Semi-Supervised Disentanglement of Tactile Contact Geometry from Sliding-Induced Shear

Anupam K. Gupta and Nathan F. Lepora .

***Abstract*—The sense of touch is fundamental to human capabilities of exploration and manipulation. When mimicked in robotic touch, particularly by use of soft optical tactile sensors, it suffers from distortion caused by motion-dependent shear. This complicates tactile tasks like shape reconstruction and exploration that require information about contact-geometry. In this work, we pursue a semi-supervised approach to remove shear while preserving contact-only information. We validate the approach by showing a match between the model-generated unsheared images with their discrete tap counterparts. In addition, the model-generated unsheared images give faithful reconstruction of contact-geometry otherwise masked by shear, along with robust estimation of object pose later used for sliding exploration and full shape reconstruction of several planar shapes. We show that our semi-supervised approach achieves comparable performance to its fully supervised counterpart across all validation tasks with an order of magnitude less supervision. The semi-supervised method is thus more computational and data efficient, and we expect it will have broad applicability to wide range of complex tactile exploration and manipulation tasks performed via a shear-sensitive sense of touch.**

***Index Terms*—Robotic Touch, Disentanglement, Shear, Semi Supervision, Object Reconstruction.**

I. Introduction

The sense of touch is ubiquitous in everyday life. It allows us to interact, explore and manipulate the environment around us using physical contact. The physical contact allows us to measure the environment directly, which is not possible with other sensory modalities like vision. However, this physical contact also results in entangling of relevant information like contact geometry with the manner in which the sensor-stimuli contact is established. For example, when sliding or rubbing fingers across tactile stimuli, soft tactile sensors are inevitably distorted by motion-dependent shear, making their response history-dependent and generalization hard across tactile tasks. This necessitates the removal of motion-induced global shear while preserving the sensor distortions due to the contact-geometry or any other attribute of interest. This issue plagues all shear-sensitive tactile sensors, in particular camera-based soft tactile sensors that use the displacement of markers to encode stimuli attributes [1]–[4], complicating tactile-dependent tasks like shape modelling and exploration.

Most recently, this shear problem was tackled in [5], [6] for sliding exploration of complex 2D and 3D objects by using a supervised deep learning model to learn mappings between shear-distorted data and the target pose. This approach to learn insensitivity to shear, however, requires training of separate models for each tactile task; for example contact reconstruction or shape exploration. A recent study, by the authors of this paper, instead used a supervised deep learning model to remove shear distortion from the original sheared tactile image so that quantities of interest (e.g. pose) can be predicted from unsheared tactile data in a task-agnostic manner, which is more computational and data efficient [7]. However, because these approaches are supervised, they suffer from costly burden of requiring annotated data for learning. This limitation can be overcome by the use of self- and semi-supervised approaches that can learn rich representations from uncurated datasets, which are thus more *sample-efficient*. This mode of learning is closer to human learning, as we likewise learn with only small amounts of labelled data and a vast resource of self-supervised experiences.

In this work, we propose a semi-supervised learning approach to remove shear from shear-distorted tactile images that uses only a small subset of annotated data (10%) to achieve comparable performance to the previous fully supervised approaches [7]. Self-supervision with tactile data differs from computer vision [8]–[10] because it lacks attributes like colour, texture and the large-scale structure of natural image data. Furthermore, the present work relies only on tactile data without recourse to techniques that correlate data across multiple sensory modalities to generate training signals without external supervision [11]–[13] (see Background). In the presence of these limitations, we achieved self-supervision by first disentangling sliding data into its contact-only and shear components, followed by recombining the shear component with paired and novel canonical (non-sheared or vertical tap) data to generate training signals without external supervision (see Methods).

To validate our approach, we: 1) demonstrate a good match (92.5%) between model-generated unsheared tactile images and their non-sheared (vertical tapping) counterparts using the structural similarity index measure; 2) reconstruct the approximate stimulus contact geometry from model-generated unsheared images that would otherwise be masked by sliding-induced shear (Fig. 3, top row); 3) predict the local object pose from a pose-prediction network [7] trained on tapping data with minimum global shear (Table I); 4) this pose-prediction model was then used for sliding exploration of multiple planar objects (Fig. 3, second row), combining 2) and 3) to reconstruct full object shapes for several planar objects (Fig. 3, third row).

II. Background

Camera-based soft optical tactile sensors like the TacTip [1] and other marker-based sensors [2], [4] are vulnerable to sliding-induced global shear that distorts the geometry of contact. Local

This work was supported by an award from the Leverhulme Trust on "A biomimetic forebrain for robot touch" (RL-2016-39)

AKG and NFL are with the Faculty of Engineering and Bristol Robotics Laboratory, University of Bristol, U.K.

