EF-Net: End-to-End Friction Estimation Using Intrinsic Imaging and Deep Networks

Sarvesh Prajapati,[1] Abhinav Kumar[1] and Rupesh Pathak[1]

***Abstract*—Material recognition and friction estimation although overlooked is an important aspect of computer vision especially for robotic applications. Domains like grasping, motion planning, footstep planning, etc., are some of the few applications where it is important for the agent/robot to be aware of the material or friction coefficient associated with the material. In this paper we build up on why friction estimation is important, go over the current state-of-the-art material recognition and friction estimation methods existing datasets and address the problems by formulating an end-to-end friction estimation network. At last we present a novel architecture *EF-Net* that relates material patches to their respective spectral values. The proposed EF-Net is able to estimate the spectral value associated with a material both with trained and untrained materials.**

***Index Terms*—Intrinsic Imaging, Material Recognition, Friction Estimation, Robotics**

I. Introduction

With the rapid rise in industries and advancement in technology there has been rapid rise of robots in workspace and industries. To address topics like Human-Robot Interaction either for industry or household purposes, it becomes necessary for robots to be aware of their surrounding and know what materials they are interacting with. Fundamental challenges in robotics include grasping, path-planning, and day-to-day interaction with humans. For us humans, it comes naturally how much force to apply when holding a plastic cup as opposed to a ceramic cup, but for a robot, differentiating between a paper cup and a ceramic cup can be challenging, and excess force may deform the paper cup. For mobile robots as well, sometimes state estimation or localization also depends on material it is travelling on (1), also we want to avoid materials where a mobile robot may get stuck. At last for legged robots, it is important to know the friction values of terrain such that the constraints of a convex optimization problem are respected and the system remains stable.

Our work is motivated by the quadruped locomotion problem, existing work in quadruped doesn't account for friction values and assumes that the knowledge of the world is given beforehand. Figure 1 shows the importance of friction values for a quadruped locomotion and (2) formulations of MPC make it clear that friction estimation is an important aspect of quadruped locomotion as the normal forces should stay in the friction cone in order for a legged system to remain stable (3). Building up on all this, it becomes important to have material recognition and friction estimation that depends on physics and not just some such that we can utilize deep networks to their full potential (4). We propose a Deep Convolution Neural Network EF-Net that maps RGB images to domain, which can be further utilized for friction estimation, material recognition, etc.

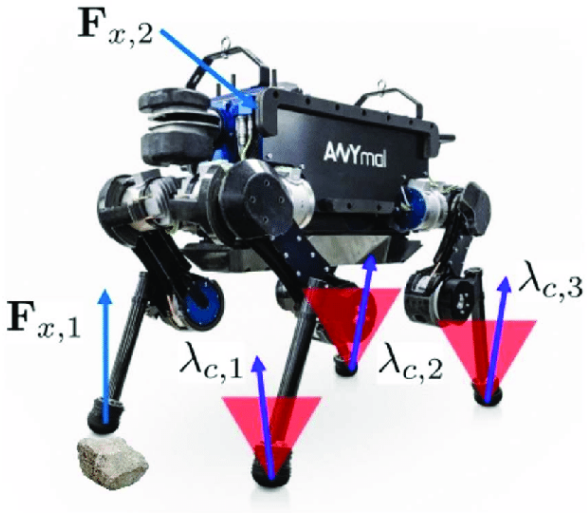

Fig. 1. Friction Cone constraints of a quadruped (5)

II. Related Work

The task of material recognition and friction estimation has always excited Roboticists now and then and there have been several attempt to solve the problem of material recognition, either by using single modality or fusing multiple modalities. The current State of the Art are heavily dependent on dataset they are being trained (6) and there have different attempts to create generalized dataset (7) but still the performance of SOTA is very low compared to other tasks in intrinsic imaging domain.

The simplest and one of the first work for friction estimation tailored for legged robot was proposed by (8) (9) that involved passing RGB images through CNNs, segmenting them and using a look-up table to map the class labels to friction values.

