2024 9th International Conference on Automation, Control and Robotics Engineering (CACRE) | 979-8-3503-5030-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/CACRE62362.2024.10635041

Helium Loss for Robust Keypoints Detection with Convolutional Neural Networks

1st Haode Huo, 2nd Tianran Wang, 3rd Linjie Zhao
State Key Laboratory of Robotics and System
Harbin Institute of Technology
Harbin, China
hdh339484@gmail.com, 529708894@qq.com, zhaolinjie@hit.edu.cn

***Abstract*—In this paper, a novel loss function for key point detection, named Helium Loss is proposed. All our experiments are based on Ultralytics' YOLOv8n-pose architecture. Initially, this paper analyze various existing loss functions, including L1 loss, L2 loss, and smooth L1 loss. There is a problem that current loss functions do not account for issues related to the quality of dataset annotations and unclear boundaries of targets. To address this challenge, this paper introduce a loss function that incorporates partial trust in annotated key points and utilizes Gaussian distribution to mitigate the impact of manual annotations.**

***Index Terms*—distribution loss, pose estimation, keypoint detection, deep learning**

I. Introduction

In recent years, with the advancement of robot technology, computer vision control systems have increasingly played a pivotal role in various aspects of robot navigation, target recognition, and obstacle avoidance. These technologies are also extensively employed in robot competitions such as boxing and soccer.

RoboMaster, being a national robotics competition for college students, necessitates participants to independently develop robots and compete against each other. During the past decades, this competition has served as a catalyst for the emergence of numerous research achievements [1], [2]. Given the complexity of the competition environment, traditional computer vision techniques struggle to achieve desired outcomes; hence computer vision based on deep learning becomes particularly crucial in this context.

As an integral component of deep learning technology used for model training guidance, loss function assumes significant importance. Traditional object detection methods tend to be hard to reach a satisfying effect and rely on pre-defined templates; whereas object detection methods based on deep learning offer simple implementation but still require improvements in accuracy. Based on conditions in the RoboMaster competition setting, this paper proposes a novel loss function called Helium loss that primarily focuses on three key aspects:

- Keypoint detection and pose estimation: These two techniques within the realm of deep learning aim at identifying specific keypoints associated with target objects depicted in images. Numerous bottom-up network architectures like SSD [3] have been proposed to accurately locate these keypoints; however, this paper introduces a new method for calculating the loss function based on YOLO [4] architecture which aims at addressing challenges encountered when dealing with samples featuring unclear boundaries.
- The labeled keypoints in traditional pose estimation and keypoint detection tasks are typically represented as Dirac delta distribution. However, this paper proposes a loss function based on Gaussian distribution to better capture the uncertainty in labeling, aiming to facilitate smooth convergence of the model during training.
- The idea of incorporating the loss function proposed in this paper, which simulates uncertainty using a Gaussian distribution, can be combined with existing loss functions to achieve improved results.

The dataset used in this study is derived from recent RoboMaster competitions. It consists of images capturing several targets under low exposure conditions, with the main target points being the ends of two light bars on the robot's armor plate.

This paper also compare Helium loss with commonly used L1, L2, and smooth L1 losses. Experimental results demonstrate that models trained using this novel loss function exhibit higher accuracy compared to those trained using other conventional loss functions.

II. Related work

A. Keypoints detection & pose estimations

Pose estimation and keypoints detection aim to identify specific points (keypoints) on an object in an image. These keypoints typically represent joints, landmarks, or other distinctive features.Presently, the prevailing trend in pose estimation and keypoint detection tasks leans towards bottom-up architectures [5], [6] based on Convolutional Neural Network, exemplified by prominent models like HRnet [7] and the YOLO series [4], [8]–[12]. HRnet features a high-resolution feature map and improves the performance of keypoint detection by combining information from multiple resolutions. YOLOv8n-pose's head processes the input through the convolutional layer sequences and concatenates the results to get the keypoints. All these models are able to process input images and precisely locate desired keypoints on the output image. In this study, leveraging the extensive development efforts by Ultralytics [12], all our experiments are conducted utilizing their robust architectures and meticulously crafted codebase.

