MonStereo: When Monocular and Stereo Meet at the Tail of 3D Human Localization

Lorenzo Bertoni  Sven Kreiss  Taylor Mordan  Alexandre Alahi  
EPFL, VITA lab  
CH-1015 Lausanne, Switzerland  
first-name.last-name@epfl.ch

Abstract: Monocular and stereo vision are cost-effective solutions for 3D human localization in the context of self-driving cars or social robots. However, they are usually developed independently and have their respective strengths and limitations. We propose a novel unified learning framework that leverages the strengths of both monocular and stereo cues for 3D human localization. Our method jointly (i) associates humans in left-right images, (ii) deals with occluded and distant cases in stereo settings by relying on the robustness of monocular cues, and (iii) tackles the intrinsic ambiguity of monocular perspective projection by exploiting prior knowledge of human height distribution. We achieve state-of-the-art quantitative results for the 3D localization task on KITTI dataset and estimate confidence intervals that account for challenging instances. We show qualitative examples for the long tail challenges such as occluded, far-away, and children instances.

Keywords: 3D localization, autonomous vehicles, social robots, long tail

1 Introduction

Recently, human 3D localization for autonomous vehicles or social robots has been addressed with cost-effective vision-based solutions [1, 2, 3, 4]. All the approaches strive to improve state-of-the-art results in popular metrics. Yet these solutions do not necessarily convey trust in real-world applications, and the long tail of 3D perception opens a Pandora’s box of undetected challenges. While many methods perform very well “on average”, can they still be trusted on the most challenging cases? The long tail of 3D object localization, i.e., the share of instances where methods struggle the most, is crucial for safety but rarely evaluated in standard benchmarks [5]. This is especially relevant for pedestrians, arguably the most crucial category from a safety point of view.

Stereo-based methods have the potential for accurate 3D human localization, as they are free from the perspective projection ambiguity, inevitable in the monocular case [2]. Pseudo-LiDAR [4] drastically reduced the discrepancy between camera and LiDAR performances by converting a stereo-based dense depth map into 3D point clouds and directly applying LiDAR-based object detectors [6, 7]. However, computing depth from disparity poses two main challenges. Instances can be located out of the field-of-view or be occluded in one of the two images and an association may not be available. Furthermore, a small disparity error (e.g., a pixel shift) for far-away objects leads to unacceptable errors of several meters, as the error grows quadratically with depth [8]. We identify occluded and far instances as the largest share of the stereo-based long tail of predictions. On the contrary, monocular images are less error-prone for far instances and do not depend on accurate detections on both images. Bertoni et al. [2] have achieved competitive performances in 3D human localization by exploiting the known prior distribution of human heights. However, this approach fails in the presence of children or very tall people, connecting the long tail of monocular 3D localization with the distribution of human heights.

In this work, we want to leverage the best of both worlds, i.e., stereo and monocular methods, in a unified learning framework tailored for pedestrian 3D localization. Our method, referred to as Mon-Stereo, jointly associates detections in left-right images and implicitly learns to leverage monocular and/or stereo cues. Moreover, it also learns to communicate uncertainty driven by the cues (again without direct supervision at training time). Our approach uses an off-the-shelf pose detector [9] on left-right images to obtain 2D keypoints, a low-dimensional representation of humans. A simple feed-forward network estimates if each input pair is formed by the same person from left-right im-
ages and, concurrently, estimates the 3D location of pedestrians with their corresponding uncertainty (accounting for stereo disparity and/or monocular cues).

The popular KITTI dataset [5] has limited variation of instances, oversimplifying the monocular task. We address the long tail of height distribution by injecting prior knowledge from the real world. Leveraging the simplicity of manipulation of 2D keypoints, we create instances of people from a broader spectrum of heights. This conveys information about the real challenge of the task in the data domain, thus, increases the network performance and calibrates the estimated confidence intervals without the need of hand-crafted architectures.

In summary, we propose a unified learning framework that jointly matches detections in left-right images and estimates the 3D localization of each pedestrian. We focus on the limitations of monocular and stereo vision, referred to as the long tail challenge, by jointly exploiting stereo and monocular cues with a measure of uncertainty. We also design a data augmentation procedure to tackle the long tail of the human heights distribution. Our network achieves state-of-the-art results on 3D localization metrics and provides reliable confidence intervals even for challenging cases. Our code is available online at https://github.com/vita-epfl/monstereo.