Recently fine-tuning and transfer learning (10) really shined in extracting fine features from images using a deep pre-trained network and using the fine-tuned value for friction estimation. (11) used ResNet50 (12) and fine-tuned on reflectance disks of different materials and proposes to use transfer learning. It is worth noting that the results were not good as SOTA,

[1]Institute for Experiential Robotics, Northeastern University, Boston, Massachusetts, USA. {prajapati.s, kumar.abhina, pathal.r}@northeastern.edu
Code available at - https://github.com/prajapatisarvesh/EF-Net

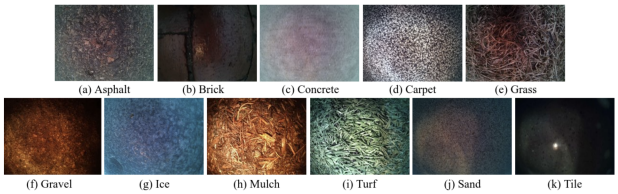

Fig. 2. Vast Dataset by Hanson et al. (15)

also the reflectance disc dataset is no longer available and the paper also proposes to release a friction dataset, that hasn't been done yet.

Different modalities for material estimation also included spectroscopy (13) (14). The results were getting better and better with the use of spectroscopy, however due to the nature of the sensor and non plug-and-play nature of these sensors, it is often quite difficult to use these sensors a lot in conjunction with the system. Recently, this idea of spectroscopy informed material recognition was extended by (15), which used multiple-modality i.e., RGB images, spectral data and IMU data, and used all of them together for material recognition.

The majority of work in material recognition and/or friction estimation formulated the problem differently, and used different approaches to solve the same with different dataset, which also becomes difficult to have comparison between existing work and extending it to future work or improving upon it.

III. Dataset

There aren't a lot of datasets available out there. Some of the widely used datasets are Material Recognition in the Wild (7), Flickr Material Database (6), etc. Another limitations of these datasets are either they are not labelled properly or are not suitable for real world implementation. Another problem in friction estimation is there's no dataset available that maps pixel values to friction coefficient, some papers (16) did create a dataset, but haven't oponsourced it yet. We used two datasets for our approach, Apple DMS Dataset (17), one of the challenge was downloading this dataset as the dataset only contains the labels and images needs to be downloaded separately, and there's no proper information available for the same. We also used VAST dataset (15) for our proposed EF-Net that maps RGB values into corresponding spectral data.

A. Apple DMS Dataset

The Dense Material Segmentation Dataset (DMS) consists of 3 million polygon labels of material categories (metal, wood, glass, etc) for 44 thousand RGB images. It consist 3.2 million dense segments on 44,560 indoor and outdoor images. Figure 3 shows the dataset. Segmentation and Regression analysis by U-Net an SRCNN was done on this dataset. For regression analysis, the labels were mapped to friction coefficients (9) (11) and we didn't consider shift in friction coefficient based on physical properties like temperature, weather, etc.

Fig. 3. Apple dms segmentation dataset (17)

B. VAST Dataset

The extensive dataset for material recognition is modular, comprising three distinct types of information: images, spectral data, and inertial measurement unit (IMU) data. The author utilized material recognition using the multi-modal approach, we just use the RGB images and estimate the spectral values. Figure 2 shows the dataset. Each image from a camera mounted under a robot, closed to ground plane gives a high resolution image, which is further divided into 12 tiles. Each tile has an associated spectral data, and different classes have a different distinguishable spectral profile.

IV. Methodology

Our methodology follows a structured progression, starting with the implementation of existing work with better deep network (U-Net) in case of segmentation and regression, and is further explained in this section. This initial step serves as a benchmark for our subsequent efforts. Expanding on this, our approach evolves to tackle the challenge by eventually finding relationship between spectral data and RGB image. A significant contribution comes in the form of EF-Net, a novel architecture designed specifically for our unique problem domain. The intricacies of EF-Net are explained in the dedicated subsection, highlighting the innovative features and tailored design choices that set it apart in the landscape of neural network architectures. This multi-stage approach emphasizes our commitment to a comprehensive exploration of the problem space—from replication to adaptation and, ultimately, innovation.

A. Semantic Segmentation

We utilize U-Net (18) for semantic segmentation on Apple DMS dataset and further mapping these pixel to friction is done through a LUT as proposed by (9). These classes encompass a spectrum of materials such as wood, paper, asphalt, stone, and more. Each of the 57 segmented classes is methodically mapped to specific friction. This mapping is informed by insights gleaned from a diverse set of sources (9).

B. Regression

Deep Neural Networks are very powerful in learning underlying representations and match the desired output based on some loss function, so instead of additional lookup-table, we simply try to find relation between an RGB image and corresponding friction value at a pixel. We made several changes to U-Net and adaptations were specifically geared towards transforming these architectures into potent tools for regression tasks. The foundational step involved acquiring

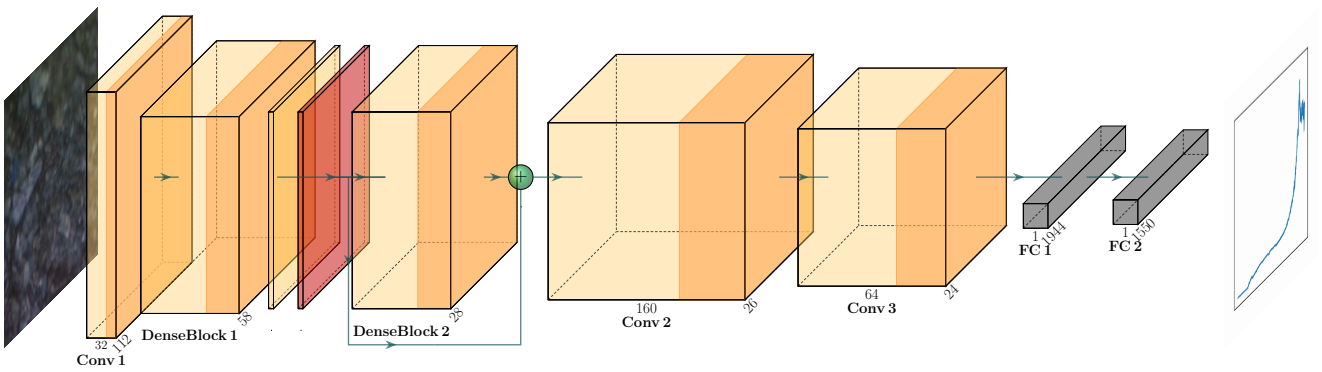

Fig. 4. EF-Net: End-to-End model for friction estimation. The input is a RGB image, that is passed through DenseNet's 1^{st} and 2^{nd} DenseBlock, output from 1^{st} transitional layer is concatenated with output of second DenseBlock, the feature maps are concatenated, passed through conv and fc layers.

ground truth masks from the Apple DMS Dataset, and subsequently utilizing the earlier-established mapping to generate masks with friction values assigned to every pixel across the entire image.

1) SRCNN (19): Building on insights gained from a previous mini project and inspired by findings in the literature (20), we test the performance of SRCNN model into our regression framework. SRCNN originally used for super-resolution takes in a bicubic interpolated image and is able to increase the resolution of image upto 2-4 times, and this network is learning some regression mapping, which gave us motivation to try this model. Drawing from our prior experience with SRCNN in a different context, we harnessed its capabilities for the nuanced challenges posed by our formulated problem.

2) U-Net: U-Net populary used for semantic segmentation, we add some modifications to tailor its functionality for regression purposes. The transformation involved changing the output channel from 57 (number of classes) to 1. By transitioning from 57 output channels to a singular channel, U-Net model was reconfigured to generate a single-channel output, aligning with the requirements of a regression task.

Recognizing the distinct demands of a regression task, we replaced the Cross-Entropy Loss with the more fitting Mean Squared Error (MSE) Loss. This adjustment fine-tuned the training process to better align with the regression-based objective.

C. EF-Net

In the development of our EF-Net architecture (Figure 4), a dedicated model crafted for RGB-to-spectral mapping, we executed a meticulous design process leveraging the VAST dataset. Here is a detailed breakdown of the architecture:

Network Design:

DenseNet: The input image undergoes processing through DenseNet (21), specifically passing through denseblock1 and denseblock2. This sequential passage through different dense blocks enables the network to learn fine-grained details in denseblock1 and high-level abstract representations in denseblock2. Features extracted from these layers contribute to capturing the texture of the surface, thereby enhancing our ability to achieve better spectral correspondence. Features from max-pooling of 1^{st} transition layer and 2^{nd} denseblock is passed to the Integration Layer.

Integration Layer: The outputs from the max pool layer after denseblock1 and denseblock2 are concatenated, creating a fused representation of global features from both the dense blocks.

CNN Layers:

- The fused representation is then fed through the first CNN layer, reducing input channels from 160 to 64, which in-turn helps in reducing the spatial dimension.
- Subsequently, the output from the first CNN layer is further processed by a second CNN layer, with input channels as 64 and output channels as 9. This step is done for regularization, and reducing the dimensionality such that when the network is passed through FC layers, it is more memory efficient and learns the representation quickly.

Fully Connected Layers:

- The output from the second CNN layer is flattened and passed through a fully connected layer with input channels as 1944 and output channels as 1550.
- This is followed by another fully connected layer, maintaining input and output channels at 1550.

This intricately designed architecture reflects our commitment to precision in RGB-to-Spectral mapping, utilizing a fusion of DenseNet169 and customized CNN layers to capture and process the nuanced variations present in the spectral data.

V. Experiments

A. Implementation

1) U-Net Segmentation: In the training of our U-Net model tailored for segmentation tasks on the Apple DMS dataset, a carefully chosen configuration was employed to optimize performance. The training settings are detailed below:

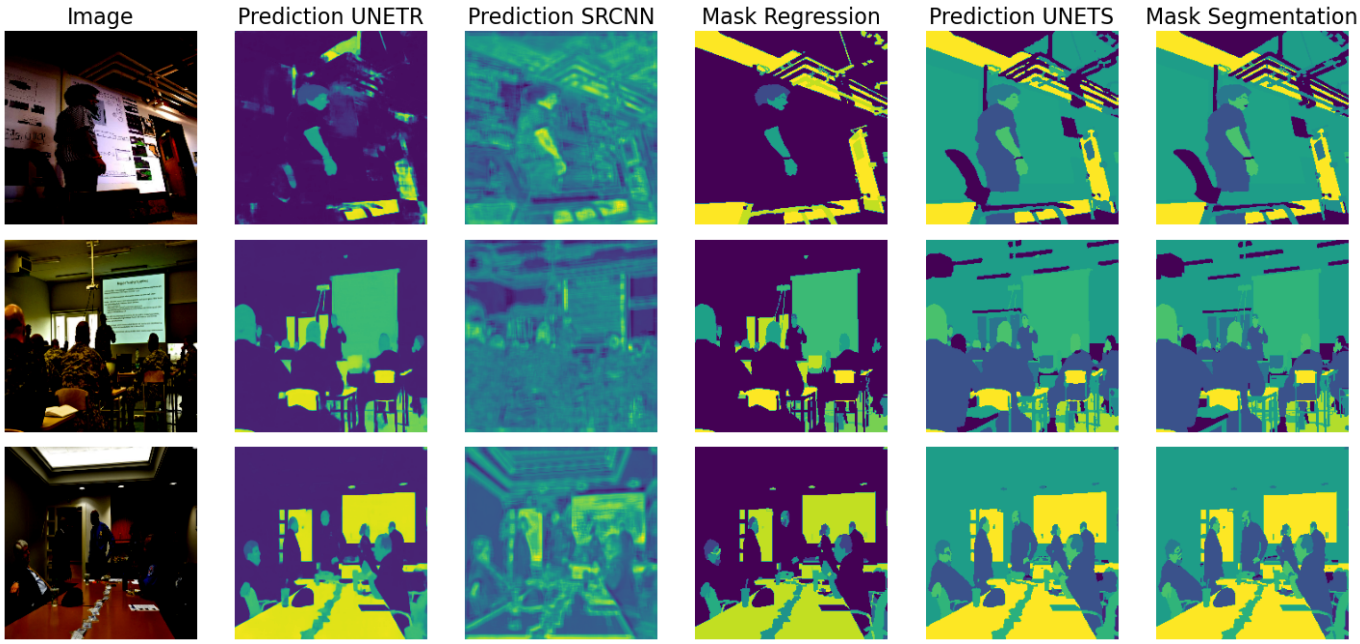

Fig. 5. Comparison between different results

- **Batch Size:** 32
- **Learning Rate:** 3e-4
- **Number of Epochs:** 300
- **Loss Function:** Cross Entropy Loss
- **Optimizer:** Adam Optimizer

