



UNIVERSITÀ DI TRENTO

Dipartimento di Ingegneria e Scienza dell'Informazione

Corso di Laurea in Ingegneria informatica, delle comunicazioni ed elettronica

ELABORATO FINALE

DIRECT REGRESSION FOR TACTILE SIGNALS LOCALIZATION ON 3D GAUSSIAN-SPLATTING SCENES

Relatori

Francesco De Natale

Nicola Conci

Laureando

Filippo Marcatili

Anno accademico 2025/2026

Ringraziamenti

Con la conclusione di questo importante percorso durato tre anni, si conclude anche una fase della mia vita, che ha portato con sé gioie e grandi soddisfazioni ma anche sconfitte e delusioni. Chi mi conosce, però, dovrebbe sapere bene che sono una persona che tende a guardare il bicchiere mezzo pieno, a volte anche più di quanto non lo sia davvero, ed è per questo che ci tengo a fare i dovuti ringraziamenti a tutte le persone che hanno fatto parte di questo mio percorso. Non sono bravo con le parole, ma non riesco a fare a meno di menzionare ogni singola persona che ha contribuito, anche minimamente, a questa avventura.

Innanzitutto mi sembra doveroso ringraziare il professor **Francesco De Natale**, per avermi guidato nella scelta della direzione da intraprendere per sigillare al meglio questa triennale. Mi preme soprattutto ringraziare il neo Dottore **Antonio Stefani**, che è stato paziente e disponibile nonostante fosse già indaffarato nel conseguimento del titolo di Dottorato di ricerca. Ringrazio anche il corpo docente che ha contribuito alla mia formazione, e i miei colleghi, nonché compagni di merende, con cui ho avuto il piacere di condividere successi e sofferenze negli ultimi tre anni.

Ringrazio quindi, in ordine, **Anas**, il primo pezzo del puzzle, il primo sconosciuto che è poi diventato parte della mia quotidianità; subito dopo **Andrea**, primo compagno di serata insieme ad **Angelo**, senza il quale non sarei sopravvissuto allo stress delle sessioni.

Poi **Baia**, che mi ha colpito immediatamente per via del suo bel nome (grande Fillo) e che purtroppo ha portato con sé **Saif**, che è stato un ottimo esempio da seguire per prendere la direzione sbagliata, ma anche protagonista di gran parte dei miei bei ricordi di questi anni.

Ringrazio anche i colleghi di ICE ING & Co.: **Rebbo, Daniel, Vic, Gaetano, Davide, Chiara, Marco, Leo, Ouss, Riccardo** e **Micky**.

Un ringraziamento speciale va a **Sam** e **Gio**, con cui ho condiviso 4 mesi di paradiso in Norvegia. **Sam**, spirito guida, sempre pronto per qualsiasi cosa, soprattutto se è una pessima idea: il direttore d'orchestra, TDH.

Giovanni, invece, mi ha sempre fatto paura, ma poi ho avuto il piacere di conoscere il patatone che muove il suo corpo, e ora faccio fatica a separarmene: grande dedizione, testa sulle spalle, punto di riferimento.

Ora posso iniziare a tornare in madre patria, e ringraziare quelli che fanno il tifo per me già da molto tempo.

Ringrazio **FG**: per primo il mio perito elettrotecnico di fiducia, **Adriano**. Non basta un manuale intero per elencare le situazioni in cui hai fatto la differenza e in cui ti sei riconfermato come pilastro della mia esistenza; sei stato il primo vero amico che ho abbandonato, e anche il primo a tornare da me quando ero in difficoltà: uomo dall'orgoglio d'acciaio, ma col cuore più grande dell'orgoglio, sempre covetto.

Stegi, uomo di estremo gusto, infatti è amico mio, compagno di avventure; quanto tempo ho perso a fare niente con te, lo rifarei comunque da capo.

Pietro, non so nemmeno dove trovare le parole che facciano onore al grande uomo che sei, sempre sorridente, sempre pronto ad aiutare senza chiedere nulla in cambio.

Valerio, dinamite pura, il quarto degli uomini con cui sono cresciuto e che reputo mio fratello: voglio quindi sentire le mani battere.

Grazie ragazzi, siete una parte indimenticabile del mio passato, presente e futuro.

Sto andando cronologicamente, o almeno ci provo, e teoricamente è ora del secondo gruppo, quello che mi sta riempiendo l'iCloud di foto.

Partirei dal Grande Gigante Gentile, **Massi**: inutile farci tanti giri attorno, senza di te avrei fatto molti più errori, ma mi hai comunque dato la possibilità di commettere quelli che mi hanno aiutato a crescere e maturare, e per questo non finirò mai di ringraziarti.

Gerry, dai SGDBR: condividiamo così tanto, sia a livello emotivo che mentale, peccato non sia la stessa cosa per i capelli.

Diego, il manifesto della gioia, mi ricordi sempre che non devo abbandonare il bambino che è in me.

Valerio, non hai potuto fare il pilota fin da subito, quindi ti sei costruito le ali da solo, letteralmente sì enorme, meriti di volare davvero ora.

Frassì, sei arrivato dopo, ma in qualche modo ti sei fatto spazio nella folla, probabilmente con lo sfollagente.

Un pensiero anche a **Mattia**, **Morgan**, **Matteo**, **Seve** e **Coccia** che, nonostante non siano più parte attiva della mia quotidianità, rimangono comunque parte di me.

Non posso non fare un ringraziamento speciale ad **Alessandro**, in arte **Puddu**, che reputo ormai un fratello: una persona con cui mi confronto per le cose più stupide e con cui avrò il piacere di proseguire questo percorso lontano da casa. Sono sicuro che continueremo a essere grandi compagni di avventure.

Grazie anche a **Nicholas**: nonostante ci fossimo conosciuti di persona per poche ore, si è rivelato una persona di grandissimo spessore ma soprattutto di animo buono, e ha continuato a fare il tifo per me anche a distanza.

Come dimenticare la famiglia? Grazie a tutti e quattro i miei nonni, che ho la fortuna di poter ringraziare di persona e che continuano ad esserci per me; così come tutti gli zii, le zie e i cugini.

Ringrazio i miei genitori, che mai una volta hanno dubitato di me e che non smettono mai di fare il possibile perché io e mio fratello possiamo inseguire i nostri sogni. Grazie per darmi sempre la possibilità di provare, di sbagliare e di imparare dai miei errori. Grazie per la fiducia e per avermi fatto capire il significato di ricchezza, che non ha niente a che fare con i soldi.

Grazie **Babbo** per avermi insegnato a portare rispetto e non dare nulla per scontato, per avermi dimostrato che piangere non è da deboli e per essere sempre stato il mio mentore. Come dico sempre a tutti, potessi arrivare ad essere un decimo del grande uomo che sei tu, mi riterrei già più che soddisfatto.

Grazie **Mamma** per avermi ascoltato, consolato e sgridato quando ne ho avuto bisogno, per avermi insegnato la pazienza e per aver dimostrato non solo a me che non è mai troppo tardi. Sei sempre stata e sempre sarai la scintilla che mi dà forza ogni giorno per dare il massimo.

Grazie a **Riccardo**, che ha paura di non riuscire a fare la parte di fratello maggiore ma, inconsapevolmente e spesso indirettamente, mi ha insegnato a tenere la testa alta, a non sottostare all'opinione che le persone hanno di me e a portare avanti le mie passioni e i miei progetti. Spero di poter essere sempre anch'io, per te, la persona su cui puoi contare in qualsiasi occasione, nel bene e nel male. Ti voglio bene.

