Estimation of Robotic Interaction Forces: A Stereo Vision Approach with Unsupervised Disparity Learning

Tim Reinhart
Stanford University
Departement of Computer Science
rtim@stanford.edu

Abstract

This project aims to develop a neural network to predict interaction forces in robot-assisted surgery, as part of ongoing research at the CHARM Lab at Stanford. The proposed approach leverages stereo vision to estimate interaction forces, eliminating the need for direct force sensors, which are not feasible for use within the human body. The model integrates depth information by first predicting disparity maps from stereo images, which are then processed by a convolutional neural network to predict force values. The model's performance was evaluated using both visual and quantitative methods, demonstrating that incorporating depth information significantly increases force estimation accuracy. The developed network with a ResNet-50 backbone achieved the lowest RMSE of 0.0415N and an NRMSE of 2.75%, outperforming baseline models and previous approaches.

1. Introduction

This project is a component of my ongoing research at the CHARM Lab at Stanford, supervised by Dr. Alaa Eldin Abdelaal and Prof. Allison Okamura, with advisory support from Prof. Jeannette Bohg. The primary objective of the project at the CHARM Lab is to develop a force-aware imitation learning algorithm capable of operating the daVinci Surgical System[1]. By leveraging force data from expert demonstrations and integrating it into a neural network model, the algorithm enhances the automation of the third arm of the robot, facilitating safer and more efficient surgeries. The developed algorithm relies on force feedback from a sensor, that measures the interaction force between the robot and the tissue. Since having a force sensor in a human body is not feasible, it is ultimately the idea to replace the sensor with another exterior system that can also work for a human body. The goal for this final project is to develop a model capable of estimating the interaction forces from stereo images, where the stereo images record the end effector manipulating the tissue. An example image of the system setup is displayed in figure 1.

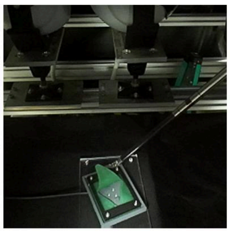


Figure 1. Example image of the setup. The end effector of the robot first pushes into the tissue, then lifts it up.

Current models for estimating interaction forces exerted by robots from images typically utilize convolutional neural networks (CNNs), such as ResNet [4] or DenseNet [5], followed by fully connected layers to predict force values in three-dimensional space, as noted in [1]. More recent advancements have enhanced performance by integrating temporal dynamics into the network, allowing for prediction of force values across sequences [6]. Despite these advancements, these approaches have been less effective in our specific research setup. A reason for this shortfall could be that the model is not able to extract meaningful features from the images by simply using convolutional networks as feature extractors. A limiting factor is that the tissue and the end effector are relatively small in the images and that the background of the setup has a very similar color. A key hypothesis for this project is that if the model can interpret complex spatial relationships and get depth information from stereo images, it will perform better at the task of force estimation, because the deformation of the tissue might be better 'visible' in a depth map then in a normal image.

Therefore, this project proposes the development of a neural network model that integrates depth information into the network architecture to improve the accuracy and reli-

[1] https://www.intuitive.com/en-us/products-and-services/da-vinci

ability of force estimation from stereo images. By adopting a strategy inspired by recent advancements in disparity prediction from stereo images, such as those described in the methodologies of [2] and [7], the model aims to leverage depth cues more effectively to predict interaction forces. This is done by first estimating the disparity map between the two stereo images, before feeding this map into a convolutional feature extractor and then predicting the force values.

2. Prior Work

Previous approaches to force estimation from images, such as [1], utilized a ResNet [4] CNN to extract feature vectors from images, which were then processed by a feed forward neural network to predict three-dimensional force vectors. This method achieved an average root mean squared error (RMSE) of $0.31N$ on a held-out test set. While their approach showed promising results, there is still room for improvement. The experimental setup allowed for a better observation of tissue deformation, when compared the setup given in this project, because the camera angle and background colors were a lot better. Also, the end effector was only pushing into the tissue without lifting it, which simplifies the task of force estimation.

Incorporating depth information into neural networks has been approached in various ways. For instance, the work [2] focused on generating disparity maps from single images to improve monocular depth estimation. This unsupervised approach ensured that disparity predictions remained consistent across different views, enhancing accuracy. The model was trained using an image reconstruction loss, leveraging the consistency between left and right view disparities generated from a single input image. Given that I do not have ground truth depth available for my dataset, I will use a very similar strategy for my approach.

SimNet [7] is another example where disparity images are used to enhance neural network performance. SimNet employs a multi-headed neural network trained exclusively on simulated data, integrating a stereo sub-network to predict disparity maps from stereo image pairs using a cost volume for stereo matching. This approach allows SimNet to handle optically challenging objects under various lighting conditions. Inspired by SimNet, my project aims to integrate a stereo sub-network within the neural architecture to improve the downstream task of estimating interaction forces from stereo images.

3. Technical Approach

To incorporate three-dimensional information, such as depth, into the network architecture, I adopt an approach similar to the one from SimNet [7]. The proposed network architecture is displayed in Figure 2. I use a ResNet-based Feature Pyramid Network (FPN) to extract features ϕ_l from the left image and features ϕ_r from the right image. This approach is similar to the feature extractor used in [3]. The input images each have dimensions $3 \times H_0 \times W_0$. Unlike the approach from SimNet, I do not down-sample ϕ_l and ϕ_r; instead, they both have dimensions $C_\phi \times H_0 \times W_0$, where $C_\phi = 16$.

Next, the stereo cost volume network (SCVN) from SimNet [7] is used to estimate disparity from the extracted features ϕ_l and ϕ_r. The SCVN searches horizontally in the features for correspondences and outputs a disparity map of dimension $1 \times H_0 \times W_0$. A significant difference from SimNet is that the disparity predicted from the network has the same height and width as the input image. This is essential because I do not have ground truth depth and thus need to employ a self-supervised approach to train the network to predict the disparity.

After the disparity estimation, the disparity map is downsampled by factor of 4 using two convolutional layers (with a kernel size of 3 and a stride of 2). The output of the downsampling step has dimensions $4 \times H_1 \times W_1$, where $H_1 = \frac{H_0}{4}$ and $W_1 = \frac{W_0}{4}$. An additional set of features are extracted from the left image using a ResNet-18-FPN network, where the output has dimensions $256 \times H_1 \times W_1$, which are concatenated with the downsampled disparity map. This is then fed into the backbone network, where I trained a version using a ResNet-50-FPN and a ResNet-101-FPN network, similar to the approach in SimNet. However, I use a different head network after the backbone to predict three-dimensional interaction forces, instead of keypoints, bounding boxes, and segmentation masks. In this version of the network I flatten the last layer of the FPN network and feed these features through a two layer feed forward network, with hidden layer sizes of 256 and 128, which predicts the force values.

To train the network, I use two types of loss functions. First, I use the mean-squared error (MSE) between the predicted and ground truth forces. Additionally, to train the network to learn the disparities, I use a self-supervised approach as described in [2]. Specifically, I reconstruct the left image $\tilde{I}_l$ from the right image I_r and the predicted disparity, and the right image $\tilde{I}_r$ from the left image I_l. The reconstructed images are created using backward mapping with a bilinear sampler. The total loss ℓ is a combination of the appearance matching loss ℓ_{ap} and the disparity smoothness loss ℓ_{ds}:

$$\ell = c_{ap}\ell_{ap} + c_{ds}\ell_{ds}$$

where $c_{ap} = 1$ and $c_{ds} = 0.1$ are scalar weights. The appearance matching loss ℓ_{ap} is a combination of the $L1$-loss and the structural similarity index measure [8]. The loss between an input image I and the reconstructed image $\tilde{I}$ is given by:

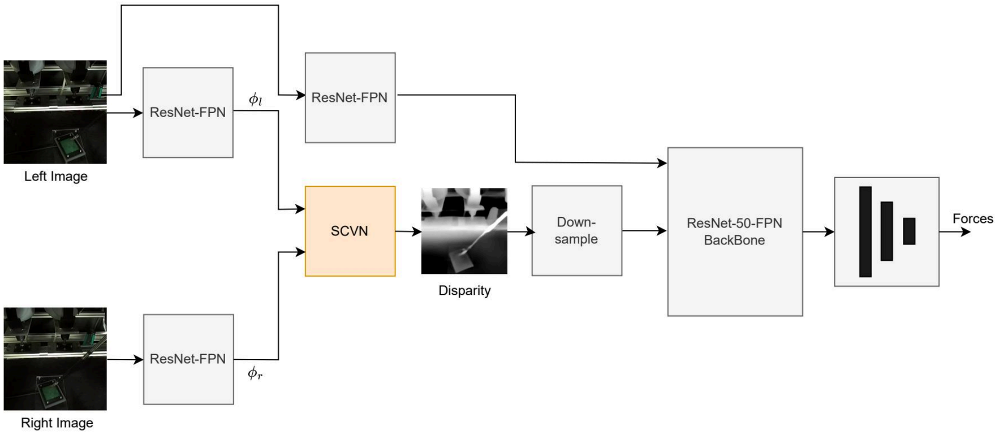


Figure 2. Proposed network architecture to estimate 3D forces from stereo image pairs.

