2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Learning an Efficient Terrain Representation for Haptic Localization of a Legged Robot

Damian Sójka Michał R. Nowicki Piotr Skrzypczyński

***Abstract*— Although haptic sensing has recently been used for legged robot localization in extreme environments where a camera or LiDAR might fail, the problem of efficiently representing the haptic signatures in a learned prior map is still open. This paper introduces an approach to terrain representation for haptic localization inspired by recent trends in machine learning. It combines this approach with the proven Monte Carlo algorithm to obtain an accurate, computation-efficient, and practical method for localizing legged robots under adversarial environmental conditions. We apply the triplet loss concept to learn highly descriptive embeddings in a transformer-based neural network. As the training haptic data are not labeled, the positive and negative examples are discriminated by their geometric locations discovered while training. We demonstrate experimentally that the proposed approach outperforms by a large margin the previous solutions to haptic localization of legged robots concerning the accuracy, inference time, and the amount of data stored in the map. As far as we know, this is the first approach that completely removes the need to use a dense terrain map for accurate haptic localization, thus paving the way to practical applications.**

I. Introduction

Recent years have brought legged locomotion from labs to real-world applications, focusing on inspection or search-and-rescue tasks in harsh environments like industrial facilities, disaster sites, or mines [1]. So far, few works have demonstrated the possibility of localizing a walking robot without visual or LiDAR-based SLAM, employing haptic sensing, using signals from IMUs, force/torque (F/T) sensors in the feet, and joint encoders [2], [3], [4]. Whereas these papers demonstrated a possibility of solving the pose tracking problem employing the Monte Carlo Localization (MCL) algorithm with particle filtering, the representation of the terrain and foot/terrain interactions extracted from haptic information remained an open problem. This representation is essential for haptic localization, as interactions between the robot's feet and the terrain are the only source of exteroceptive information in this problem formulation. Hence, haptic information representation must be descriptive enough to distinguish between the steps taken at different locations, even if these footholds are located on a similar surface. Moreover, this representation needs to be compact to allow quick retrieval of the data from terrain map and efficient comparison of the locations. The practical aspect of the representation problem is how the terrain map is obtained. A dense 2.5D elevation map used in [3] has to be surveyed using an external LiDAR sensor. In contrast, a map of terrain types encoded as classes on a grid map [4] needs tedious manual labeling. Both approaches confine the operation of a walking robot to small-scale pre-surveyed environments making the haptic localization concept rather impractical in real-world applications.

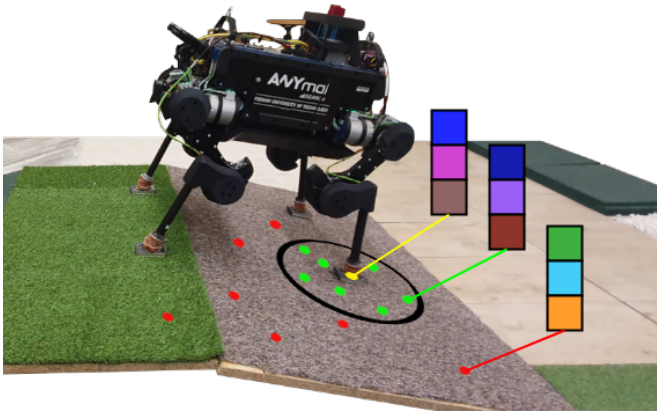

Fig. 1: Haptic localization requires a distinctive representation of the foot/terrain interaction to distinguish between locations. We propose to train a transformer-based neural network with triplet loss to minimize the difference between embeddings for steps close to each other while maximizing this difference for steps further away.

In contrast, this research uses the building blocks of machine learning methods already proven in computer vision [5] and place recognition [6] problems to create a sparse map of highly descriptive signatures in the locations touched by the robot's feet. This concept leverages the possibility of training a neural network with triplet loss to extract the features that differentiate the neighboring footholds from the collected haptic signals while suppressing those irrelevant for localization (Fig. 1). Interestingly, this approach addresses both the mentioned challenges, creating embeddings (latent vectors of the signatures) that are simultaneously highly descriptive and extremely compact. Our approach uses a new neural network architecture based on parameter-efficient transformer layers to build embeddings for a sparse terrain map, which ensures short inference times to achieve real-time operation of the haptic MCL. The contribution of our work can be summarized as follows:

- The first adaptation of the triplet loss training paradigm for learning local terrain representations from haptic information that lacks explicit class labels for the positive and negative examples.
- An efficient transformer-based neural network architec-

Institute of Robotics and Machine Intelligence, Poznan University of Technology, Poznan, Poland michal.nowicki@put.poznan.pl
*M. R. Nowicki is supported by the Foundation for Polish Science (FNP).

ture for computing the embeddings.
- A novel variant of the MCL method for legged robots that employs a sparse map of haptic embeddings and 3D positions of steps. It allows the robot to self-localize using only haptic information without any map that needs to be created or annotated manually.