Ci tengo a ringraziare anche **Emy**. Le nostre strade si sono separate da tempo, ma se sono dove sono ora è anche grazie a te. Siamo praticamente cresciuti insieme e abbiamo imparato molte lezioni di vita, alcune molto dolorose, ma sicuramente tutto ciò mi ha spinto a impegnarmi a essere la migliore versione di me stesso.

Spero di non aver dimenticato nessuno, se così fosse chiedo scusa.

E per tutti quelli che mi chiedono perché sono arrivato fino a Trento per sostenere la mia triennale, o addirittura perché io abbia deciso di partire per la Danimarca per la mia magistrale, cito una frase che ho letto in un libro 6 anni fa e che non mi ha più lasciato:

*“Un'imbarcazione è più sicura quando si trova in porto;
tuttavia non è per questo che le barche sono state costruite.”*

Grazie di cuore a tutti.

Contents

Abstract	2
1 Introduction	3
1.1 The visuo-tactile chain and localization	3
1.2 Problem statement and contributions	3
1.3 Thesis outline	4
2 Background and Related Work	5
2.1 3D scene representation	5
2.2 Visuo-tactile sensing and datasets	5
2.3 Cross-modal tactile generation	5
2.4 Normalized Object Coordinate Space (NOCS)	6
2.5 Tactile contact localization	6
3 Method	8
3.1 Problem formulation and overview	8
3.2 NOCS position encoding	8
3.3 Dataset and ground-truth extraction	8
3.4 Localization architecture	10
3.5 Spatial attention pooling	10
3.6 Training	11
3.7 Evaluation metrics	11
4 Experiments and Results	13
4.1 Experimental setup	13
4.2 Main results	13
4.2.1 Random split versus SplatTouch split	14
4.2.2 Comparison with baselines	14
4.2.3 Comparison with SplatTouch	14
4.3 Ablation study	15
4.4 Qualitative results	15
4.5 Limitations	16
5 Conclusions	18
5.1 Contributions and impact	18
5.2 Future work	19
Bibliography	19

Abstract

Digital experience has, so far, been almost entirely visual: screens, images and video have made sight the main sense through which people interact with digital content, while touch has been left largely outside it. Bringing touch into these environments, alongside vision, would widen their possibilities, from more immersive virtual reality to robots that can both see and feel an object. Combining the two senses is the goal of multimodal perception. Recent work represents scenes as 3D Gaussian Splatting reconstructions and learns to translate between the visual and tactile modalities along two complementary directions: generating the tactile signal expected at a 3D location, and localizing a tactile observation back to its position on the scene. SplatTouch addresses both directions, performing localization with a contrastive, SimSiam-style network that maps a touch sample to a Normalized Object Coordinate Space (NOCS) colour and hence to a 3D position.

This thesis focuses on the localization direction and studies an alternative model for it. Instead of contrastive representation learning, localization is treated as a direct supervised regression from the tactile image to the NOCS colour. The proposed architecture is lightweight, consisting of a ResNet-18 backbone, a learned spatial pooling stage and a small regression head. The method is evaluated on two real scenes, reporting both a normalized NOCS error and a physically meaningful error in centimetres, under a random split and under the harder periodic split used by SplatTouch; the main design choices are ablated.

Under the periodic split, taken as the honest reference setting, the model localizes contacts with a mean NOCS error of 0.040 and a metric error of 7.06 cm on the indoor scene, and 0.107 / 11.5 cm on the harder outdoor scene; averaged over the two scenes (9.3 cm) it improves over the localization figures reported by SplatTouch (with the caveats of a different-method, fewer-scenes comparison) and clearly outperforms naive baselines. Under the easier random split the indoor error drops to 0.019 and 3.58 cm. The ablation shows that the accuracy is largely insensitive to the pooling aggregation, while a 1×1 pooling kernel and a 10% NOCS crop are the preferable configuration. Overall, the results indicate that a simple direct-regression model is a competitive alternative to contrastive localization within the visuo-tactile chain.

1 Introduction

People rarely rely on a single sense. Picking up an unfamiliar object, the eyes judge its shape and distance while the hand, almost in the same instant, reports its weight, its texture, the warmth or coldness of its surface and how much it yields under a light press. Vision works at a distance and takes in a whole scene at once; touch works only on contact, but precisely where it matters, and exactly where the hand itself tends to hide the object from view. The two senses are not redundant but complementary, and human perception constantly fuses them without any conscious effort.

The systems that reconstruct and display the world digitally, however, have so far developed along a single one of these channels. Screens, images and video have made sight the sense through which people interact with digital content, while touch has been left almost entirely outside it. Closing this gap, by giving artificial systems a sense of touch alongside vision, would widen what they can do: virtual environments could be felt and not only seen, and robots could reason about a contact the way a person does, combining what is in front of them with what is under their fingers. Teaching machines to relate these two modalities, so that one can inform the other, is the goal of *multimodal* visuo-tactile perception [12].

A first requirement for relating vision and touch is a common ground in which both can be placed: a model of the scene that is three-dimensional, so that a tactile reading can be attached to the exact point it came from. Recent advances in 3D reconstruction provide exactly this. Neural Radiance Fields (NeRF) [10] learn a continuous volumetric function of a scene from a handful of images and render it from new viewpoints, and 3D Gaussian Splatting [7] achieves comparable quality far more efficiently by describing the scene explicitly, as a large cloud of small 3D Gaussians each carrying a position and a colour. Because every Gaussian sits at a known location, this explicit form offers a natural shared space in which visual and tactile observations can be registered together. Tactile-Augmented Radiance Fields [3] take a first step in this direction, pairing real tactile readings with the precise 3D points they were collected from and so producing aligned visuo-tactile data.

1.1 The visuo-tactile chain and localization

Within this shared 3D space, vision and touch are related by two inverse mappings: tactile *generation*, which predicts the tactile signal at a given 3D location, and tactile *localization*, which estimates the 3D location of a contact from a tactile observation (Figure 1.1). SplatTouch implements both on Gaussian-Splatting scenes. Its localization stage relies on a contrastive, SimSiam-style objective [2] that aligns tactile features with a colour-coded encoding of 3D position, so that a query touch can be matched to a coordinate at inference time.

The accuracy of this contrastive stage decreases on weakly textured surfaces and under reconstruction noise. Localization is therefore the component of the chain with the most room for improvement, and it is the subject of this thesis. In place of a contrastive objective, the localization mapping is formulated here as a supervised regression, trained directly to predict the 3D coordinate of a contact; the guiding question is whether such a direct model can localize contacts more accurately than the contrastive one while remaining simpler to train.

1.2 Problem statement and contributions

This thesis addresses the localization mapping introduced above. The task is defined as follows: given a tactile observation of a contact on a Gaussian-Splatting scene, estimate the 3D coordinates of the contact point. Following SplatTouch, positions are expressed in a Normalized Object Coordinate Space (NOCS) [14], which encodes each Gaussian’s coordinates as a normalized colour and thereby provides a bounded regression target. The proposed model predicts this target directly from the tactile image, through a convolutional backbone, a learned spatial aggregation stage and a regression