2 Related Work

Monocular 3D Detection. Estimating depth from a single RGB image is an ill-posed task for non-rigid human bodies. To the best of our knowledge, few methods have explicitly tackled vulnerable road users in contrast to the large body of works related to rigid vehicles [10, 11, 12, 13, 14]. The recent MonoLoco method [2] predicted confidence intervals of pedestrians to address the task ambiguity for 3D localization, while MonoPSR [1] learned local shapes of objects with privileged signal at training time. Yet, they fail to address the long tail of 3D human localization.

Stereo 3D Detection. Stereo-based 3D detectors can be grouped into instance-level and pixel-level depth estimators. The instance-level approach consists in detecting instances in the image plane and comparing features of proposals in left and right frames to correctly associate objects and estimate their location [15, 16, 17, 18, 3, 19]. Among them, PSF [3] was designed for human localization and, similarly to our method, leverages 2D keypoints to solve the association task. However, their 3D output is simply the median depth calculated from a set of disparities. The pixel-level approach consists in estimating a dense disparity map for every pixel and transforming the dense map into a 3D point cloud [4, 20]. The pseudo point cloud can then be used to detect vehicles and pedestrians by applying LiDAR-based algorithms [6, 7]. The underlying task of all previous methods is to compute disparity from pixels, either locally to associate and align pairs of left-right instances, or globally to find dense correspondences between pixels. Qin et al. [18] have recently proposed to extend a monocular baseline to predict 3D locations of car instances with a triangulation network. Our work goes beyond the concept of “depth from disparity” and is not limited by the discrete nature of pixels, but can exploit together monocular and stereo cues to directly estimate a continuous depth.
3 Vision-based 3D Localization Ambiguity

In this section, we quantify the limitations of monocular and stereo modalities for the 3D human localization task. Estimating the 3D location of objects from a single RGB image is a fundamentally ill-posed problem due to the ambiguous projections to the 2D image. This is particularly true for humans due to their variation of height and non-rigid body structure. Bertoni et al. [2] quantified this ambiguity as a function of the distance from the camera, assuming that the distribution of human stature follows a Gaussian distribution for male and female populations [21]. The expected localization error \( \hat{e}_{\text{mono}} \) due to height variations of people can be obtained by
\[
\hat{e}_{\text{mono}} = C \cdot r_{\text{gt}},
\]
where the constant \( C \) is modelled from the distribution of human heights and \( r_{\text{gt}} \) is the ground-truth distance. On the other side, even if stereo methods do not suffer from the intrinsic ambiguity of perspective projection, the error grows quadratically with depth, making disparity estimation very sensitive to pixel resolution. The depth error \( e_z \) can be expressed as a function of the disparity error \( e_d \) [8] as
\[
e_z \approx \frac{z^2}{bf} e_d,
\]
where \( z \) is the depth, \( b \) the camera baseline and \( f \) the focal length. With the goal of comparing monocular and stereo limitations, we analyze what we call the pixel error: the depth error due to a disparity error of one pixel. Its value depends on the characteristics of the camera and we use the camera parameters of KITTI dataset [5], a popular dataset for 3D object detection with stereo imaging at a resolution of 1240 × 380 pixels. The results, shown in Figure 4b, highlight that the stereo depth error can become more challenging than the monocular one for humans just over 20 meters far. For example, a disparity error of 1 pixel at 40 meters corresponds to 4.5 meters of depth error. These conclusions depend on the precision of the disparity estimation and the image resolution, but highlight the importance of monocular estimation for 3D perception.

4 Method

The goal of our approach is to detect, associate and estimate the 3D positions of pedestrians in a pair of stereo images. We identified two main challenges for a stereo network: (i) when a person is not identified in both images, there is no disparity information and (ii) disparity estimation for faraway objects leads to poor predictions. We propose a simple yet effective way to tackle both issues.

Architecture. Our method consists of two steps. First, we reduce the input dimensionality by predicting 2D keypoints for each person in left-right images. Keypoints are a low dimensional representation which is invariant to many nuisances, is suitable in the low data regime and is prone to easy manipulations (for data augmentation). Second, we analyze pairs of keypoints from left-right images in an “all-vs-all” setting to predict 3D location, and confidence interval of every person in the scene. Our simple architecture is shown in Figure 2 and consists of few fully-connected layers with batch-normalization, residual connections [22], and dropout [23].

