Per-Object Distance Estimation: A Quantitative Comparison

L. Barsellotti, D. A. Coccomini, J. Fink, E. Martini, A. Panariello, Bin Ren
luca.barsellotti@unimore.it, davidealessandro.coccomini@isti.cnr.it, jerome.fink@unamur.be,
enrico.martini@univr.it, aniello.panariello@unimore.it, bin.ren@unitn.it

Abstract

Accurate distance estimation for individual objects is crucial in applications with stringent safety requirements, such as autonomous driving, surveillance, and robotics. Existing methods can be broadly categorized into two groups: one that relies on local information related to the target object, including factors like its bounding box proportions or object pose, and another that leverages visual features, namely, local semantic scene knowledge. However, these methods have not yet integrated both approaches, which could prove advantageous for the task at hand. In this study, we extend the work of Zhu et al. [24] by combining information about the pedestrian's pose to enhance the accuracy of our distance estimation system.

1. Introduction

In today's high-stakes domains like autonomous driving, surveillance, and robotics, achieving precise object distance estimation is crucial. Accurate measurements of distances between objects and their surroundings are essential for ensuring the highest safety and operational efficiency in these critical domains.

Consequently, significant advancements have been made in the field of distance estimation. This project is dedicated to enhancing the outcomes achieved using a well-known architecture proposed by Zhu et al. [24]. Such an architecture consists of multiple modular models, providing flexibility for modifications at various stages. The key questions addressed in this report are as follows:

- **Which feature extractor outperforms the baseline proposed by Zhu et al. [24]?**
- **What is the influence of different pooling methods on the results?**
- **Can we enhance baseline results by incorporating additional meta-information into the input data?**
- **Is there potential for improvement by introducing a graph-based global refinement module into the architecture?**

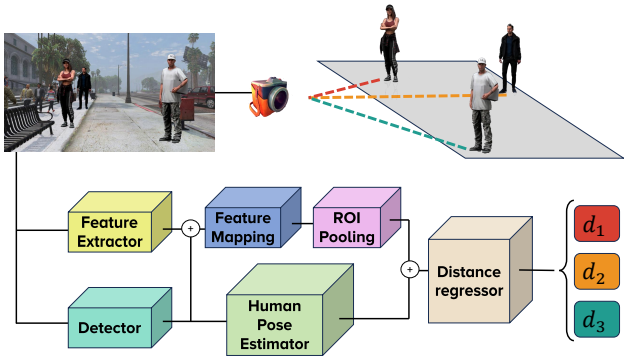

Figure 1. Overview of the proposed pipeline.

2. Related Works

The task of per-object distance estimation from monocular images is approached through two methods: a geometric approach [8, 9, 21], which relies solely on the bounding box dimensions of the object (*i.e.*, height and width), and a feature-based approach [12, 15, 24], where a convolutional backbone is utilized to extract visual features.

As a baseline, we adopted the methodology proposed by Zhu et al. [24], in which they introduce a straightforward architecture. This architecture extracts object-specific features using RoIPool [7] from the features generated by a ResNet [11] backbone, and subsequently, the distances are regressed using a Multilayer Perceptron (MLP).

3. Method

We conducted a series of experiments involving modifications to the architecture proposed by Zhu et al. [24]. Specifically, we began by exploring different backbone models, including a ResNet with an FPN [16] and the Transformer-based backbone DINO-ViT [1]. Once we identified the most suitable backbone, we proceeded to assess various architectural variations, which included:

- Substituting RoIPool [7] with RoIAlign [10].
- Incorporating additional information related to human pose obtained from a human pose estimator (Figure 1).

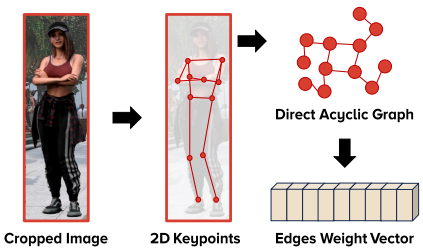

Figure 2. HPE pipeline.

- Exploiting global features through the usage of GATv2 [3].

3.1. HPE Integration

We integrate a Human Pose Estimation (HPE) module to feed the distance regressor with additional information about the size of the principal body segments. As reported in Figure 2, the process begins with cropping the images using bounding boxes. These bounding boxes are generated through object detection or localization techniques, outlining each person in the image.

Once cropped, the image is fed into an HPE. This deep learning model is trained to locate keypoints positions from images. These joint positions correspond to anatomical landmarks like shoulders, elbows, and knees. To capture the spatial relationships and dependencies between these keypoints, a subset of the extracted ones is used to construct a Weighted Directed Acyclic Graph (DAG). Within this graph representation, each keypoint corresponds to a node, and the connections correspond to edges between nodes, that are directed and acyclic, ensuring no loops in the graph. Each edge in the DAG is equipped with a weight equivalent to the spatial distance between the vertices it connects. Finally, we transform the weights associated with the edges in the DAG into a vectorial representation. Finally, such a vector is processed by a regressor model to obtain the final prediction.