$$\ell_{ap} = \frac{1}{N}\sum_{i,j} \alpha \frac{1 - SSIM(I_{ij}, \tilde{I}_{ij})}{2} + (1-\alpha)\|I_{ij} - \tilde{I}_{ij}\|$$

Where I_{ij} denotes the pixel-value at pixel i, j. The appearance matching loss ℓ_{ap} is computed for both the left and right image and then summed up. The disparity smoothness loss ℓ_{ds} encourages smooth transitions in disparity values while preserving edges and is defined as:

$$\ell_{ds} = \frac{1}{N}\sum_{i,j} |\partial_x d_{ij}| e^{-\|\partial_x I_{ij}\|} + |\partial_y d_{ij}| e^{-\|\partial_y I_{ij}\|}$$

Where d is the predicted disparity map. The network was fully implemented in PyTorch and the code for the model is available at github.com/TimReX-22/stereo_force_estimation. A NVIDIA GeForce RTX 4090 GPU was used to train the model. I used the hyperparameters displayed in table 1 to train the network.

Hyperparameter	Value
Nr. Layers Backbone (ResNet)	50
Pretrained Weights Backbone	True
Batch Size	8
Number of Epochs	20
Learning Rate	1×10^{-5}

Table 1. Hyperparameters used to train the model

4. Evaluation and Results

4.1. Dataset

The dataset used for training and testing the network consists of stereo image pairs and corresponding interaction force values. The data was recorded while rolling out the policy from the force-aware imitation algorithm mentioned in section 1. An example of force values recorded during a single policy rollout is displayed in figure 3.

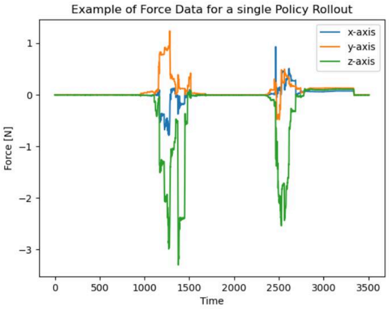


Figure 3. Example of force measurements during a policy rollout.

The training set comprises 8,000 samples, while the validation set includes 2,000 samples. For the training images, I apply data augmentation by randomly flipping the images and randomly changing the lighting conditions. The unused images were used to evaluate the model on unseen data.

4.2. Training Results

The model, depicted in figure 2, was trained end-to-end. First training the disparity prediction, then freezing the weights of the part of the network that predicts the disparity, and then training the force prediction part, did not yield any improvement. During the training process, both the training and validation loss decreased consistently, indicating that the model is learning effectively and generalizing well to unseen data. Also, both the MSE loss and the disparity loss (consisting of the appearance matching loss and the disparity smoothness loss) decreased during training, as displayed in figure 4 and figure 5, indicating that the network is able to learn both the disparity prediction and the force prediction at the same time.

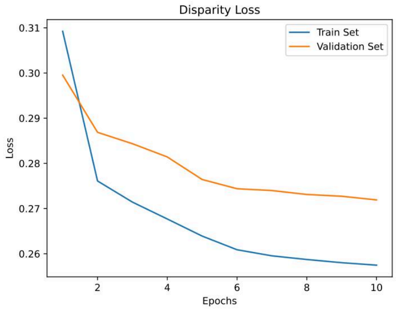


Figure 4. Disparity loss per epoch for training and validation sets.

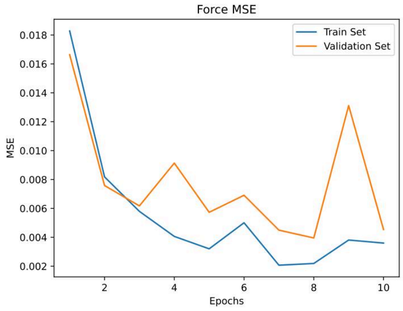


Figure 5. MSE per epoch for training and validation sets.

Because the root mean squared error (RMSE) is of great importance for the task, as it has the same unit as the force values, I also investigate the RMSE during training. The results are displayed in figure 6. As seen in the plots of the MSE and RMSE, the model starts to overfit slightly after 8 epochs of training, therefore training was limited to total of 10 epochs and the weights of the model with the lowest RMSE score on the validation set is saved.

