.9906

The below result was obtained when the model was trained with Cauchy loss and changed unnormalization method using Monoloco. It performed better than Monoloco trained with Laplacian loss and old unnormalization technique in certain distance clusters.

Model output with Gev Loss:

The below result was obtained when the model was trained with Gev (Generalized extreme value) loss and changed unnormalization method using Monoloco.

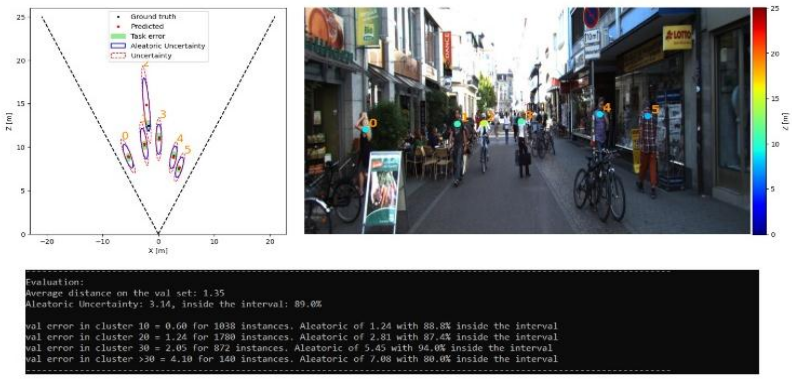

Fig 5: The top image shows the predicted location of the pedestrians and the bird's eye view with associated aleatoric and total uncertainty. The bottom image shows the average aleatoric uncertainty in different clusters with certain confidence intervals.

Comparison between models trained with different losses:

When three models were generated using Monoloco with changed training and validation dataset sizes, unnormalization method and different losses, the following results were obtained. It should be noted that the original ground truth distance of the person in the below images is 4.5 meters from the camera.

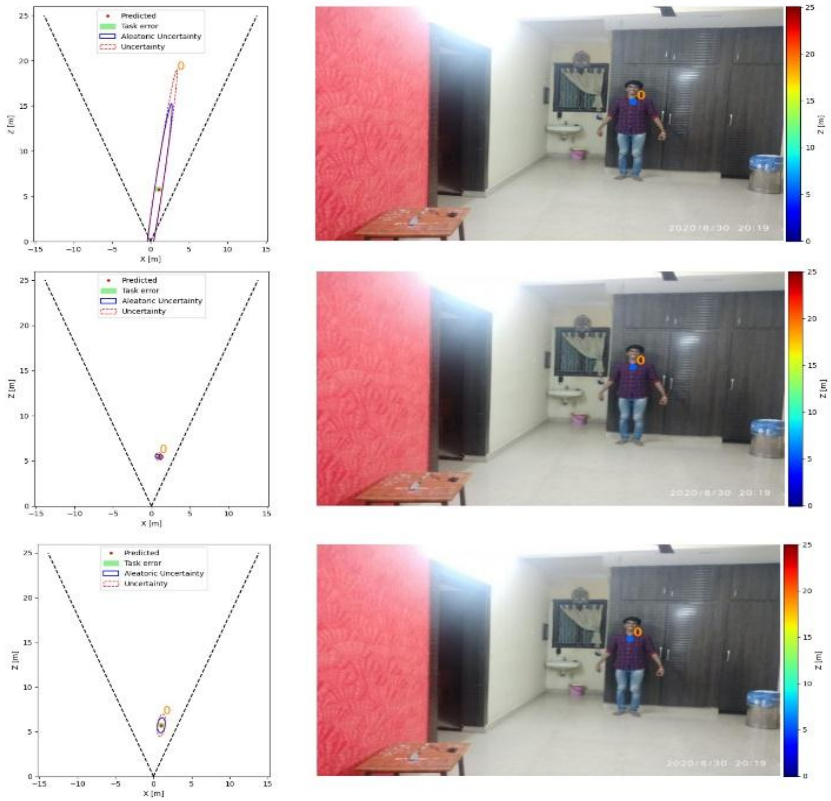

Fig 6: The results when models were generated using Laplacian, Cauchy and Gev losses (top to bottom) are shown in the above image. The changes in the aleatoric uncertainty with change in the loss function can be observed.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 9, September 2020

DOI 10.17148/IJARCCE.2020.9906

TABLE I: IMPACT OF DIFFERENT LOSS FUNTCIONS ON THE MODEL.

Losses	**ALP [%]** <0.5m	<1m	<2m	**ALE [m]** Easy	Moderate	Hard
Laplace	22.97	40.3	57.5	1.25	1.25	1.95
Cauchy	22.92	38.99	56.62	1.26	1.27	1.97
Gev	23.48	39.91	56.86	1.24	1.28	1.84

Table 1 shows the influence of using different loss function on the same model. Cauchy and Gev are one of the many possible loss functions that can be used. After inspect, these found to be effective in predicting the 3D distance from 2D images.

V. CONCLUSION

We have proposed a new model with updated loss functions such that the prediction is better in certain intervals. These models can be used in autonomous driving vehicles along with other stereo cameras and models. When combined with other technologies, these models can work in tandem with other systems.