Thanks to years of research in the field of keypoint and pose recognition, significant advancements have been made, leading to substantial progress in this domain. However, there remains a notable gap in addressing the handling of annotations for samples with unclear boundaries. In this paper, the proposed method aim to bridge this gap by proposing a novel computational approach for a new loss function. This approach is designed to effectively handle samples with unclear boundaries, thereby contributing to the further advancement of keypoint and pose recognition research.

B. Distribution Loss

Traditionally, in pose estimation and keypoint detection tasks, annotated keypoints have been treated as Dirac delta distribution. However, some advancements have explored alternative approaches, incorporating Gaussian assumptions [13]–[15] to better model the uncertainty inherent in keypoint annotations. For instance, Mean-Squared Error (MSE) [15] loss, facilitates the quantitative evaluation by juxtaposing the predicted heatmap against the corresponding ground-truth heatmap. This ground-truth heatmap is generated using a 2D Gaussian distribution centered precisely on the joint location, with a standard deviation of 1 pixel. This method ensures effective supervision by quantifying the differences between predicted and ground-truth heatmaps, thus facilitating precise keypoint detection during training. Noteworthy among the object detection methods is the introduction of the Distribution Focal Loss (DFL) [16], which accurately depict the flexible distribution in real data, potentially leading to enhanced model performance and robustness.

In the proposed approach, it introduces a novel loss function based on Gaussian distribution, aiming to mitigate convergence challenges caused by unclear sample boundaries and coarse annotations. By leveraging Gaussian distribution, our proposed loss function effectively addresses the issue of uncertain boundaries, thus facilitating smoother convergence of the model during training.

III. Helium loss

A. Analysis between different loss functions

In recent years, there has been a growing recognition of the limitations of the commonly used L2 loss function in various machine learning tasks. While L2 loss has historically been the default choice due to its simplicity and computational efficiency, empirical evidence suggests that alternative loss functions such as L1 and smooth L1 can offer superior performance in many scenarios.

The L1 loss function, also known as the mean absolute error, has gained attention for its robustness to outliers compared to L2 loss. By computing the absolute difference between predicted and target values, L1 loss reduces the impact of outliers that may disproportionately influence the training process. Furthermore, the smooth L1 loss function presents an appealing compromise between the robustness of L1 loss and the smoothness of L2 loss.

Smooth L1 loss [17]–[19], introduced as a piecewise function combining quadratic and linear components, offers a balanced approach to handling errors of different magnitudes. Unlike L2 loss, which penalizes large errors quadratically, smooth L1 loss behaves linearly for smaller errors and quadratically for larger errors. This property makes it particularly suitable for tasks where outliers or extreme values are prevalent, as it provides a more nuanced treatment of error contributions.

In the specific context of facial landmark localization, where variability in pose and expression can lead to significant deviations from ground truth, the choice of loss function becomes critical. Here, a loss function that mimics L1 behavior for larger errors while incorporating a logarithmic function with an offset for smaller errors has been proposed [18], [19]. This formulation aims to strike a balance between sensitivity to outliers and the ability to capture subtle variations in facial features.

Formally, the smooth L1 loss is defined as follows [17]:

$$smoothL1(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| < 1 \\ |x| - \frac{1}{2} & \text{otherwise} \end{cases} \tag{1}$$

Where x represents the error between predicted and target values. This formulation ensures that errors below a certain threshold are treated linearly, while larger errors are penalized quadratically, thereby offering a robust and versatile loss function for tasks such as facial landmark localization.

Despite the efficacy of L1, smooth L1, and L2 loss functions in various machine learning tasks, they all exhibit a common limitation: the inability to accommodate labeled samples resulting from the presence of a labelable point as a region. In scenarios where a single point serves as a representative label for an entire region, these conventional loss functions may fail to capture the inherent uncertainty associated with such cases. For instance, in facial landmark localization, where certain facial features may be represented by a single landmark point but actually encompass a broader region, the traditional loss functions may struggle to adequately account for this ambiguity. Therefore, while L1, smooth L1, and L2 loss functions offer valuable contributions to machine learning, there remains a need for further research into loss functions capable of effectively handling labeled samples that represent regions rather than single points.Figure 1 demonstrates that the Helium loss exhibits a cautious approach towards the annotated points, suggesting a nuanced level of trust in their accuracy.