II. Related work

Walking robots commonly use haptic information from their legs in terrain classification and gait adaptation. The approaches to terrain classification [7], [8], [9], [10] demonstrated that haptic information, like IMU or force/torque signals, can be used to determine the class of the terrain the robot is walking on. These methods achieve similar, high accuracies [11] but the supervised approaches to terrain classification are data-hungry and offer poor generalization to unseen classes. Moreover, the choice of these classes needs too be known upfront and is based on human perception. To overcome these challenges, Bednarek et al. [12] proposed an efficient transformer-based model resulting in fast inference and a decreased need for samples to train the network while improving the robustness of the solution to noisy or previously unseen data. Another attempt to decrease the required number of samples is [13], which uses a semi-supervised approach to the training of a Recurrent Neural Network (RNN) based on Gated Recurrent Units.

Gait adaptation is commonly mentioned as one of the applications of terrain classification. Lee et al. [14] proved that end-to-end learning can generate walking policies that adapt well to the changing environment. In this approach, driven by simulation, the terrain information is represented internally without explicit classes, while more recently [15] demonstrated a learned robust controller that combines terrain map data and proprioceptive locomotion when needed. Gangapurwala et al. [16] employ reinforcement learning policies to prioritize stability over aggressive locomotion while achieving the desired whole-body motion tracking and recovery control in legged locomotion. Despite progress with end-to-end approaches, we also see works that benefit from combining the trained and classical model-based approaches. One example is the work of Yuntao et al. [17], who shows that model-based predictive control can predict future interactions to increase the robustness of training policies.

Our haptic localization system takes inspiration from both of the presented domains. It follows the general processing scheme introduced by Buchanan et al. [3], who proposed an MCL algorithm for walking robots based on the measured terrain height and a dense 2.5D map of this terrain built with an accurate external 3D LiDAR. Their approach uses the relative height of the leg touching the ground to compute the updated particle positions in MCL, thus reducing the accumulation of the localization drift. Moreover, their follow-up work [4] introduced a haptic localization system that additionally utilizes the terrain classification information to further improve the localization accuracy, as long as the dense 2.5D map has prior class labels for each cell of the map. While this work demonstrated that geometric information is complementary to tactile sensing, it also revealed the limitations of terrain classification employing a discrete number of terrain classes that might not have strict borders when applied in a real-world scenario. In this context, Łysakowski et al. [18] showed that localization could be performed with compressed tactile information using Improved AutoEncoders, thus avoiding explicit terrain classification. Moreover, their approach demonstrated the possibility of working with a sparse map of latent signal representations, making it possible to learn the terrain map by the robot itself without tedious manual labeling. But the terrain representation from [18] just compresses the haptic signals, without selecting valuable features.

In this work, we propose a new HL-ST approach to haptic localization using a sparse geometric map and a latent representation of the haptic information that benefits from a training scheme with triplet loss, which has not yet been used in training on haptic signals for localization problems. This stands in contrast to [18], where the training process is fully unsupervised, thus giving no control over the learned representation. Similarly to [12], we also employ a transformer-based architecture to achieve a parameter-efficient network, but we train it to generate embeddings rather than class labels. Moreover, inspired by works in gait adaptation [14], the critical ingredient of our localization solution (latent representation/embedding) is trained to benefit from a large number of collected samples.

III. Proposed Method

Our problem statement is driven by an application of a walking robot performing repetitive tasks over a known route. We would like to quickly explore the desired path with the robot and then operate solely based on the legged odometry and haptic signals, even in challenging environments. In contrast to [3], [4], we assume no prior dense map, only requiring accurate localization for the first walk along the given route.

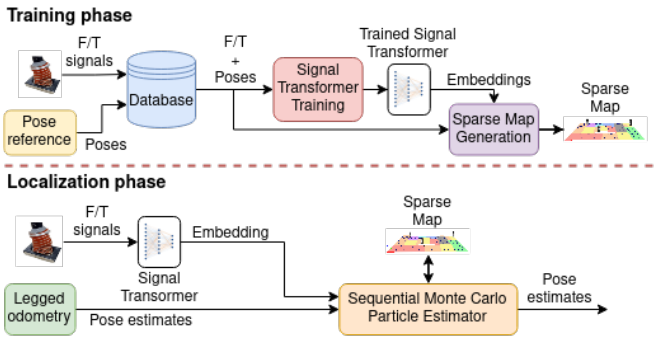

Fig. 2: Overview of the training and processing pipelines for our Haptic Localization utilizing Signal Transformer.