Input/output. We use an off-the-shelf pose detector (e.g., PifPaf [9]) to obtain a set of keypoints \( [\vec{x}_i, \vec{y}_i]_l^T \) for every person \( i \) in the left and right images. Each keypoint is projected into normalized image coordinates \( [\vec{x}^*, \vec{y}^*, 1]^T_i = I(i) \) to prevent overfitting to a specific camera. To construct the network inputs, we associate in an “all-vs-all” way the keypoints \( I^{(l)} \) from each person \( l \) in the left image with the one \( I^{(r)} \) from each person \( r \) in the right image, to form the associated pair \( I^{(l,r)} \):
\[
I^{(l,r)} = I^{(l)} \parallel (I^{(l)} - I^{(r)}) \quad \forall \ l \in N_L, r \in N_R,
\] (1)
where $||$ is a concatenation operation, and $N_L$, $N_R$ denote the sets of detected instances in the left-right image pair. If the sets of keypoints $I^{(l)}$ and $I^{(r)}$ belong to the same person, we call the input $I^{(l,r)}$ stereo pair, otherwise monocular pair. We treat this problem as a binary classification task and use binary cross entropy loss to train our network. We refer to this association task as Instance-based stereo matching (ISM) and to its loss as ISM loss. Concurrently, we predict 3D localization, 3D bounding box, and viewpoint angle with a regressive model in a multi-task setting. Estimating depth is arguably the most critical component due to intrinsic limitations of monocular and stereo modalities described in Section 3. To disentangle the depth ambiguity from the other components, we use a spherical coordinate system $(r, \beta, \psi)$, namely radial distance $r$, azimuthal angle $\beta$, and polar angle $\psi$. Another advantage of using a spherical coordinate system is that the size of an object projected onto the image plane directly depends on its radial distance $r$ and not on its depth $z$ [2].

Uncertainty. We model aleatoric uncertainty for the depth estimation task following Bertoni et al. [2] and using a relative Laplace loss based on the negative log-likelihood of a Laplace distribution as $L_{Laplace}(\mu, b) = \frac{1}{2} \frac{||x - \mu||_2}{b} + \log(2b)$, where $x$ is the ground-truth and $(\mu, b)$ the predicted distance and the spread, respectively, making this training objective an attenuated $L_1$-type loss via spread $b$. At inference time, the model predicts a radial distance $\mu$ and a spread $b$ which indicates its confidence about the predicted distance. The use of spherical coordinates allows to convey all the 3D localization uncertainty into the radial component $r$. The spherical angles $\beta$ and $\psi$ can be derived from the projection of the object onto the image plane.

Inference. The network performs 3D localization as well as ISM by predicting whether each pair of keypoints belongs to the same person (stereo pair) or to different ones (monocular pair). The ISM component is also used to filter multiple results for the same person. At inference time, the network predicts $N_R$ outputs for each person in the left image (one for each associated pair) and selects the one with the highest predicted stereo matching. In fact, a stereo pair always contains more information about the left instance than a monocular pair. For a single image pair, the number of pairwise combinations grows quadratically as $N_L * N_R$ but, as the inputs are low-dimensional, the computation is parallelizable by including all the pairs in same batch.

Knowledge Injection. Monocular estimates are essential to address the long tail of stereo-based 3D localization, but they present their own issues. A typical dataset for 3D object detection, such as KITTI [5] is not representative of the real world as it only contains few scenes recorded from a single city. For instance, we identified only five images with children in the entire dataset and, in the case of a child, any monocular estimate of depth will either fail or rely solely on the ground plane estimation [24]. These settings make the network over-confident toward monocular estimates, creating two issues: (i) the predicted confidence intervals do not reflect the real distribution of human heights and the model can drastically fail in case of children or tall people; (ii) the training phase becomes ineffective as the network relies on monocular estimates even when a stereo association is available. To tackle both issues, we inject knowledge in the training data by augmenting it with relevant examples from the long tail of human height distribution. We augment KITTI dataset with synthetic 2D keypoints of people of heights ranging from 1.2 meters to 2 meters. We rely on the mild assumption that the aspect ratio between children and adults is unchanged and for each set of keypoints $I^{(l,r)}$, we sample a height $h$ from the uniform distribution $\mathcal{U}(1.2, 2)$ and we derive a new ground-truth distance from the triangle similarity relation of human heights and distances. Then, we create a new input $I^{(l,r)}_*$ updating the disparity and the ground-truth distance. We repeat this procedure for every stereo pair and monocular pair with double-sided advantages. The network benefits from augmented stereo pairs as it learns that disparity estimates correspond to correct depth whereas the monocular assumption of average height breaks down. It also benefits from augmented monocular pairs, becoming receptive to more realistic human height variations, including children or very tall people. This knowledge is reflected in more calibrated confidence intervals.