2) SRCNN Regression: In the training of our SRCNN model, adapted for regression tasks on the tweaked Apple DMS dataset, a carefully chosen configuration was implemented to optimize model performance. The training settings are outlined below:

- **Batch Size:** 100
- **Learning Rate:** 3e-4
- **Number of Epochs:** 300
- **Loss Function:** Mean Squared Error (MSE) Loss
- **Optimizer:** Adam Optimizer

3) U-Net Regression: In the training of our U-Net model, specifically adjusted for regression tasks on the tweaked Apple DMS dataset, a meticulous configuration was applied to optimize model performance. The training settings are detailed below:

- **Batch Size:** 100
- **Learning Rate:** 3e-4
- **Number of Epochs:** 300
- **Loss Function:** Mean Squared Error (MSE) Loss
- **Optimizer:** Adam Optimizer

4) EF-Net: In the training of our proposed EF-Net, tailored for spectral-to-friction mapping using the VAST dataset, a carefully configured set of parameters was employed to optimize model performance. The training settings are as follows:

- **Number of Epochs (NUM EPOCHS):** 100
- **Batch Size:** 100
- **Learning Rate (lr):** 0.001
- **Loss Function:** Mean Squared Error (MSE) Loss
- **Optimizer:** Adam Optimizer

These parameters were thoughtfully selected to strike a balance between computational efficiency and the model's capacity to capture complex relationships within the output.

B. Evaluation

The assessment metric for the prediction model in segmentation relies on the validation data and is compared against the original mask. In the context of segmentation, the loss is calculated using the Dice score loss. Conversely, for other methods, we employ Mean Squared Error (MSE) loss for error evaluation.

MODEL	EVALUATION	SCORE
SRCNN	MSE	0.10±0.01
U-Net Regression	MSE	0.03±0.01
U-Net Segmentation	DICE	0.9402±0.01
EF-NET	MSE	0.002±0.001

TABLE I
RESULTS OF OUR IMPLEMENTATION

APPROACH	EVALUATION	SCORE
Brandão et al.(9)	Accuracy	79.29%±0.3
Kicki et al. (11)	Accuracy	66.3%±0.3
Hanson et al.(15)	Accuracy	99.98%±0.01

TABLE II
RESULTS OF OTHER IMPLEMENTATIONS

VI. RESULTS AND DISCUSSIONS

After training, the test set was evaluated (30% of the dataset), and the mean MSE loss for regression and DICE