An overview of our approach is presented in Fig. 2. Each localization event is triggered by a foot placement on the ground that captures 160 consecutive samples from the F/T sensors mounted at that foot. In the initial phase, each step event has an associated localization estimate (e.g., from SLAM), and the whole sequence is used to gather

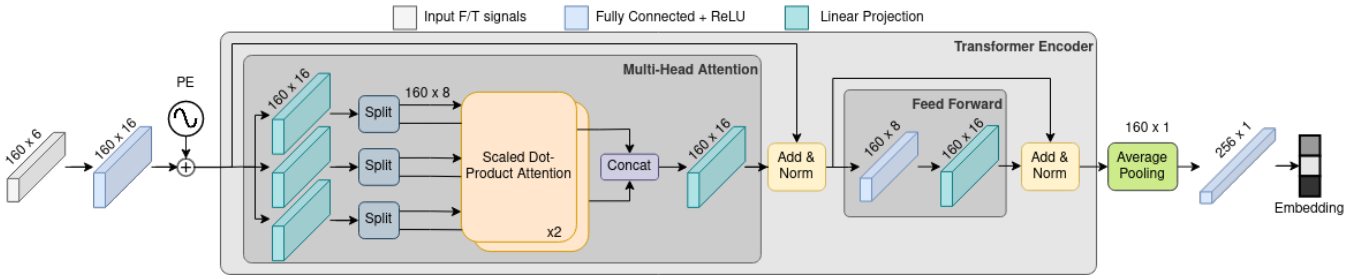

Fig. 3: The proposed Signal Transformer network that processes time sequence of force/torque signals to generate a location descriptive embedding.

a database of signals. This database is used to train our transformer-based network on triplets of samples to find the latent representation that best suits the localization purposes. The trained network processes the entire database to create a sparse map of embeddings at the measured locations.

During the localization phase, raw measurements from step events are fed to the trained neural network to obtain embeddings that can be compared to those already stored in the sparse terrain map. These comparisons are used to update our localization estimates represented by particles in the MCL framework.

A. Learning Terrain Representation

The critical component and our main contribution is the network to determine the embeddings that encode relevant features from the raw haptic input. The network is called Signal Transformer, based on the original transformer architecture from [19] and is presented in Fig. 3. The 6-dimensional sensor input (from 3-axis force and 3-axis torque sensors) for 160 consecutive measurements is converted into a 16-dimensional feature space with a fully-connected layer. We apply layer normalization and augment the sequence with learnable positional encoding. Augmented data are passed to the encoder of the reference transformer architecture from [19], with $h = 2$ attention heads, the dimensionality of model $d_m = 16$, and the size of inner feed-forward layer $d_{ff} = 8$. The use of average pooling flattens output from the encoder. The final latent representation is generated by applying batch normalization and feeding normalized data to the dense feed-forward layer with the ReLU activation function. The final layer has the number of neurons equal to the length of the embedding, which by default is set to 256, as in [18]. Implementation of the proposed network is publicly available[1].

The network is trained with triplet loss presented in Fig. 4. For efficiency, an online triplet mining technique is used to reduce the number of inactive triplets and to improve the convergence and speed of training. We form mini-batches randomly, with every example considered an anchor during the training. We need an associated location where this sample was captured for each training sample. Positive examples for a specific anchor are those sampled closer to the anchor than the defined constant distance threshold d_{thr}.

[1]https://github.com/dmn-sjk/signal_transformer

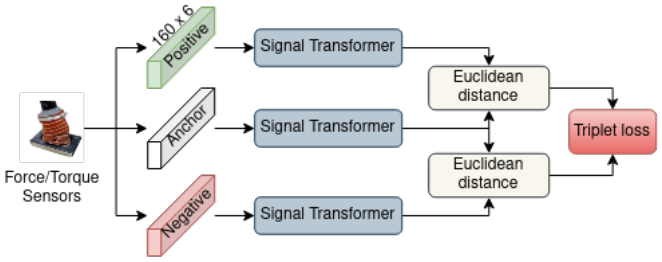

Fig. 4: The Signal Transformer network is trained using a triplet of data samples to achieve the desired similarity of embeddings for the anchor and positive sample while increasing the difference for the anchor and negative sample.