Corresponding author: anupamkg@vt.edu

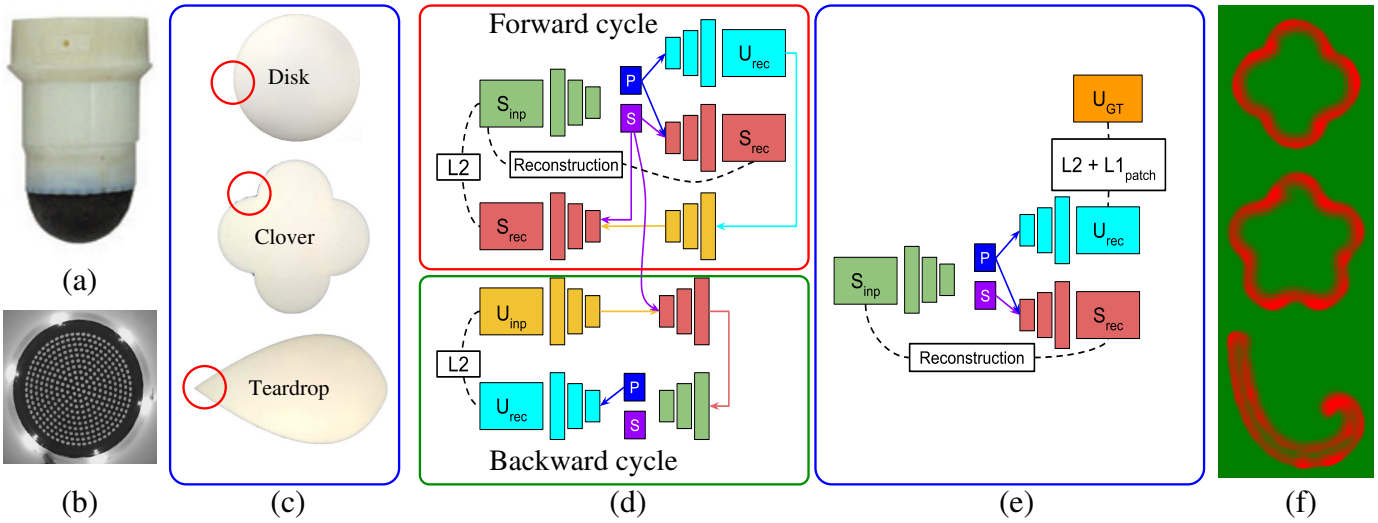

Fig. 1. **Sensor, stimuli, model schematic and object shape reconstructions.** (a) TacTip sensor. (b) TacTip internal sensing surface. (c) 3D printed tactile stimuli. Red circles highlight approximate location of data collection for training. (d) & (e) Schematic of the model architecture for self-supervised and supervised training phases respectively. S_{inp} and U_{inp} represent unpaired sheared and unsheared (vertical tap) input samples. S_{rec} and U_{rec} represent sheared and unsheared (vertical tap) reconstructed samples. U_{GT} represents unsheared (vertical tap) ground truth samples used for fine-tuning sheared-unsheared mapping in the supervised phase. P & S represent pose (contact-only) and shear latent codes respectively. Detailed architectures in Fig. 2. (f) Representative examples of full object shape reconstructions. For full set of results, see Fig. 3.

shear caused by stimulus imprint on the sensor is represented by the lateral displacement of the markers, which is then superimposed upon a global shear caused by contact motion such as sliding. This distortion, by masking contact geometry, complicates challenging tactile tasks like shape reconstruction and tactile exploration, which require spatial information about the local geometry of an object free from motion-induced distortion.

Recent works that use TacTip sensor (also used here) addressed this shear problem by using a shear-insensitive pose-prediction deep network to slide around complex planar shapes [5] and complex 3D objects [6], [14]. Shear-insensitive training was achieved by mapping the shear-distorted samples directly to target pose outputs. While effective, this approach necessitated separate training for each new tactile task, for example surface- and contour-following [14], greatly complicating the hyper-parameter tuning for accurate network performance [6], and could not be applied to shape reconstruction. To overcome these drawbacks, the authors of the present paper instead focused on learning a mapping between shear-distorted tactile images and their unsheared (vertical discrete tap) counterparts with a fully-supervised convolutional neural network [7]. Once learned, this mapping can be reused to remove shear for any number of downstream tasks, making this approach computational and data efficient. However, the supervised nature of that approach necessitated complex labelled-data collection. Here, we improve upon that work by proposing a semi-supervised model that achieves comparable performance to the previous supervised baseline, by training first in self-supervised phase followed by a brief supervised phase (with one-tenth of labelled data used in [7]).

Tactile data differs from natural image data in being low-dimensional due to absence of attributes like color and texture. In addition, tactile data also lacks the large-scale global structure of natural image data. This is primarily due to the localised nature of tactile sensing unlike vision or audition which are global sensing modalities with much wider field of view. This increases the complexity of generating a supervisory signal in the absence of external supervision; for example, in computer vision, supervisory signals in prior studies have been generated by: 1) learning to predict withheld data dimensions at the output, such as in [8] two sub-networks are trained that are each tasked with predicting one subset of the data channels (color or lightness channel) from another (lightness or color channel); 2) image completion conditioned on the immediate surroundings, as in [10] in which the input image has a missing part that needs to be predicted at the output based on the surroundings; and 3) predicting the correct arrangement of image subsections, as in [9] in which the input image is first split into N rectangular tiles whose locations are shuffled to break spatial structure, and the network is then tasked to predict the original image from a shuffled input image. These methods cannot be readily extended to tactile data due to its low-dimensional structure and lack of global structure as discussed previously.

Another approach to generate supervisory signals in absence of external supervision is by correlating data from multiple sensory modalities like audio and vision [11] or touch and vision [12], [13]. In this work, however we rely on the information encoded by the tactile sensor alone that excludes above approaches.