5 Experiments

Pedestrian 3D localization is a safety-critical task for self-driving cars and social robots where it is not sufficient to be accurate “on average”. In parallel to standard metrics, we evaluate the long-tail by analyzing box plots and predicted confidence intervals. In addition, we critically review official KITTI 3D metrics for pedestrians and propose a practical 3D localization metric for pedestrians.
| Method                 | Easy | Mod. | Hard | All  | Easy | Mod. | Hard | All  |
|-----------------------|------|------|------|------|------|------|------|------|
| **Monocular**         |      |      |      |      |      |      |      |      |
| Mono3D [25]           | 2.26 | 3.00 | 3.98 | 2.62 | 9.21 | 1.26 | 0.21 | 7.22 |
| MonoPSR [1]           | 0.89 | 2.00 | 2.40 | 1.51 | 48.87| 12.54| 0.47 | 35.35|
| MonoLoco [2]          | 0.83 | 1.12 | 1.15 | 0.93 | 49.01| 19.44| 1.89 | 38.76|
| **Stereo**            |      |      |      |      |      |      |      |      |
| E2E-PL [20]           | 0.12 | 0.17 | 0.60 | 0.15 | 49.32| 4.43 | 0.44 | 31.31|
| OC [19]               | 0.10 | 0.14 | 0.75 | 0.13 | 65.58| 26.38| 1.46 | 41.30|
| 3DOP [15]             | 0.67 | 1.19 | 1.93 | 0.93 | 57.88| 22.70| 3.85 | 45.92|
| PSF [3]               | 0.55 | 0.65 | 0.80 | 0.56 | 57.27| 19.94| 4.82 | 46.15|
| P-LiDAR [4]           | **0.16** | 0.72 | 1.59 | 0.46 | **88.94** | 42.91 | **10.41** | 66.33 |
| B-ReID                | 0.73 | 0.78 | 1.02 | 0.77 | 73.81| 39.44| 4.48 | 58.23 |
| B-Pose                | 0.65 | 0.77 | 1.18 | 0.72 | 73.92| 39.10| 4.82 | 58.28 |
| B-Median              | 0.57 | 0.69 | 0.78 | 0.61 | 80.19| 50.38| 8.17 | 64.00 |
| Our MonStereo         | 0.29 | **0.41** | **0.50** | **0.34** | 85.54| **54.27** | 8.92 | **67.60** |

Table 1: Comparing our proposed method against baselines on KITTI dataset [5]. We use PifPaf [9] as off-the-shelf network to extract 2D poses. On the RALP metric, our MonStereo achieves state-of-the-art results. On the ALE metric, the confidence threshold of methods has been set to 0.5 and we show the recall between brackets to insure fair comparison. Italics entries are not directly comparable as they achieve a lower recall even when no threshold is set. Our method performs better on hard instances while maintaining 2-5 times higher recall. The improvement of jointly solving the ISM and the 3D localization tasks is shown by the three baselines (B-).

Figure 3: Box plots of Average Localization Error (ALE). Circles identify outliers. Our MonStereo achieves very robust performance in the long tail with a maximum error of 7 meters for far instances and less than 5 meters in all other cases. Every other stereo method has few catastrophic estimates even for very close people. MonStereo monocular component stabilizes the performances as shown by the performances of the monocular MonoLoco [2]: on average not as accurate as a stereo method but more robust.

### 5.1 Baselines

Our learning framework jointly solves the instance-based stereo matching (ISM) and the 3D localization tasks in an end-to-end manner. As baselines, we analyze performances when solving these two tasks separately with deterministic approaches.