Consequently, the negative examples were sampled further away than d_{thr}:

$$\begin{cases} d(\mathbf{s}_{b_a}, \mathbf{s}_{b_i}) > d_{\text{thr}} \rightarrow b_i \in N_a \\ d(\mathbf{s}_{b_a}, \mathbf{s}_{b_i}) \leq d_{\text{thr}} \rightarrow b_i \in P_a, \end{cases} \tag{1}$$

where $d(\mathbf{s}_{b_a}, \mathbf{s}_{b_i})$ is the Euclidean distance between step position of an anchor $\mathbf{s}_{b_a}$ and step position $\mathbf{s}_{b_i}$ of the i-th data sample b_i. P_a and N_a denote a set of positives and negatives concerning the a-th anchor. The distance threshold d_{thr} is a hyperparameter that can be adjusted. We used $d_{\text{thr}} = 25$ cm since it provided the best localization accuracy. During training, positive and negative examples depend strictly on the spatial dependencies between step positions without any terrain class labels. Inspired by [20], we use Batch All triplet loss variation, but without special mini-batch sampling, considering the lack of class annotations, and calculate it as:

$$\mathcal{L} = \sum_{a=1}^{B} \sum_{p=1}^{|P_a|} \sum_{n=1}^{|N_a|} [d(f(b_a), f(b_p)) - d(f(b_a), f(b_n)) + m]_+ , \tag{2}$$

where B is the size of the mini-batch, m indicates the margin, and $|\cdot|$ denotes the cardinality of sets. The distance function $d(\cdot)$ is implementing an Euclidean distance. The average of Batch All triplet loss is calculated considering only these triplets, which have a non-zero loss as in [20]. The training process was performed using AdamW optimizer from [21]. The learning rate was exponentially decreased with an initial value of 5×10^{-4}. The initial value of weight decay was equal to 2×10^{-4} and was reduced with cosine decay. Mini-batch size was set to 128. The training lasted for 200 epochs.

B. Sparse Haptic Map Generation

The proposed network is used to build a sparse haptic map of embedding vectors visualized in Fig. 5. During the initial run, raw measurements taken at the contact of each foot with the ground are recorded along with the reference robot's position. The network is then trained with the triplet loss using these data. Once training is completed, data from each step are passed through the network to obtain the reduced latent representation (embedding), which is added to the map at the exact location of this step's foothold. The resulting map is sparse and unevenly distributed. The map generation phase requires an independent source of 6 DoF robot's pose estimates to properly train the network (generation of positive and negative examples) and to place the inferred embeddings in the map accurately. In practical applications, an onboard LiDAR-based localization subsystem of the robot [22] can be utilized as this independent source.

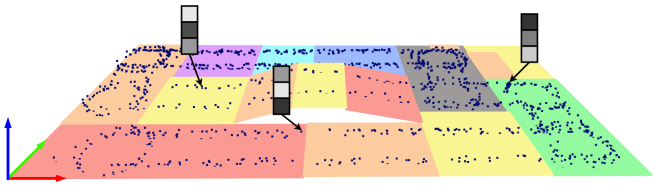

Fig. 5: Sparse haptic map visualization. Dark blue points indicate footholds recorded during the mapping run. Each 2D foothold holds an embedding and terrain elevation value. Color patches distinguish terrain classes in the PUTany dataset, but class labels are not used in our new approach.

C. Sequential Monte Carlo Localization

The proposed neural network generates highly descriptive embeddings – a latent representation of tactile sensing. To verify its impact on the localization, we use the sequential MCL algorithm proposed in [4] that was also used in the follow-up works on unsupervised terrain localization [18]. We present the general idea of this method while noting that this part of the processing is not a contribution of this paper.

Given a history of measurements $\boldsymbol{z}_0, \ldots, \boldsymbol{z}_k = \boldsymbol{z}_{0:k}$, the MCL algorithm estimates the most likely pose $\boldsymbol{x}_k^* \in SE(3)$ at time k:

$$p\left(\boldsymbol{x}_k|\boldsymbol{z}_{0:k}\right) = \sum_i w_{k-1}^i p\left(\boldsymbol{z}_k|\boldsymbol{x}_k\right) p\left(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}^i\right) \tag{3}$$

where w^i is the *importance weight* of the i-th particle, $p\left(\boldsymbol{z}_k, \boldsymbol{x}_k\right)$ is the measurement likelihood function, and $p\left(\boldsymbol{x}_k|\boldsymbol{x}_{k-1}^i\right)$ is the motion model for the i-th particle state.

The system gets the measurements whenever the robot's foot touches the ground while using a statically stable gait. Once measurements are taken, the probability of the particles is updated, particles are resampled, and a new pose estimate is available.

D. Haptic Measurement Model

For each taken step, the proposed neural network generates an embedding $\mathbf{v}$, based on signals from a single foot f placed at the ground at the location $\mathbf{d}_f$, in the base coordinates:

$$\mathbf{d}_f^i = (d_{xf}^i, d_{yf}^i, d_{zf}^i) = \boldsymbol{x}_k^i \mathbf{d}_f. \tag{4}$$

The position of the foot $\mathbf{d}_f^i$ is used to retrieve the haptic map entry $\mathbf{m} = \mathcal{S}(s_{xf}^i, s_{yf}^i)$ located in the map at (s_{xf}^i, s_{yf}^i), which is the nearest neighbor in 2D. For the sake of speed, search is based on 2D Euclidean distance with a map matching error $d_{2D} = \sqrt{\left(d_{xf}^i - s_{xf}^i\right)^2, \left(d_{yf}^i - s_{yf}^i\right)^2}$. The map entry $\mathbf{m}$ contains both the embedding of the haptic signal $\mathbf{w}$ and the elevation of the original measurement e.

Although we search for the map entries in 2D, we compare 3D positions for the purpose of the MCL measurement model taking advantage of the elevation values stored in the sparse map. We compare the latent representations using the L_2 norm:

$$d_l = f_{L_2}(\mathbf{v}, \mathbf{w}) = \sqrt{\sum_{j=1}^{n} (v_j - w_j)^2}, \tag{5}$$

where v_j, w_j encoding j-th component of the embeddings and d_l is the latent representation distance. To include the elevation information, d_e is defined as the difference between z-axis component of estimated foot position d_{zf}^i and the elevation e saved in the closest map entry:

$$d_e = d_{zf}^i - e. \tag{6}$$

We use univariate Gaussian distribution centered at the cell matched in the sparse latent map:

$$p(z_k|\boldsymbol{x}_k^i) = \begin{cases} p_{min} & d_{2D} > d_t \\ \mathcal{N}(d_l, \sigma_l)\mathcal{N}(d_{2D}, \sigma_{2D})\mathcal{N}(d_e, \sigma_e) & d_{2D} \leq d_t, \end{cases} \tag{7}$$

where $d_t = 25$ cm is the Euclidean threshold when a step is considered to have a proper match in the sparse map, otherwise the probability value is set to $p_{min} = 0.001$. The impact of haptic, 2D geometric component, and elevation is weighted by the experimentally determined sigma values $\sigma_l = 0.4$, $\sigma_{2D} = 0.4$, and $\sigma_e = 0.01$, respectively.

We denote our haptic localization method as HL-ST when all source of information are read from a sparse map, HL-T when elevation is not considered, and HL-GT when elevation comes from a dense geometric map and is used in a separate measurement model as in [4].

IV. Experimental Results

A. Dataset and Ground Truth

In the presented evaluation, we use the PUTany dataset proposed in [4] that was already applied to evaluate other haptic localization systems. The dataset consists of three trials (walks) of a quadruped robot ANYmal B300 robot equipped with F/T sensors in the feet conducted over a 3.5×7 m area containing uneven terrain with eight different terrain types ranging from ceramic to sand. The total distance traveled by the robot equals 715 m, with 6658 steps and a duration of 4054 s. We used a different sequence gathered on the same route to train the neural network proposed in this

work. The ground truth for the robot motion is captured with the millimeter-accuracy motion capture system (OptiTrack), providing 6 DoF robot poses at 100 Hz.

B. Accuracy measures for haptic localization

The robot generates a trajectory of 6 DoF poses that can be compared to the ground truth trajectory to determine the accuracy of the localization method. We use the Absolute Pose Error (APE) metric that is computed for a single relative pose $\mathbf{T}$ between the estimated pose $\mathbf{P}$ and the ground truth pose $\mathbf{G}$ at the time stamp t [23]:

$$\mathbf{T} = \mathbf{P}_t^{-1}\mathbf{G}_t, \tag{8}$$

where $\mathbf{P}_t, \mathbf{G}_t \in SE(3)$ are poses either available at time stamp t or interpolated for the selected time stamp. In our evaluation, we follow the error metrics reported in the previous articles concerning haptic localization [3], [4], [18] using the 3D translational part of the error $\mathbf{T}$ calling it $\mathbf{t}_{3D}$. The accuracy of our method is compared to the previously published results taking these results directly from the respective papers [4], [18].

As already observed in [18], the earlier haptic terrain recognition methods are not constraining the localization of particles in MCL in the vertical direction (parallel to the gravity vector), because the latent information is stored in a 2D array and lacks an elevation component. Therefore, in our comparisons, we also include the 2D translation error $\mathbf{t}_{2D}$ that ignores the error in the elevation. The $\mathbf{t}_{2D}$ results for the state-of-the-art methods were computed based on the publicly available source code of these systems.

C. Haptic localization with latent map (without geometry)

Using a dense and accurate terrain map of the environment for robot localization is impractical, as creating such a map usually requires deploying a survey-grade LiDAR in the target environment. Therefore, we first consider a scenario when only haptic signals and accurate localization are available for the robot's training run, with the following runs relying solely on haptics for localization. With these conditions, neither state-of-the-art methods for haptic localization (HL) can use elevation information. As other methods cannot use an elevation map, we also constrain our HL-T system to ignore the elevation data stored in the sparse map, using only haptic data. We compare our work to HL-C [4] utilizing terrain classification and HL-U [18], which uses unsupervised haptic latent representation.