B. Robust keypoints detection based on CNN

Keypoint detection in real-world environments poses significant challenges, particularly in scenarios like the RoboMaster competition, where varying light levels are prevalent. Traditional threshold-based algorithms struggle to adapt to these conditions effectively. However, YOLOv8-pose, leveraging convolutional neural networks (CNNs), offers a robust detection method capable of handling diverse lighting conditions and angles, crucial in environments where vehicles can rotate at high speeds. The inherent feature learning capabilities of CNNs enable them to effectively navigate these challenges by automatically learning and extracting relevant features from the input data.

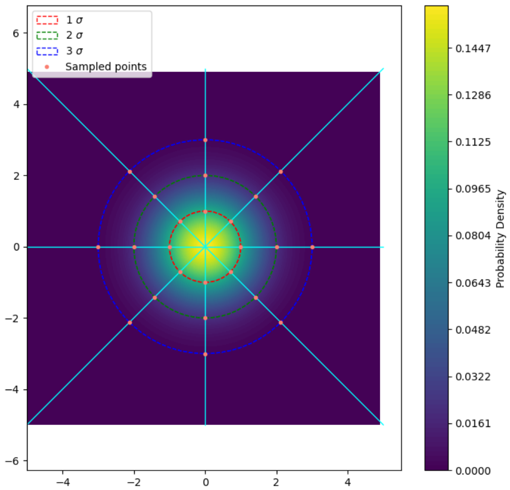


Fig. 1: Example of the sampling process.

C. Our proposed loss function

During the annotation of keypoints, the precise location of the annotated points is often not uniquely defined. Even if there exists an absolute ideal location for annotation, the actual annotated points may fall in the vicinity of this ideal point as Figure 2. In some cases, especially in the annotation process of self-made datasets, practical constraints in human resources may result in annotated points being significantly distant from the ideal locations. This can potentially impact the training effectiveness of the model, leading to early stagnation in gradient descent and consequently consuming more training time with inferior training outcomes. In this approach, the actual annotated points follow a Gaussian distribution around the ideal points. While this concept has been used in some papers, they often simplify it by constructing a Gaussian distribution based solely on pixel distances.

First, considering a single-target scenario with N keypoints, the objective is to select the minimum distance beween the keypoints in a single target. This can be formally expressed as follows:

$$d_{min} = \min_{0 \leq m < n < N} ||p_m - p_n||_2 \tag{2}$$

Where p_m and p_n are points of a single target and d_min is the minimum distance between the keypoints of a single target.

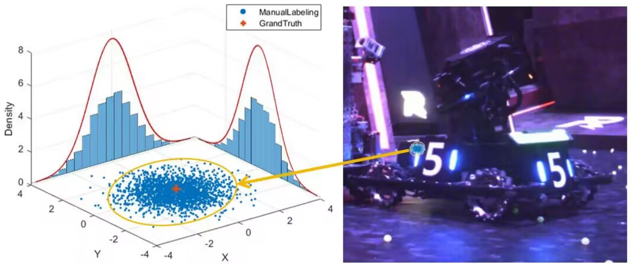
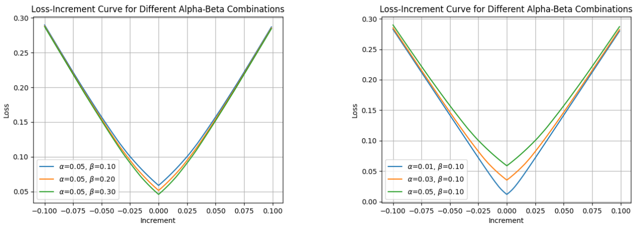


Fig. 2: Illustration depicting the distribution of ideal keypoints versus actual manually annotated keypoints. The ground truth keypoint is assumed as ideal keypoint.



(a)

(b)

Fig. 3: Plots of the Helium loss under different parameters of α and β.

