Force-EvT: A Closer Look at Robotic Gripper Force Measurement with Event-based Vision Transformer

Qianyu Guo[1], Ziqing Yu[2], Jiaming Fu[3], Yawen Lu[4], Yahya Zweiri[5], Dongming Gan[6]

***Abstract*— Robotic grippers are receiving increasing attention in various industries as essential components of robots for interacting and manipulating objects. While significant progress has been made in the past, conventional rigid grippers still have limitations in handling irregular objects and can damage fragile objects. We have shown that soft grippers offer deformability to adapt to a variety of object shapes and maximize object protection. At the same time, dynamic vision sensors (e.g., event-based cameras) are capable of capturing small changes in brightness and streaming them asynchronously as events, unlike RGB cameras, which do not perform well in low-light and fast-moving environments. In this paper, a dynamic-vision-based algorithm is proposed to measure the force applied to the gripper. In particular, we first set up a DVXplorer Lite series event camera to capture twenty-five sets of event data. Second, motivated by the impressive performance of the Vision Transformer (ViT) algorithm in dense image prediction tasks, we propose a new approach that demonstrates the potential for force estimation and meets the requirements of real-world scenarios. We extensively evaluate the proposed algorithm on a wide range of scenarios and settings, and show that it consistently outperforms recent approaches.**

***Index Terms*— Event-based Vision, Vision Transformer, Dynamic Vision Sensor, Soft Robotic Gripper.**

I. INTRODUCTION

The robotic hand represents a critical component of a robot, typically mounted on the robot's arms. A key part of the robotic hand is the gripper, which facilitates interaction with the environment and manipulation of target items. Gripping mechanisms find extensive application across various industrial sectors, including the food industry [17], healthcare [16], and agriculture [8]. Traditional grippers often employ rigid metals as their primary material. However, such rigid grippers lack flexibility in handling objects with irregular shapes and may inadvertently damage items made from fragile materials. In contrast, innovative soft robotic grippers can deform to accommodate the shape and size of the target object, thereby enhancing the object's protection. Consequently, soft grippers emerge as the preferred option for many manipulation tasks [11], [10].

[1]Qianyu Guo is a Ph.D. student in the School of Engineering Technology at Purdue University, USA. guo716@purdue.edu

[2]Ziqing Yu is a Ph.D. student in the School of Engineering Technology at Purdue University, USA. yu1154@purdue.edu

[3]Jiaming Fu is a Ph.D. student in the School of Engineering Technology at Purdue University, USA. fu330@purdue.edu

[4]Yawen Lu is a Ph.D. student in the Computer Graphics Technology Department at Purdue University, USA. lu976@purdue.edu

[5]Dr. Yahya Zweiri is a professor of Aerospace Engineering at Khalifa University, UAE. yahya.zweiri@ku.ac.ae

[6]Dr. Dongming Gan is an associate professor in the School of Engineering Technology at Purdue University, USA. dgan@purdue.edu

Using vision-based methods to predict the deformation and force applied to the robotic gripper is a popular topic, while the use of traditional RGB cameras is always the first choice. However, when the experimental environment is dark or the gripper is moving very fast, RGB cameras cannot clearly capture the trajectory of objects, or even get images with significant motion blur, where thermal imaging [19] and LiDAR [20] are often utilized as compensation. Event cameras, known as bio-inspired sensors, are able to detect very slight brightness changes at the pixel-level and output them as events, as shown in Fig. 1. The events include four important information: timestamps, x coordinates, y coordinates, and their polarities. Compared to standard cameras, event cameras have four remarkable advantages. *High Temporal Resolution*: Capture slight brightness changes fast, and output events in the order of μs. *Low Power Consumption*: Due to their efficient design, which only transmits brightness changes and does not output redundant data, event cameras achieve low power consumption. *Wide Dynamic Range*: Event cameras can acquire visual information by over 120 dB, exceeding standard cameras by over 60 dB, thanks to logarithmic-scale photoreceptors and asynchronous pixel operation. *Low Latency*: Event cameras have ultra-low latency since pixels detect and transmit changes independently without global exposure timing [12]. Therefore, event cameras have a strong ability to capture gripper motion, even if the experimental environment is not perfect.

The cooperation of machine learning and event cameras is a new way of solving computer vision problems, which achieves great performance [32], [24], [25]. For example, object detection [26], object tracking [28], 3D reconstruction [30], steering prediction for self-driving cars [22], optical flow and intensity estimation [2], etc. These methods typically use a continuous stream of asynchronous events, which allows for efficient processing. Nevertheless, due to the sparse and unstructured nature of the event streams, it is a challenge to directly observe and process the event data [37]. To better adapt to the traditional frame-based computer vision algorithms, most event data is converted into event frames based on timestamp or polarity.