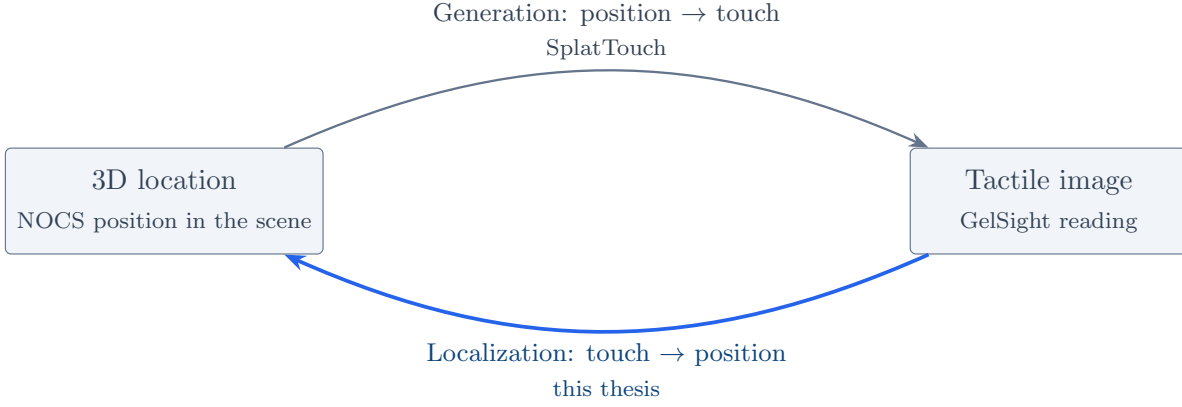


Figure 1.1: The visuo-tactile chain. Generation maps a 3D location to a predicted tactile image; localization performs the inverse mapping, recovering the 3D location from a tactile observation. This thesis concerns the localization direction, for which it studies an alternative to the contrastive model of SplatTouch.

head trained end-to-end under a supervised loss. Compared with the two-stage contrastive pipeline of SplatTouch, this single-stage formulation is simpler to train and, on the scenes evaluated, attains a lower localization error.

The main contributions of this work are the following:

- a localization model for the SplatTouch visuo-tactile setting that, unlike its contrastive formulation, treats localization as a direct supervised regression in NOCS space;
- a lightweight architecture combining a ResNet-18 [4] backbone, a learned spatial aggregation module (termed *spatial attention pooling*, with the term clarified in Section 3.5) and a compact regression head, together with an analysis of the contribution of each component;
- an evaluation on two real scenes that reports both a normalized NOCS error and a metric error in centimetres, under a random split and under the periodic split of SplatTouch, with comparisons against naive baselines and against the figures reported for SplatTouch;
- an ablation of the principal design choices (spatial aggregation, convolution kernel size, NOCS crop fraction) and an interactive 3D visualizer for the qualitative inspection of the predictions.

1.3 Thesis outline

The remainder of the thesis is organized as follows. Chapter 2 reviews the related work on which this thesis builds (3D scene representation, visuo-tactile sensing, cross-modal tactile generation, the NOCS encoding and contact localization) and positions the proposed approach. Chapter 3 details the method: the NOCS representation, the dataset, the localization architecture and its spatial aggregation module, the training procedure and the evaluation metrics. Chapter 4 presents the experimental evaluation, including the main results, the comparison with baselines and with SplatTouch, the ablation study and a qualitative analysis, and closes with a discussion of limitations and future work.

2 Background and Related Work

This chapter surveys the work on which the thesis builds and situates its contribution. Section 2.1 reviews 3D scene representations, which provide the geometric substrate for grounding touch in space; Section 2.2 covers visuo-tactile sensing and the datasets that pair touch with vision and 3D geometry; Section 2.3 discusses cross-modal tactile generation, the forward direction of the visuo-tactile mapping; Section 2.4 introduces the NOCS encoding adopted as the position representation; and Section 2.5 reviews tactile contact localization, the inverse direction, where the proposed model is positioned.

2.1 3D scene representation

Reconstructing a scene in three dimensions from images is a prerequisite for attaching a tactile reading to the exact point it came from. Two families of methods are relevant here. Neural Radiance Fields (NeRF) [10] encode a scene *implicitly*, as a multilayer perceptron that maps a 3D coordinate and a viewing direction to colour and volumetric density, and synthesize images by integrating these quantities along camera rays through differentiable volume rendering. This formulation yields high reconstruction quality but is costly to train and to render, a limitation that later work mitigated with specialized data structures such as multiresolution hash encodings [11].

3D Gaussian Splatting [7] adopts the opposite, *explicit* strategy: the scene is represented as a large set of anisotropic 3D Gaussians, each parameterized by a position, a covariance, an opacity and a view-dependent colour, and is rendered by rasterizing and α -blending these primitives, which makes both training and rendering substantially faster. The explicit representation is particularly convenient for the task considered in this thesis. Because every Gaussian carries an absolute 3D position, the primitives can be queried directly, recoloured according to their coordinates, and rendered into auxiliary maps that are spatially aligned with the visual views. These positions serve as the spatial ground truth for localization, as detailed in Chapter 3. Figure 2.1 shows the indoor scene used in this work in these terms: the bare Gaussian centres as a point cloud, the full splat reconstruction, and the same primitives recoloured by their normalized coordinates (Section 2.4).

2.2 Visuo-tactile sensing and datasets

Vision-based tactile sensors turn contact into an image. A sensor such as GelSight [16] observes, with an internal camera, the deformation of a soft elastomeric pad pressed against a surface, recovering fine local geometry and contact forces; DIGIT [8] follows the same optical principle in a compact, low-cost form factor designed for in-hand manipulation. Because the output is an image, tactile signals can be processed with the same convolutional machinery used for vision, which is the property exploited by the model in this thesis.

Several datasets pair this tactile modality with vision and, increasingly, with 3D geometry. Early collections aligned tactile readings with synchronized images to enable cross-modal learning between the two senses [9], and in-the-wild datasets later broadened the range of materials and objects covered [15]. Most relevant here, Tactile-Augmented Radiance Fields [3] register tactile readings to specific 3D locations of a NeRF-reconstructed scene, producing data in which each touch is annotated with the precise point of the scene it was collected from. The data used in this work is of this kind: tactile observations paired with their 3D contact locations on a reconstructed scene.

2.3 Cross-modal tactile generation

Cross-modal tactile generation is the forward direction of the visuo-tactile mapping: predicting the tactile signal that a sensor would measure at a chosen location, from visual information alone. The problem was first approached as a direct image-to-image translation between the two modalities [9],

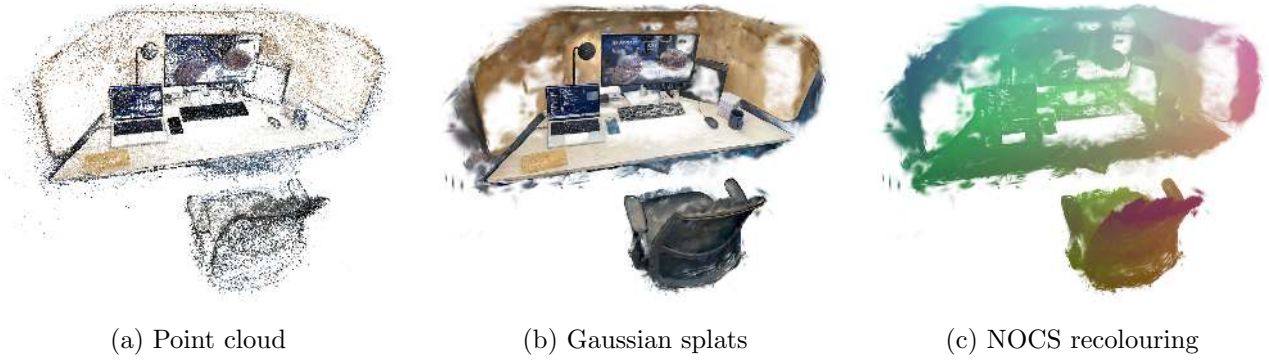


Figure 2.1: The same reconstructed indoor scene in three representations. (a) The Gaussians rendered as a bare point cloud. (b) The full 3D Gaussian-Splatting reconstruction, whose anisotropic primitives fill the surfaces. (c) The same Gaussians recoloured by their normalized coordinates, which turns the geometry into a renderable NOCS map (Section 2.4) and provides the spatial ground truth used throughout this thesis.

and benefited from the subsequent progress of generative modelling, in particular denoising diffusion models [5], which generate samples by progressively reversing a noising process and have become a standard tool for conditional synthesis. SplatTouch [12] instantiates this idea on Gaussian-Splatting scenes, using a diffusion model conditioned on aligned RGB, depth and NOCS maps to synthesize the tactile sample at a queried location. This direction is reviewed for completeness, as the localization task addressed in this thesis is its inverse; generation itself is not a contribution of this work.