REFERENCES

[1] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. *Multi-view 3d object detection network for autonomous driving.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1907– 1915, 2017
[2] Yin Zhou and Oncel Tuzel. Voxelnet: *End-to-end learning for point cloud based 3d object detection.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4490–4499, 2018.
[3] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. *Frustum pointnets for 3d object detection from rgbd data.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 918– 927, 2018.
[4] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. *3d object proposals for accurate object class detection.* In Advances in Neural Information Processing Systems, pages 424–432, 2015.
[5] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. *Stereo r-cnn based 3d object detection for autonomous driving.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7644–7652, 2019.
[6] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. *Deep continuous fusion for multi-sensor 3d object detection.* In Proceedings of the European Conference on Computer Vision (ECCV), pages 641–656, 2018.
[7] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. *Monocular 3d object detection for autonomous driving.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2147–2156, 2016.
[8] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. *3d bounding box estimation using deep learning and geometry.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7074–7082, 2017.
[9] Thomas Roddick, Alex Kendall, and Roberto Cipolla. *Orthographic feature transform for monocular 3d object detection.* In the British Machine Vision Conference (BMVC), 2019.
[10] Alex Kendall and Yarin Gal. *What uncertainties do we need in bayesian deep learning for computer vision?* In Advances in Neural Information Processing Systems, pages 5574–5584, 2017.
[11] Armen Der Kiureghian and Ove Ditlevsen. *Aleatory or epistemic? does it matter?* Structural Safety, 31(2):105–112, 2009.
[12] Lorenzo Bertoni, Sven Kreiss, and Alexandre Alahi. *Monoloco: Monocular 3d pedestrian localization and uncertainty estimation.* In ICCV, 2019.
[13] Feller, William (1971). *An Introduction to Probability Theory and Its Applications*, Volume II (2 ed.). New York: John Wiley & Sons Inc. pp. 704. ISBN 978-0-471-25709-7.
[14] Muraleedharan. G, C. Guedes Soares and Cláudia Lucas (2011). "*Characteristic and Moment Generating Functions of Generalised Extreme Value Distribution (GEV)*". In Linda. L. Wright (Ed.), Sea Level Rise, Coastal Engineering, Shorelines and Tides, Chapter-14, pp. 269–276. Nova Science Publishers.
[15] Zengyi Qin, Jinglu Wang, and Yan Lu. *Monogrnet: A geometric reasoning network for monocular 3d object localization.* In the AAAI Conference on Artificial Intelligence, volume 33, pages 8851–8858, 2019.
[16] Hou-Ning Hu, Qi-Zhi Cai, Dequan Wang, Ji Lin, Min Sun, Philipp Krhenbhl, Trevor Darrell, and Fisher Yu. Joint *monocular 3d vehicle detection and tracking.* arXiv, abs/1811.10742, 2018.
[17] Bin Xu and Zhenzhong Chen. *Multi-level fusion based 3d object detection from monocular images.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2345–2353, 2018.
[18] Fabian Manhardt, Wadim Kehl, and Adrien Gaidon. *Roi10d: Monocular lifting of 2d detection to 6d pose and metric shape.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2069–2078, 2019.
[19] Michael D. Richard and Richard Lippmann. *Neural network classifiers estimate bayesian a posteriori probabilities.* Neural Computation, 3:461–483, 1991.
[20] Radford M Neal. Bayesian learning for neural networks, volume 118. Springer Science & Business Media, 2012.
[21] Alex Graves. Practical variational inference for neural networks. In Advances in Neural Information Processing Systems, pages 2348–2356, 2011.
[22] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. *Weight uncertainty in neural network.* In the International Conference on Machine Learning, Proceedings of Machine Learning Research, pages 1613–1622. PMLR, 2015.
[23] Tim Salimans, Diederik Kingma, and Max Welling. *Markov chain monte carlo and variational inference: Bridging the gap.* In the International Conference on Machine Learning, pages 1218–1226, 2015.
[24] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: *Representing model uncertainty in deep learning.* In the International Conference on Machine Learning, pages 1050–1059, 2016.
[25] Yarin Gal, Jiri Hron, and Alex Kendall. *Concrete dropout.* In Advances in Neural Information Processing Systems, pages 3581–3590, 2017.
[26] Jishnu Mukhoti and Yarin Gal. *Evaluating bayesian deep learning methods for semantic segmentation.* arXiv preprint arXiv:1811.12709, 2018.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 9, September 2020

DOI 10.17148/IJARCCE.2020.9906

[27] Di Feng, Lars Rosenbaum, and Klaus Dietmayer. *Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3d vehicle detection.* In the IEEE International Conference on Intelligent Transportation Systems (ITSC), pages 3266–3273, 2018.

[28] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Christoph Bregler, and Kevin P. Murphy. *Towards accurate multi-person pose estimation in the wild.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3711–3719, 2017.

[29] Haoshu Fang, Shuqin Xie, and Cewu Lu. Rmpe: *Regional multi-person pose estimation.* Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 2353–2362, 2017.

[30] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross B. ´ Girshick. *Mask r-cnn.* Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 2980–2988, 2017.

[31] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. *Realtime multi-person 2d pose estimation using part affinity fields.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1302–1310, 2017.

[32] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. *A simple yet effective baseline for 3d human pose estimation.* In Proceedings of the IEEE International Conference on Computer Vision (ICCV), pages 2659–2668. IEEE, 2017.

[33]] Francesc Moreno-Noguer. *3d human pose estimation from a single image via distance matrix regression.* Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1561–1570, 2017.

[34] Sergey Ioffe and Christian Szegedy. Batch normalization: *Accelerating deep network training by reducing internal covariate shift.* arXiv preprint arXiv:1502.03167, 2015.

[35] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. *Pifpaf: Composite fields for human pose estimation*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 11977–11986, 2019