The advantage of exploring the Vision Transformer on force measurement via event frames has become evident. Notably, in a real-world scenario (Fig. 2), we achieve a 13.0% percentage error and 0.13 N RMSE compared to the ground truth force sensor measurement, benefiting from the *event frame representation* and *event transformer architecture*. In the application of our previous proposed variable stiffness robotic gripper [11], it is an important part of outputting

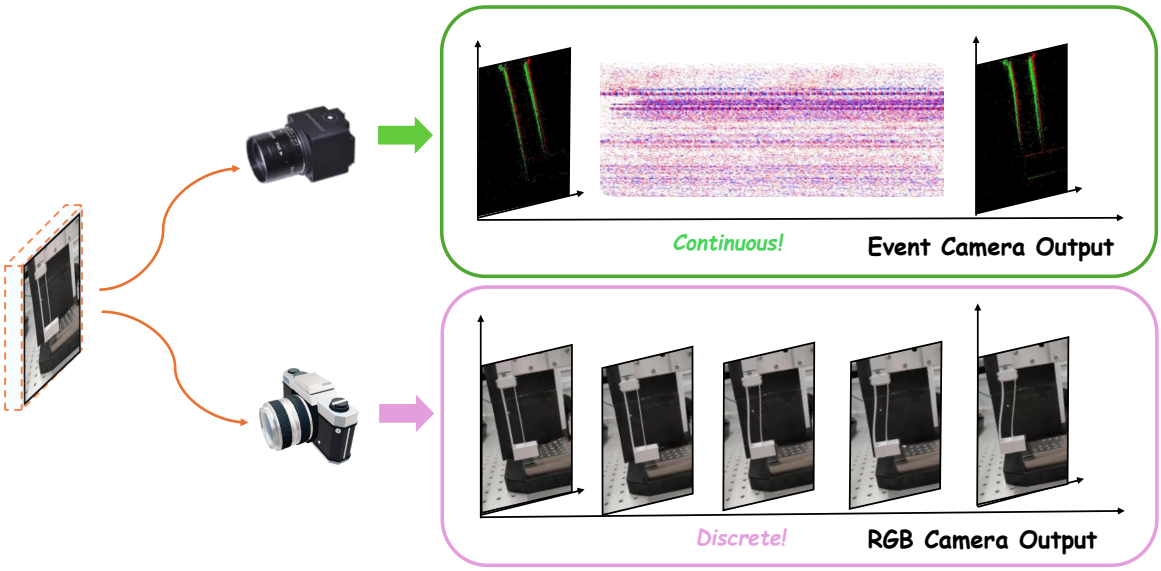


Fig. 1. Comparison of RGB Camera and event camera outputs in experimental scenarios. While traditional cameras output frame-based images, event cameras asynchronously output detected events. The comparison reveals significant distinctions in how these two types of cameras capture and process visual information.

required force to the control end. ArUco marker detection strategy was used to monitor the deformation of the gripper, enabling the control end to read the force applied to the grippers efficiently. However, while applying this technique in a less-than-ideal environment, for example, if the illumination is not enough, the marker detection would become intermittent. Therefore, in the paper, we propose a vision-based Vision Transformer in force measurement task, which is able to capture the slight deformation of the robotic gripper and make force predictions with high accuracy.

The main contributions of this work are summarised as:

- We propose a novel approach to estimate the force applied to a robotic gripper using a Dynamic Vision Sensor.
- We collect a dataset using our custom-designed robotic gripper, a force sensor and an event camera, named RG-Event, which contains 1000 event frames and their corresponding force labels.
- We utilize state-of-the-art Vision Transformer architecture as a backbone to train the data collected using an event camera. We show that Vision Transformer performs well in regression tasks.

II. Related Work

In this section, we explore various recent advancements in event-based vision and Vision Transformer techniques, as well as existing methodologies for force measurement in robotics, encompassing traditional contact-type force sensors, and sensorless approaches.

A. Event-based Vision

Event-based vision is a developing technique and has great potential. Due to their advantages of high temporal resolution, low power consumption, low latency and high dynamic range, these bio-inspired visual sensors are able to be used in various complex environments. In [22], the authors proposed a deep neural network for steering angle prediction using event cameras, showcasing superior performance in challenging conditions like low light and fast motion compared to traditional cameras. Weng et al. first presented a novel Transformer-based network called ET-Net for event-based video reconstruction [37]. GSCEventMOD was an approach for detecting moving objects based on events, which had great performance in challenging scenarios such as fast movements and sudden changes in lighting conditions [23]. In addition, a lot of computer vision applications such as optical flow estimation, depth estimation, motion segmentation, and visual-inertial odometry all achieved excellent performances using event-based methods [12].

B. Vision Transformer in Prediction

Transformer refers to a type of neural network architecture that was initially utilized in natural language processing (NLP) tasks, such as machine translation and text generation [9]. