PÁZMÁNY PÉTER CATHOLIC UNIVERSITY

ROSKA TAMÁS DOCTORAL SCHOOL OF SCIENCES AND TECHNOLOGY

KOVÁCS Lóránt

3D Change Detection and Human Pose Estimation In Lidar Perception

Theses of PhD Dissertation

Thesis Supervisor:
Prof. Dr. BENEDEK Csaba DSc

Budapest, 2024

1 Introduction

In recent years, advancements in 3D perception technology have significantly enhanced the understanding of complex environments. This dissertation covers two research areas of 3D perception using terrestrial mobile laser scanners, specifically focusing on Lidar point clouds.

The first research topic is change detection in Lidar point clouds. Change detection is crucial for various applications, including urban planning, environmental monitoring, and infrastructure maintenance.

The second research topic is human pose estimation using only Lidar point clouds. Human pose estimation, which involves detecting and predicting the positions of various body parts, is traditionally performed using visual data. However, this research investigates the feasibility and advantages of utilizing Lidar data.

Together, these topics highlight the potential of Lidar technology in advancing 3D perception capabilities, paving the way for innovative applications and improved methodologies in various fields.

I introduce the two research topics in Sections 1.1 and 1.2, followed by the general introduction to the Lidar sensor (Section 2) and the description of the two types of Lidar sensors used for my research in Sections 2.1 and 2.2.

1.1 Change detection

Due to the increasing population density, and the rapid development of smart city applications and autonomous vehicle technologies, growing demand is emerging for automatic public infrastructure monitoring and surveillance applications. Detecting possibly dangerous situations caused by e.g., missing traffic signs, and damaged street furniture is crucial. Expensive and time-consuming efforts are required therefore by city management authorities to continuously analyze and compare multi-temporal recordings from large areas to find relevant environmental changes.

From the perspective of machine perception, this task can be formulated as a change detection problem. In video surveillance applications [16, 17], change detection is a standard approach for scene understanding by estimating the background regions and by comparing the incoming frames to this background model. Change detection is also a common task in many remote sensing applications, which require the extraction of the differences between aerial images, point clouds, or other

measurement modalities [18, 19]. However, the vast majority of existing approaches assume that the compared image or point cloud frames are precisely registered since either the sensors are motionless, or the accurate position and orientation parameters of the sensors are known at the time of each measurement.

1.2 Human pose estimation

The main task of pose estimation is to localize the anatomical keypoints of the human body. Human pose estimation is an essential task in machine perception and has several real-world applications, among others in robotics [20], security and surveillance [21, 22], and autonomous driving [23].

Human pose estimation is usually solved by camera-based methods [24–26] in the image space. However, such solutions are inherently limited by the camera's incapability to directly measure distance, the high sensitivity of the captured images to various lighting and weather conditions, and the varying visual appearances of real-world objects.

The consideration of additional depth information can increase the pose estimation robustness, as shown in [20], which uses an RGBD camera for 3D human pose estimation, outperforming camera-based 3D estimators and depth-only methods.

In applications, where privacy is a serious concern, Lidar-based human surveillance can be efficiently applied as the observed people cannot be identified by an observer in the sparse point cloud.

2 Lidar sensor

Lidar is an active sensor that illuminates the surroundings by emitting laser beams. Distances are measured precisely by processing the received laser reflections from the surfaces.

A general Lidar operates by scanning its Field of View (FoV) with one or several near-infrared (NIR) laser beams.

The laser beam is reflected to the scanner from the environment, the returned signal is received by a photodetector. Fast electronics filter the signal and measure the time difference between the transmitted and received signals, which is proportional to the reflecting object's distance. The range is estimated from the sensor model based on this calculated time difference. The Lidar outputs 3D point clouds that correspond to

the scanned environment and the intensities that correspond to the reflected laser energies [27]. The Lidar's maximum range is limited by the eye-safe transmission power regulations.

The scanning system of a Lidar sensor is responsible for the rapid exploration of the observed space. A few scanning methodologies at different Lidar types are introduced below. In the *mechanical spinning-type* sensors (rotating multi-beam (RMB) Lidar) the laser beams are steered through a rotating sensor head, having a moving mirror and optics inside. The Lidar I used for my change detection research (Thesis 1) works following this principle, described in detail in Section 2.1.

Another mechanical approach uses *rotation of prisms* to direct the laser beams. The Lidar I used for my Lidar-only human pose estimation research (Thesis 2) works following this scanning method, described in detail in Section 2.2.

The scanning can also be achieved by moving a "mirror" in a chip with elastic and electromagnetic forces in a *Micro-electromechanical system* (MEMS) [28].

Flash Lidars does not have any rotating component [29]. A single emitted laser beam is spread by an optical diffuser to illuminate the whole scene, and the reflections are detected on an array of photodiodes.

2.1 Velodyne HDL-64 rotating multi-beam Lidar sensor

The Velodyne HDL-64 sensor (shown in Figure 1a) is a high-resolution and high-performance RMB Lidar sensor, that is designed to help the real-time perception of autonomous robots and vehicles. It captures high-definition and real-time 3D measurements from its surrounding environment. The sensor has 64 laser beams, determining a 26.9° vertical FoV. Due to the rotating head of the sensor, its horizontal FoV is 360°. The measured data's spatial accuracy is 1-2 cm. Due to the sensor characteristics, the point density quickly decreases with the distance from the sensor.

The Velodyne HDL-64 is a pioneer of the RMB Lidars. Recent RMB Lidar sensors are available on the market (e.g., produced by Ouster) having similar characteristics, but their size and consumption have decreased significantly, making the measurements and the research conducted with the Velodyne Lidar in this research still relevant [30].

