CMSC426 Final Project:
Using U-Net for Depth Estimation from Visuo-Tactile Sensors

Anh N. Nhu
University of Maryland - College Park
UID: 119 385 173
anhu@terpmail.umd.edu

Ansh Viswanathan
University of Maryland - College Park
UID: 117 863 239
anshvis@terpmail.umd.com

Kamal Narra
University of Maryland - College Park
UID: 118 007 359
snarra@umd.edu

Nimay Vyas
University of Maryland - College Park
UID: 117127522
nvyas@umd.edu

Abstract

A visuo-tactile sensor is a multifunctional sensing device that combines visual and tactile capabilities. It integrates technologies to capture visual information, such as images or video, along with tactile sensing components that detect touch, pressure, or other tactile feedback. It provides an additional modality for a more comprehensive perception to solve various robotic tasks, allowing robots to not only see but also physically interact with the environment. In this report, we investigate the efficiency of U-Net-based architectures, a state-of-the-art Convolutional Neural Network-based framework for segmentation and various image translation tasks, in depth estimation from visuo-tactile images. Furthermore, we find that depth prediction conditioned on predicted contacts, which are simply binary masks of pixels with positive depth, help improve the accuracy of depth prediction models. Specifically, using the same U-Net architecture, we obtained a test score of 14.486 using contact-conditioned depth prediction, remarkably higher than the test score of 9.298 obtained when using direct depth prediction from only visuo-tactile images. Our approach generates visually realistic depth prediction and is currently placed as the second best solution among all other teams.

1. Introduction

Depth estimation in perception is a rapidly growing domain in rotobitcs research. A novel development in this sector is the integration of visuo-tactile sensors, which generate richer, multimodel datasets by combining visual and tactile feedback, and in doing so enhance the depth perception capabilities of robotic systems, enabling them to interact with their environments more effectively.

Unlike traditional vision-based systems, visuo-tactile sensors can capture not only the appearance but also the physical interaction of objects, such as texture and pressure. This combination is particularly beneficial for understanding complex environments where visual data alone might be insufficient due to occlusions, varying lighting conditions, or transparent surfaces. Moreover, recent Convolutional Neural Network frameworks like U-Net [11] have proven to be extremely successful in segmentation and image translation due to their ability to capture both local and global contexts.

In this study, we explore the potential of U-Net-based architectures for depth prediction, leveraging their segmentation prowess to interpret the rich data provided by visuo-tactile sensors. Our work is motivated by the rising need for sophisticated depth-estimation techniques in robotics [2, 5, 6], and we aim to bridge the gap in literature by introducing a novel approach to integrate tactile information for depth estimation.

2. Related Works

There have been several works investigating different models for the depth estimation in the context of robotics and scene representations. These prior works mostly focused on CNN-based architecture, including but not limited to ResNet [9], U-Net [1, 4], Vision Transformer [10], 3D CNN [3], and CNN [6, 7]. These approaches have obtained state-of-the-art performance on depth prediction by learning scene representations from natural input images. However, there is limited work explicitly attempts to predict depth from tactile images, which is inherently more challenging

compared to natural images due to its irregular shapes and color space.

A work that have exclusively focused on monocular depth prediction from soft visuo-tactile sensor [7]. This work also leverage U-Net as the backbone architecture in the framework, reaching high accuracy and generalizable to different object types and sensor configurations. However, the core limitation of this work is that it requires a large number of data available, limiting its usablity in small dataset like in this work.

Given such limitations, we proposed an U-Net-based framework to estimate depth from tactile images in the following section.

3. Methodology

In this section, we describe in details the complete learning framework and models we used to map visuo-tactile images to depth prediction. Overall, our framework consists of two separate training stages:

1. **Stage 1:** Training Contact Net. Contact Net is the model that learns to map the input visuo-tactile images to the corresponding contact masks, which are basically the binary masks of pixels where depth is greater than a certain threshold.

2. **Stage 2:** Training Depth Net. Depth Net is the final model that we want to obtain in this work. It learns to map the input visuo-tactile images to the corresponding depth maps. In our work, the prediction of depth maps is conditioned on the predicted contact mask in Stage 1. We will elaborate more on the details of Depth Map below.