To generate a supervisory signal given the above-discussed limitations, here we use an approach of disentangling the sheared sensor response into a contact geometry (pose code) component and a second component due to motion-induced shear (shear code), as in [7]. Then we recombine the shear code either with the corresponding pose code (shear code and pose code extracted from same sheared sample) or novel pose codes to synthesize input or novel sheared samples. In consequence, we can generate a supervisory signal for the self-supervised

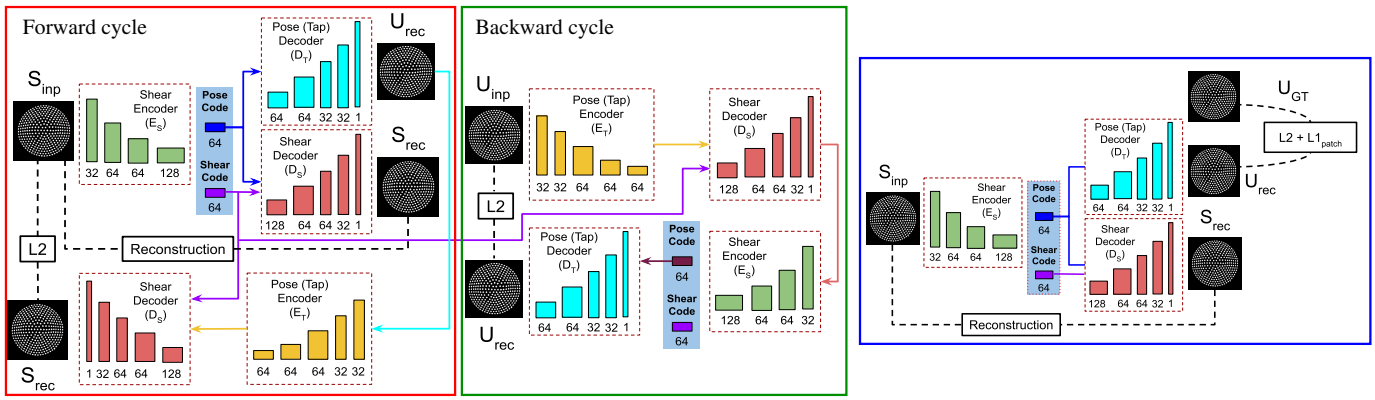

Fig. 2. **Network architecture.** (a) Self-supervised training: The sheared (S_{inp}) – unsheared (U_{rec}), unsheared (U_{inp}) – sheared (S_{rec}) and sheared (S_{inp}) – sheared (S_{rec}) mappings M, K & N are learned via cycle consistency loss [15] and reconstruction loss. The disentanglement of sensor response into its pose (contact-only) component and that due to motion-induced shear enables synthesizing of novel sheared sensor responses by combination of shear component with novel canonical (non-sheared) sensor responses U_{inp}. This combined with cycle consistency loss [15] is used to train mappings M & K through self-supervision. (b) Supervised training: The mappings M & N previously learned in (a) are further fine-tuned via $L2$ loss (M), $L1_{patch}$ (M) and reconstructions loss (N). $L1_{patch}$, computed between 100 random patches (20×20 pix.) of model-generated unsheared images UI_{out} and its canonical (non-sheared) counterparts U_{GT} (ground truth vertical tap images), is used to enforce local image compliance.

model training phase (see Fig. 2 (a)) using cycle consistency loss [15]. This semi-supervised phase was then followed by a brief supervised training phase (see Fig. 2 (b)) that used paired non-sheared (vertical tap) and sheared data to fine tune the representations.

III. Methods

A. Experimental Setup

The setup used in this work is similar to that of [7], where a soft biomimetic optical tactile sensor – the TacTip [1] – is mounted on an industrial robot arm (ABB IRB120) as an end effector. The tactile sensor encodes tactile information in the shear displacement of marker pins on the inner side of the sensing surface caused by the soft sensor skin deformation. The morphology of this TacTip consisted of a 3D-printed soft rubber-like hemispherical dome (40 mm dia., TangoBlack+) with its inner surface covered by 331 hard white tip pins arranged in a concentric circular grid fashion (Figs 1a,b). The soft dome was filled with an optically-clear silicon gel to mimic the compliance of a human fingertip. An RGB camera (ELP 1080p module) captures the inner surface of the sensor dome. For more details, we refer to Ref. [1].

For test stimuli, we used seven distinct planar shapes with various morphologies, including two acrylic spiral shapes that differ in frictional properties to the other five 3D-printed ABS plastic shapes (Fig. 3, second row). For model training we used three shapes: the circular disk, clover and teardrop shown in Fig. 1 (c) (red circles show approximate location of data collection). All shapes were securely fastened to the workspace to prevent accidental motion during experiments.

B. Data Collection

This work proposes a semi-supervised deep learning model, as an alternative to the fully-supervised model proposed in [7] to remove the motion-induced shear distortion from tactile images. This motion-induced shear is governed by the manner of contact between the sensor and the object, distorting the sensor response induced by the stimulus geometry during the contact.

Our semi-supervised model is trained in two phases. In the first phase, the model is trained in a self-supervised fashion on a set of unpaired canonical (non-sheared) tactile images UI_{inp} (collected by vertically tapping on the stimulus) and a set of (sheared) tactile images SI_{inp} with random global shear (collected by sliding across the stimulus). This is then followed by a second training phase in which the model is trained on a subset (10%) of paired canonical and sheared data that was previously used to train the fully-supervised counterpart [7]. To ease data collection, we reused the paired data collected in [7] to train the fully-supervised model. However, we randomized the pairings (every epoch) between the canonical and sheared data for all but a small subset (10%) of data held back for a brief supervised training phase.

To collect non-sheared canonical tap data, the sensor is brought in the contact with the stimuli vertically to minimise global shear. For sliding data, the sensor is brought to the target pose while in contact with the object from a random offset location along either the x- or y-axis or both lateral directions simultaneously. To obtain paired data, the tap and sliding data are collected for same relative target poses between the sensor and stimulus (Fig. 1c). In total, a set of 200 random poses were collected, with the sensor location relative to the initial contact location randomly sampled from a uniform distribution spanning a range $[-5,5]$ mm in the two lateral directions (x- and y-axes), $[-45,45]$ deg in yaw ($\theta \boldsymbol{n}_z$) and $[-6,-1]$ mm in depth (z-axis). Finally, the tactile images captured are cropped and subsampled to a 256×256-pixel region and binarized to minimise any effect of lighting changes inside the sensor.

*1) **Canonical Data**:* The canonical tapping dataset had 30,000 samples in total: 50 instances of each of the 200 random poses for the three stimuli shapes used for training (Fig. 1 (c)). The instances were generated by varying the indentation depth (z-axis) randomly between $[-1,1]$ mm from a set indentation

depth. This dataset was used as the target (reference) to which the sliding data should be restored.

*2) **Sheared Data**:* The sheared data had 90,000 samples in total: 150 instances for each of the 200 random poses for each of the three stimuli shapes (Fig. 1 (c)). To generate instances, the sensor was first brought in contact with the stimulus at a lateral offset location sampled randomly between $[-2.5, 2.5]$ mm laterally (along the x- or y-axes or both), then sliding to the target pose.

*3) **Training and Test Data**:* The canonical and sheared data were first partitioned randomly into training, validation and test sets in the ratio 60:20:20. Thus, for each stimulus shape (Fig. 1), 120 poses were aligned to training set and 40 poses each to validation and test sets.

The paired training set (canonical-sheared) was further randomly split in the ratio 90:10. The larger subset (90%) was shuffled to generate random paired data for the first self-supervised phase of model training. The shuffling was repeated before every training epoch to increase variability in training data. The other subset (10%) kept the pairing between sheared data and a canonical instance of the same pose, for the supervised phase of training.

C. Formulation

Our goal is to learn mapping functions between two domains, tap (T) and sheared (S), given training samples comprising canonical (tap) data $\{x_t\}_{i=1}^{N}$ where $x_t \in X_t$ and shear-transformed training samples $\{x_s\}_{i=1}^{N}$ where $x_s \in X_s$. Our model includes three mappings, $M : X_s \to X_t$, $N : X_s \to X_s$ and $K : X_t \to X_s$. Our objective has four loss terms: *reconstruction loss* to match the input S_{inp} and reconstructed sheared samples S_{rec} to learn the mapping N, an *L2 loss* to match the reconstructed U_{rec} and canonical samples U_{GT} in the supervised training phase to fine tune the learned mapping M, an *L1 patch loss* to enforce image similarity locally between the reconstructed U_{rec} and canonical samples U_{GT} in the supervised training phase to fine tune the learned mapping M and *cycle consistency loss* to learn mappings M, N and K and prevent mappings M and K from contradicting each other. The detailed mathematical formulation is given below.

The encoder E_S (Fig. 2) disentangles the latent representation ($\mathbf{z} \in \mathbf{Z}$) of input sheared samples S_{inp} into pose code ($\mathbf{z_t} \in \mathbf{Z_t}$) and shear code ($\mathbf{z_s} \in \mathbf{Z_s}$) respectively. This disentanglement of sensor response not only aids in learning of robust mapping M as shown in [7], but also allows recombination of unpaired shear and pose codes to synthesize novel sheared samples necessary for generating training signals via *cycle consistency loss* [15] without external supervision.

a) Cycle Consistency Loss: We apply the cycle consistency loss $\mathbb{L}_{\text{cyc}}$ to train the model (Fig. 2) in the self-supervised phase of training. This loss aids learning of mappings M and K, and prevents them contradicting each other. Enforcing cycle-consistency reduces the space of possible mapping functions that can be learned. For each image x_s from the domain S, the forward cycle should be able to bring the input image x_s back to original image i.e. $x_s \to M(x_s) \to K(M(x_s)) \approx x_s$. Similarly, the backward cycle should satisfy $x_t \to K(x_t) \to M(K(x_t)) \approx x_t$ for each $x_t \in T$. This behaviour is enforced via the *cycle consistency loss*:

$$\begin{aligned}\mathbb{L}_{\text{cyc}}(M,K) = \mathbb{E}_{x_s \sim X_s}[\|x_s - K(M(x_s))\|_2] \\ + \mathbb{E}_{x_t \sim X_t}[\|x_t - M(K(x_t)\|_2].\end{aligned} \tag{1}$$