Following this, the standard deviation of the Gaussian distribution for annotation is computed as following:

$$\sigma = \alpha \cdot e^{-\beta \cdot d_{min}} \tag{3}$$

Where α is a value to assure the distributions of keypoints in a single target don't overlap (typically recommended to be less than 0.05); β is value of approximately 0.1 to adjust to have mapping relation on different tasks. The rationale behind this computation is to calculate the standard deviation using the previously determined minimum distance; additionally, with the help of this convex function, the smaller targets have a wider range of distribution area, since smaller targets which are hard to be annotated tend to have coarser annotation quality compared to larger objects which can be annotated easily. Figure 3 reveals a notable trend: as the value of beta decreases and the value of alpha increases, there is a discernible increase in the degree of skepticism towards the ground truth during the loss computation.

In light of the significant computational demands associated with the standard Gaussian distribution, a sampling method is necessary for simplifying computation. The sampling is conducted on the simplified distribution according to Equation 4 and Equation 5:

$$P_k \in R^{8 \times 3} \tag{4}$$

$$p_{ij} = p_k + j\sigma(\sin\frac{i\pi}{4}, \cos\frac{i\pi}{4}) \tag{5}$$

Where P_k denotes the set of sampled points for the k^{th} keypoint of a single target; p_k is the ground truth of k^{th} keypoint of a single target; p_{ij} means i^{th} point on the j^{th} sampling circle , visually depicted in Figure 1. 8 points were sampled evenly on the margins of each distribution area as expressed in Equation 4. Owing to its visual resemblance to the sun, the method is named ”Helium”. Following sampling, the positional probability distribution of the sampled points to derive weights for loss computation is used. These weights correspond to the probability of the region surrounding the sampled points and are instrumental in the following computation, as delineated by Equation 7:

$$loss_k = (\sum_{j=1}^{3} \beta_j \sum_{i=1}^{8} \cdot ||\hat{y}_k, \, p_{ij}||_2 + ||\hat{y}_k, \, p_k||_2)/(\sum_{j=1}^{3} \beta_j + 1) \tag{6}$$

$$loss = (\sum_{k=1}^{L} loss_k) \tag{7}$$

Where $loss_{\text{k}}$ is the loss of the k^{th} keypoint of a single target; $\hat{y}_k$ is the k^{th} predicted keypoint of a single target; β_j is the weights for sample points on j^{th} sampling circle; L is the number of keypoints of a single target. The weights, determined based on the properties of the Gaussian distribution, are defined as [0.68, 0.27, 0.047]. The final computation incorporates the L1 distance between the predicted points and the ground truth points, weighted by these coefficients. In Figure 4, it is evident that the Helium loss imposes penalties even at the ground truth points, indicating a partial trust in the accuracy of the ground truth annotations.

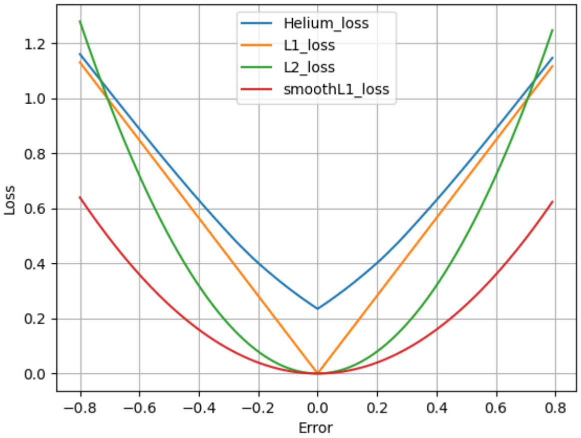


Fig. 4: Plots of the Helium, L1, L2, smooth L1 loss functions

D. Combination with other loss functions

Essentially, Helium Loss offers a novel approach to computing distances. Therefore, the resulting loss values can be directly utilized as inputs in certain loss functions that rely on L1 distance. For instance, in Equation 7, the $loss$ obtained can serve as the input variable x of smooth L1 in Equation 1. This approach facilitates convenient integration of our method with other loss functions.

IV. Experimental results

In this section, the Helium method is validated on our proprietary dataset. Initially, the dataset is introduced, followed by a detailed explanation of our training methodology and conditions. Finally, our approach is compared with other existing methods.

A. Task-specific data augmentation techniques

To prevent the model from relying on semantic information from lower-level features outside the annotation boundaries during recognition (which often leads to misclassification, such as classifying based on the shape of the vehicle rather than focusing solely on the numbers on the armor plates), a crop-paste method of armor plates from one part of the dataset onto others is implemented. The effect is displayed in Figure 5.

