A Global-Local Feature Space for Per-Object Distance Estimation from Monocular Images

Castagnolo Giulia*, De Marchi Mirco*, Fincato Matteo*, Mancusi Gianluca* and Vasile Federico*
*Summer school of European Laboratory for Learning and Intelligent Systems (ELLIS)

***Abstract*—Per-object distance estimation is a critical task in computer vision, particularly in the context of autonomous systems and object recognition. We explore state-of-the-art techniques, challenges, and advancements in per-object distance estimation and tackle the intricate task of estimating distances of objects within urban outdoor settings using single images. Traditional approaches to distance estimation often rely on geometric principles and sensor data, which may have limitations in complex real-world scenarios. Recent developments in deep learning have revolutionized this field, enabling the creation of models capable of predicting distances directly from image data. These models leverage convolutional neural networks (CNNs) and other deep learning architectures to learn intricate relationships between object appearance and distance, leading to impressive results. The study considers two reference methods for assessment: DisNet [1] and an architecture introduced by Zhu et al. [2] Nonetheless, contemporary state-of-the-art models focus on extracting local features through bounding boxes and region-of-interest (ROI) pooling, inadvertently neglecting the broader visual context of the entire image. The principal goals of this project revolve around assessing the effectiveness of the existing baseline methods and proposing an architecture that incorporates both local and global spatial reference frames to improve the accuracy of per-object distance estimation. To address privacy concerns, we have opted to utilize synthetic data in our approach.**

***Index Terms*—Distance Estimation**

I. Introduction

Among past and novel challenges, the Computer Vision community has a long-standing commitment for 3D visual perception, i.e., seeing the surrounding environment in all three spatial dimensions (including depth). In this respect, humans continuously practice such a capability in everyday life: for example, when approaching a stop sign, the driver visually assesses the remaining distance to the sign and adjusts the car's velocity accordingly. Such a simple scenario suggests the importance of a reliable 3D automatic perception, especially for AI systems dealing with autonomous driving and video surveillance applications.

While many researchers focused on popular tasks such as object detection and segmentation, a few efforts have been spent on object-specific distance estimation. Herein, the goal is to estimate the distance of a target object from the camera when it is projected onto the image plane. Pioneer works [3], [4] accomplished it with the pinhole camera model and the standard projective transformation; unfortunately, these approaches are viable only in static scenarios and suffer from radial lens distortion, hindering the estimation for objects located far from the center.

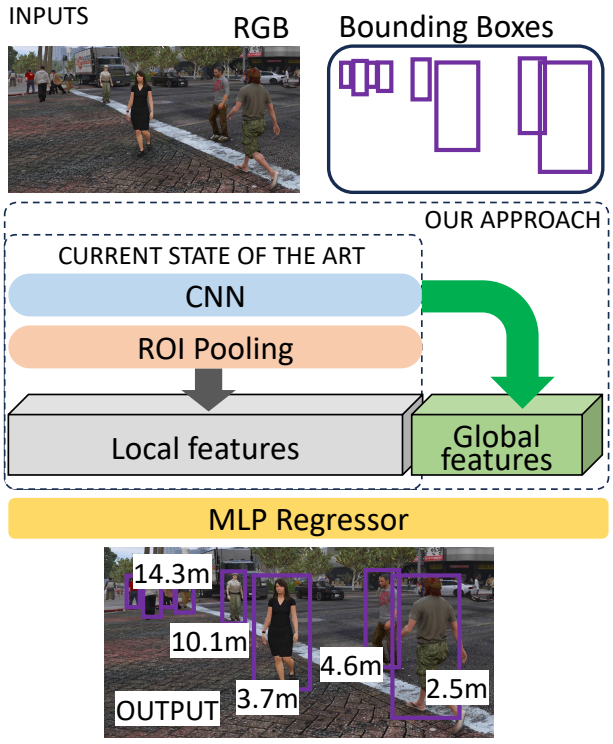

Fig. 1. Methodology overview: our contribution differs from the state of the art by integrating a global space to the features used by the distance regressor

Instead, modern approaches [5], [6] rely either on geometric or feature-based strategies. The former treat the problem as a regression task and attempt to learn the relationship, expected to be roughly linear, between the visual size of an object (such as the height of its bounding box) and its distance. In contrast, feature-based approaches exploit deep learning architectures, e.g., Convolutional Neural Network (CNN), and incorporate visual cues of the target object, as well as global information related to the scene.