Ring patterns can be observed in the recorded point clouds, as can be seen in Figure 1b, as the laser beams are rotated along the vertical axis

(a) (b)

Figure 1. Velodyne HDL-64 rotating multi-beam Lidar sensor and its recorded point cloud in urban environment

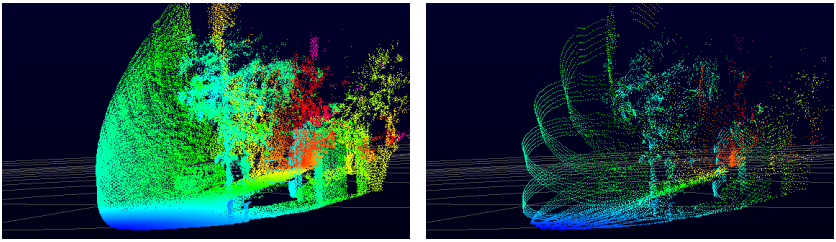
of the sensor. The sensor continuously streams the 3D measurements, which are collected to point cloud frames, where the term *frame* refers to a single horizontal turnaround of the sensor head.

2.2 Livox Avia Lidar sensor with Non-repetitive Circular Scanning

The Livox Avia sensor [31], shown in Figure 2, is a lightweight Lidar sensor that has a unique, Non-repetitive Circular Scanning (NRCS) technique. The sensor has six Lidar beams organized in a linear beam array, which is moved and rotated inside the sensor to scan its FoV (horizontal: 70°, vertical: 77°).

Figure 2. Livox Avia Lidar sensor

Unlike most RMB Lidars, which boost a repetitive scanning pattern, the Avia does not repeat the exact scanning paths in every frame, but instead, the lasers cover new parts of the FoV. This key difference is both beneficial and implicates some disadvantages. NRCS Lidars cover the complete FoV over time, providing rich spatial



(a) Point cloud recorded with 1s integration time

(b) Point cloud recorded with 100 ms integration time

Figure 3. Livox Avia point clouds, colored by the points' distance

information, especially in static scenarios. On the other hand, because the same region is scanned less frequently than by using "regular" RMB Lidars, dynamic objects, such as humans may cause challenges as they induce heavy motion blur in the recorded NRCS point clouds.

The sensor continuously records distance measurements with corresponding timestamps following its non-repetitive circular pattern in its FoV. By setting a fixed integration time, the consecutively collected points can be grouped into separate Lidar time frames. The main challenge is to efficiently balance between the spatial and the temporal resolution of the recorded range data.

While allowing larger integration time, the laser beams cover a higher proportion of the FoV yielding high spatial measurement resolution of the measurement frame, the object movements of dynamic objects in the observation area induce various artifacts (e.g., blurred pedestrian silhouettes), which do not allow efficient dynamic event analysis. For example, the Livox Avia sensor collects 240000 points within a time window of $1s$, as can be seen in Figure 3a. On the other hand, if the measurements are collected in a narrow time window (e.g., in 100 ms) the resulting point clouds are very sparse, which phenomenon yields a loss of details across the spatial dimension of the FoV:a sample frame of 24000 points is shown in Figure 3b.

3 New scientific results

1. Thesis: I proposed a novel change detection approach for coarsely registered RMB Lidar point clouds in complex, street-level urban environments. The input point clouds are represented by range images, the result of the method is a pair of binary masks showing the change regions on each input range image, which can be back-projected to the input point clouds without loss of information. I have evaluated the proposed method in various challenging scenarios, and I have shown its superiority against state-of-the-art change detection methods.

The method was published in a journal [1] and a submitted patent application [3]. In the initial phase of this research, a method was published in a conference paper [5] for multi-object detection in urban scenes utilizing 3D background maps and tracking. It uses a dense 3D city map to increase the accuracy of object detection on a sparse point cloud from a Lidar sensor. This method can extend the camera-based machine perception of a road vehicle, described in [6]. For the evaluation of the results considering the object trajectories, a track-to-track evaluation method can be used [7].

The need to solve the point-based detection of changed regions due to object displacements between initially unmatched (coarsely registered) pairs of point clouds can be emphasized with practical cases, where reliable registration and therefore the change detection cannot be achieved with currently available methods. I have identified a new way of posing a problem: I described the differences among a coarsely registered pair of point clouds without exactly matching the available input point cloud measurements.

As a key feature, the proposed method does not require precise registration of the point cloud pairs. Based on my experiments, the proposed method is more efficient than existing solutions, and it can efficiently handle up to 1 m translation and 10° rotation misalignment between the corresponding 3D point cloud frames. Figure 4 shows the input point clouds recorded with a Velodyne HDL-64 rotating multi-beam Lidar and the results of the proposed method.

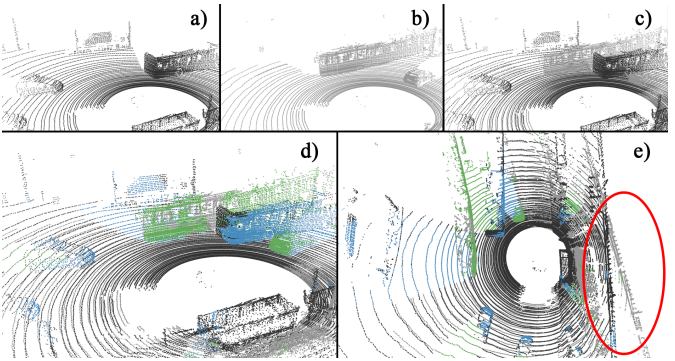


Figure 4. Changes detected by *ChangeGAN* for a coarsely registered point cloud pair. (a) and (b) show the two input point clouds, (c) displays the coarsely registered input point clouds in a common coordinate system. (d), (e) present the change detection results: blue and green colored points represent the objects marked as changes in the first- and second point cloud, respectively. The red ellipse draws attention to the global alignment difference between the two coarsely registered point clouds.

1.1. Subthesis: I have defined a deep neural network structure, capable of learning and robustly extracting changes between coarsely registered 3D sparse point clouds obtained in a complex street-level environment. For the training of this neural network I proposed a semi-automatic method to create a change detection dataset with coarsely registered point cloud pairs using simulated registration errors.