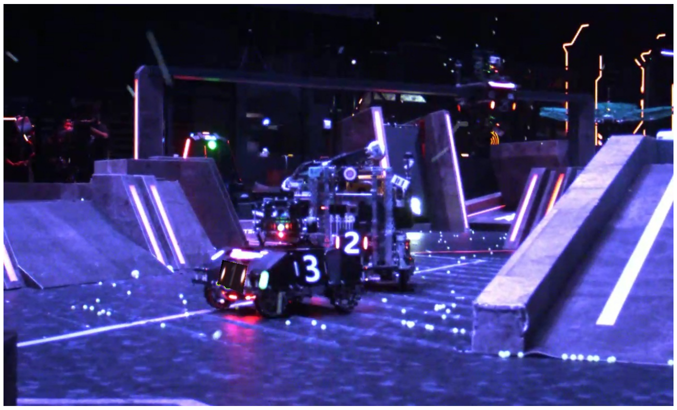
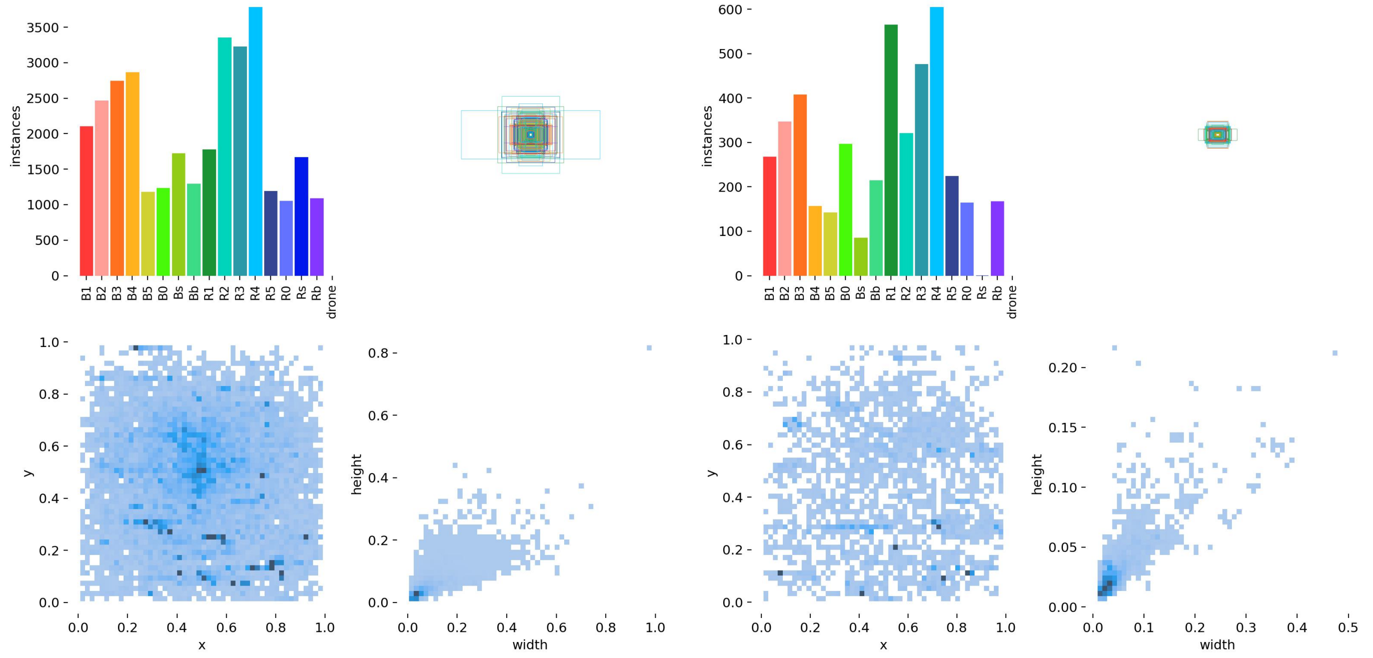


Fig. 5: An example of our crop-paste method. The red armor 1 and 2 are cropped from other images. All of our experiments have used this method.