This project aims to investigate and propose an approach for per-object distance estimation given a single image. In greater detail, the focus of the project is on outdoor urban scenarios: namely, the images come from a security camera filming a public place (e.g., a square or a street) with moving pedestrians and cars. Due to privacy concerns, the project relies on synthetic data only and specifically on MOTSynth [7], a recently released synthetic dataset for pedestrian detection, tracking, and segmentation. To help the students, the bounding

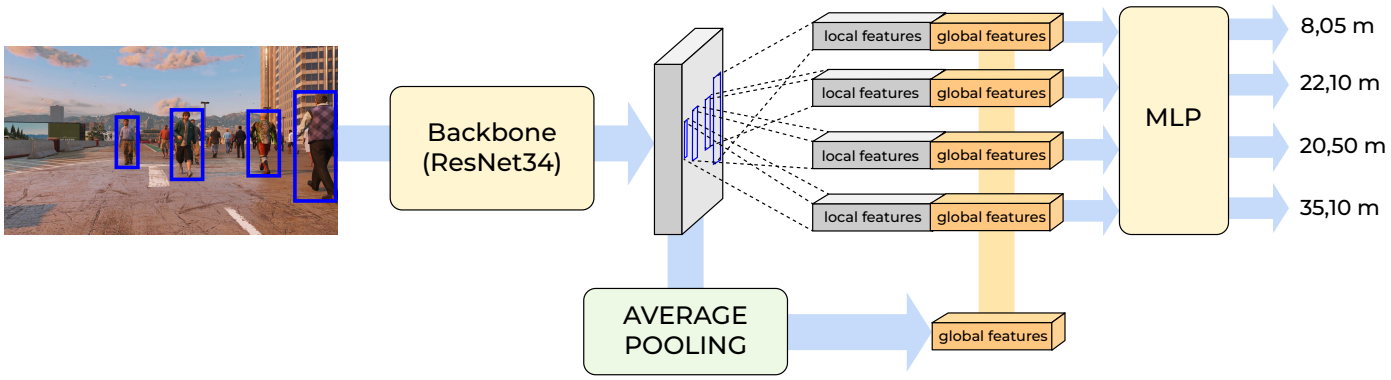

Fig. 2. Core architecture: Local features are extracted using RoI pooling, while global features are obtained through Average pooling.

boxes of the region of interest (ROI) will be provided at both training and evaluation time, along with the entire original image.

We compare our approach with two baselines: DisNet [1], a simple regression-like approach based on Multi-Layer Perceptrons (MLPs) that is fed with the relative width, height, and diagonal of bounding boxes. The architecture proposed by Zhu et al. [2] relies on ResNet as a feature extractor and RoI pooling to build object-level representations.

Figure 1 illustrates the architecture upon which our approach is based, as presented in Zhu et al. [2]. Nonetheless, the regression model employed relies solely on local features, specifically those associated with the input bounding boxes. This methodology could potentially introduce inaccuracies, particularly for objects situated farther away, as the bounding boxes are quite small, containing limited information. Consequently, we incorporate global features that encompass the entire image, capturing the broader visual patterns that extend over the individual bounding boxes.

We extend Zhu et al. [2] in two ways:

- The local representation is combined with a global image representation in order to reason about the surrounding environment
- A recurrent version that processes bounding boxes from highest to lowest, as it helps coherence among similar boxes.

II. Methodology

In light of the limitations of the traditional geometrical algorithm for distance estimation, we propose a novel learning-based approach for robust distance estimation specific to objects. Our model provides direct estimations of physical distances using RGB images and object bounding boxes. The core components of our base model consist of four key elements: a feature extractor backbone, a local feature pooling, a global feature pooling and a distance regressor (as illustrated in Figure 2).

Feature extractor: In our model, an RGB image is processed through an image feature learning network, extracting feature maps that encompass the entire RGB image. We leverage the well-established ResNet34 as our feature extractor.

Local feature pooling: We feed the feature map extracted by the feature extractor, along with the object bounding boxes (which pinpoint object locations within the image), into an ROI pooling layer. This process yields a standardized feature vector, encapsulating image characteristics within the designated bounding box regions. Consequently, this operation furnishes localized information to each individual object.