The proposed deep learning approach takes as input two coarsely registered 3D point clouds recorded with an RMB Lidar sensor $\mathcal{P}_1$ and $\mathcal{P}_2$ represented by range images I_1 and I_2, respectively (shown in Figures 6a and 6b). The proposed architecture assumes that the images I_1 and I_2 are defined over the same pixel lattice and have the same spatial dimensions.

For this purpose, I propose a new generative adversarial neural network-like architecture, more specifically a discriminative method, with an additional adversarial discriminator as a regularizer, called *ChangeGAN*, which is shown in Figure 5.

Since the main goal is to find meaningful correspondences between the input range images I_1 and I_2, I have adopted a Siamese style [32] architecture to extract relevant features from the input range image pairs.

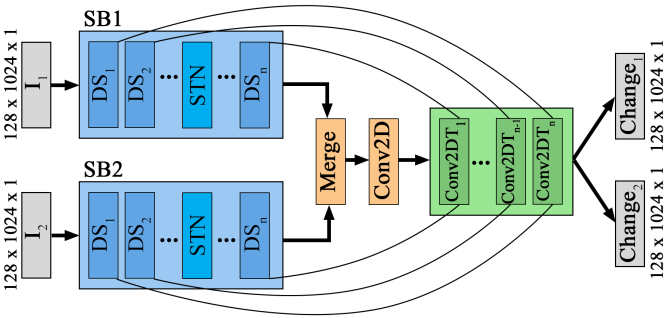


Figure 5. Proposed *ChangeGAN* architecture. Notations of components: SB1, SB2: Siamese branches, DS: downsampling, STN: spatial transformer network, Conv2DT: transposed 2D convolution

The Siamese architecture is designed to share the weight parameters across multiple branches, allowing us to extract similar features from the inputs and to decrease the memory usage and training time. Each branch of the Siamese network consists of fully convolutional downsampling blocks.

The second part of the proposed model contains a series of transposed convolutional layers to upsample the signal from the lower-dimensional feature space to the original size of the 2D input images. Finally, a 1×1 convolutional layer, activated with a sigmoid function, generates the two binary change maps Λ_1 and Λ_2.

To regularize the network and prevent over-fitting, I use the dropout technique after the first two transposed convolutional layers. To improve the change detection result, I have adapted an idea from U-net [33] by adding higher resolution features from the downsampling blocks to the corresponding transposed convolutional layers.

In this case, as the point clouds are coarsely registered, the same regions of the input range images might not be correlated with each other. To achieve more accurate feature matching, I have added Spatial Transformer Network blocks [34] for both Siamese branches. STN can learn an optimal affine transformation between the input feature maps to reduce the spatial registration error between the input range images. Furthermore, STN dynamically transforms the inputs, also yielding an advantageous augmentation effect.

For the training of the *ChangeGAN* neural network I have created a new Lidar-based urban dataset called *Change3D*.The measurements

were recorded over two days in downtown Budapest using a Velodyne HDL-64 RMB Lidar mounted on a car.

As the main purpose of the proposed *ChangeGAN* method is to extract changes from coarsely registered point clouds, for model training and evaluation a large and annotated set of point cloud pairs collected in the same area with various spatial offsets and rotation differences is needed. The annotation accurately marks the point cloud regions of objects or scene segments that appear only in the first frame, only in the second frame, or which ones are unchanged, thus observable in both frames.

The manual annotation of point cloud differences is very challenging, even if the point clouds originate from the same coordinate system. To ensure the accuracy of the ground truth, I performed the change labeling for registered point cloud pairs captured from the same sensor position and orientation, then I randomly transformed the reference positions and orientations of the second frames yielding a large set of accurately labeled coarsely registered point cloud pairs.

The steps of our proposed dataset generation process are as follows. First, pairs of Lidar frames are taken in the same global coordinate system, thus they can be considered as *registered*.

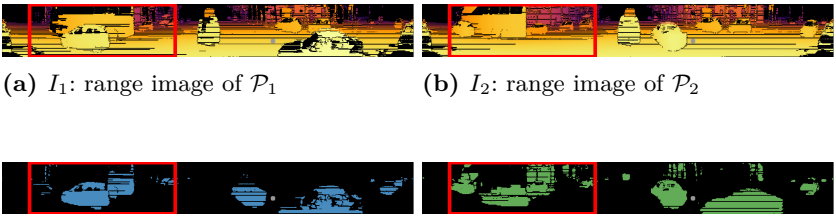
To simulate coarsely registered point cloud pairs, I applied randomly an up to ± 1 m translation and an up to $\pm 10°$ rotation transform around the z-axis for the second frame ($\mathcal{P}_2$) of each point cloud pair both in the training and test datasets. The performance of the method was evaluated on sub-datasets with defined rotation and translation offsets. The ground truth labels remained attached to the $p \in \mathcal{P}_2$ points and were transformed together with them, as shown in Figures 6c and 6d.

In the next step spatial cropping was applied, only the points below 5 m and closer than 40 m were kept. In the remaining point cloud, the distances were normalized to the $[0 - 1]$ range.

The transformed 3D point clouds were projected to 2D range images I_1, and I_2, as shown in Figures 6a and 6b. The Lidar's horizontal 360° FoV was mapped to 1024 pixels and the 5 m vertical height of the cropped point cloud was mapped to 128 pixels, yielding that the size of the produced range image is 1024×128.

The training database contains 20000 point cloud pairs from 50 locations, while the test set was composed of 2000 point cloud pairs from completely different measurement locations.

In summary, I have created a new dataset suitable for training and evaluating new change detection methods where accurate registration of



(a) I_1: range image of $\mathcal{P}_1$

(b) I_2: range image of $\mathcal{P}_2$

(c) Λ_1: ground truth mask of changes in range image I_1

(d) Λ_2: ground truth mask of changes in range image I_2

Figure 6. Input data representation. (a), (b): range images I_1, I_2 from a pair of coarsely registered point clouds ($\mathcal{P}_1$, $\mathcal{P}_2$). (c), (d): binary ground truth change masks Λ_1, Λ_2 for the range images I_1 and I_2, respectively.

the compared point clouds is not required.

1.2. Subthesis: I have proposed a novel, competitive classifier - discriminator-based adversarial training method for the change detection task on a coarsely registered pair of 3D point clouds.

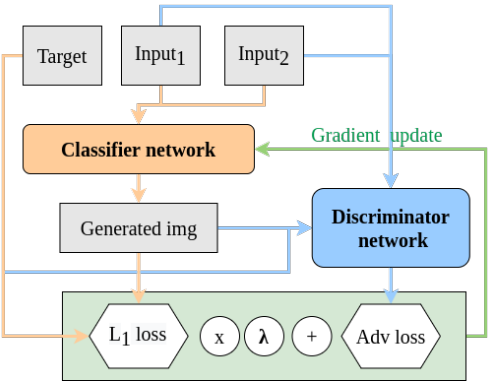


Figure 7. Proposed adversarial training strategy of the *ChangeGAN* architecture.

The *classifier* network is responsible for learning and predicting the changes between the range image pairs. In each training epoch, the clas-

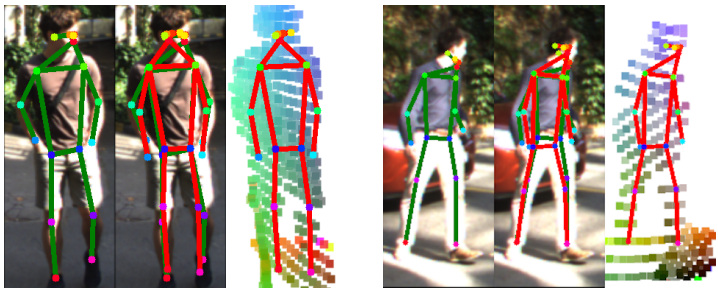
sifier model is trained on a batch of data. The actual state of the classifier is used to predict validation data, which is fed to the discriminator model.

The *discriminator* network is a fully convolutional network that classifies the output of the classifier network. The discriminator model divides the image into patches and decides for each patch whether the predicted change region is real or fake. During training, the discriminator network forces the classifier model to create better and better change predictions, until the discriminator cannot decide about the genuineness of the prediction.

Figure 7 demonstrates the proposed adversarial training strategy. I calculate the L1 Loss (L_{L1}) as the mean absolute error between the generated image and the target image, and I define the Adversarial Loss (L_{Adv}), which is a sigmoid cross-entropy loss of the feature map generated by the discriminator and an array of ones. The final loss function of the method (L) is the weighted combination of the Adversarial Loss and the L1 Loss: $L = L_{\text{Adv}} + \lambda * L_{L1}$.

2. Thesis: I proposed a novel, end-to-end method for real-time foreground-background segmentation and human pose estimation, solely based on point cloud measurements of a Non-repetitive Circular Scanning Lidar sensor.

The proposed method is based on the ViTPose architecture [35], which is a transformer-based [36, 37] human pose estimation method using optical camera images.



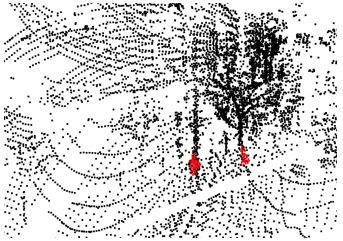
(a) **(b)**

Figure 8. *LidPose–2D* predictions are shown in red, overlaid on the input Lidar point cloud (*right*). The ground truth is shown in green, drawn over the corresponding camera frame (*left*). The prediction in red and the ground truth in green are shown together in the input Lidar point cloud (*middle*).

I introduced a modified ViTPose approach, which is adapted to the 3D point clouds, and it can efficiently handle the sparsity and the unusual rosetta-like scanning pattern of the NRCS Lidars. The proposed method's [2] first step utilizes a foreground-background segmentation technique [8] for the NRCS Lidar sensor to select foreground points. In the next step, the *LidPose* human pose estimator network estimates the human pose in the filtered NRCS Lidar point cloud segments. The proposed method is a complete and end-to-end approach to human pose estimation from raw NRCS Lidar measurement sequences, captured by a static sensor for surveillance scenarios. To evaluate the method, I have created a novel, real-world, multi-modal dataset, containing camera images and Lidar point clouds from a Livox Avia sensor with annotated 2D and 3D human skeleton ground truth.

The method was published in a journal [2] and a conference paper [8]. Figure 8 shows the predictions of the proposed *LidPose* method in 2D.

2.1. Subthesis: I proposed a point-level foreground-background segmentation technique for NRCS Lidar point cloud sequences recorded in a static sensor configuration. I proved that the proposed method can handle the sparsity of the NRCS Lidar measurements in a surveillance scenario. I created a database for the testing and evaluation of the proposed approach and demonstrated its efficiency [8].



(a) Detected foreground (red) in a single time frame of the NRCS Lidar point cloud

(b) Detected foreground regions (red) displayed with the high-resolution background model point cloud

Figure 9. Foreground detection results (by red) in the *City Center* scene, displayed in 3D point cloud representation.

To solve the point-wise foreground-background segmentation task, it is required to efficiently balance between the spatial and the temporal resolution of the recorded NRCS Lidar data, shown in Figure 3. For this reason, I create and maintain a very high-resolution background model of the sensor's FoV using a Mixture of Gaussians-based method [8], displayed in Figure 9b. On the other hand, to enable real-time analysis of dynamic objects, I use low integration time to extract the consecutive Lidar frames. As a result, the laser reflections from foreground objects reflect sparse, but geometrically accurate samples of the silhouettes (shown in Figure 9a) providing valuable input for higher-level shape description, object detection, and pose estimation, as described in Subthesis 2.3. I demonstrated the efficiency of the new approach in different realistic NRCS Lidar measurement sequences.

2.2. Subthesis: I have proposed a semi-automatic method to create a human pose dataset with camera images and NRCS Lidar measurements.

Ground truth annotation of Lidar point clouds is a challenging process, since the visual interpretation of sparse 3D Lidar point clouds is difficult for human observers, and the inhomogeneous NRCS pattern makes this task even harder. In the experiments, a camera was mounted near the NRCS Lidar sensor to record optical images as well, besides the point clouds, as shown in Figure 10. The camera images were only used for creating the ground truth information for human pose estimation, and for helping the visual evaluation of the results of *LidPose*. During annotation, the operator used the camera images to mark, validate, and verify the skeleton joint positions.

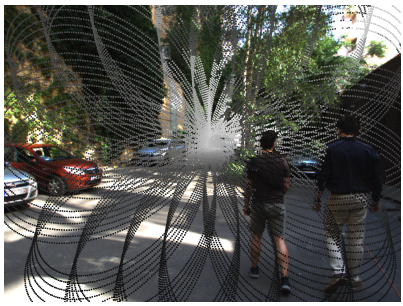


Figure 10. NRCS Lidar point cloud with 100 ms integration time represented as a 2D range image overlaid on a sample camera image.

Since the experimental configuration uses both camera and Lidar data for creating the ground truth human poses and validating the results, the spatial transformation parameters between the two sensors' coordinate systems need to be determined by a calibration process. The camera's extrinsic and intrinsic parameters were calibrated using OpenCV libraries and a Livox-specific, targetless calibration method [38]. Thereafter, the camera images and the Lidar range images were transformed into a common coordinate system. The camera and the Lidar data were properly timestamped using the *Precision Time Protocol daemon (PTPd)* [39].

The camera and the Lidar data were recorded with different, sensor-specific data acquisition rates, at 30 Hz on the camera and at 10 Hz in the case of the Lidar. The data collection speed was adjusted to the Lidar's slower frame rate.

Ground truth generation has been implemented in a semi-automatic way, exploiting established camera-based person detection and pose-fitting techniques.

1. In each data sample, the YOLOv8 [40] was run to detect the per-

sons in the camera images.
2. The initial pose estimation was created on the cropped camera images by the state-of-the-art 2D human pose estimator ViTPose [35] network.
3. The camera images were used to manually check, validate, filter, and fine-tune each 2D human pose, resulting in the 2D ground truth of human poses.
4. The filtered camera-based human pose model was directly used as the ground truth of the 2D human poses in the co-registered Lidar's range image domain.
5. The 3D human pose ground truth is created by the extension of the 2D human skeleton dataset, so I attempted to assign to each joint a depth value, based on the depth measurements of the Lidar sensor around the joint's 2D position.
6. Spatio-temporal interpolation was applied on joints without direct range measurements from the depth values of other nearby joints, and nearby frames.

In total, the created new dataset contains 9500 skeletons, and 161000 joints. The dataset was split into independent training, validation, and test sets, having 5500, 500, and 3400 skeletons.

In summary, I have created a new dataset suitable for training and evaluating a new human pose estimation method that uses only the NRCS Lidar point cloud as an input. To prove the usability of the dataset, I have proposed a vision transformer-based neural network to perform human pose estimation, the details of which are described in Subthesis 2.3.

2.3. Subthesis: I proposed a novel, visual transformer-based method for real-time human pose estimation from inhomogeneous and sparse Lidar point clouds recorded with an NRCS Lidar sensor.

The published *LidPose* method [2] solves the human pose estimation task using only NRCS Lidar measurements, in a surveillance scenario, where the sensor is mounted in a fixed position. The *LidPose* method's workflow is shown in Figure 11.

The proposed method is based on the state-of-the-art ViTPose [35] human pose estimation method working on camera images, based on a Vision Transformer (ViT) architecture [37], which was trained on the COCO dataset [41].

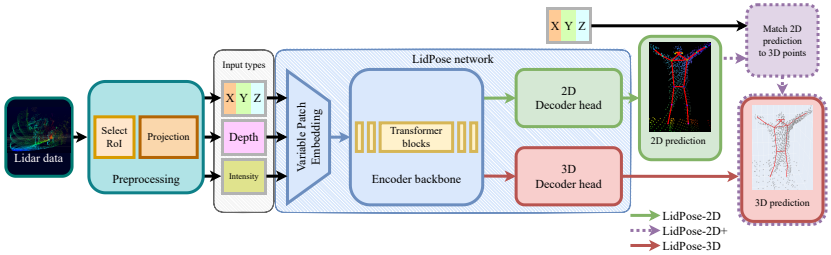


Figure 11. *LidPose* end-to-end solution:
Lidar data: full Lidar point cloud. Select ROI: selects the 3D points in the vicinity of the observed human. Projection stores the 3D point cloud in a 2D array. Input types: 3D XYZ coordinates (`XYZ`), Depth (`D`) and Intensity (`I`). *LidPose* network: Both *LidPose–2D* and *LidPose–3D* use our patch embedding module and the encoder backbone, visible in blue. *LidPose–2D* and *LidPose–3D* use the corresponding Decoder head and *LidPose–2D+* is calculated from the 2D prediction and the input point cloud.

First, the moving objects are separated from the static scene regions in the NRCS Lidar point clouds, as described in Subthesis 2.1 and [8].

Second, the foreground point regions are segmented to separate individual moving objects, and the footprint positions of the detected pedestrian candidates are estimated. The result of this step is a set of bounding boxes for the detected people, which can be represented both in the 3D space and in the 2D range image domain.

In the next step, the NRCS Lidar point cloud and the range image are cropped with the determined bounding boxes.

To jointly represent the different available measurement modalities, I proposed a new 2D data structure that can be derived from the raw Lidar measurements straightforwardly and can be efficiently used to train and test our proposed *LidPose* model. I construct from the input point cloud a five-channel image over the Lidar sensor's 2D range image lattice, where two channels directly contain the depth and intensity values of the Lidar measurements, while the remaining three layers represent the `X,Y,Z` coordinates of the associated Lidar points in the 3D world coordinate system.