2.4 Normalized Object Coordinate Space (NOCS)

The Normalized Object Coordinate Space (NOCS) [14] was introduced for category-level 6D pose and size estimation. Its key idea is to map the surface of the objects of a category into a shared, canonical unit cube and to predict, for each pixel, the corresponding normalized coordinate encoded as an RGB colour; pose and size then follow from aligning these predicted coordinates with the observed geometry. The SplatTouch line of work repurposes the same encoding as a *renderable representation of absolute 3D position*: the Gaussians of a scene are recoloured by their normalized coordinates, so that rendering the recoloured scene produces NOCS maps aligned with the visual views, and a predicted colour can be decoded back into a 3D position. This thesis adopts the encoding for the same reason it is convenient elsewhere: it provides a target that is bounded in $[0, 1]^3$, smooth, and directly interpretable as an image, which makes it well suited to a regression formulation (Chapter 3).

2.5 Tactile contact localization

Tactile contact localization is the inverse of generation: given a tactile observation, recover the 3D location of the contact on the scene. The localization component of SplatTouch [12] is built on contrastive representation learning, a paradigm in which an encoder is trained so that related inputs are close in an embedding space and unrelated ones are far apart, as popularized by methods such as SimCLR [1] and SimSiam [2]. Concretely, a SimSiam-style network jointly embeds tactile samples and the NOCS encoding of 3D position; at inference, a query touch is mapped to its NOCS embedding, and hence to a colour and a 3D coordinate. Localization is therefore obtained *indirectly*, as a by-product of a learned cross-modal embedding, with an accuracy that degrades on weakly textured surfaces and under reconstruction noise.

This thesis takes a deliberately different route to the same target. Rather than learning an embedding and reading the position out of it, the proposed model is trained to predict the NOCS coordinate of a contact *directly* from the tactile image, as a single-stage supervised regression. This casts localization as a coordinate-regression problem of the kind common in vision, where spatial quantities are predicted directly by a convolutional network, and removes the contrastive pre-training stage entirely. The two formulations share the NOCS output space, which makes them directly comparable; the

experiments of Chapter 4 show that, on the scenes considered, the simpler direct-regression model attains a lower localization error than the figures reported for the contrastive approach. To the best of the available knowledge, direct regression to a NOCS target has not previously been studied for tactile localization on Gaussian-Splatting scenes, which is the gap this work addresses.

3 Method

This chapter describes the localization method in full. It states the problem and gives an overview of the pipeline (Section 3.1), defines the NOCS encoding used as the prediction target (Section 3.2), details the dataset and how the ground truth is obtained (Section 3.3), and then presents the network (Section 3.4) and its spatial attention pooling module (Section 3.5). The chapter closes with the training procedure (Section 3.6) and the evaluation metrics (Section 3.7). The guiding design principle is simplicity: localization is solved by a single feed-forward network, trained end-to-end by supervised regression, with no contrastive pre-training and no multi-stage pipeline.

3.1 Problem formulation and overview

Let a scene be reconstructed as a set of 3D Gaussians $\{g_i\}_{i=1}^N$, where each Gaussian carries an explicit 3D position $\mathbf{p}_i \in \mathbb{R}^3$. A contact is described by a tactile observation $\mathbf{x}$, an RGB image produced by a vision-based tactile sensor (Section 3.3). Localization consists in learning a function f_θ that maps the tactile observation to the location of the contact, expressed in normalized NOCS coordinates,

$$\hat{\mathbf{y}} = f_\theta(\mathbf{x}), \quad \hat{\mathbf{y}} \in [0, 1]^3, \quad (3.1)$$

which are decoded into metric coordinates only at evaluation time. The function f_θ is realized as a convolutional network and trained by direct regression against the ground-truth position. End to end, the pipeline maps the tactile image through a ResNet-18 backbone to a spatial feature map, aggregates that map into a single descriptor with a learned spatial attention pooling stage, and regresses the NOCS coordinate with a small multilayer perceptron (Figure 3.3).

3.2 NOCS position encoding

The prediction target is a position written in the Normalized Object Coordinate Space (NOCS) [14] introduced in Section 2.4. Given the axis-aligned bounding box of the scene, with per-axis minimum $\mathbf{p}_{\min}$ and maximum $\mathbf{p}_{\max}$ taken from the Gaussian positions, a position $\mathbf{p}$ is normalized componentwise into a colour $\mathbf{c} \in [0, 1]^3$:

$$\mathbf{c} = \frac{\mathbf{p} - \mathbf{p}_{\min}}{\mathbf{p}_{\max} - \mathbf{p}_{\min}}. \quad (3.2)$$

The inverse mapping recovers the metric position from a predicted colour:

$$\mathbf{p} = \mathbf{c}(\mathbf{p}_{\max} - \mathbf{p}_{\min}) + \mathbf{p}_{\min}. \quad (3.3)$$

The network regresses the colour $\mathbf{c}$ directly, and the inverse transform (Eq. 3.3) is applied only when a physical error is computed. Encoding positions in this way has two practical advantages: the target is bounded in $[0, 1]^3$, which matches a sigmoid output and keeps the regression well conditioned, and positions become renderable as images, since each pixel of a NOCS render is itself a normalized coordinate (Figure 3.1).

3.3 Dataset and ground-truth extraction

Each sample corresponds to a contact and provides several spatially aligned modalities: a tactile image (**touch**), the background reading of the sensor with no contact (**touch.bg**), the visual image (**vision**) and the NOCS image (**nocs**) at the contact region (Figure 3.2). It is worth stating precisely what the model sees: the localization network takes *only* the tactile image as input, while **touch.bg** and **vision** are not used, and the NOCS image is used solely to derive the ground-truth position and is never given to the network.



Figure 3.1: The NOCS encoding. Left: the close-up **vision** image of a contact region. Centre: the corresponding **nocs** image, where each pixel’s RGB colour is the normalized 3D position (Eq. 3.2) of the scene point it observes. Right: the unit cube illustrating the mapping: the x, y, z axes are tinted by the matching RGB channels and the object is coloured by its position inside the normalized scene box.

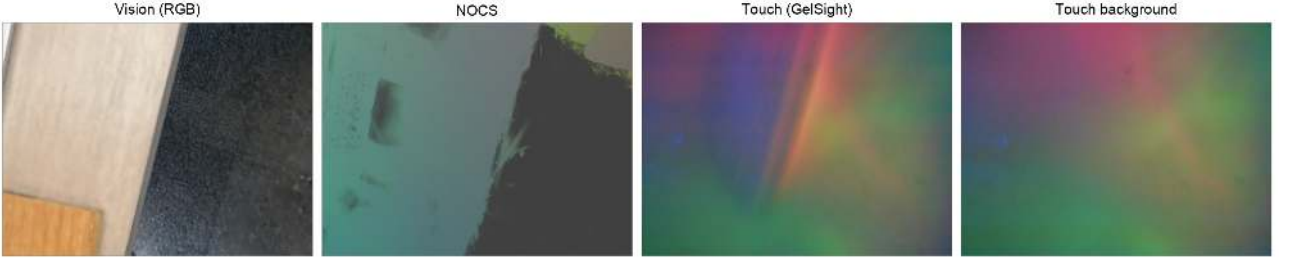


Figure 3.2: Example of a dataset sample showing the aligned modalities: visual image, NOCS encoding, tactile image, and tactile background. Only the tactile image is given to the localization network; the NOCS image is used only to obtain the ground-truth position.