Global feature pooling: Our contribution deals specifically with the integration of global features. We apply a global average pooling on the feature map coming from the feature extractor. The resulting vector is concatenated with the local features in order to fuse global and local information. The feature map extracted by the feature extractor follows an additional branch that ends up in an average pooling, act to compact the information of the whole image. Then, this global feature space is concatenated with all local feature vectors.

Distance regressor: The global-local feature vector F_i passes through the distance regressor to predict the distance value for each object. The distance regressor comprises three fully connected (FC) layers, with layer sizes of 1024, 512, 1. To ensure that the predicted distance (denotated as $D(F_i)$) is positive, a softplus activation function is applied to the output of the final fully connected layer. Our loss for the distance regressor L_{dist} can be written as:

$$L_{dist} = \frac{1}{N}\sum_{i=1}^{N} smooth_{L1}(d_i^* - D(F_i)) \quad (1)$$

III. Results

We train the network with the ADAM optimizer, using a beta value (β) of 0.5 for 15 epochs. The learning rate begins at 0.001 and undergoes exponential decay after 10 epochs. After training, our base model is capable of directly predicting object-specific distances when provided with RGB images and object bounding boxes as input.

The metrics considered in the evaluation error are the threshold distance ($\delta_{<d}$ where d is the distance from ground truth), the Root Mean Squared Error (RMSE), the RMSE logarithmic (RMSE_{log}), the Average Localization Precision (ALP), the Average Localization Error (ALE), the Squared Relative difference (Squa.Rel.) and the Absolute Relative difference (Abs.Rel.).

Metrics	$\delta_{<1.25} \uparrow$	$\delta_{<1.25^2} \uparrow$	$\delta_{<1.25^3} \uparrow$	RMSE $\downarrow$	$\text{RMSE}_{log} \downarrow$	ALP @$1m \uparrow$	ALE $\downarrow$	Squa. Rel. $\downarrow$	Abs. Rel. $\downarrow$
DisNet [1]	87.6%	98.7%	99.4%	2.930	0.162	31.99%	2.263	0.480	0.124
Zhu et al. [2]	47.4%	84.4%	94.3%	6.800	0.395	12.74%	5.159	1.642	0.229
Ours	**68.2%**	**92.8%**	**98.6%**	**5.085**	**0.245**	**18.47%**	**3.806**	**1.233**	**0.205**

TABLE I

COMPARISON OF DISNET [1], ZHU ET AL. [2] AND OUR DISTANCE ESTIMATORS ON THE SMALL VERSION OF MOTSYNTH [7] (TEST SEQUENCES 2 AND 6). OUR METHOD EXHIBITS SUPERIOR PERFORMANCE ACROSS ALL THE METRICS REPORTED.

Metrics	$\delta_{<1.25} \uparrow$	$\delta_{<1.25^2} \uparrow$	$\delta_{<1.25^3} \uparrow$	RMSE $\downarrow$	$\text{RMSE}_{log} \downarrow$	ALP @$1m \uparrow$	ALE $\downarrow$	Squa. Rel. $\downarrow$	Abs. Rel. $\downarrow$
DisNet [1]	96.5%	99.8%	99.9%	1.958	0.105	49.56%	1.418	0.160	0.077
Zhu et al. [2]	**77.26%**	**87.67%**	90.93%	12.42	0.494	**45.96%**	5.737	2.982	**0.167**
Ours	59.9%	87.25%	**95.43%**	**9.462**	**0.297**	24.75%	**5.435**	**2.136**	0.224

TABLE II

COMPARISON OF DISNET [1], ZHU ET AL. [2] AND OUR DISTANCE ESTIMATORS ON THE SMALL VERSION OF MOTSYNTH [7] (SEQUENCES 8 AND 9). OUR APPROACH DEMONSTRATES EXCEPTIONAL PERFORMANCE ACROSS GLOBAL METRICS THAT DO NOT FOCUS ON SPECIFIC INTERVALS, INCLUDING RMSE, RMSE_{log}, ALE, AND Squa. Rel..

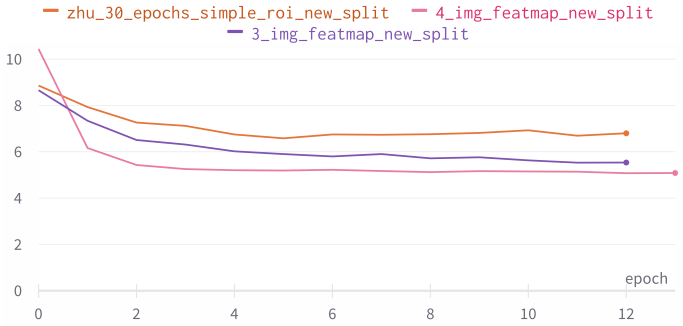

Fig. 3. The descent of RMSE over increasing epochs in Zhu's model and our approach in two distinct training setups. Our approach demonstrates a significantly faster convergence of RMSE, surpassing Zhu's model with a notably lower error rate.

Due to data availability constraints, our model's training was *limited to only 10 sequences* from the extensive MOTSynth dataset. For this reason, we conducted two separate experiments using distinct training and test set choices to determine if the model's performance indicates its actual learning capabilities or if the specific composition of the training and test sets influences it.

In the context of result analysis, Table I showcases models evaluated on sequences 2 and 6, designated as the test set, with the remaining sequences comprising the training set. In this configuration, the Figure 3 shows a graph of the RMSE descent over epochs. On the other hand, Table II focuses on sequences 8 and 9. The results in Table I and Figure 3 show that our method outperforms Zhu et al. [2] model in all the metrics, performing better results. However, the results worsen if the method is evaluated in 8 and 9 sequences as in Table II.

This limitation arises from the fact that the model underwent training and evaluation using a mini version of the MOTSynth dataset. As a result, our model faces challenges when attempting to acquire scene-level features from the training subset consisting of sequences 0-7. In contrast, it is apparent that DisNet, built solely on an MLP regressor and geometric features, demonstrates robust learning and generalization capabilities, even when confronted with a small dataset.

IV. CONCLUSION

In conclusion, this study investigates and proposes a novel approach for per-object distance estimation in outdoor urban settings using synthetic data. The proposed approach ultimately aims to bridge the gap on local information, introducing global features, improving accuracy, and enhancing the capabilities of AI systems in real-world applications like autonomous driving and video surveillance. The proposed model achieves better results in accuracy, but by varying the test dataset there may be some large performance differences. While DisNet, being a lighter and simpler model, performs well in any case. The proposed approach should be tested in the future on the total amount of MOTSynth dataset.

REFERENCES

[1] M. A. Haseeb, J. Guan, D. Ristic-Durrant, and A. Gräser, "Disnet: a novel method for distance estimation from monocular camera," 2018.

[2] J. Zhu, Y. Fang, H. Abu-Haimed, K.-C. Lien, D. Fu, and J. Gu, "Learning object-specific distance from a monocular image," 2019.

[3] F. Gökçe, G. Üçoluk, E. Şahin, and S. Kalkan, "Vision-based detection and distance estimation of micro unmanned aerial vehicles," *Sensors*, vol. 15, no. 9, pp. 23 805–23 846, 2015. [Online]. Available: https://www.mdpi.com/1424-8220/15/9/23805

[4] S. Tuohy, D. O'Cualain, E. Jones, and M. Glavin, "Distance determination for an automobile environment using inverse perspective mapping in opencv," in *IET Irish Signals and Systems Conference (ISSC 2010)*, 2010, pp. 100–105.

[5] L. Jing, R. Yu, H. Kretzschmar, K. Li, C. R. Qi, H. Zhao, A. Ayvaci, X. Chen, D. Cower, Y. Li, Y. You, H. Deng, C. Li, and D. Anguelov, "Depth estimation matters most: Improving per-object depth estimation for monocular 3d detection and tracking," 2022.

[6] Y. Li, T. Chen, M. Kabkab, R. Yu, L. Jing, Y. You, and H. Zhao, "R4d: Utilizing reference objects for long-range distance estimation," 2022.

[7] M. Fabbri, G. Braso, G. Maugeri, O. Cetintas, R. Gasparini, A. Osep, S. Calderara, L. Leal-Taixe, and R. Cucchiara, "Motsynth: How can synthetic data help pedestrian detection and tracking?" 2021.