The ViTPose [35] network structure was used as a starting point in the research and development of the proposed *LidPose* methods' pose estimation networks. My main contributions to the proposed *LidPose* method:

- A new patch embedding implementation was applied to the network backbone to handle efficiently and dynamically the different input channel counts.
- The number of transformer blocks used in the *LidPose* backbone is increased to enhance the network's generalization capabilities by having more parameters.
- The output of the LidPose–3D configuration has been modified as well by extending the predictions' dimension to be able to predict the joint depths alongside the 2D predictions.

As Figure 11 demonstrates, the *LidPose* network structure can deal with different input and output configurations, depending on the considered channels of the above-defined five-layer image structure. The optimal channel configuration is a hyperparameter of the method, that was selected upon experimental evaluation, described in [2].

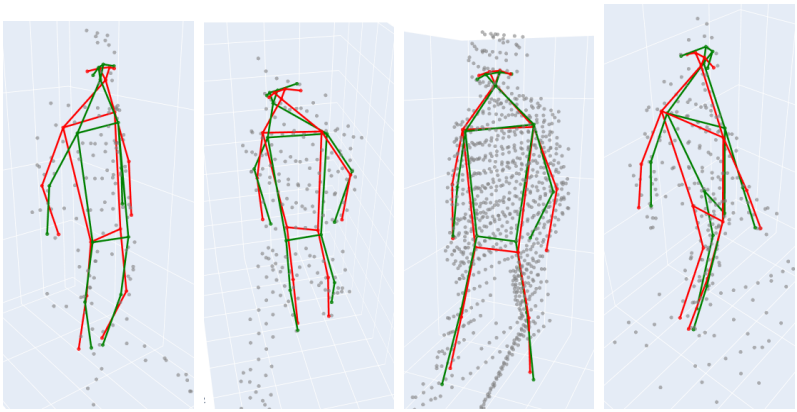
For the *LidPose–3D* network, the training loss is composed of two components: the Mean Squared Error responsible for the joints' 2D prediction accuracy (L_{joint2D}), and the other component reflecting the depth estimation accuracy (L_{depth}). The total training loss is a weighted sum of the position and depth losses:

$$L_{\text{LidPose–3D}} = W_{\text{joint2D}} \cdot L_{\text{joint2D}} + W_{\text{depth}} \cdot L_{\text{depth}} \tag{1}$$

Regarding the depth loss (L_{depth}), I tested three different formulas: *L1 loss*, *L2 loss* and *Structural Similarity Index Measure (SSIM)* [42]. Based on the evaluations and considering training runtime, the *SSIM* was selected for the depth loss measure in the proposed *LidPose–3D* network.

The quantitative evaluation of the method was done by calculating multiple metrics. The best model achieved a 0.694 score with the Area Under the Percentage of Correct Keypoints Curve. The Average Distance Error of each predicted skeleton was calculated as well, where the best model achieved a 0.158 m. For qualitative evaluation, the 3D human poses predicted with the proposed method are shown in Figure 12. The obtained results confirm, that the proposed method can detect human skeletons in sparse and inhomogeneous NRCS Lidar point clouds.

The approach gives accurate human pose estimation results in real-time in the 3D world coordinate system of the scene, which can be used in higher-level scene analysis steps of surveillance systems.



(a) (b) (c) (d)

Figure 12. *LidPose3D* predicted skeletons. Red skeleton: 3D prediction. Green skeleton: ground truth. Gray points: NRCS Lidar points.

4 Application and dissemination of the results

4.1 ChangeGAN

The proposed *ChangeGAN* [1], [3] can robustly extract changes between sparse point clouds obtained in a complex street-level environment. As a key feature, the proposed method does not require precise registration of the point cloud pairs. Based on my experiments, it can efficiently handle up to 1 m translation and 10° rotation misalignment between the corresponding 3D point cloud frames. This makes the proposed method suitable for real-world applications, where the precise registration of the point clouds is not feasible due to the complexity of the environment or the limitations of the sensors. The method can be applied in automatic public infrastructure monitoring, where detecting possibly dangerous situations caused by e.g., missing traffic signs, and damaged street furniture is crucial. The method can be used for the efficient update of high-resolution 3D maps for autonomous vehicles. Expensive and time-consuming efforts can be reduced in city management offices by applying this method to automatically and continuously analyze and compare multi-temporal recordings from large areas to find relevant environmental changes.

4.2 LidPose

In the *LidPose* paper [2] I gave evidence, that the Livox Avia [31] NRCS Lidar can be widely adopted in real-life scenarios due to its low price, can be used for solving complex human pose estimation tasks, while the process highly respects the observed people's privacy as the people are barely recognizable by human observers from the recorded sparse point clouds.

The change detection accuracy can be increased by applying a novel depth image completion technique, which eliminates the uneven sparseness of the NRCS Lidar data, as described in a submitted patent application [4].

4.3 Publications and dissemination

The research results were published mainly in prestigious journals and conferences, as cited in the theses. [1, 2] [3, 4] [5–8]

On top of those I presented my research progress at the biannual *Conference of the Hungarian Association for Image Analysis and Pattern Recognition (KÉPAF)* [9–11] and in the *PhD proceedings, annual issues of the Doctoral School, Faculty of Information Technology and Bionics* [12–15].

I demonstrated my results among others at the *Researcher's Night*[1], and at various events organized by the *Artificial Intelligence National Laboratory (MILAB)* and *National Lab for Autonomous Systems (ARNL)*, including the AI & Aut Expo 2023[2].

5 Acknowledgements

This thesis could not have been done without the support of people around me. I would like to express my deepest gratitude and appreciation to my PhD supervisor, **Csaba Benedek**. His insightful guidance and feedback have challenged me to think critically, push the boundaries of knowledge, and strive for excellence in my work.