b) Reconstruction Loss: We apply the reconstruction loss $\mathbb{L}_{\text{rec}}$ to learn the mapping $N : X_s \to X_s$ between input sheared samples S_{inp} and reconstructed output sheared samples S_{rec} through a decoder D_S, in both self-supervised and supervised phases of training. This loss aids the disentanglement of sheared-input samples into pose and shear codes:

$$\mathbb{L}_{\text{rec}}(N) = \mathbb{E}_{x_s \sim X_s}[\|x_s - N(x_s\|_2]. \tag{2}$$

c) L2 Loss: We apply an L2 loss $\mathbb{L}_{\text{sup}}$ to fine tune the mapping $M : X_s \to X_t$:

$$\mathbb{L}_{\text{sup}}(M) = \mathbb{E}_{x_s \sim X_s, x_t \sim X_t}[\|x_t - M(x_s)\|_2]. \tag{3}$$

d) $L1_{patch}$ Loss: We apply $L1_{patch}$ loss $\mathbb{L}_{\text{patch}}$ to fine tune the mapping $M : X_s \to X_t$. Unlike $\mathbb{L}_{\text{sup}}$, this loss helps enforce image similarity locally, because the change in tactile sensor response with contact conditions is predominantly local in nature. The loss was computed between 100 random crops of size 20×20 of model-generated unsheared images $U_{rec}(M(x_s))$ and its canonical (tap) counterparts $U_{GT}(x_t)$.

$$\mathbb{L}_{\text{patch}}(M) = \mathbb{E}_{x_s \sim X_s, x_t \sim X_t}[\|\text{crops}(x_t) - \text{crops}(M(x_s))\|_1]. \tag{4}$$

e) Full Objective: Our full objectives for Encoders E_S, E_T and Decoders D_S, D_T are:

$$\mathbb{L}_{E_S} = \mathbb{L}_{\text{cyc}}(M,K) + \mathbb{L}_{\text{rec}}(N) + \mathbb{L}_{\text{sup}}(M) + \lambda \cdot \mathbb{L}_{\text{patch}}(M) \tag{5}$$

$$\mathbb{L}_{E_T} = \mathbb{L}_{\text{cyc}}(M,K) \tag{6}$$

$$\mathbb{L}_{D_T} = \mathbb{L}_{\text{cyc}}(M,K) + \mathbb{L}_{\text{sup}}(M) + \lambda \cdot \mathbb{L}_{\text{patch}}(M) \tag{7}$$

$$\mathbb{L}_{D_S} = \mathbb{L}_{\text{rec}}(N) \tag{8}$$

where $\lambda = 0.1$ is the relative loss scale factor.

D. Implementation

1) Network Architecture:

***Shear Encoder**:* The shear encoder E_S compresses the sheared input S_{inp} (256×256-pixel tactile image) using four convolutional (Conv) layers, each followed by batch normalization (BN) and rectified linear unit (ReLU) activation layers respectively (architecture in Fig. 2). The output of the last convolutional layer is passed to two additional Conv layers whose outputs represent the two latent codes: Pose ($8 \times 8 \times 64$) and Shear ($8 \times 8 \times 64$).

***Pose (Non-sheared) Decoder**:* The pose decoder D_T upsamples the pose code to reconstruct unsheared output U_{rec} (256×256-pixel tactile image). It had five transposed convolutional layers (Tconv), each followed by a BN and ReLU activation layers respectively except the output layer which used sigmoid activation function.

***Shear Decoder**:* The shear decoder D_S upsamples the pose and shear code to reconstruct sheared input S_{rec} (256×256-pixel tactile image). It had the same overall architecture as the pose decoder D_T except the number of filters used in Tconv layers.

***Pose (Tap) Encoder**:* The pose encoder E_T compresses the unsheared output of D_T to $8 \times 8 \times 64$ using five Conv layers each followed by a BN and ReLU activation layer.

All encoders and decoders used a kernel of 4×4 with stride of 2. For detailed architecture, see Fig. 2.

2) Training Details: The input and output tactile data (binary 256×256-pixel images) were scaled to the range $[0, 1]$. The network weights were initialized from a zero-centered normal distribution with standard deviation 0.02.

For training, a batch size of 32 was used. All network layers used L1/L2 regularization along with random image shifts of 1-2% to prevent overfitting. We used ADAM optimizer [17] with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and learning rate of 5×10^{-5} for self-supervised (SS) and 2.5×10^{-5} for supervised (S) training phases respectively. The model was trained for 50 epochs in the SS phase and for another 50 epochs in the S phase. We used learning rate scheduling for S training phase with training rate reduced to one-tenth after 25 epochs. The training and optimization of the networks was implemented in the Tensorflow 2.0 library on a NVIDIA GTX 1660 (6 GB memory) hosted on a Ubuntu machine.

IV. Experimental Results

A. Disentanglement of Latent Representations

We did an ablation study to verify the successful disentanglement of the latent representation into pose (non-sheared) and shear codes respectively. To do this, we first reconstructed the unsheared sample using shear code instead of pose code. This resulted in an approximately 5-fold drop in mean-squared error (MSE) between test set model-generated unsheared images U_{rec} and its canonical counterparts U_{GT} (0.11 from 0.024). Similarly, a significant drop from 92.5% to 18% was observed in image similarity index measure (SSIM). We then reconstructed the sheared input S_{rec} only using the pose code to result in a 60-fold drop in MSE error between the input S_{inp} and reconstructed S_{rec} sheared samples, and over a 70% drop in image similarity between $S_{\text{inp}}$and S_{rec} (18% from 93%). These studies demonstrate the successful disentanglement of the latent representation by our model (Fig. 2).

B. Motion-induced Shear Removal from Tactile Images

We used multi-scale structural similarity index measure (SSIM) [20] as a first test of effectiveness of our approach in removing motion-induced global shear. To do this, we computed the average image similarity between the model-generated unsheared images U_{rec} and their test set sheared S_{inp} and canonical (non-sheared) U_{GT} counterparts. On comparison, we found a closer match between U_{rec} and U_{GT} (92.5% (10% supervision) drops to 74% in absence of any supervision) than U_{rec} and S_{inp} (32%). The corresponding values for fully supervised baseline [7] were 93% and 32% respectively. This clearly shows that our model not only successfully removes global shear from tactile images but is also economical with labelled data (using just 10% of the supervised method from [7]). A degradation in performance was observed in both supervised baseline [7] and semi-supervised approach proposed in this work on reduction in the labelled data used for training. For example, a 20% reduction in labelled training data lead to a drop of approximately 2% (from 93%) and 0.5% (from 92.5%) in image similarity between the model-generated unsheared images and their test set canonical (non-sheared counterparts) for supervised baseline [7] and proposed semi-supervised approaches respectively.

C. Local Contact Geometry Reconstruction

Motion-induced shear masks the contact geometry requiring removal of global shear for its faithful reconstruction (Fig. 3, top row). To reconstruct the approximate indentation field, we used a Voronoi-based method [16] that transforms the inner sensor surface with tactile markers into hexagonal cells, thus tessellating the grid of the markers [16, Fig. 4]. The change in cell areas from an undeformed reference represent the magnitude of local skin deformation that correlates with the indentation due to contact geometry.

For display, a 3D surface was fitted to the (x, y) centroid coordinates of the Voronoi cells with cell areas as the corresponding height values (Fig. 3, top row). The fitted surface from the test set sheared data S_{inp} is visibly distorted by shear (Fig. 3, top row). Once global shear is removed in the model-generated unsheared images U_{rec}, the fitted surface closely resembles the true representation obtained from the original canonical (non-sheared) data U_{GT} and the unsheared images obtained via supervised baseline [7] across multiple stimuli (Fig. 3, top row). This again demonstrates the effectiveness of our approach in removing global shear *vis-á-vis* the supervised baseline [7].

TABLE I
Mean Absolute Error (MAE) of Pose Prediction

	Pose Prediction Error			
	Tap Data	**Sheared Data**	**Unsheared Data (Supervised)**	**Unsheared Data (Semi-Supervised)**
horizontal, τ_x (mm)	0.43	2.72	0.64	0.71
yaw, θn_Z (degrees)	2.13	22.20	4.38	4.81

D. 2D Continuous Shape Exploration

We further confirmed the model performance by computing the mean-absolute error (MAE) between the predicted pose (lateral position τ_x and in-plane rotation θn_z), from test set model-generated unsheared images U_{rec}, sheared images S_{inp} and canonical (non-sheared) images U_{GT}, and target pose using a pose prediction network *PoseNet* trained on canonical (non-sheared) images that was previously used in [7]. In comparison, the prediction error with S_{inp} was over 4-fold higher than with U_{rec} (Table I), showing the successful removal of global shear. Our results also show that the model-generated unsheared images from the present model and from the the fully supervised baseline [7] have similar prediction errors (Table I).

To test the real-time performance and model generalizability to novel situations, we used the *PoseNet* [7] for sliding exploration of various planar shapes (Fig. 3, second row). This testing included four novel shapes absent from training, including two acrylic shapes (spirals 1 and 2, Fig. 3, second