3.1. Stage 1: Contact Net

As described previously, Contact Net learns to map the input tactile images to corresponding contact masks. Since contact masks are not inherently available in the dataset, we have to generate it by assigning pixels with depth larger than certain contact threshold τ to be 1 ("positive"); otherwise, the pixel is assigned 0 ("negative"). In our work, we set $\tau = 0.005$. Formally, given an input tactile image I, the contact mask $M_{contact}$ is defined as:

$$M_{contact} = J_{H,W} \odot (I > \tau) \tag{1}$$

where H, W are the height and width of the input tactile image I, $\tau = 0.005$ is the contact threshold. $J_{H,W}$ is matrix of ones with dimension $H \times W$, which is similar to the input image I. Intuitively, the contact masks are basically the masks of where depth is at least 0.005. Examples of resulting ground-truth contact masks are shown in the 3^{rd} column in Figure 3.

With generated contact targets $M_{contact}$, we now define the Contact Net $f_{contact} : \mathbb{R}^{H\times W\times 3} \rightarrow \mathbb{R}^{H\times W}$ that learn to map tactile image I to contact mask $M_{contact}$:

$$\hat{M}_{contact} = f_{contact}(I) \tag{2}$$

The architecture of our Contact Net $f_{contact}$ is a modified U-Net [11] for smaller dataset in this work. Figure 1 shows the details of this modified U-Net for Contact Net.

We chose U-Net for depth prediction from tactile images due to its well-suited architecture for semantic segmentation tasks. In the context of depth prediction, where assigning depth values to individual pixels is akin to segmentation, the encoder-decoder structure allows capturing both global and local information essential for understanding spatial relationships. As shown in Figure 1, The incorporation of skip connections facilitates the preservation of intricate tactile details during the upsampling process, mitigating information loss. Moreover, U-Net's adaptability to varying input sizes and its effectiveness with limited training data are particularly advantageous in the domain of depth prediction from tactile images.

3.2. Stage 2: Contact-conditioned Depth Net

The backbone architecture of Depth Net is exactly the same as Contact Net, which is the modified U-Net shown in Figure 1. Only one modification is made: instead of inputting only RBG images, we concatenate the predicted contact mask from the first stage to the input RGB tactile image as an additional channel. In other words, there are 4 channels: Red, Green, Blue, and Contact Mask. We call this modification "contact-conditioned" depth prediction. The formal definition of Depth Net is $f_{depth} : \mathbb{R}^{H\times W\times 4} \rightarrow \mathbb{R}^{H\times W}$, where the predicted depth $\hat{D}$ is:

$$\hat{D} = f_{depth}(I \oplus \hat{M}_{contact}) = f_{depth}(I \oplus f_{contact}(I)) \tag{3}$$

where $\oplus$ denotes concatenation operator, and $f_{contact}$'s weights are trained and frozen.

The core novelty in our work is the combination of modified U-Net and contact-conditioned depth prediction. The rationale is that U-Net is the state-of-the-art model for various image translation tasks, especially image segmentation which is closely similar to the depth prediction task. In the next section, we briefly mention quantitative results showing U-Net outperforms Multi-scale CNN approach in [6].

Inspired by [6], where the authors use predicted contact mask $\hat{M}_{contact}$ as an additional input to their Stack CNN, we also adopt the same approach. We empirically tested that depth models conditioned on contact mask consistently outperform models with exactly similar architecture but directly predict depth maps from only tactile images. Specifically, the test scores of contact-conditioned and direct depth model on leaderboard are 14.386 and 7.364, respectively.