All these methods are evaluated using the $\mathbf{t}_{2D}$ error as the elevation component is unconstrained, following the odometry drift. The obtained results are presented in Tab. I.

The worst results are produced by TSIF, the legged odometry estimator, which is both a baseline solution and a component of the MCL method in the remaining systems. Among the compared solutions, the proposed HL-T outperforms previous approaches by a large margin, reducing the APE values by almost 50%. We believe it stems from the fact that our method is not constrained to a limited number of discrete classes like HL-C, while unlike HL-U, it can learn an internal representation of the haptic signals that promote discriminative features. What is important, the error did not exceed 10 cm despite the lack of other sensing modalities, which may be sufficient to let an autonomous robot continue its operation despite a vision-based sensor failure or a sudden change in the environmental conditions.

Trial	**TSIF** [24] $\mathbf{t}_{2D}$	**HL-C** [4] $\mathbf{t}_{2D}$	**HL-U** [18] $\mathbf{t}_{2D}$	**HL-T** $\mathbf{t}_{2D}$
1	0.34	0.39	0.17	**0.07**
2	0.92	0.22	0.14	**0.06**
3	0.51	0.29	0.18	**0.08**

TABLE I: Comparison of the 2D Absolute Pose Error (APE, in [m]) for localization with haptic sensing only. The new HL-T method achieved the lowest error on all sequences.

D. Haptic localization with dense geometric and sparse latent map

Let's consider scenarios when an accurate 2.5D map of the environment is available for localization purposes. For such scenarios, we have HL-G [3] utilizing the dense height map of the terrain for pose correction, HL-GC [4] utilizing both the geometry and terrain classification, and HL-GU [18] which uses the geometry and unsupervised haptic latent representation. In these experiments, our HL-GT method is configured to use a dense elevation map and a sparse latent map. The results of the legged odometry estimator TSIF [24] were omitted as it has already been proven that HL-G, HL-GC, and HL-GU outperform it in these trials. The results for both types of errors ($\mathbf{t}_{2D}$, $\mathbf{t}_{3D}$) are presented in Tab. II.

Trial	**HL-G** [4]		**HL-GC** [4]		**HL-GU** [18]		**HL-GT**	
	$\mathbf{t}_{3D}$	$\mathbf{t}_{2D}$	$\mathbf{t}_{3D}$	$\mathbf{t}_{2D}$	$\mathbf{t}_{3D}$	$\mathbf{t}_{2D}$	$\mathbf{t}_{3D}$	$\mathbf{t}_{2D}$
1	0.23	0.23	0.14	0.12	0.15	0.09	**0.09**	**0.08**
2	0.25	0.20	**0.11**	0.11	0.18	0.12	**0.11**	**0.10**
3	0.21	0.18	0.18	0.17	0.13	0.13	**0.09**	**0.09**

TABLE II: Comparison of the 3D and 2D Absolute Pose Error (APE, in [m]) for localization solutions utilizing both prior dense geometric map and haptic terrain recognition solutions. Proposed HL-GT provides the best results using both $\mathbf{t}_{3D}$ and $\mathbf{t}_{2D}$ error metric.

The obtained results for all considered methods show that the $\mathbf{t}_{3D}$ and $\mathbf{t}_{2D}$ errors almost match each other, proving that there is no significant drift in the elevation direction due to the availability of the dense elevation map. The proposed HL-GT outperforms other solutions, which suggests that the latent representation trained with triplet loss is more suited for distinguishing between terrain locations than terrain classification (HL-GC) or unsupervised terrain representation/signal compression (HL-GU). Moreover, the haptic signal information encoding in HL-GT is complementary to the dense elevation map as the method improves the performance of the bare geometric approach (HL-G).

E. Haptic localization with sparse geometric map

One of the advantages of the proposed solution is the ability to use elevation information even if only localization

ground truth was available for training. This improvement significantly impacts the solution's practicality as no survey-grade LiDAR is required to take advantage of the elevation data. Therefore, we decided to compare three solutions: HL-T, which uses solely haptic signals for localization, HL-GT which uses dense geometric map and haptic signals, and HL-ST, which uses sparse geometric and latent map. The results are presented in Tab. III.