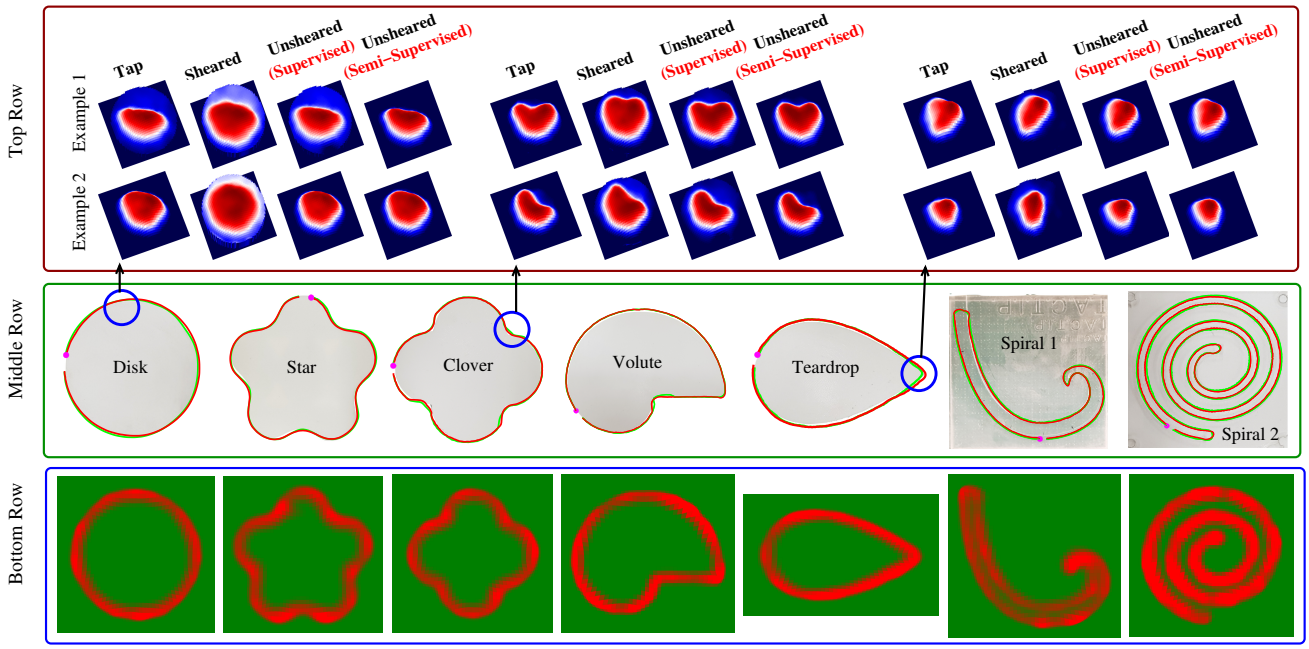

Fig. 3. **Contact geometry reconstructions, contour following and full object shape reconstructions.** Top row: Local contact geometry reconstruction: The approximate contact geometry can be recovered from sheared data using the model-generated unsheared images but not from the original sheared data. Middle row: Trajectories, overlaid on the objects, under robust sliding using pose estimation from model-generated unsheared images. Red and green trajectories correspond to semi-supervised and supervised [7] approaches respectively. Red dot is the starting point of trajectory. For both, contact reconstruction and contour following, supervised [7] & semi-supervised approaches achieve comparable performance. Bottom row: Full object reconstructions from combining the unsheared local contact geometry reconstructions along the sliding trajectories.

row) with distinct frictional properties to the other stimuli that were 3D-printed. The robot successfully traced contours around all test shapes (red, Fig. 3, second row) by a combination of the model-generated unsheared images, *PoseNet* for prediction of pose parameters and a simple servo control policy [7], demonstrating successful removal of global shear. Furthermore, comparison between the contours using the current approach (red) and supervised baseline (green) [7] show near-identical performance (Fig. 3, second row).

E. 2D Object Reconstruction

As a final test, we demonstrate our model's utility in the removal of global shear from sheared images by reconstruction of full object shape. This required: 1) faithful contact-geometry reconstruction and 2) successful contour tracing, with both tasks adversely impacted by motion-induced global shear. To obtain full object reconstructions, we first fused together the contact information extracted from unsheared images of sliding contacts recorded during shape exploration and later interpolated them on a rectangular grid (Fig. 3, third row) to obtain smooth object reconstructions from discreet sliding contacts. These results demonstrate not only the successful removal of global shear but also the effectiveness of learning the sheared-unsheared mapping once, for later reuse on multiple downstream tasks. The faithful object reconstructions with successful shape exploration across multiple shapes also that our model generalizes to novel situations (Fig. 3, second and third row).

V. Discussion

This work proposed a semi-supervised approach that preserved the sensor deformations due to stimulus contact geometry while removing the shear-distortion caused by motion-dependent shear. We showed that the proposed approach achieves comparable performance to its fully supervised counterpart [7] with an order-of-magnitude less annotated data, simplifying data collection considerably.

We validated our approach, similar to [7] by: 1) demonstrating a good match (92.5%) between the model-generated unsheared images (from sliding contacts) and their non-sheared counterpart taken from vertical tapping, using the structural similarity index measure; 2) reconstructing an approximate stimulus contact geometry from model-generated unsheared images that was previously masked by sliding-induced shear (Fig. 3, top row); 3) predicting local object pose from a pose prediction network [7] trained on non-sheared (tapping) data (Table I). This pose prediction was then used for sliding exploration of multiple planar objects (Fig. 3, second row), and so result 4) combined 2) and 3) to reconstruct full object shapes for several planar objects (Fig. 3, third row).

One limitation of our study is that it considered only planar objects. We expect our approach will extend to 3D objects, like recent work that has successfully extended *PoseNet* to moving over complex 3D surfaces (3 pose components) and edges (5 pose components) for sliding exploration [5], [6]. An extension of this work to 3D objects for sliding shape exploration and full object reconstruction, an open problem under active investigation in the field [18], [19], would open interesting avenues for future research, such as: 1) the fusion of visual and tactile sensory modalities for robust object representations and 2) tactile object recognition by exploiting prior visual experience of the object. An additional limitation of work is the requirement of annotated data, although much less than

the fully supervised counterpart [7]. Exploration of techniques to completely eliminate external supervision is left to future work. Finally, in this work, training was limited to translation shear only even though servo control introduces rotational shear due to sensor rotation while sliding over objects. The fact that the methods worked well both for sliding exploration and contact geometry reconstruction even in the absence of explicit compensation for rotational shear, shows that some unshearing carries over from translation to rotational shear, at least for planar objects. Our expectation is that bringing rotational shear into the training will become important when extending the methods to 3D objects.