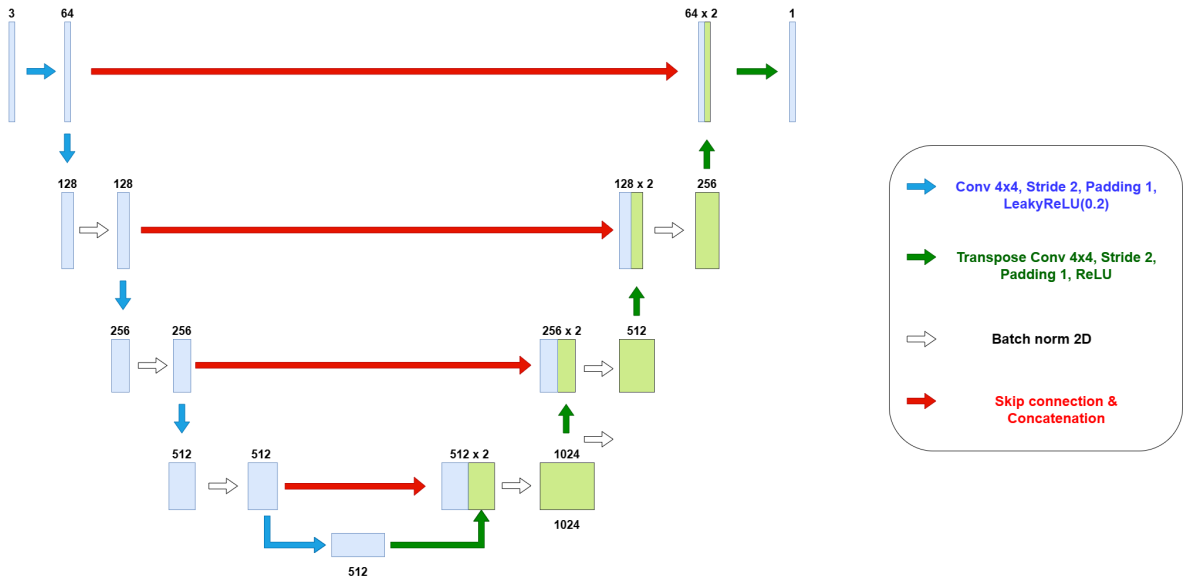

Figure 1. Modified U-Net architecture for Contact Net and Depth Net

4. Experimental Setup

4.1. Training-validation split

In this work, there are 3 options to split the training and validation set:

1. Split randomly from all images
2. Split by randomly sampling a subset of each object type as the training set and the rest as the validation set (e.g: take 70% of long cylinder samples, 70% of cube samples, 70% of cable samples, etc. as the training set; the rest is validation set)
3. Split based on object types (e.g: long cylinders, cubes, cables, etc.)

In our work, we follow the 3^{rd} strategy, that is split based on object types. Specifically, we consider all samples of "big decagon", "cables", "cube", "digit mount", "hafez" and "ring" to be training set, while samples of "long cylinder" and "med decagon" as validation set. The motivation is that we want our model to be generalizable to new objects with unseen shape and/or interactions, so monitoring validation loss based on unseen object is a reasonable approach.

4.2. Preprocessing steps

For our preprocessing step, we only have 3 steps: (1) resize to (256, 320); (2) random horizontal flip; (3) random vertical flip.

Since our U-Net implementation requires input height and width to be multiple of 32, we interpolate original image size from (240, 320) to (256, 320). Once the U-Net made prediction from this interpolated input, we interpolate the output back from (256, 320) to (240, 320) as final prediction.

We used random horizontal and vertical flip to increase the dataset size. Typical color space augmentation is not used because it does not fit well for this problem. This is because unlike natural image tasks, in visuo-tactile images, exact color is important to depth prediction. Therefore, if the color space is distorted, it will impact the underlying signals, reducing convergence optimality and performance of the trained model.

4.3. Hyperparameters

In our experiments, the hyperparameters used to train both Contact Net and Depth Net are as follow:

- batch size: 64
- total epochs: 200
- initial learning rate: 1e-3
- optimizer: Adam [8]

4.4. Loss function

For contact prediction task in Stage 1, since the contact are binary mask, we use Binary Cross-Entropy (BCE) loss for Contact Net. On the other hand, for Depth Net, we use Mean Squared Error (MSE). In both BCE and MSE loss, we use mean as the reduction technique to aggregate the loss across all pixels from all images in the batch.

5. Results and Discussion

5.1. Results

In this section, we present the prediction results that we obtained from our proposed approach described in the previous section. Qualitatively, Figure 2 show visualization of input tactile images, predicted depths, and ground-truth depths sampled from the validation set. Furthermore, Figure 3 show contact prediction results on valiation set, which is the intermediate step in our contact-conditioned U-Net framework.

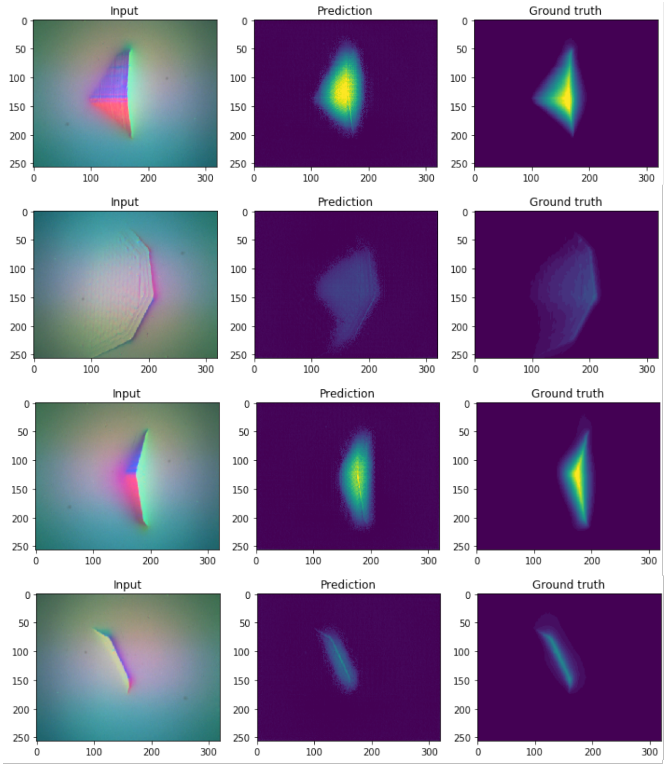

Figure 2. Input Tactile image, Predicted Depth, and Ground-truth Depth from the validation set

Additionally, we also qualitatively investigated the depth prediction on test set, which is shown in Figure 4. Since the test set is provided without ground-truth depth, only the input tactile images and the predicted depths are available for test set in the results.

As requested by the assignment, we also include the prediction of the test image on Piazza. The provided tactile image posted on Piazza and our model's depth prediction from that image is shown in Figure 5.

In order to provide a additional insights into the learning process of our model, Figure 6 and Figure 7 are included to show the loss curves of the contact model and depth model, respectively, during the optimization process.

Based on the depth predictions results shown above,

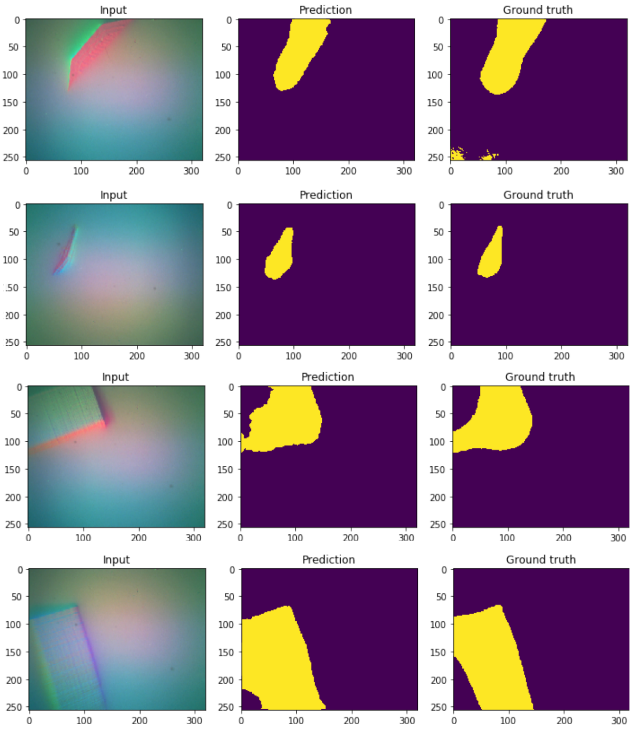

Figure 3. Input Tactile image, Predicted Contact Mask, and Ground-truth Contact Mask from the validation set

qualitatively, we see that the predicted depth are visually convincing and realistic. The quantitative evaluation metric for depth prediction is Mean Squared Error (MSE), which is also the evaluation metric used in the leaderboard. Compared to other teams, our team obtained a score of **14.386**, ranking as 2^{nd} place at the moment, outperforming the approaches of most other teams and being only marginally below the current leading team, which has a score of 14.737.

5.2. Discussion

Here, we analyze and interpret the experimental results obtained in the last subsection. We have the following comments regarding the performance of our proposed model:

1. Qualitatively, based on Figure 2, our model outputs visually accurate and realistic depth prediction compared to the ground-truth depths. This indicate that the model has learned useful representation to extract signals from the data, being able to generalize to unseen tactile images in the validation set.