B. Dataset

The dataset is collected from various RoboMaster competitions held over the years. It comprises 16 categories, with each target annotated with four key points in 2D YOLO format for pose labeling. This dataset consists of 32,744 (7693 images in it are augmented using crop-paste method) images for training and 7,740 images for testing, with varying numbers of targets per image. The statistical information of the dataset is depicted in the Figure6. Targets are annotated at the endpoints of the two light bars of the quadrilateral armor plate. This annotation scheme is designed to facilitate the subsequent computation of the target’s pose and position. The actual scenario being addressed pertains to low-light conditions, where challenges such as halation of light bars and imaging quality variability introduce ambiguity in annotating point locations. Consequently, these challenges have motivated the development of our proposed method.



(a) The statistical information of the train dataset.

(b) The statistical information of the test dataset.

Fig. 6: The comprehensive dataset information automatically collected by the Ultralytics library.

C. Implementation specifics

In our experiments, all the trainings are processed on the PyTorch platform using the Ultralytics codebase [12]. Both training and testing processes were executed on a single machine with Ubuntu 20.04, equipped with a 9th generation Intel Core i7 CPU and a TITAN Xp GPU. The system also utilized CUDA version 11.4. All experiments were performed on the same dataset using the identical model architecture (Ultralytics' YOLOv8n-pose) for training and testing. Consistent parameters were employed across all training sessions, with a patience value set to 80 for automatic termination. The input image size for both training and testing was set to 640x640 pixels. During training, a batch size of 8 was used, while testing employed a batch size of 1. Additionally, the Non-Maximum Suppression (NMS) [20] threshold for testing was set to 0.3 for optimal results.

D. Comparison between other designs

Mean Average Precision (mAP) as Table I, mAP at IoU threshold of 0.5 (mAP50), mAP at IoU threshold of 0.75 (mAP75), mAP between IoU thresholds of 0.5 and 0.95 (mAP50-95), and Root Mean Squared Error (RMSE) as Table IIand Figure 7 as validation metrics are employed. These metrics are computed using our dedicated test set.Our approach demonstrates a reduction in RMSE errors and a enhancement in AP accuracy. Furthermore, the utilization of our method facilitates expedited model convergence. Moreover, it is notable that incorporating the Helium method to calculate distances as inputs for Smooth L1 loss leads to a significant improvement in the model's performance.

TABLE I: Comparison between commonly used loss functions for keypoint detection and our proposed loss function. SmoothL1-H means use Helium method to calculate the distance instead of L1 as the input of smooth L1 loss.

Method	Epoch	mAP	mAP50	mAP75	mAP$_{50-95}$
L1	335	0.891	**0.933**	**0.918**	0.891
L2	318	0.872	0.927	0.904	0.872
smoothL1	322	0.861	0.927	0.903	0.867
smoothL1-H	321	0.889	0.929	0.914	0.889
Helium	**303**	**0.892**	0.929	**0.918**	**0.897**

TABLE II: The average Root Mean Squared Error (RMSE) for each method. SmoothL1-H means use Helium method to calculate the distance instead of L1 as the input of smooth L1 loss.

Loss function	L1	L2	smooth L1	smoothL1-H	Helium
RMSE($\times 10^{-2}$)	0.2785	0.2887	0.2995	0.2669	**0.2568**

V. Conclusion

This paper proposes that existing keypoints detection and pose estimation architectures' loss functions are not specifically designed to address the challenges posed by rough annotations and unclear boundaries commonly encountered in practical datasets. To address this, a novel distribution loss that incorporates a unique sampling approach and utilizes Gaussian distributions is proposed. Our method also accounts for differences in annotation quality between large and small objects.

Furthermore, the design principles of this loss function can have broad applications, as it introduces a novel method for distance computation. For example, the Helium loss can serve as an input for other loss functions.

To prove the effectiveness of our approach, extensive experiments have been conducted using YOLOv8n-pose. Our experiments on a custom dataset demonstrate that our proposed loss function not only results in shorter training times, higher accuracy, and improved model performance but also enhances the effectiveness of other loss functions when used in combination as a distance computation method.

It is worth noting that this design concept can be broadly applied to other computer vision tasks and certain control algorithms. However, due to space constraints, these topics will be discussed in future work.