Overall, we expect that our methods should apply to various other exploration and manipulation tasks performed using soft optical tactile sensors that are adversely impacted by motion-induced shear.

REFERENCES

[1] N. F. Lepora, "Soft Biomimetic Optical Tactile Sensing with the TacTip: A Review," in IEEE Sensors Journal, doi: 10.1109/JSEN.2021.3100645.

[2] X. Lin, L. Willemet, A. Bailleul and M. Wiertlewski, "Curvature sensing with a spherical tactile sensor using the color-interference of a marker array," 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020, pp. 603-609, doi: 10.1109/ICRA40945.2020.9197050.

[3] K. Kamiyama, K. Vlack, T. Mizota, H. Kajimoto, K. Kawakami and S. Tachi, "Vision-based sensor for real-time measuring of surface traction fields," in IEEE Computer Graphics and Applications, vol. 25, no. 1, pp. 68-75, Jan.-Feb. 2005, doi: 10.1109/MCG.2005.27.

[4] C. Sferrazza and R. D'Andrea. "Design, Motivation and Evaluation of a Full-Resolution Optical Tactile Sensor" Sensors, vol. 19, no. 4, pp. 928, 2019, doi: https://doi.org/10.3390/s19040928

[5] N. F. Lepora, A. Church, C. de Kerckhove, R. Hadsell and J. Lloyd, "From Pixels to Percepts: Highly Robust Edge Perception and Contour Following Using Deep Learning and an Optical Biomimetic Tactile Sensor," in IEEE Robotics and Automation Letters, vol. 4, no. 2, pp. 2101-2107, April 2019, doi: 10.1109/LRA.2019.2899192.

[6] N. F. Lepora and J. Lloyd, "Optimal Deep Learning for Robot Touch: Training Accurate Pose Models of 3D Surfaces and Edges," in IEEE Robotics & Automation Magazine, vol. 27, no. 2, pp. 66-77, June 2020, doi: 10.1109/MRA.2020.2979658.

[7] A. K. Gupta, L. Aitchison and N. F. Lepora, "Tactile Image-to-Image Disentanglement of Contact Geometry from Motion-Induced Shear," 5th Conference on Robot Learning (CoRL), 2021, url: https://arxiv.org/abs/2109.03615.

[8] R. Zhang, P. Isola and A. A. Efros, "Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 645-654, doi: 10.1109/CVPR.2017.76.

[9] M. Noroozi and P. Favaro, "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles," In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9910. Springer, Cham. https://doi.org/10.1007/978-3-319-46466-4_5.

[10] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell and A. A. Efros, "Context Encoders: Feature Learning by Inpainting," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2536-2544, doi: 10.1109/CVPR.2016.278.

[11] A. Owens, A. A. Efros, "Audio-Visual Scene Analysis with Self-Supervised Multisensory Features," In: Ferrari V., Hebert M., Sminchisescu C., Weiss Y. (eds) Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol 11210. Springer, Cham. https://doi.org/10.1007/978-3-030-01231-1_39.

[12] M. Zambelli, Y. Aytar, F. Visin, Y. Zhou and R. Hadsell, "Learning rich touch representations through cross-modal self-supervision," 4th Conference on Robot Learning (CoRL), 2020, url: https://arxiv.org/abs/2101.08616.

[13] M. A. Lee *et al.*, "Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks," 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 8943-8950, doi: 10.1109/ICRA.2019.8793485.

[14] N. Lepora and J. Lloyd, "Pose-Based Tactile Servoing: Controlled Soft Touch Using Deep Learning," in IEEE Robotics & Automation Magazine, doi: 10.1109/MRA.2021.3096141.

[15] J. Zhu, T. Park, P. Isola and A. A. Efros, "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2242-2251, doi: 10.1109/ICCV.2017.244.

[16] L. Cramphorn, J. Lloyd and N. F. Lepora, "Voronoi Features for Tactile Sensing: Direct Inference of Pressure, Shear, and Contact Locations," 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, pp. 2752-2757, doi: 10.1109/ICRA.2018.8460644.

[17] D.P. Kingma and J.L. Ba," Adam: A method for stochastic optimization," CoRR, vol. abs/1412.6980, 2015, url: https://arxiv.org/abs/1412.6980.

[18] S. Suresh, M. Bauza, K. T. Yu, J, G. Mangelson, A. Rodriguez and M. Kaess, "Tactile SLAM: Real-time inference of shape and pose from planar pushing," 2018 IEEE International Conference on Robotics and Automation (ICRA), 2018, url: https://arxiv.org/abs/2011.07044.

[19] M. Bauza, O. Canal and A. Rodriguez, "Tactile Mapping and Localization from High-Resolution Tactile Imprints," 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 3811-3817, doi: 10.1109/ICRA.2019.8794298.

[20] Z. Wang, E. P. Simoncelli and A. C. Bovik, "Multiscale structural similarity for image quality assessment," The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, 2003, pp. 1398-1402 Vol.2, doi: 10.1109/ACSSC.2003.1292216.