2. Although we do not have access to ground-truth depths for the test set, results in Figure 4 shows that our model have outputted visually realistic depth estimation compare to patterns in visuo-tactile images. Specifically, the shape, patterns, and depth intensity visually match

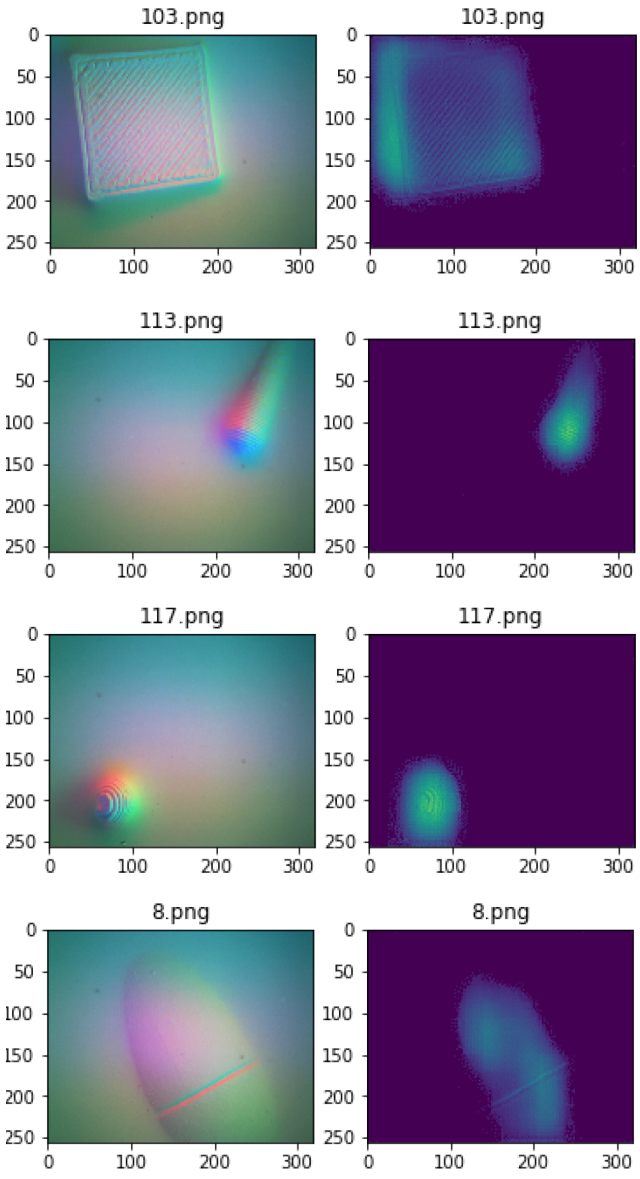

Figure 4. Input Tactile image and Predicted Depth from the test set

each other between input tactile images and the corresponding predicted depths. Quantitatively, our model perform well on the test set, reaching a performance of **14.386** on the leaderboard and ranked 2^{nd} in terms of score.

3. We find that U-Net performs much the baseline Multi-scale CNN model [6] on the test set. Specifically, with the same input, preprocessing steps, and optimal hyper-parameters, U-Net reaches a score of 14.386 on the leaderboard while Multi-scale CNN only reached 7.364.