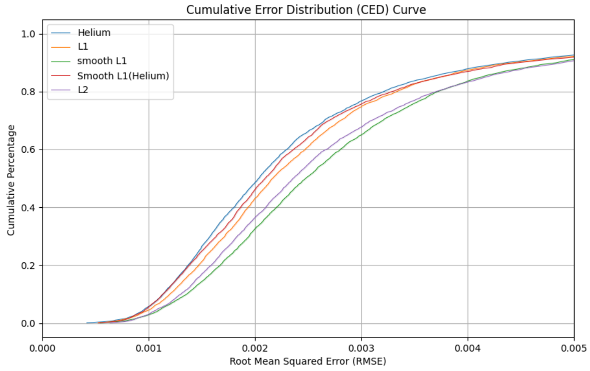


Fig. 7: Comparison of Cumulative Error Distribution (CED) curves illustrating the performance of different loss functions on the custom dataset. SmoothL1-H means use Helium method to calculate the distance instead of L1 as the input of smooth L1 loss.

Acknowledgment

This work was supported by the Heilongjiang Postdoctoral Fund (No. LBH-Z22136); New Era Longjiang Excellent Master and Doctoral Dissertation Fund (No. LJYXL2022-057); Self-Planned Task Foundation of State Key Laboratory of Robotics and System (HIT) of China (No. SKLRS202305C);

References

[1] Haoran Li, Zicheng Duan, Jiaqi Li, Mingjun Ma, Yaran Chen, and Dongbin Zhao. Neurons perception dataset for robomaster ai challenge. In 2022 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE, 2022.

[2] Xinyang Tang, Chuntao Leng, Yiheng Guan, Li Hao, and Shukun Wu. Development of Tracking and Control System Based on Computer Vision for RoboMaster Competition Robot. In 2020 5th International Conference on Advanced Robotics and Mechatronics (ICARM), pages 442–447, Shenzhen, China, 2020. IEEE.

[3] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pages 21–37. Springer, 2016.

[4] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. pages 779–788, 2016.

[5] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-Up Human Pose Estimation via Disentangled Keypoint Regression. pages 14676–14686, 2021.

[6] Duncan Zauss, Sven Kreiss, and Alexandre Alahi. Keypoint communities. In Proceedings of the IEEE/CVF International conference on computer vision, pages 11057–11066, 2021.

[7] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep HighResolution Representation Learning for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(10):3349–3364, 2021. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[8] Glenn Jocher. YOLOv5 by Ultralytics, 2020.

[9] Munawar Muhammad, Rizwan, Jocher Glenn, Chaurasia Ayush, and Laughing-q. Pose - Ultralytics YOLOv8 Docs, 2023.

[10] Joseph Redmon and Ali Farhadi. YOLO9000: Better, Faster, Stronger. pages 7263–7271, 2017.

[11] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement, 2018. arXiv:1804.02767 [cs].

[12] Dillon Reis, Jordan Kupec, Jacqueline Hong, and Ahmad Daoudi. Real-Time Flying Object Detection with YOLOv8, 2023. arXiv:2305.09972 [cs].

[13] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked Hourglass Networks for Human Pose Estimation. In Computer Vision – ECCV 2016, pages 483–499, Cham, 2016. Springer International Publishing.

[14] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative Embedding: End-to-End Learning for Joint Detection and Grouping. In Advances in Neural Information Processing Systems. Curran Associates, Inc., 2017.

[15] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. In Advances in Neural Information Processing Systems. Curran Associates, Inc., 2014.

[16] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. pages 21002–21012, 2020.

[17] Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[18] Ross Girshick. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 1440–1448, 2015.

[19] Chao Liu, Shuai Yu, Min Yu, Baole Wei, Boquan Li, Gang Li, and Weiqing Huang. Adaptive smooth l1 loss: A better way to regress scene texts with extreme aspect ratios. In 2021 IEEE Symposium on Computers and Communications (ISCC), pages 1–7, 2021

[20] Alexander Neubeck and Luc Van Gool. Efficient nonmaximum suppression. In 18th international conference on pattern recognition (ICPR'06), pages 850–855. IEEE, 2006.