**ISM Baselines.** We develop two baselines to associate people in left-right images. **B-Pose:** we use pose similarity based on the detected 2D keypoints, i.e., calculating how similar two poses are. We zero-center reference and target poses and we calculate the L2 norm between our reference vector...
and all the target vectors and save the scores. **B-RelID:** We associate the same person in left-right pairs of images by looking at the appearance of the person and the scene around him. We use a state-of-the-art Re-Identification model [26] trained on Market-1501 [27] to make the association from cropped images. Both methods provide the best similarity score for each person in the left image with respect to all the people in the right image.

**Median Baseline.** Our network not only solves the ISM task, but also estimates the depth from a set of keypoints. As a baseline **B-Median** we apply our network and, if a match between two people is found, we calculate the depth as the median value of a set of disparities.

**Other Baselines.** We also compare our method with several state-of-the-art monocular and stereo baselines in Table 1. All of them provide results on the KITTI validation set.

### 5.2 Implementation Details

We train and evaluate our model on KITTI Dataset [5] using the train/val split of Chen et al. [25]. To detect 2D keypoints, we use the off-the-shelf pose detector PifPaf [9] and we upscale the images by a factor of two to match the minimum dimension of 32 pixels of COCO instances. We train our network for 400 epochs using Adam optimizer [28], a learning rate of $10^{-3}$, mini-batches of 512 and gradient clipping. We use a Laplace loss [2] for the radial distance, binary cross entropy loss for stereo matching, and L1 loss for all other components. Losses are not weighted. KITTI dataset [5] does not provide pairwise matching information, thus, we extend the ground-truth by associating each person in the left image with the corresponding one in the right image. We include details in Appendix. We also perform horizontal flipping and switch left and right instances.

### 5.3 3D Metrics for Pedestrians

The majority of previous works for vision-based 3D object detection only reports results on the car category [12, 17]. We argue that KITTI official metrics, i.e., bird’s eye view and 3D average precision [5], are not be appropriate for pedestrians, as a pedestrian 3D bounding box has average width and length of 60 cm and 75 cm. Considering perfect orientation and dimensions, a distance error of 18 cm already leads to an intersection over union lower than 0.5. This requirement is unnecessarily strict and shift the attention of the community from the challenging instances to the easy ones, where obtaining results with a precision of few centimeters may still be possible. Furthermore, KITTI official metric assigns to each instance a difficulty regime based on bounding box height, level of occlusion and truncation: *easy, moderate and hard*. Each category includes instances from the simpler categories, and, due to the predominant number of easy instances (1240 “easy” pedestrians and 300 “hard” ones), the metric can underestimate the impact of challenging instances. To address the current limitations, we propose to consider a safety-critical area around a pedestrian. We aim to recognize a prediction as correct if the localization error between the predicted and ground-truth box is less than a threshold error. Differently from the metric proposed by Xiang et. al. [29], we...
Figure 5: Illustration of the long tail of height distribution in case of occlusion. In case A, children 2 and 3 are visible in both left and right images. The network associates each left instance with the right one and predicts a stereo confidence interval. In case B, we simulate an occlusion in the right image by removing the two instances at the image level. Performance drops as only monocular cues are available. However, due to our knowledge injection in the training data, the confidence is sufficiently large to include the children. For clarity, only instances closer than 12 meters that match a ground-truth are shown.

5.4 Results

Table 1 summarizes our 3D localization results with the ALE and RALP metrics. Our method outperforms every other stereo method in the ALE metric for Moderate, Hard and All instances. Solving jointly the ISM task and the 3D localization one is a crucial component, as shown by the three baselines. We make in-depth comparisons with the ALE as a function of the ground-truth distance in Figure 4a. On the ISM task, we obtain 98.2% accuracy.

Outliers. To go beyond “average-based metrics”, we analyze the entire distribution of predictions through the box plots in Figure 3. Our MonStereo is drastically more reliable for the long tail of predictions, especially when compared with other stereo methods. MonStereo’s maximum error is lower than 5 meters, while Pseudo-LiDAR [4] and 3DOP [15] have maximum errors of 24 and 17 meters respectively.

Long Tail-aware Confidence Intervals. The spread $b$ is the result of a probabilistic interpretation of the model. We introduced a distribution shift in the training data by including the long tail of the height distribution. In addition, in the training data the number of monocular pairs and stereo pairs is balanced by design. During validation, for the majority of instances a stereo match is present in the right image. As a consequence, the confidence intervals are calibrated for the training data distribution and they are more conservative for the validation data, where 86.0% of training instances lie inside the confidence interval. On the other side, the length of each side is only 3.9% of the predicted distance, making the confidence intervals useful for practical purposes.