Ground truth. The target position of a contact is read from its NOCS image. The image is loaded, the alpha channel is discarded and the RGB values are scaled to $[0, 1]$. A central crop covering 10% of the width and of the height is taken, and the colours inside it are averaged into a single value $\mathbf{y} \in [0, 1]^3$. Averaging over a small central window, rather than reading a single pixel, makes the target robust to the noise present at the centre of a contact.

Filtering. Not every acquired sample corresponds to a meaningful contact: some NOCS crops are essentially achromatic, with the three channels almost equal, which corresponds to empty or background regions of the recoloured scene. Such samples are detected by the standard deviation of the averaged colour across its three components and discarded when this value falls below 10^{-4} , so that only chromatic, position-bearing contacts are kept. The extracted positions and the resulting splits are cached, so that this preprocessing is performed once and reused across runs.

Preprocessing. Before entering the network, each tactile image is resized to 224×224 , converted to a tensor, and normalized with the ImageNet channel statistics (mean $[0.485, 0.456, 0.406]$, standard deviation $[0.229, 0.224, 0.225]$) expected by the pretrained backbone. No data augmentation is applied; the only transformations are the resize and this normalization.

Scenes. Two real scenes are used, an indoor and an outdoor table. The data is split either by holding out a random subset of contacts or by the periodic protocol of SplatTouch; the exact sizes and the rationale for the two splits are reported in Chapter 4.

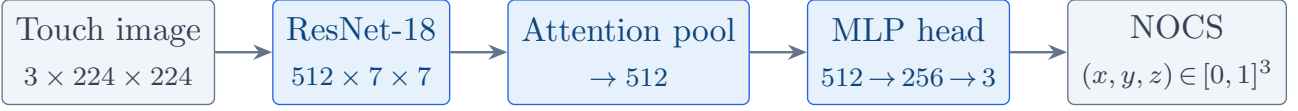


Figure 3.3: Overview of the localization architecture. A ResNet-18 backbone produces a $512 \times 7 \times 7$ feature map; the spatial attention pooling module aggregates it into a single 512-dimensional descriptor, which a small MLP maps to the predicted NOCS coordinate.

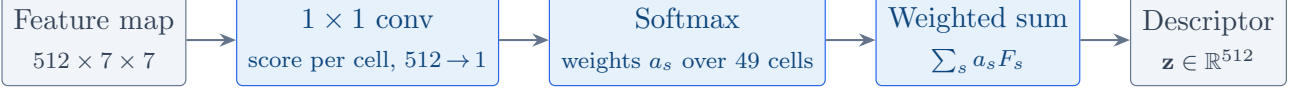


Figure 3.4: Spatial attention pooling in detail. A 1×1 convolution assigns a score to each of the 49 locations of the feature map; a softmax turns the scores into non-negative weights a_s that sum to one, and the descriptor $\mathbf{z}$ is the weighted sum of the per-location feature vectors F_s (Eq. 3.4).

3.4 Localization architecture

The network f_θ has three stages (Figure 3.3).

Backbone. A ResNet-18 [4] pretrained on ImageNet is used as the feature extractor. ResNet-18 is a residual convolutional network whose skip connections ease the training of deeper models; here its final global average pooling and fully connected classifier are removed, and only the convolutional stages up to the last residual block are kept. For a 224×224 input these map the image to a feature map $F \in \mathbb{R}^{512 \times 7 \times 7}$, that is, 512 channels over a 7×7 grid of 49 spatial locations.

Spatial attention pooling. The feature map is collapsed into a single 512-dimensional descriptor by the module described in Section 3.5, which weights the 49 locations by a learned, content-dependent attention map instead of averaging them uniformly.

Regression head. The descriptor is passed to a small multilayer perceptron: a linear layer $512 \rightarrow 256$, a ReLU non-linearity, a linear layer $256 \rightarrow 3$, and a final sigmoid that constrains the output to $(0, 1)^3$, matching the bounded NOCS target. The three outputs are the predicted NOCS coordinate, decoded to metric units with Eq. 3.3 when needed.

3.5 Spatial attention pooling

A convolutional backbone produces a feature map over a spatial grid, but the regression head expects a single vector, so the grid must be aggregated. The common choices are global average pooling, which treats all locations equally, and global max pooling, which keeps only the strongest response per channel. Neither adapts to the input. The module used here instead learns *where* to look.

Given the feature map $F \in \mathbb{R}^{512 \times 7 \times 7}$, a 1×1 convolution maps the 512-dimensional vector at each location to a single score. The 49 scores are normalized with a softmax into weights a_s that are non-negative and sum to one, forming a single attention map shared across all channels. The pooled descriptor is the weighted sum of the per-location feature vectors F_s :

$$\mathbf{z} = \sum_{s=1}^{49} a_s F_s, \quad a_s = \frac{\exp(\phi(F_s))}{\sum_{s'} \exp(\phi(F_{s'}))}, \quad (3.4)$$

where ϕ denotes the 1×1 convolution that produces the score. The weights act as a learned saliency map over the spatial grid, concentrating the descriptor on the locations the network finds informative (Figure 3.4).

On the term “attention”. The module is called *spatial attention pooling*, and the term deserves a precise qualification. The mechanism is a content-dependent, softmax-normalized weighted average of

features: each location scores itself through a shared linear map and the scores compete through the softmax. This is exactly the notion of attention used as a pooling operator in attention-based multiple-instance learning [6], which justifies the name. It is, however, *not* the scaled dot-product self-attention of Transformers [13]: there are no queries, keys or values, no pairwise interaction between locations, and no context mixing; the weights form a single learned saliency map rather than a token-to-token attention matrix. The term is therefore used in the broad sense of a learned soft weighting, not in the Transformer sense.

Design choices. Two choices characterize the module and are examined empirically in Section 4.3. First, the aggregation is a learned weighted *sum*; on this task its accuracy matches that of a plain average or a max aggregation, so it is preferred for the interpretable saliency map it provides rather than for an accuracy gain. Second, the scores are produced by a 1×1 convolution, which considers each location independently; enlarging the kernel to 3×3 does not improve accuracy, because the large receptive field of the backbone already makes neighbouring locations redundant. As a useful side effect, the attention map is directly inspectable, and the distribution of its weights is analysed in Section 4.3 as evidence that the module is spatially selective.

3.6 Training

The network is trained by direct regression of the NOCS coordinate. The loss is the L_1 distance between the predicted and the ground-truth coordinate,

$$\mathcal{L}(\theta) = \|f_\theta(\mathbf{x}) - \mathbf{y}\|_1, \quad (3.5)$$

preferred over an L_2 loss for its lower sensitivity to occasional outliers in the extracted ground truth. Optimization uses Adam with a learning rate of 1×10^{-4} and a batch size of 32, for 50 epochs, with a cosine-annealing schedule that decays the learning rate smoothly to zero over training. The backbone is not frozen: starting from its ImageNet weights, it is fine-tuned end to end together with the pooling module and the regression head. The model is evaluated on the validation set after every epoch, and the checkpoint achieving the lowest validation error is retained. A fixed random seed is used so that the data split and the optimization are reproducible.

3.7 Evaluation metrics

Localization is assessed with several complementary metrics, all computed on the validation set of N contacts.

Mean Euclidean error. The principal metric is the mean over the validation set of the Euclidean (L_2) distance between predicted and ground-truth NOCS coordinates,

$$\text{med} = \frac{1}{N} \sum_{i=1}^N \|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2. \quad (3.6)$$

It is bounded, because both coordinates lie in $[0, 1]^3$, and independent of the scene scale. (In the code and result files this quantity is stored under the key `med`; despite the name it is the mean, not the median.) The mean absolute error along each axis is also reported, to reveal any asymmetry between the x , y and z directions.

Threshold accuracy. The fraction of predictions whose error is below a threshold τ ,

$$p@{\tau} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[\|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2 < \tau], \quad (3.7)$$

reported for $\tau \in \{0.05, 0.10, 0.20\}$. This summarizes how often the prediction is close enough rather than how large the average error is.

Physical error. To obtain an error in centimetres, two routes are available. The first scales the normalized coordinates by the per-axis physical size of the scene before recomputing the statistics above. The second, used for the main physical results, is a Gaussian nearest-neighbour metric: the predicted and the ground-truth colours are each matched to the Gaussian whose normalized colour is closest, and the Euclidean distance between the 3D positions of the two matched Gaussians is measured. Its mean (`gauss_med`) and the fractions of contacts localized within 2, 5 and 10 units ($gp@2$, $gp@5$, $gp@10$) are reported; for the scenes considered one reconstruction unit corresponds to approximately one centimetre.

Bias diagnostic. Finally, the per-axis mean and standard deviation of the predictions are compared with those of the ground truth. This is not an accuracy metric but a diagnostic: a predicted standard deviation far below the true one signals that the model is collapsing towards a constant output, a failure mode that the proposed model avoids (Section 4.2).

4 Experiments and Results

This chapter reports the experimental evaluation of the proposed localization method. It describes the setup (Section 4.1), presents the main quantitative results together with the comparison against naive baselines and against SplatTouch (Section 4.2), analyses the design choices through an ablation study (Section 4.3), and provides a qualitative analysis (Section 4.4). It closes by discussing the limitations of the evaluation (Section 4.5).

4.1 Experimental setup

Evaluation is carried out on two real scenes, an indoor table and an outdoor table. Two data splits are considered. A *random* split sets aside a random subset of samples for validation; this is the easier setting used in the early experiments. The second, hereafter the *SplatTouch split*, follows the periodic protocol introduced by SplatTouch [12]: for every 50 consecutive samples, 40 are used for training, 5 for validation and 5 for test (an 80/10/10 pattern). This periodic split is harder and more honest, because the held-out samples are spread across the whole acquisition rather than drawn at random, and it is used for the headline results unless stated otherwise. Table 4.1 summarizes the two scenes under the SplatTouch split.

Under the SplatTouch split the outdoor scene is markedly harder than the indoor one. With roughly half of the training data it reaches a mean NOCS error of 0.107, against 0.040 indoors, and only 36% of its predictions fall within 0.05 (11.5 cm against 7.1 cm indoors in physical terms). The gap narrows considerably under the easier random split, where the outdoor error drops to 0.041, close to the indoor random-split regime: the approach does transfer to a second scene, but the honest periodic split shows that its accuracy there does not yet match the indoor case. The outdoor validation set is also small (55 samples), so its threshold accuracies carry a larger uncertainty.

Implementation. The model is implemented in PyTorch, with the ResNet-18 backbone and its ImageNet weights taken from torchvision. All experiments use the training configuration of Section 3.6 (Adam with learning rate 1×10^{-4} , batch size 32, 50 epochs, cosine annealing). Runs execute on a single GPU when available and fall back to CPU otherwise. The extracted ground-truth positions and the data splits are cached on disk and reused across runs, and a fixed random seed makes both the split and the optimization reproducible.

4.2 Main results

Table 4.2 reports the main results on the two scenes under the SplatTouch split. On the indoor scene the model achieves a mean NOCS error of 0.040, with 84% of predictions within 0.05 and 86% within 0.10, corresponding to a physical mean error of 7.06 cm. The outdoor scene is markedly harder: a mean NOCS error of 0.107, only 36% of predictions within 0.05, and an (indicative) physical error of 11.5 cm.

Beyond the aggregate error, two finer checks support these numbers. The mean absolute error is of a similar magnitude along the three axes, largest along the y direction, so no single coordinate dominates the residual. The bias diagnostic of Section 3.7 shows that the mean of the predictions matches that of the ground truth, and that their per-axis standard deviations are comparable, with the predicted spread only slightly below the true one. The model therefore covers the actual range of contact positions rather than regressing towards their average, which is the behaviour expected of a working regressor.

Table 4.1: Scenes used in the evaluation: split sizes and error under the SplatTouch split, as a normalized NOCS error and a physical error in centimetres (`gauss_med`).

Scene	Train	Val	NOCS	cm (<code>gauss_med</code>)
Indoor table	891	114	0.040	7.06
Outdoor table	462	55	0.107	11.54

Table 4.2: Main localization results on the two scenes (SplatTouch split; indoor 114 / outdoor 55 validation samples). The outdoor physical error is indicative (Section 4.5).

Metric	Indoor	Outdoor
Mean NOCS error	0.040	0.107
$p@0.05$	84.2%	36.4%
$p@0.10$	86.0%	60.0%
$p@0.20$	98.2%	85.5%
Physical mean (<code>gauss_med</code>)	7.06 cm	11.54 cm
$gp@2$	52.6%	14.5%
$gp@5$	70.2%	32.7%
$gp@10$	79.8%	50.9%

4.2.1 Random split versus SplatTouch split

The choice of split has a large effect on the reported accuracy. Under the random split, the same model attains a mean NOCS error of 0.019 with 96% of predictions within 0.05 and a physical mean of 3.58 cm. Under the periodic SplatTouch split these figures become 0.040, 84% and 7.06 cm (Table 4.3). The same pattern holds on the outdoor scene, where the error grows from 0.041 / 5.05 cm under the random split to 0.107 / 11.54 cm under the SplatTouch split. The gap does not reflect a regression of the model but the higher difficulty of the periodic holdout, in which validation contacts are systematically interleaved with, rather than randomly drawn from, the training distribution. The SplatTouch-split numbers are therefore regarded as the honest estimate of localization accuracy, and the random-split numbers are reported only for completeness and for comparison with the earlier experiments.

4.2.2 Comparison with baselines

To confirm that the model learns a genuine mapping rather than exploiting a constant prior over contact positions, it is compared against two naive baselines that always predict the mean or the median training position (Table 4.4). The proposed model is about $3.7\times$ more accurate than these baselines in mean error on the indoor scene and about $1.9\times$ on the outdoor scene; the gap on threshold accuracy is even larger, as the constant baselines rarely fall within the tightest threshold whereas the proposed model reaches 84% at $p@0.05$ indoors and 36% outdoors.

4.2.3 Comparison with SplatTouch

Table 4.5 places the physical mean error of the proposed model next to the localization figures reported for SplatTouch [12]. Since the SplatTouch figures are themselves *averages over 16 scenes*, the most directly comparable quantity is the mean of the proposed model over its two evaluated scenes, 9.30 cm: this is already about 20% below the best reported SplatTouch configuration (11.65 cm) and about 29% below its non-augmented NOCS variant (13.02 cm). On the indoor scene alone the model reaches 7.06 cm, and on the harder outdoor scene 11.54 cm. The comparison nonetheless remains indicative rather than conclusive: the proposed average is over two scenes against sixteen, the two methods differ (contrastive representation learning versus direct regression), and the outdoor centimetre figure is itself indicative (Section 4.5).

Table 4.3: Effect of the data split on both scenes. The periodic SplatTouch split is harder and is used as the reference setting; outdoor physical errors are indicative.