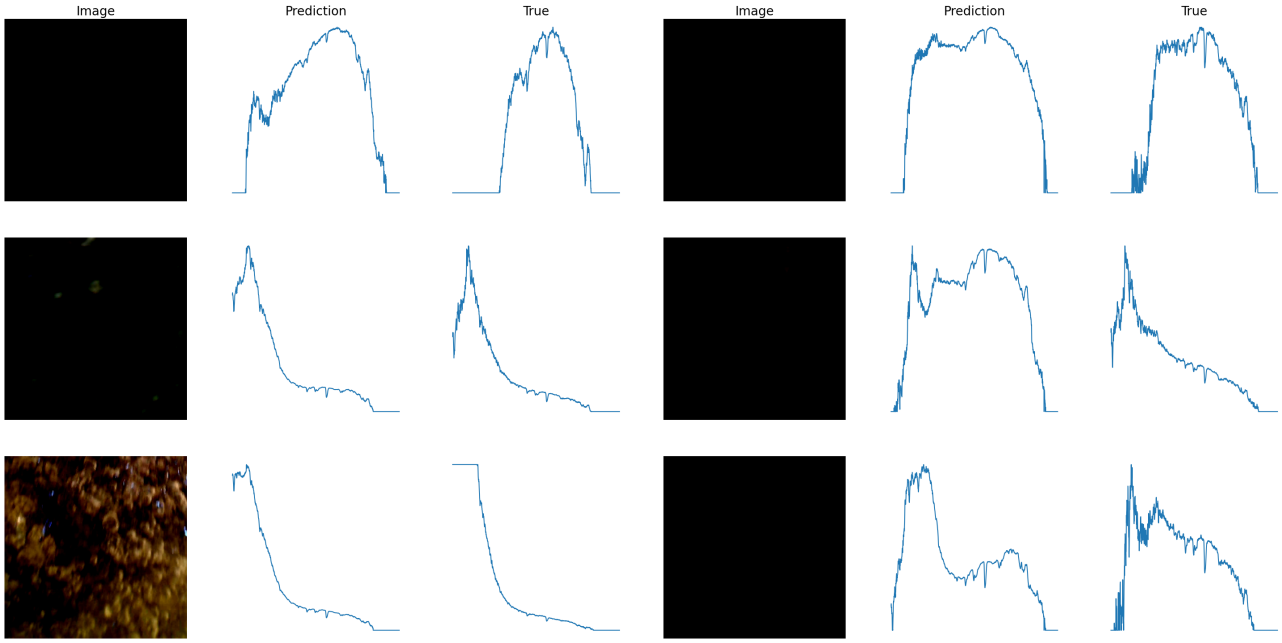

Fig. 6. Results from test set of seen data (left) and unseen data (right)

score for segmentation was computed, Table V-B shows the results. From Figure 5, SRCNN didn't estimated the friction correctly at all because of only convolution layers, it is able to learn high level features but fine features are still missing, however these higher level features can be fused with EF-Net in future and ablation studies can be done on how the the network performs when adding or removing global features.

Moving to U-Net, the classification performed better than existing work, however there were still certain scenarios where model was getting confused between material of same type or matching appearances, for example snow and wall. This is where U-Net regression outperformed U-Net segmentation, though the model was getting confused between classes, we were getting a gradient instead of one fix class, and theoretically these gradient values when combined with variance, can give better estimates compared to U-Net segmentation.

For the segmentation part, it is difficult to compare with prior work as the dataset utilized in both the cases were different, thus we just add the results reported by other authors of state of the art methods just for comparison. Table V-B accuracy for material classification.

At the end, EF-Net was able to find a good correspondence between RGB image and spectral profile of a material. Figure 6 shows the performance of the model on both on training on 11 classes and prediction on test set as well as training on 10 classes and prediction on 11^{th} class. From the results, and MSE score, the model is performing better than expected, we didn't expect the model to learn representations for an unseen material but the results were astonishing.

VII. Conclusion

The challenge of friction estimation, often overlooked in its complexity, has been a focal point of our research. In addressing this gap, we underscore the significance of acquiring a robust dataset tailored for this specific purpose. Our devised neural network demonstrates remarkable capabilities by successfully mapping RGB images to their corresponding spectral values, showcasing its adaptability to both seen and unseen data.

Notably, the versatility of our proposed methodology extends beyond friction estimation. The same framework can be harnessed for diverse applications, including but not limited to material recognition. This broader utility highlights the potential impact of our work on advancing not only the field of friction analysis but also contributing to a broader spectrum of material characterization tasks. As we bridge the gap between spectral data and real-world friction coefficients, our findings pave the way for enhanced understanding and application in multiple domains.

VIII. Future Work

In our ongoing research endeavors, the upcoming focus revolves around the meticulous creation of a specialized dataset, specifically geared towards mapping spectral data to friction coefficients. The primary objective is to establish a robust and nuanced correlation between the intricate spectral features and their corresponding friction values. Concurrently, we plan to design a dedicated neural network architecture with the overarching goal of achieving precise end-to-friction mapping. This novel network will take spectral data as its input and produce friction coefficients as output, seamlessly integrating into our existing framework.

Furthermore, recognizing the necessity for continuous improvement, an essential component of our future work involves the fine-tuning of our End-to-Friction Network (EF-Net). This refinement process aims to optimize the network's parameters, ensuring enhanced performance and accuracy in the spectral-to-friction mapping task. The culmination of these efforts will result in a complete End-to-Friction Network capable of not only processing images but also providing detailed friction values for each pixel. This strategic integration of dataset creation, network design, and EF-Net fine-tuning aims to elevate the sophistication and efficacy of our overall network architecture.