I express my sincere gratitude to the present and past leaders of *Roska Tamás Doctoral School of Sciences and Technology*, **Gábor Szederkényi** and **Péter Szolgay**. And also, to the deans of the *Faculty*

[1] https://sztaki.hun-ren.hu/kutatok-ejszakaja-2022#xr

[2] https://www.facebook.com/photo/?fbid=877979741876158 2&set=pcb.8780186178722706

of Information Technology and Bionics at Pázmány Péter Catholic University (PPKE), **György Cserey** and **Kristóf Iván**. And I must not forget **Mrs. Vida Tivadarné Katinka néni**, who was helpful and patient with all my administrative issues and challenges.
Furthermore, I express my deep appreciation to **Tamás Szirányi** and the *Machine Perception Laboratory (MPLAB)* at the *Institute for Computer Science and Control (HUN-REN SZTAKI)* for providing me with all the research equipment, the sensors, the computers, the servers, and the stimulating research environment.
I extend my heartfelt appreciation to my co-authors **Balázs Nagy, Balázs Bódis, Marcell Kégl, Örkény H. Zováthi** for their valuable contributions to my publications. I am grateful to my colleagues and friends for their unwavering support and invaluable contributions throughout my journey.
Special thanks to those whom I may have not mentioned here by name but who supported me directly or indirectly in accomplishing my research.
For financial support, thanks to the European Union within the framework of the National Laboratory for Autonomous Systems (RRF-2.3.1-21-2022-00002) and of the Artificial Intelligence National Laboratory (RRF-2.3.1-21-2022-00004) programs. Further support was provided by the TKP2021-NVA-27 and TKP2021-NVA-01 grants and by the OTKA #143274 project of the Hungarian National Research, Development and Innovation NRDI Office.
I would like to thank my parents, my grandfather, and my siblings for their unconditional support and for everything they did for me not only during my PhD but also until then.
Last but not least, I would like to thank my family, and most importantly my beloved *Klári*, who provided me with a solid background, and a home where I could always refresh and rest. You always supported me, regardless of the obstacles, and understood that I really wanted to achieve this. Finally, I am thankful to my daughter *Bíbor*, and to my sons *Özséb* and *Donát*, who accepted that in the tough periods, I worked more than they wanted. Their love and understanding were always there, and they were the ones who made me smile and laugh even in the darkest moments.

6 Bibliography

Journal publications of the thesis

[1] B. Nagy, **L. Kovács**, and C. Benedek, "ChangeGAN: A deep network for change detection in coarsely registered point clouds," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8277–8284, 2021, **IF: 4.3, Scimago Q1/D1**. (Cited on pages 7, 19, and 20.)

[2] **L. Kovács**, B. M. Bódis, and C. Benedek, "LidPose: Real-time 3d human pose estimation in sparse lidar point clouds with non-repetitive circular scanning pattern," *Sensors*, vol. 24, no. 11, 2024, **IF: 3.9, Scimago Q1**. (Cited on pages 13, 16, 18, and 20.)

Patents related to the thesis

[3] B. Nagy, **L. Kovács**, C. Benedek, T. Szirányi, Ö. Zováthi, and L. Tizedes, "Training method for training a change detection system, training set generating method therefor, and change detection system," WO Patent application, WO/2023/007198, International Filing Date: 08.07.2022, Priority data: P2100280, 27.07.2021, HU. (Cited on pages 7, 19, and 20.)

[4] Ö. Zováthi, Z. Rózsa, B. Pálffy, Z. Jankó, C. Benedek, T. Szirányi, **L. Kovács**, and M. Kégl, "Methods for spatial and temporal densification of Lidar measurements," Patent application, Priority data: P2300075, 01.03.2023, HU. (Cited on page 20.)

Conference publications of the thesis

[5] Ö. Zováthi, **L. Kovács**, B. Nagy, and C. Benedek, "Multi-object detection in urban scenes utilizing 3d background maps and tracking," in *2019 International Conference on Control, Artificial Intelligence, Robotics and Optimization (ICCAIRO)*, 2019, pp. 231–236. (Cited on pages 7 and 20.)

[6] A. Horvath, I. Horvath, A. Kiss, D. Huszar, A. Palffy, **L. Kovács**, D. Babicz, B. Farkas, G. Majoros, and C. Rekeczky, "Cellular vision based adas applications," in *CNNA 2016; 15th International Workshop on Cellular Nanoscale Networks and their Applications*, 2016. (Cited on pages 7 and 20.)

[7] **L. Kovács**, L. Lindenmaier, H. Nemeth, V. Tihanyi, and A. Zarandy, "Performance evaluation of a track to track sensor fusion algo-

rithm," in *CNNA 2018; The 16th International Workshop on Cellular Nanoscale Networks and their Applications*, 2018. (Cited on pages 7 and 20.)

[8] **L. Kovács**, M. Kégl, and C. Benedek, "Real-time foreground segmentation for surveillance applications in nrcs lidar sequences," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLIII-B1-2022, pp. 45–51, 2022. (Cited on pages 13, 14, 17, and 20.)

Other publications of the author

[9] **L. Kovács**, B. M. Bódis, and C. Benedek, "LidPose: Real-time 3d human pose estimation in sparse lidar point clouds with non-repetitive circular scanning pattern," in *15th Conference of the Hungarian Association for Image Analysis and Pattern Recognition*, Hévíz, 2025. (Cited on page 20.)

[10] **L. Kovács**, M. Kégl, and C. Benedek, "Real-time foreground segmentation for surveillance applications in sequences from a non-repetitive circular scanning lidar," in *14th Conference of the Hungarian Association for Image Analysis and Pattern Recognition*, Gyula, 2023. (Cited on page 20.)

[11] **L. Kovács**, B. Nagy, and C. Benedek, "Demonstration of changegan: change detection in unregistered point clouds using neural networks," in *13th Conference of the Hungarian Association for Image Analysis and Pattern Recognition*, Budapest, 2021. (Cited on page 20.)

[12] **L. Kovács**, "Change detection in unregistered 3d point clouds," in *PhD proceedings, annual issues of the Doctoral School, Faculty of Information Technology and Bionics*, vol. 16, 2021, p. 67. (Cited on page 20.)

[13] **L. Kovács**, "Change detection in lidar point clouds," in *PhD proceedings, annual issues of the Doctoral School, Faculty of Information Technology and Bionics*, vol. 15, 2020, p. 68. (Cited on page 20.)

[14] **L. Kovács**, "Challenges in track to track sensor fusion using neural networks," in *PhD proceedings, annual issues of the Doctoral School, Faculty of Information Technology and Bionics*, vol. 14, 2019, p. 60. (Cited on page 20.)

[15] **L. Kovács**, "Challenges in sensor fusion," in *PhD proceedings, annual issues of the Doctoral School, Faculty of Information Technology and Bionics*, vol. 13, 2018, p. 55. (Cited on page 20.)

References

[16] C. Benedek, B. Gálai, B. Nagy, and Z. Jankó, "Lidar-based gait analysis and activity recognition in a 4d surveillance system," *IEEE Trans. Circuits Syst. Video Techn.*, vol. 28, no. 1, pp. 101–113, 2018. (Cited on page 2.)

[17] F. Oberti, L. Marcenaro, and C. S. Regazzoni, "Real-time change detection methods for video-surveillance systems with mobile camera," in *European Signal Processing Conference*, 2002, pp. 1–4. (Cited on page 2.)

[18] C. Benedek, X. Descombes, and J. Zerubia, "Building development monitoring in multitemporal remotely sensed image pairs with stochastic birth-death dynamics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 33–50, 2012. (Cited on page 3.)

[19] S. Ji, Y. Shen, M. Lu, and Y. Zhang, "Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples," *Remote Sensing*, vol. 11, no. 11, 2019. (Cited on page 3.)

[20] C. Zimmermann, T. Welschehold, C. Dornhege, W. Burgard, and T. Brox, "3d human pose estimation in rgbd images for robotic task learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 1986–1992. (Cited on page 3.)

[21] M. Cormier, A. Clepe, A. Specker, and J. Beyerer, "Where are we with human pose estimation in real-world surveillance?" in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2022, pp. 591–601. (Cited on page 3.)

[22] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 34, no. 3, pp. 334–352, 2004. (Cited on page 3.)

[23] A. Zanfir, M. Zanfir, A. Gorban, J. Ji, Y. Zhou, D. Anguelov, and C. Sminchisescu, "Hum3dil: Semi-supervised multi-modal 3d human

pose estimation for autonomous driving," in *Proceedings of The 6th Conference on Robot Learning*, vol. 205, 2022, pp. 1114–1124. (Cited on page 3.)

[24] Z. Cao, G. Hidalgo, T. Simon, S. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 01, pp. 172–186, 2021. (Cited on page 3.)

[25] H.-S. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y.-L. Li, and C. Lu, "Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7157–7173, 2023. (Cited on page 3.)

[26] P. Lu, T. Jiang, Y. Li, X. Li, K. Chen, and W. Yang, "RTMO: Towards high-performance one-stage real-time multi-person pose estimation," 2023. (Cited on page 3.)

[27] Y. Li and J. Ibanez-Guzman, "Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems," *IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 50–61, 2020. (Cited on page 4.)

[28] H. W. Yoo, N. Druml, D. Brunner, C. Schwarzl, T. Thurner, M. Hennecke, and G. Schitter, "Mems-based lidar for autonomous driving," *e & i Elektrotechnik und Informationstechnik*, vol. 135, no. 6, pp. 408–415, Oct 2018. (Cited on page 4.)

[29] F. Amzajerdian, V. E. Roback, A. Bulyshev, P. F. Brewster, and G. D. Hines, "Imaging flash lidar for autonomous safe landing and spacecraft proximity operation," 2016. (Cited on page 4.)

[30] A. Palffy, E. Pool, S. Baratam, J. Kooij, and D. Gavrila, "Multi-class road user detection with 3+1d radar in the view-of-delft dataset," *IEEE Robotics and Automation Letters*, pp. 1–1, 2022. (Cited on page 4.)

[31] "Livox avia specifications," https://www.livoxtech.com/avia/specs, accessed: 2024-03-11. (Cited on pages 5 and 20.)

[32] J. Bromley, J. Bentz, L. Bottou, I. Guyon, Y. Lecun, C. Moore, E. Sackinger, and R. Shah, "Signature verification using a "siamese"

time delay neural network," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 7, p. 25, 1993. (Cited on page 8.)

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Int. Conf. Medical Image Computing and Comp.-Ass. Intervention*, 2015, pp. 234–241. (Cited on page 9.)

[34] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," *Advances in Neural Information Processing Systems (NIPS)*, 2015. (Cited on page 9.)

[35] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 38 571–38 584. (Cited on pages 13, 16, and 17.)

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, 2017, p. 6000–6010. (Cited on page 13.)

[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021. (Cited on pages 13 and 16.)

[38] C. Yuan, X. Liu, X. Hong, and F. Zhang, "Pixel-level extrinsic self calibration of high resolution lidar and camera in targetless environments," *CoRR*, 2021. (Cited on page 15.)

[39] K. Lao and G. Yan, "Implementation and analysis of ieee 1588 ptp daemon based on embedded system," in *2020 39th Chinese Control Conference (CCC)*, 2020, pp. 4377–4382. (Cited on page 15.)

[40] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics YOLOv8," https://github.com/ultralytics/ultralytics, 2023. (Cited on page 15.)

[41] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft

coco: Common objects in context," in *Computer Vision – ECCV 2014*, 2014, pp. 740–755. (Cited on page 16.)

[42] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. (Cited on page 18.)