Trial	HL-T		HL-GT		HL-ST	
	$\mathbf{t}_{3D}$	$\mathbf{t}_{2D}$	$\mathbf{t}_{3D}$	$\mathbf{t}_{2D}$	$\mathbf{t}_{3D}$	$\mathbf{t}_{2D}$
1	0.51	**0.07**	**0.09**	0.08	**0.09**	0.09
2	0.77	**0.06**	**0.11**	0.10	**0.11**	0.11
3	0.44	**0.08**	**0.09**	0.09	0.10	0.09

TABLE III: Comparison of the 3D and 2D Absolute Pose Error (APE, in [m]) for localization without geometry (HL-T), with dense geometric map (HL-GT), and sparse geometric map (HL-ST). HL-ST performs similarly to HL-GT in 2D and 3D without a tedious mapping phase.

The results show that the HL-T approach provides the best results in 2D. Still, the 3D error is unbounded following the legged odometry's general drift, making it impractical for any autonomous operation. On the other hand, HL-GT provides the most accurate 3D localization due to the dense geometric map. The proposed HL-ST is a good trade-off between these approaches as the 2D and 3D errors are comparable with HL-T and HL-GT while only using the haptics and localization for the first trial. We believe HL-ST is, therefore, a unique solution that may support legged robot autonomy in challenging, real-world applications.

F. Inference time evaluation

Autonomous operation requires real-time processing with short inference times. We compared the average inference times of the proposed Signal Transformer network with the classification neural network from HL-C [4] and the unsupervised network HL-U [18]. Inference times were measured on over 10 000 samples on NVIDIA GeForce GTX 1050 Mobile GPU matching a similarly capable GPU that can fit in a walking robot. Table IV contains the obtained results.

HL method	HL-C	HL-U	HL-T
Inference time [ms]	21.20 $\pm$ 2.94	30.72 $\pm$ 4.84	2.20 $\pm$ 0.28

TABLE IV: Inference time comparison between models used to process the haptic signal for localization purposes.

The inference time of the Signal Transformer being a part of the HL-T/HL-ST solutions is one magnitude smaller when compared to the inference times of neural networks used in HL-C and HL-U. The observed gains stem from a reduced number of parameters for our networks (45992 parameters) compared to over 1 million parameters for networks used in HL-C and HL-U. The transformer-based architecture proved to be more compact and suitable for on-board deployment in a robot than previously used solutions.

G. The size of the latent representation

The transformer-based architectures are known to be efficient, needing almost a fraction of the resources (parameters and inference time) of other known architectures to achieve comparable results. Moreover, learning with triplet loss can train a representation with desired characteristics. Therefore, we wanted to verify the required size of the embeddings necessary to achieve good localization results using a more challenging HL-T approach. The obtained APEs depending on the chosen size of the embeddings are presented in Fig. 6.

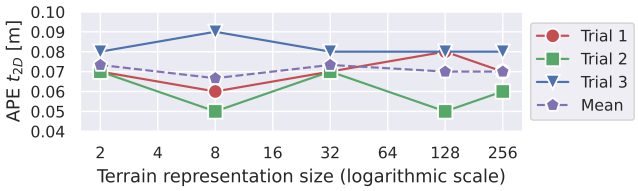

Fig. 6: The graph of the HL-T method's 2D localization error $\mathbf{t}_{2D}$ as a function of the trained embedding size for the proposed Signal Transformer architecture. The Mean is the average outcome of all three trials. We see no major difference in APE, even with a small embedding size.

Decreasing the embedding size from the original size of 256 did not affect the localization accuracy, suggesting that an embedding vector with a length as low as 2 contains enough information to distinguish between embeddings from multiple positions in a given environment. This result is of practical importance, as for a possible map containing 10 000 steps, the original size of 75 MB for embeddings with length 256 can be reduced to 1.4 MB using embeddings of size 2, which means a substantial reduction in the amount of stored data, a possibility to operate over a larger area, and the reduction of map matching times.

V. Conclusions

This work investigated how to employ triplet loss to train a Signal Transformer network to compute descriptive embeddings from haptic signals. The experiments indicate that the novel localization method employing these embeddings outperforms state-of-the-art haptic localization solutions (HL-C and HL-U) when only haptics are used in a controlled environment. At the same time, the HL-GT variant achieves the lowest localization error (3D APE, compared to HL-GC and HL-GU) when a dense 2.5D map is used due to an efficient representation of the haptic data. In contrast to previous works, we can build and use a sparse geometric map (HL-ST), resulting in a practical solution requiring only reference poses for the first robot run while achieving results on par with solutions utilizing dense 2.5D maps. Moreover, we show that our network can process data 10 times faster than previous approaches and can reduce the embeddings from a size of 256 to 2 without a significant impact on the localization error. As a result, we achieve a haptic localization method that is more practical than state-of-the-art solutions. Our future work will concern outdoor experiments in scenarios without clear terrain type boundaries.

REFERENCES

[1] R. Zimroz, M. Hutter, M. Mistry, P. Stefaniak, K. Walas, and J. Wodecki, "Why should inspection robots be used in deep underground mines?" in *Proceedings of the 27th International Symposium on Mine Planning and Equipment Selection - MPES 2018*, E. Widzyk-Capehart, A. Hekmat, and R. Singhal, Eds. Cham: Springer International Publishing, 2019, pp. 497–507.