Scene	Split	Mean NOCS	$p@0.05$	Physical mean
Indoor	Random	0.019	96%	3.58 cm
Indoor	SplatTouch	0.040	84%	7.06 cm
Outdoor	Random	0.041	73%	5.05 cm
Outdoor	SplatTouch	0.107	36%	11.54 cm

Table 4.4: Comparison with naive baselines on both scenes (SplatTouch split, same validation samples per scene).

Scene	Method	Mean	$p@0.05$	$p@0.10$	$p@0.20$
Indoor	Mean predictor	0.149	0.0%	26.3%	74.6%
	Median predictor	0.147	8.8%	34.2%	71.9%
	Proposed model	0.040	84%	86%	98%
Outdoor	Mean predictor	0.198	0.0%	14.5%	49.1%
	Median predictor	0.198	0.0%	14.5%	54.5%
	Proposed model	0.107	36%	60%	85%

4.3 Ablation study

The effect of the main design choices of Chapter 3 is studied in Table 4.6, on the indoor scene. The pooling-aggregation and crop-fraction rows are measured under the SplatTouch split; the kernel-size rows were measured under the random split (hence the lower absolute errors) and should be read for the relative comparison only.

Pooling aggregation. The spatial aggregation has little effect on accuracy. Trained under identical settings, a uniform average (0.038) and a max aggregation (0.053) match the attention pooling used in the main model (0.040) to within run-to-run variance, and none collapses towards a constant prediction. The attention pooling is therefore retained not because it lowers the error, but because it yields an inspectable, content-dependent saliency map over the tactile image (analysed below), which a uniform average cannot provide.

Kernel size. Computing the pooling scores with a 1×1 convolution performs on par with a 3×3 convolution (0.019 versus 0.018 mean, within noise). The simpler 1×1 kernel is therefore kept: the large receptive field of the backbone already makes neighbouring locations redundant.

NOCS crop fraction. Increasing the central crop used to compute the ground-truth position from 10% to 20% does not help (0.043 versus 0.040), so 10% is kept.

A direct inspection of the learned weights confirms the spatial selectivity of the pooling module: across the dataset the maximum weight per sample is about 0.10 on average and exceeds 0.30 for the most peaked inputs, in every case well above the uniform value of $1/49 \approx 0.02$, while the smallest weights are close to zero. The module genuinely concentrates on specific spatial regions rather than pooling uniformly, and the attended region shifts with the input (Figure 4.1).

4.4 Qualitative results

An interactive 3D visualizer was developed to assess the predictions qualitatively. It renders the reconstructed Gaussian scene together with the ground-truth contact points, the predicted points, and the error segments connecting each prediction to its target (Figure 4.2). The tool makes it possible to inspect, scene by scene, where the model is accurate and where it fails, and it proved useful

Table 4.5: Comparison with SplatTouch [12] (Euclidean error in cm). The SplatTouch figures are averages over 16 scenes; the proposed model is evaluated on two scenes, and its scene-averaged value (in bold) is the most directly comparable quantity. The outdoor figure is indicative (Section 4.5).

Method	Error (cm)
SplatTouch: Random baseline	56.47
SplatTouch: RGB + Touch	22.44
SplatTouch: NOCS + Touch	13.02
SplatTouch: NOCS + Touch (Augmented)	11.65
Proposed model: indoor	7.06
Proposed model: outdoor	11.54
Proposed model: mean of the two scenes	9.30

Table 4.6: Ablation study on the indoor scene. Pooling and crop rows use the SplatTouch split; the kernel-size rows use the random split (relative comparison only).

Design choice	Variant	Mean	$p@0.05$
Pooling aggregation	attention (used)	0.040	84%
	uniform average	0.038	83%
	max	0.053	68%
Kernel size (random split)	1 (used)	0.019	95%
	3	0.018	94%
NOCS crop fraction	10% (used)	0.040	84%
	20%	0.043	78%

for diagnosing failure cases during development, such as systematic offsets between predictions and the reconstructed geometry.

4.5 Limitations

The evaluation has limits that temper its conclusions. The physical (centimetre) metric depends on the Gaussian reconstruction used to decode the NOCS colours, and this dependence is not negligible: recomputing the indoor metric with an independently trained reconstruction of the same scene changes it from 7.06 to 4.75 cm, a swing of about 30%, even though the predictions are identical. The indoor figure reported here uses the exact reconstruction that defined the indoor NOCS ground truth and is therefore frame-consistent; the outdoor figure (11.5 cm) instead relies on a reconstruction trained separately, since the original outdoor splat was not available, and is consequently indicative rather than exact. The normalized NOCS errors are unaffected by this, as they are computed purely in colour space. The validation sets are small, especially outdoors (55 contacts), so the threshold accuracies carry a non-negligible uncertainty. The comparison with SplatTouch is indicative rather than rigorous, for the reasons given in Section 4.2.3: a two-scene average against a sixteen-scene one, and two different methods. Finally, the ground truth is itself derived from the NOCS encoding (Section 3.3), so any inaccuracy in the recoloured reconstruction propagates into both training and evaluation.

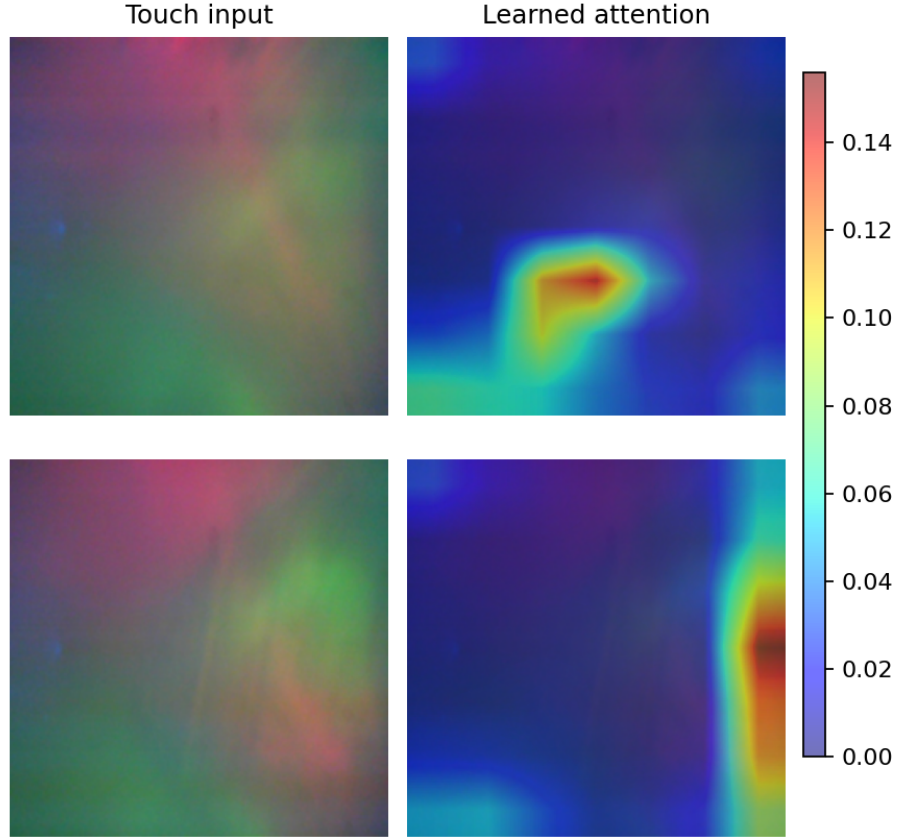


Figure 4.1: Learned attention maps for two tactile inputs (indoor scene). For each input the softmax weights a_s over the 7×7 grid are upsampled and overlaid on the touch image. The weights concentrate on a small region (peak ≈ 0.15 against the uniform $1/49 \approx 0.02$) and the attended region moves with the input, evidence that the pooling is spatially selective and content-dependent.