Monocular and Stereo Limitations. We compare in Figure 4b the predicted aleatoric uncertainty $b$ with the 3D localization ambiguity for monocular and stereo modalities through the monocular task error and the stereo pixel error, respectively. As described in Section 3, the stereo ambiguity is very small at close distances but grows quadratically, while the monocular one grows linearly. Our MonStereo intrinsically learns to predict 3D location of instances combining stereo and monocular
Table 2: Impact of the ISM loss with mean and standard deviation of localization error. S simulates a standard stereo method by training a model only with stereo pairs; the network could learn monocular cues but is not guided by the ISM loss. S-x is as S, but without providing y-coordinates of input keypoints to remove information on human heights. M+S is trained with the same set of monocular and stereo pairs of Monstereo but without the ISM Loss. The long tail of far instances is the most impacted by the ISM loss.

| Method     | ALE (m) [σ (m)] | d < 10 | 10 < d < 20 | 20 < d < 30 | 30 < d < 50 |
|------------|-----------------|-------|-------------|-------------|-------------|
| S          | 0.24 [0.6]      | 0.47 [0.9] | 1.38 [1.4] | 3.95 [2.4] |
| S-x        | 0.52 [1.5]      | 0.61 [1.4] | 1.72 [1.5] | 5.50 [3.0] |
| M+S        | 0.25 [0.4]      | 0.50 [0.7] | 1.08 [1.2] | 2.24 [1.9] |
| MonStereo  | **0.20 [0.4]**  | **0.38 [0.7]** | **0.73 [1.0]** | **1.63 [1.8]** |

Table 3: Impact of knowledge injection (KI). We trained a monocular baseline M and a stereo one without KI. I. Recall measures the % of instances inside the intervals and I. Size is the interval ratio between the spread b and the ground-truth distance. KI increases both performances on challenging instances and recall. The intervals do not grow proportionally, as the spread b is reduced by better exploiting stereo cues.

|        | Easy       | Mod.       | Hard      | I. Recall | Easy       | Mod.       | Hard      | I. Size | Mod.       | Hard      |
|--------|------------|------------|-----------|-----------|------------|------------|-----------|---------|------------|-----------|
| M      | 0.77 0.82  1.35 | 58.0 58.9  32.5 | 4.6 5.0  4.9 |
| W/o KI | 0.51 0.68  0.87 | 76.2 72.7  42.0 | 4.1 4.5  4.4 |
| With KI | **0.29 0.41 0.50** | **91.2 81.9 65.4** | **3.8 4.1 4.1** |

5.5 Ablation Studies

The ensemble of monocular and stereo cues requires a delicate balance. The ISM loss prevents our method from overfitting to stereo disparity, while KI prevents it from overfitting to monocular cues.

ISM Loss. This loss encourages the use of monocular cues when stereo ones are not available or more convenient (e.g., faraway people where pixel disparity is not accurate enough). Without explicit guidance, the network over-relies on stereo cues, as shown in Table 2.

Knowledge Injection (KI). Without KI, the network over-relies on monocular cues. We illustrate it by training a monocular baseline, and a stereo baseline without KI. We analyze ALE metric, recall (% of instances inside the intervals) and relative size of the intervals in Table 3. KI improves results and calibrates the confidence intervals including the long tail of the height distribution. Recall increases, yet the interval size decreases, as KI helps to exploit stereo cues and reduce the spread b.

6 Conclusions

We have proposed a vision-based approach tailored for the long tail of 3D human localization. We have presented a neural network architecture that jointly matches detected body poses in left-right images and estimates the 3D localization of each pedestrian (regardless of whether there is a match). Our neural network implicitly learns to leverage monocular and/or stereo cues. Moreover, it also learns to communicate uncertainty driven by the cues. Our method goes beyond providing competitive results “on average” and shows reasonable estimates on challenging scenarios. We hope to direct the attention of the community towards the long tail for autonomous driving and social robot applications, still an unchartered research territory.
A Additional Results and Discussions

In this section, we show qualitative examples of challenging and critical cases from a safety point of view, and we shed light on the key challenges of our approach in these scenarios. Finally, we analyze the quality of 2D keypoints for disparity estimation using two off-the-shelf pose detectors and show that our method is agnostic to the choice of the detector.