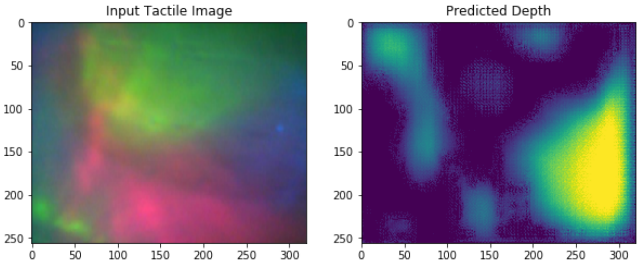

Figure 5. Input Tactile image and Predicted Depth of the Piazza Tactile Image

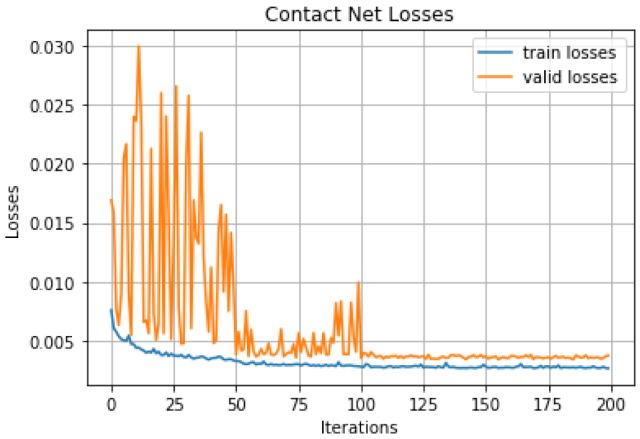

Figure 6. Loss Curve of the Contact Model

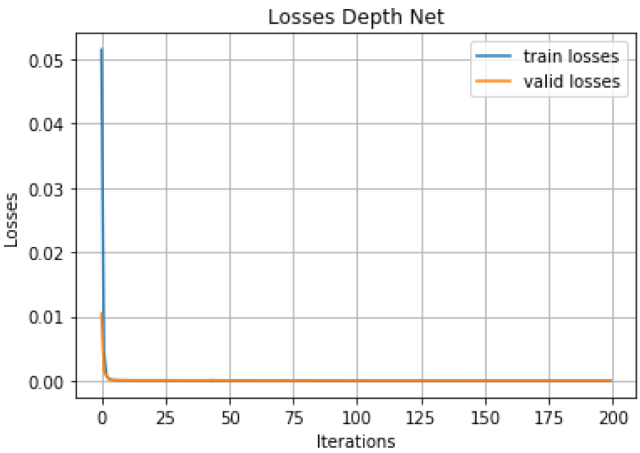

Figure 7. Loss Curve of the Depth Model

4. The learning curves of Contact Net and Depth Net, shown in Figure 6 and Figure 7, respectively, shows that the both model converge to optimal solution. However, Contact Net seems to have very noisy validation loss curve despite smooth training loss curve in the first 100 iterations, indicating that the learning rate might be too large. The learning curve of Contact

Net might be smoother with a smaller initial learning rate.

5. The most important detail that we observe is that when conditioned on contact predictions, accuracy of depth prediction become much higher (14.486 vs 9.298 on the leaderboard). Therefore, we decided to use contact-conditioned U-Net as our final primary model.

6. Conclusion and Future works

In this work, we investigated the performance of U-Net-based architecture for depth prediction task from visuo-tactile images. Specifically, we proposed a contact-conditioned modified U-Net approach which outperforms other baselines on the test set (score = 14.386), including vanilla U-Net (score = 9.298) and Stack CNN [6] (score = 7.364). Qualitatively, our model generates visually realistic depth estimations, which are consistent with the input visuo-tactile images. Finally, our model is ranked second-best in the leaderboard in terms of score, being only marginally below the first-ranked team and significantly better than others.

In future work, we plan to extend our work by one of the following strategies:

1. Explore other state-of-the-art architecture, such as UNet++ [12].

2. Try ensemble prediction of model trained on different training and validation folds using k-fold ensembling.

7. Takeaways

These are what we learned from this project:

1. We learned more about state-of-the-art segmentation models, specifically U-Net, and how to implement them for depth prediction from tactile images.

2. We learned that conditioning the Depth Model, which can either be U-Net-based or other backbone architecture, improve the performance on test set significantly.

3. From the experiments and hyperparameter search, we find that batch size and initial learning rate selection significantly impact both the optimality and rate of convergence of the learning process. This subsequently impact the performance of the learned model. For example, the batch size should be at least 32 for a stable optimization process, and the learning rate should be around 1e-3 in most cases.

The implications of our work includes but is not limited to:

1. A novel approach showing that U-Net is a better approach compared to Multi-scale CNN in depth prediction from tactile images. This will guide future research on using U-Net-based backbone network and conditioning on contact mask and/or other novel masks.

2. A pretrained depth model with decent performance which can be used to fine-tune and/or augment the learning of new model in similar tasks.

References

[1] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning, 2019. 1

[2] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(11):3174–3182, 2018. 1

[3] Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network, 2018. 1

[4] Rohit Choudhary, Mansi Sharma, and Rithvik Anil. 2t-unet: A two-tower unet with depth clues for robust stereo depth estimation, 2022. 1

[5] Xingshuai Dong, Matthew A. Garratt, Sreenatha G. Anavatti, and Hussein A. Abbass. Towards real-time monocular depth estimation for robotics: A survey, 2021. 1

[6] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network, 2014. 1, 2, 5, 6

[7] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation, 2018. 1, 2

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. 3

[9] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks, 2016. 1

[10] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction, 2021. 1

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. 1, 2

[12] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation, 2018. 6