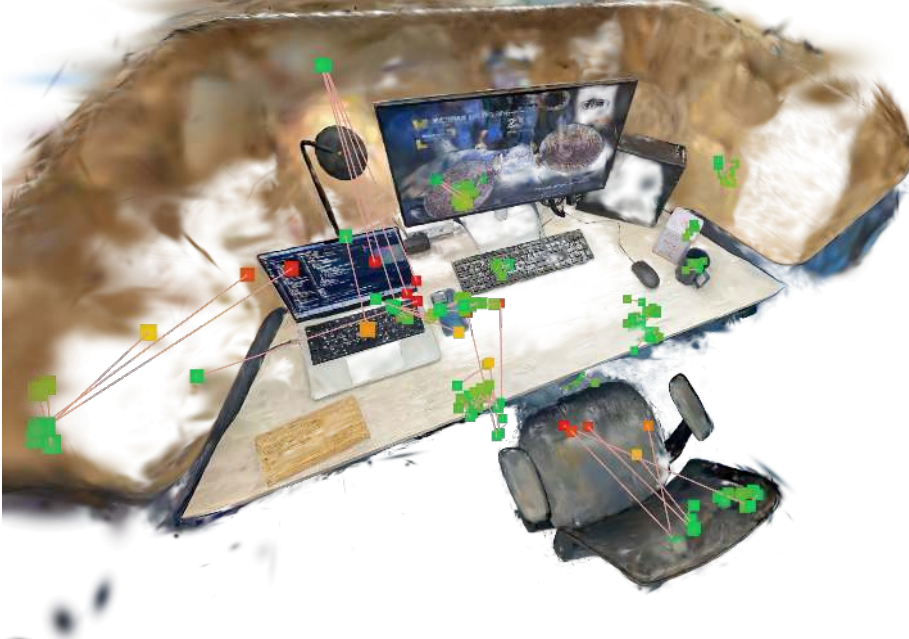


Figure 4.2: The interactive 3D visualizer on the indoor scene, showing all validation contacts at once. The textured surface is the reconstructed Gaussian-Splatting scene; the ground-truth contacts are green, while each prediction is coloured by its error, from green when it coincides with the ground truth, through yellow, to red when it is far from it; each segment links a prediction to its target, so its length is the localization error.

5 Conclusions

This thesis investigated tactile contact localization on Gaussian-Splatting scenes, the inverse direction of the visuo-tactile chain that SplatTouch [12] addresses with a contrastive model. Instead of learning a contrastive representation, the localization task was cast as a direct supervised regression in Normalized Object Coordinate Space (NOCS), using a lightweight architecture made of a ResNet-18 backbone, a learned spatial attention pooling stage and a small regression head. The method was evaluated on two real scenes, under both an optimistic random split and the harder SplatTouch split, reporting a normalized NOCS error together with a physically meaningful error in centimetres, and its main design choices were validated through an ablation study.

5.1 Contributions and impact

The contributions of this work can be summarized as follows.

- **A direct-regression formulation of tactile localization.** The contact is mapped to its NOCS colour, and hence to a 3D position, by a single supervised regression, rather than by the contrastive (SimSiam-style) representation learning of SplatTouch [12] – a conceptually simpler and computationally lighter paradigm.
- **A lightweight architecture.** A ResNet-18 backbone is followed by a learned spatial attention pooling stage, which aggregates the feature grid as a weighted sum, and by a compact regression head, keeping the model small and easy to train.
- **A reproducible evaluation protocol.** The method is assessed on two real scenes, under both the random and the SplatTouch split, with a normalized NOCS error and a physical centimetre error, against naive baselines, and with an ablation; per-scene splits are generated and cached deterministically, so the experiments are fully reproducible.
- **Empirical design insights.** An ablation shows that the localization accuracy is largely insensitive to the spatial-aggregation strategy (attention, average and max pooling perform comparably), while a 1×1 pooling kernel and a 10% NOCS crop are identified as the preferable, simplest choices.
- **An interactive 3D visualizer** that overlays the predicted and ground-truth contacts, and their error, on the reconstructed Gaussian scene, together with an analysis of how the centimetre metric depends on the reconstruction used to decode the NOCS colours.

In quantitative terms, on the indoor scene the model localizes contacts with a mean error of 0.040 in NOCS space and 7.06 cm in metric space, with more than four predictions out of five within the tightest accuracy threshold; on the harder outdoor scene the error rises to 0.107 and an indicative 11.54 cm. Averaged over the two evaluated scenes the metric error is 9.30 cm. As the SplatTouch figures are themselves averages over sixteen scenes, this scene-averaged value is the most directly comparable quantity: it is about 20% below the best localization configuration reported by SplatTouch [12] (11.65 cm) and about 29% below its non-augmented NOCS variant (13.02 cm). This comparison remains indicative, since it places a two-scene average obtained with a different method against a sixteen-scene one. Against naive constant-prediction baselines the indoor mean error is reduced by about 73% (a $3.7\times$ improvement), with the gap on threshold accuracy even larger (84% against 0% at the 0.05 threshold). The outdoor result confirms that the approach transfers to a second scene while leaving clear room for improvement. Figure 5.1 shows three representative indoor localizations spanning this range, from a near-perfect prediction to a hard, high-error case.



Figure 5.1: Three representative localizations on the indoor scene, selected to span the observed error range, from a near-perfect prediction to a high-error case. The ground-truth contacts are green, while each prediction is coloured by its error, from green when it coincides with the ground truth, through yellow, to red when it is far from it; each segment is the localization error, shown on the reconstructed Gaussian-Splatting surface.

The broader impact is that, for tactile localization on these scenes, a single supervised regression stage with an explicitly learned spatial weighting is sufficient to match, and here to surpass, a more elaborate contrastive pipeline. Localization can thus be obtained with a simpler, cheaper and more easily reproducible model, which contributes the localization half of the visuo-tactile chain whose generation half SplatTouch already provides, bringing the closed “see-and-feel” loop one step nearer.

5.2 Future work

Several directions remain open. The outdoor physical metric, obtained here from a separately trained reconstruction, should be recomputed from the same Gaussian model that renders the NOCS ground truth, so as to remove the residual frame mismatch; aligning the colour decoding of the metric with the encoding actually used to render the labels is a further, low-cost way to make the centimetre figures fully consistent. The evaluation should also be extended to a larger and more diverse set of scenes to confirm generalizability, and robustness on featureless surfaces, where localization is known to be harder, deserves dedicated study. A natural longer-term step is to couple the regression-based localization model with the generation model of SplatTouch and to compare, within the same closed visuo-tactile chain, the contrastive and direct-regression formulations of localization.

Bibliography

- [1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.
- [2] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [3] Yiming Dou, Fengyu Yang, Yi Liu, Antonio Loquercio, and Andrew Owens. Tactile-augmented radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.
- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [6] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International Conference on Machine Learning (ICML)*, 2018.
- [7] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4), 2023.
- [8] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, Dinesh Jayaraman, and Roberto Calandra. DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters (RA-L)*, 5(3), 2020.
- [9] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [10] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.
- [11] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41(4), 2022.
- [12] Antonio Luigi Stefani, Niccolò Bisagno, Nicola Conci, and Francesco De Natale. SplatTouch: Explicit 3D representation binding vision and touch. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 118–127, 2025.
- [13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

- [14] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6D object pose and size estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [15] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks*, 2022.
- [16] Wenzhen Yuan, Siyuan Dong, and Edward H. Adelson. GelSight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12), 2017.