4.3. Evaluation on Unseen Data

To evaluate the model's performance, it was tested on a held out testset. Both the disparity prediction and the force estimation was evaluated separetly. To evaluate the disparity prediction, I predict the disparity on a given pair of stereo images, and reconstruct the right image from the left one. Figure 7 shows the results of this evaluation. The three subfigures display the real right image, the reconstructed right image, and the computed disparity map, respectively.

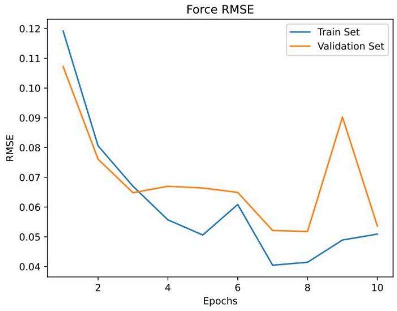


Figure 6. RMSE per epoch for training and validation sets.

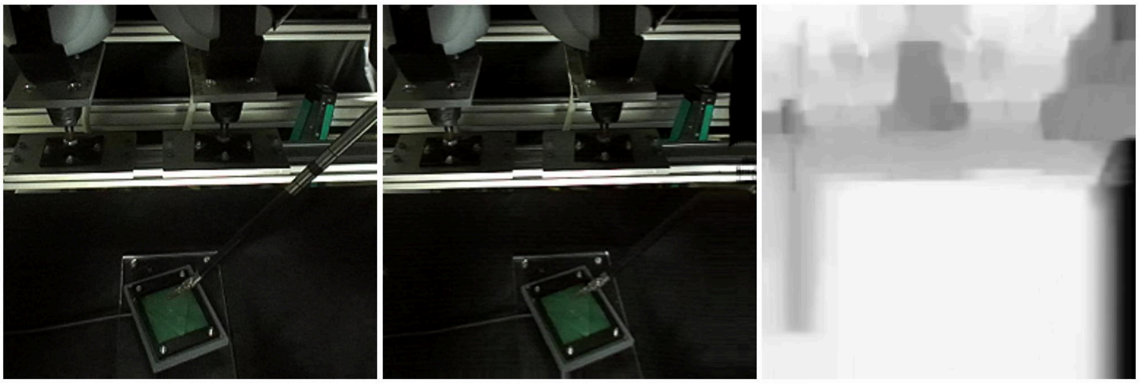
The results show that the reconstructed right image is reasonably accurate but still exhibits some artifacts, particularly a small black bar on the right side. This artifact suggests that the learned disparity map may not be perfect yet. The disparity map itself shows a range of values, indicating that the model is learning to distinguish depth variations, but the granularity and accuracy of these variations need improvement. By evaluating the disparity map visually in figure 7, one can see that especially in the center of the image, the model is not able to predict the disparity values. Since the tissue is located at the lower center of the image, this area is actually critical.

The force estimation was evaluated using both visual and quantitative methods. The visual evaluation involved comparing the predicted force values against the ground truth values over time. The model was run on a stereo image sequence, which corresponds to a full policy rollout of the imitation learning algorithm. The results are displayed in figure 8. The plot displays the smoothed predictions and ground truth values for the force in the X, Y, and Z directions during a single test run.

The quantitative evaluation was conducted using the root mean squared error (RMSE) and the normalized RMSE (NRMSE). The NRMSE is calculated by dividing the RMSE by the range of the ground truth values:

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}}$$

The NRMSE score is providing a normalized measure of the estimation error relative to the range of forces encountered. The results are displayed in table 2. The Basic-CNN in table 2 corresponds to an approach developed as baseline,



(a) Real right image (b) Reconstructed right image (c) Disparity map

Figure 7. Evaluation on unseen data. (a) Real right image, (b) Reconstructed right image, (c) Computed disparity map.

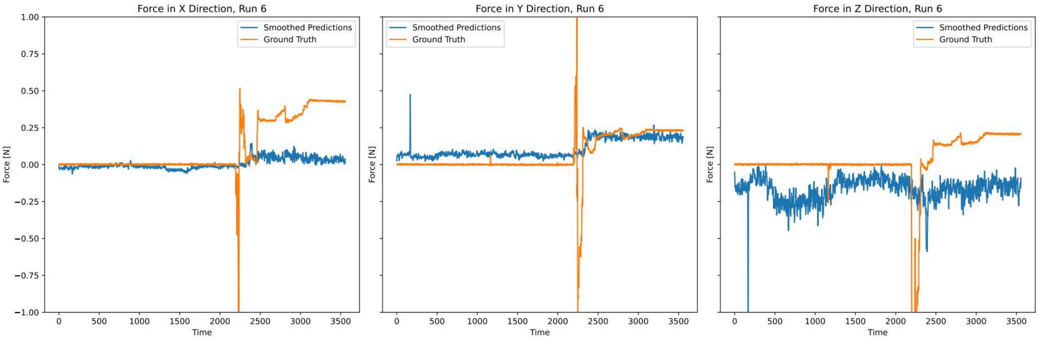


Figure 8. Predictions of force values in X, Y and Z directions over data of a full policy rollout.

which is based on [1]. It is the best working model I was able to develop that uses a ResNet-50 to extract features from the images and then estimates the force values from these features using a fully connected network. I tested two versions of the backbone network in the depth network (displayed in figure 2), where it proved that the ResNet with less layers is performing better than the one with more layers. As the results in table 2 suggest, incorporating depth into the network improves the performance of the force estimation by a factor of more than 2.

Model	**RMSE**	**NRMSE**
Basic CNN	0.0921N	0.046%
Depth Network (ResNet-101)	0.0528N	0.0264%
Depth Network (ResNet-50)	**0.0415N**	**0.0275%**

Table 2. Evaluation results of force estimation models.

5. Discussion

The disparity prediction evaluation demonstrated that the model could reconstruct the right image reasonably well from the left image and the predicted disparity map. However, some artifacts and inaccuracies in the disparity map suggest that further refinement is necessary to achieve higher precision in depth estimation. Also, the disparity map does not fully capture the whole scene, especially in the region where the tissue is located, the estimated depth is not accurate.

The results from the force estimation evaluation indicate that integrating depth information into the neural network architecture significantly improves the accuracy of force predictions. The Depth Network with ResNet-50-FPN as a backbone achieved the lowest RMSE of 0.0415N and an NRMSE of 2.75%, outperforming the other models. The basic CNN, which did not incorporate depth information, showed the highest RMSE and NRMSE, highlighting the importance of depth cues for accurate force estimation. By simply looking at the RMSE score, this work also outperforms previous work, such as [1], where the best RMSE score was $0.31N$. As seen in figure 7, the depth of the tissue is actually not yet captured by the model. My hypothesis is that if the depth of the tissue can be estimated better, then the force estimation would be even better. Therefore, a crucial next step is to improve the disparity prediction.

5.1. Future Work

Future work will focus on improving the disparity prediction accuracy to enhance the overall force estimation performance. One primary area of improvement is refining the disparity estimation network. This will involve experimenting with different architectures and loss functions to improve the quality of the disparity maps, especially in critical areas such as the tissue's center where current predictions are less accurate.

Another important direction is augmenting the dataset. Increasing the diversity and size of the training dataset by including more variations in tissue types, lighting conditions, and camera viewpoints will be crucial for improving the model's generalization capabilities. A more diverse dataset will help the model perform better under different conditions and scenarios encountered in real-world applications.

Incorporating temporal dynamics into the model is also a key aspect of future work. Integrating temporal information can capture the dynamic aspects of tissue manipulation more effectively. This can be achieved by using recurrent neural networks (RNNs) or transformers, which have shown promise in handling sequential data. Work done as part of my research at the CHARM lab suggest that using a transformer based architecture to predict over a sequence of data can improve the performance a lot.

References

[1] Zonghe Chua, Anthony M. Jarc, and Allison M. Okamura. Toward force estimation in robot-assisted surgery using deep learning with vision and robot state. *CoRR*, abs/2011.02112, 2020. 1, 2, 5, 6

[2] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *CoRR*, abs/1609.03677, 2016. 2

[3] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. *CoRR*, abs/1806.01260, 2018. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 1, 2

[5] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. 1

[6] Dae-Kwan Ko, Kang-Won Lee, Dong Han Lee, and Soo-Chul Lim. Vision-based interaction force estimation for robot grip motion without tactile/force sensor. *Expert Systems with Applications*, 211:118441, 2023. 1

[7] Thomas Kollar, Michael Laskey, Kevin Stone, Brijen Thananjeyan, and Mark Tjersland. Simnet: Enabling robust unknown object manipulation from pure synthetic data via stereo. *CoRR*, abs/2106.16118, 2021. 2

[8] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 2