3.2. GAT Integration

Although the convolutional backbone can capture certain aspects of global information, we note that, following the RoIPooling operation, each object's feature vector is processed individually. To enhance the overall contextual understanding across objects, we gather the feature vectors of all objects and process them through a Graph Attention Network [3]. This approach allows us to generate a unified and spatially consistent estimation considering the relationships between objects within the scene.

4. Experimental Results

Dataset The dataset used for evaluation is a subset of the MOTSynth dataset [5], a large synthetic dataset for pedestrian detection, tracking, and segmentation in an urban environment. All video sequences come along with bounding boxes and the distance from the camera of each person appearing in the video.

Backbone	$\delta_{<1.25}\uparrow$	Abs Rel$\downarrow$	RMSE$\downarrow$	ALP$_{@1m}\uparrow$
DINO ViT-S/8	0.259	0.475	13.430	0.066
ResNet34	0.457	0.232	6.726	0.118
VGG16	0.527	0.202	6.414	0.158
ResNet34-FPN	**0.661**	**0.191**	**6.369**	**0.248**

Table 1. Comparison of different backbones.

4.1. Experimental Setup

Feature Extractor In their paper, Zhu et al. [24] used ResNet [11] as a feature extractor. In this first experiment, we investigate the performance of various backbones. The tested feature extractors are the following:

- The widely adopted convolutional backbones VGG16 [20] and ResNet [11].
- An FPN [16] variant of ResNet to exploit the higher resolution of the feature map.
- The Transformer-based backbone DINO-ViT [1].

Metrics We rely on standard metrics of per-object distance estimation [4, 6, 17, 19, 24], such as the τ-**Accuracy** (δ_τ) [14] (*i.e.*, the maximum allowed relative error), the Average Localization Precision (**ALP**$_\tau$) [2, 23] (*i.e.*, the mean average error in true distance range) and classical error distances [24]: absolute relative error (**Abs. Rel.**), square relative error (**Squa. Rel.**), root mean squared error in linear and logarithmic space (**RMSE** and **RMSE**$_{log}$), and average localization error (**ALE**) [2]. To refer the reader to the cited works for a deeper explanation of such metrics.

Experimental Setting We train each model for 10 epochs with a fixed batch size and learning rate. We use Cosine Annealing Learning Rate scheduler, starting from 10^{-4}. We employ the AdamW optimizer. We use the MediaPipe [18] Pose Estimator to predict the poses of detected people.

4.2. Quantitative Comparison

Backbone From the experiments reported in Table 1, the DINO feature extractor seems unable to return useful features, causing a considerable decrease in performance. We argue that this is due to the size of the DINO model, which would require more training and data to learn significant patterns.

Conversely, traditional convolutional backbones yield notably superior results, with the FPN variant of ResNet standing out as the top performer. This substantial enhancement compared to the standard ResNet can be attributed to

HPE	$\delta_{<1.25}\uparrow$	Abs Rel$\downarrow$	RMSE$\downarrow$	ALP$_{@1m}\uparrow$
✗	0.661	0.191	6.369	0.248
✓	**0.683**	**0.182**	**6.059**	**0.282**

Table 2. Contribute of the keypoints features with the ResNet34-FPN backbone.

RoI Op	$\delta_{<1.25}\uparrow$	Abs Rel$\downarrow$	RMSE$\downarrow$	ALP$_{@1m}\uparrow$
Pooling	0.661	0.191	6.369	0.248
Align	**0.688**	**0.178**	**6.320**	**0.272**

Table 3. Results with RoiPool and RoIAlign with the ResNet34-FPN backbone.

GAT	$\delta_{<1.25}\uparrow$	Abs Rel$\downarrow$	RMSE$\downarrow$	ALP$_{@1m}\uparrow$
✗	0.661	0.191	6.369	0.248
✓	**0.699**	**0.176**	**4.665**	**0.290**

Table 4. Contribute of the GATv2 for the Global Context Encoding with the ResNet34-FPN backbone.

the ability of the FPN branch to produce a higher-resolution feature map. Consequently, when the RoIPool operation is applied, the selected features exhibit a finer level of detail.

HPE Contribute In Table 2, we present the results obtained by incorporating keypoints distance features into our architecture. Notably, we observe a performance improvement when using such an additional signal.

The improved performance can be attributed to using keypoints distance features, which offer a more comprehensive representation of the human pose within our model. By incorporating these features, our architecture gains access to valuable spatial information, allowing it to make more precise predictions about the absolute distance of a person.

Per-object feature pooling through RoIAlign We propose to exploit RoIAlign instead of RoIPool since it addresses the issue of spatial misalignment more effectively. Indeed, RoIPool discretizes regions of interest into a fixed grid and uses max-pooling, which can lead to inaccuracies when regions do not align perfectly with the grid.