Figure 6: A very close pedestrian who belongs to the moderate category (according to KITTI guidelines) due to the occlusion. Our MonStereo estimates accurate localization with an error of 2 cm despite the occlusion.

Figure 7: A pedestrian covered by a low wall belongs to the category hard. MonStereo performs instance-based stereo matching and 3D localization with an error of 19 cm.

A.1 Close Instances

KITTI categories are defined based on occlusions, truncations and bounding box heights. Hence, moderate and hard instances often correspond to very close but partially occluded pedestrians. These types of instances deserve a great deal of attention and, in Figure 6, we show an example of it. A bike rack partially covers the pedestrian’s legs in both images but MonStereo identifies a instance-based stereo matching and estimates depth with a localization error of 2 cm. Another example is shown in Figure 7, where the person is largely occluded by a low wall and belongs to the hard category. The person is 10 meters far and may soon cross the street. Early and accurate 3D localization is crucial for safety.

A.1.1 Challenges of Close Instances.

Humans occupy a 3D volume and estimating a single depth is not straightforward. Keypoints span over all the body and their disparities are not consistent for close instances. In Figure 8, we analyze the standard deviation of joints disparity as a function of the ground-truth distance using two pose detectors: the top-down Mask R-CNN [31] and the bottom-up PifPaf [9]. High variation of disparity for close instances can be concurrently caused by the 3D nature of humans and the quality of 2D keypoints. The two detectors lead to similar performances, highlighting the challenge of accurate 3D localization for close instances. Furthermore, the result shows that our method is agnostic to the choice of the pose detector.
Figure 8: Two sets of keypoint of each person in left-right images lead to 17 disparities. We analyze the standard deviation for keypoints obtained by two off-the-shelf pose detectors: Mask R-CNN [31] and PifPaf [9]. The resulting performances are similar for the two pose detectors, highlighting that our method is agnostic to the choice of the detector. For very close instances, the standard deviation of keypoints disparity is high as humans are 3D entities and every body joint may be located at a different depth.

Figure 9: Two far pedestrians heavily occluded by vehicles (hard category) are detected in both images and 3D localization is estimated with less than 5 cm error in both cases.

A.2 Far Instances

For far pedestrians, the standard deviation of keypoint disparities described in Figure 8 is greatly reduced, as a person only spans over few pixels. However, the depth error due to one pixel disparity grows quadratically with depth. In Figure 9, we show an example of two heavily occluded pedestrians 22 meters far, who are localized with only few centimeters of error. In Figure 10, we show a qualitative example of a “failure” as the instance 0 at 25 meters of distance is predicted with 69 cm of localization error. According to KITTI split, this instance belongs to the easy category. The performances for far instances are limited by the resolution of the camera: 69 cm of error corresponds to less than 0.5 pixel error in disparity estimation. On the contrary, all the people sitting at the cafe are closer and predicted with higher accuracy, but not evaluated by KITTI metrics as belonging to the category “person sitting”.

B Details on Ground-truth Generation

KITTI dataset [5] does not include stereo matching information for instances in left and right images, but only depth of left instances. Hence, we extend ground-truth information to train our network for the instance-based stereo matching task. We detect a set of keypoints for each person using the off-the-shelf pose detector PifPaf [9] and we compare the ground-truth depth with the one obtained through disparity estimation. For a given detection, the accuracy of every joint depends on many
Figure 10: A far pedestrian in the easy category is localized with a large error of 69 cm, while still included in the confidence interval. All the people sitting are localized with high accuracy, but not evaluated in KITTI metrics.

factors, such as occlusion. It is crucial to be able to detect and filter outlier joints that may affect our disparity calculation. Therefore, we adopt the following filters:

1. remove joints with confidence lower than a threshold;
2. remove outlier joints using Interquartile Range over the disparity estimation;
3. remove instances with large median vertical displacement.

Finally, we calculate the disparity as the median disparity of the remaining joints, we compare it with the ground-truth one and assign a binary label to the pair. We use an adaptive threshold which increases linearly with the depth, allowing for a larger error in case of far instances. This procedure does not involve the use of patches of images to analyze visual correspondences, being much faster and still very accurate. Our MonStereo reaches an accuracy of 98.2% on the validation set after being trained with binary cross-entropy loss.

References

[1] J. Ku, A. D. Pon, and S. L. Waslander. Monocular 3d object detection leveraging accurate proposals and shape reconstruction. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 11867–11876, 2019.

[2] L. Bertoni, S. Kreiss, and A. Alahi. Monoloco: Monocular 3d pedestrian localization and uncertainty estimation. In The IEEE International Conference on Computer Vision (ICCV), October 2019.

[3] W. Deng, L. Bertoni, S. Kreiss, and A. Alahi. Joint human pose estimation and stereo localization. In The International Conference on Robotics and Automation (ICRA), 2020.

[4] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger. Pseudolidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In The IEEE Conference on Computer Vision and Pattern Recognition, pages 8445–8453, 2019.

[5] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. International Journal of Robotics Research (IJRR), 2013.

[6] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas. Frustum pointnets for 3d object detection from rgb-d data. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 918–927, 2018.

[7] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander. Joint 3d proposal generation and object detection from view aggregation. In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 1–8. IEEE, 2018.

[8] D. Gallup, J.-M. Frahm, P. Mordohai, and M. Pollefeys. Variable baseline/resolution stereo. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1–8. IEEE, 2008.
[9] S. Kreiss, L. Bertoni, and A. Alahi. Pifpaf: Composite fields for human pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11977–11986, 2019.

[10] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka. 3d bounding box estimation using deep learning and geometry. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7074–7082, 2017.

[11] Z. Qin, J. Wang, and Y. Lu. Monogonet: A geometric reasoning network for monocular 3d object localization. In *the AAAI Conference on Artificial Intelligence*, volume 33, pages 8851–8858, 2019.

[12] T. Roddick, A. Kendall, and R. Cipolla. Orthographic feature transform for monocular 3d object detection. In *the British Machine Vision Conference (BMVC)*, 2019.

[13] B. Li, W. Ouyang, L. Sheng, X. Zeng, and X. Wang. Gs3d: An efficient 3d object detection framework for autonomous driving. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 1019–1028, 2019.

[14] A. Simonelli, S. R. Bulo, L. Porzi, M. López-Antequera, and P. Kontschieder. Disentangling monocular 3d object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1991–1999, 2019.

[15] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun. 3d object proposals for accurate object class detection. In *Advances in Neural Information Processing Systems*, pages 424–432, 2015.

[16] P. Li, T. Qin, et al. Stereo vision-based semantic 3d object and ego-motion tracking for autonomous driving. In *The European Conference on Computer Vision (ECCV)*, pages 646–661, 2018.

[17] P. Li, X. Chen, and S. Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7644–7652, 2019.

[18] Z. Qin, J. Wang, and Y. Lu. Triangulation learning network: from monocular to stereo 3d object detection. *arXiv preprint arXiv:1906.01193*, 2019.

[19] A. D. Pon, J. Ku, C. Li, and S. L. Waslander. Object-centric stereo matching for 3d object detection. In *The International Conference on Robotics and Automation (ICRA)*, 2020.

[20] R. Qian, D. Garg, Y. Wang, Y. You, S. Belongie, B. Hariharan, M. Campbell, K. Q. Weinberger, and W.-L. Chao. End-to-end pseudo-lidar for image-based 3d object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5881–5890, 2020.

[21] J. Freeman, T. Cole, S. Chinn, P. Jones, E. White, and M. Preece. Cross sectional stature and weight reference curves for the uk, 1990. *Archives of disease in childhood*, 73(1):17–24, 1995.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[23] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[24] T. van Dijk and G. C. de Croon. How do neural networks see depth in single images? *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[25] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun. Monocular 3d object detection for autonomous driving. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2147–2156, 2016.

[26] G. Adaimi, S. Kreiss, and A. Alahi. Rethinking person re-identification with confidence. *arXiv preprint arXiv:1906.04692*, 2019.
[27] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In The IEEE international conference on computer vision, pages 1116–1124, 2015.

[28] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[29] Y. Xiang, W. Choi, Y. Lin, and S. Savarese. Data-driven 3d voxel patterns for object category recognition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1903–1911, 2015.

[30] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems, pages 5574–5584, 2017.

[31] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick. Mask r-cnn. The IEEE International Conference on Computer Vision (ICCV), pages 2980–2988, 2017.