IX. Acknowledgements

We would like to thank Prof. Bruce Maxwell for his continuous support throughout the duration of this course and his invaluable guidance for this project work. We would like to thank Prof. Taskin Padir for providing us with invaluable guidance and his continuous support in the RIVeR Laboratory.

References

[1] A. Trivedi, S. Bazzi, M. Zolotas, and T. Padır, "Probabilistic dynamic modeling and control for skid-steered mobile robots in off-road environments," in *2023 IEEE International Conference on Assured Autonomy (ICAA)*, 2023, pp. 57–60.

[2] G. Bledt, M. J. Powell, B. Katz, J. Di Carlo, P. M. Wensing, and S. Kim, "Mit cheetah 3: Design and control of a robust, dynamic quadruped robot," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 2245–2252.

[3] R. Deits and R. Tedrake, "Footstep planning on uneven terrain with mixed-integer convex optimization," in *2014 IEEE-RAS International Conference on Humanoid Robots*, 2014, pp. 279–286.

[4] N. Thuerey, P. Holl, M. Mueller, P. Schnell, F. Trost, and K. Um, *Physics-based Deep Learning*. WWW, 2021. [Online]. Available: https://physicsbaseddeeplearning.org

[5] G. Xin, W. Wolfslag, H.-C. Lin, C. Tiseo, and M. Mistry, "An optimization-based locomotion controller for quadruped robots leveraging cartesian impedance control," *Frontiers in Robotics and AI*, vol. 7, 03 2020.

[6] L. Sharan, R. Rosenholtz, and E. H. Adelson, "Accuracy and speed of material categorization in real-world images," *Journal of Vision*, vol. 14, no. 10, 2014.

[7] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the materials in context database," *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[8] M. Brandão, K. Hashimoto, and A. Takanishi, "Friction from vision: A study of algorithmic and human performance with consequences for robot perception and teleoperation," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, 2016, pp. 428–435.

[9] M. Brandão, Y. M. Shiguematsu, K. Hashimoto, and A. Takanishi, "Material recognition cnns and hierarchical planning for biped robot locomotion on slippery terrain," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, 2016, pp. 81–88.

[10] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *CoRR*, vol. abs/1911.02685, 2019. [Online]. Available: http://arxiv.org/abs/1911.02685

[11] P. Kicki and K. Walas, "Friction from reflectance: Transfer learning approach," in *2019 4th International Conference on Robotics and Automation Engineering (ICRAE)*, 2019, pp. 79–83.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[13] Z. Erickson, N. Luskey, S. Chernova, and C. C. Kemp, "Classification of household materials via spectroscopy," *CoRR*, vol. abs/1805.04051, 2018. [Online]. Available: http://arxiv.org/abs/1805.04051

[14] Z. Erickson, S. Chernova, and C. C. Kemp, "Semi-supervised haptic material recognition for robots using generative adversarial networks," *CoRR*, vol. abs/1707.02796, 2017. [Online]. Available: http://arxiv.org/abs/1707.02796

[15] N. Hanson, M. Shaham, D. Erdoğmuş, and T. Padir, "Vast: Visual and spectral terrain classification in unstructured multi-class environments," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 3956–3963.

[16] D. Noh, H. Nam, M. S. Ahn, H. Chae, S. Lee, K. Gillespie, and D. Hong, "Surface material dataset for robotics applications (smdra): A dataset with friction coefficient and rgb-d for surface segmentation," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 6275–6281.

[17] P. Upchurch* and R. Niu*, "A dense material segmentation dataset for indoor and outdoor scene parsing," in *ECCV*, 2022. [Online]. Available: https://arxiv.org/abs/2207.10614

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015. [Online]. Available: http://arxiv.org/abs/1505.04597

[19] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *CoRR*, vol. abs/1501.00092, 2015. [Online]. Available: http://arxiv.org/abs/1501.00092

[20] W. Yao, Z. Zeng, C. Lian, and H. Tang, "Pixel-wise regression using u-net and its application on pansharpening," *Neurocomputing*, vol. 312, pp. 364–371, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231218307008

[21] G. Huang, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016. [Online]. Available: http://arxiv.org/abs/1608.06993