[2] S. Chitta, P. Vernaza, R. Geykhman, and D. Lee, "Proprioceptive localization for a quadrupedal robot on known terrain," in *IEEE International Conference on Robotics and Automation (ICRA)*, April 2007, pp. 4582–4587.

[3] R. Buchanan, M. Camurri, and M. Fallon, "Haptic Sequential Monte Carlo Localization for Quadrupedal Locomotion in Vision-Denied Scenarios," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, October 2020.

[4] R. Buchanan, J. Bednarek, M. Camurri, M. R. Nowicki, K. Walas, and M. Fallon, "Navigating by touch: haptic Monte Carlo localization via geometric sensing and terrain classification," *Autonomous Robots*, vol. 45, no. 6, pp. 843–857, 2021.

[5] K. Ho, J. Keuper, F. Pfreundt, and M. Keuper, "Learning embeddings for image clustering: An empirical study of triplet loss approaches," in *25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 87–94.

[6] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao, "Spatial pyramid-enhanced NetVLAD with weighted triplet loss for place recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 2, pp. 661–674, 2020.

[7] M. H. Hoepflinger, C. D. Remy, M. Hutter, and R. Siegwart, "Haptic Terrain Classification on Natural Terrains for Legged Robots," in *International Conference on Climbing and Walking Robots (CLAWAR)*, 2010, pp. 785–792.

[8] X. A. Wu, T. M. Huh, R. Mukherjee, and M. Cutkosky, "Integrated Ground Reaction Force Sensing and Terrain Classification for Small Legged Robots," *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 1125–1132, 2016.

[9] J. Bednarek, M. Bednarek, P. Kicki, and K. Walas, "Robotic Touch: Classification of Materials for Manipulation and Walking," in *IEEE International Conference on Soft Robotics (RoboSoft)*, 2019, pp. 527–533.

[10] H. Kolvenbach, C. Bärtschi, L. Wellhausen, R. Grandia, and M. Hutter, "Haptic Inspection of Planetary Soils With Legged Robots," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1626–1632, 2019.

[11] J. Bednarek, M. Bednarek, L. Wellhausen, M. Hutter, and K. Walas, "What am I touching? Learning to classify terrain via haptic sensing," in *IEEE International Conference on Robotics and Automation (ICRA)*, May 2019, pp. 7187–7193.

[12] M. Bednarek, M. Łysakowski, J. Bednarek, M. R. Nowicki, and K. Walas, "Fast haptic terrain classification for legged robots using transformer," in *2021 European Conference on Mobile Robots (ECMR)*, 2021.

[13] A. Ahmadi, T. Nygaard, N. Kottege, D. Howard, and N. Hudson, "Semi-Supervised Gated Recurrent Neural Networks for Robotic Terrain Classification," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1848–1855, 2021.

[14] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science Robotics*, vol. 5, no. 47, 2020.

[15] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning robust perceptive locomotion for quadrupedal robots in the wild," *Science Robotics*, vol. 7, no. 62, p. eabk2822, 2022.

[16] S. Gangapurwala, M. Geisert, R. Orsolino, M. Fallon, and I. Havoutis, "Rloc: Terrain-aware legged locomotion using reinforcement learning and optimal control," *IEEE Transactions on Robotics*, pp. 1–20, 2022.

[17] Y. Ma, F. Farshidian, T. Miki, J. Lee, and M. Hutter, "Combining learning-based locomotion policy with model-based manipulation for legged mobile manipulators," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2377–2384, 2022.

[18] M. Łysakowski, M. R. Nowicki, R. Buchanan, M. Camurri, M. Fallon, and K. Walas, "Unsupervised Learning of Terrain Representations for Haptic Monte Carlo Localization," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 4642–4648.

[19] 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 *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017.

[20] A. Hermans, L. Beyer, and B. Leibe, "In Defense of the Triplet Loss for Person Re-Identification," 2017. [Online]. Available: https://arxiv.org/abs/1703.07737

[21] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

[22] M. Ramezani, G. Tinchev, E. Iuganov, and M. Fallon, "Online LiDAR-SLAM for legged robots with robust registration and deep-learned loop closure," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 4158–4164.

[23] M. Grupp, "evo: Python package for the evaluation of odometry and SLAM." https://github.com/MichaelGrupp/evo, 2017.

[24] M. Bloesch, M. Burri, H. Sommer, R. Siegwart, and M. Hutter, "The Two-State Implicit Filter Recursive Estimation for Mobile Robots," *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 573–580, Jan 2018.