On the contrary, RoIAlign avoids misalignment by using the more accurate bilinear interpolation that produces features at the same level of granularity as the bounding box coordinates corresponding to the region of interest. Such a sub-pixel accuracy results in a more detailed and accurate representation of objects within the region.

We report the results of such a comparison in Table 3.

GATv2 for Global Context Encoding In the architecture proposed by Zhu et al. [24], pedestrians are represented individually through feature vectors derived from dense features extracted by the visual feature extractor. These individual feature vectors are subsequently fed directly into the regressor head. However, to leverage the potential benefits of interconnecting the feature vectors of different pedestrians, which can enhance distance regression, we implement a Graph Attention Network [22] (GAT) as a means to establish meaningful relationships between each pedestrian's information and that of others within the scene.

The GAT excels at capturing complex relationships and dependencies in graph-structured data. What sets the GATs apart from the GCNs [13] is their attention mechanism, which assigns different attention weights to neighboring nodes when aggregating information. This attention mechanism allows GATs to dynamically weigh the contributions of neighboring nodes during message-passing, enabling them to adaptively focus on the most relevant nodes for each node in the graph.

In particular, we employ the GATv2 architecture proposed by Brody et al. [3]. In Table 4, we report the evaluation enhanced using the GATv2. The results prove the relevance of exploiting the relationships within the features extracted for each pedestrian to improve the distance estimation performance.

5. Conclusion

In this report, we explored some architectural adaptations for distance estimation on the MOTSynth dataset, starting from the architecture proposed by Zhu et al. [24].

We evaluated several feature extractors, discovering the superiority of the convolutional-based backbones compared with a DINO-based backbone. Moreover, we embodied the information about the pose of the detected subjects by identifying their keypoints and calculating a feature vector representing the distances between connected joints. Including this information in the classification process leads to performance improvements, highlighting the relevance of these features in the distance estimation process.

Finally, we argue that combining such proposed modules could lead to an even more significant performance boost, which we could not empirically demonstrate due to time constraints.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2021. 1, 2

[2] Lorenzo Bertoni, Sven Kreiss, and Alexandre Alahi. Monoloco: Monocular 3d pedestrian localization and uncertainty estimation. In *IEEE International Conference on Computer Vision*, 2019. 2

[3] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations Workshop*, 2022. 2, 3

[4] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in Neural Information Processing Systems*, 2014. 2

[5] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Aljoša Ošep, Riccardo Gasparini, Orcun Cetintas, Simone Calderara, Laura Leal-Taixé, and Rita Cucchiara. Motsynth: How can synthetic data help pedestrian detection and tracking? In *IEEE International Conference on Computer Vision*, 2021. 2

[6] Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *Proceedings of the European Conference on Computer Vision*, 2016. 2

[7] Ross Girshick. Fast r-cnn. In *IEEE International Conference on Computer Vision*, 2015. 1

[8] Fatih Gökçe, Göktürk Üçoluk, Erol Şahin, and Sinan Kalkan. Vision-based detection and distance estimation of micro unmanned aerial vehicles. *Sensors*, 2015. 1

[9] Muhammad Abdul Haseeb, Jianyu Guan, Danijela Ristic-Durrant, and Axel Gräser. Disnet: a novel method for distance estimation from monocular camera. *10th Planning, Perception and Navigation for Intelligent Vehicles (PPNIV18), IROS*, 2018. 1

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017. 1

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016. 1, 2

[12] Longlong Jing, Ruichi Yu, Henrik Kretzschmar, Kang Li, Charles R Qi, Hang Zhao, Alper Ayvaci, Xu Chen, Dillon Cower, Yingwei Li, et al. Depth estimation matters most: Improving per-object depth estimation for monocular 3d detection and tracking. *arXiv preprint arXiv:2206.03666*, 2022. 1

[13] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 3

[14] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2014. 2

[15] Yingwei Li, Tiffany Chen, Maya Kabkab, Ruichi Yu, Longlong Jing, Yurong You, and Hang Zhao. R4d: Utilizing reference objects for long-range distance estimation. *arXiv preprint arXiv:2206.04831*, 2022. 1

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017. 1, 2

[17] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015. 2

[18] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. 2

[19] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *Proceedings of the European Conference on Computer Vision*, 2020. 2

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014. 2

[21] Shane Tuohy, Diarmaid O'Cualain, Edward Jones, and Martin Glavin. Distance determination for an automobile environment using inverse perspective mapping in opencv. In *IET Irish Signals and Systems Conference (ISSC 2010)*, 2010. 1

[22] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations Workshop*, 2018. 3

[23] Yu Xiang, Wongun Choi, Yuanqing Lin, and Silvio Savarese. Data-driven 3d voxel patterns for object category recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2015. 2

[24] Jing Zhu and Yi Fang. Learning object-specific distance from a monocular image. In *IEEE International Conference on Computer Vision*, 2019. 1, 2, 3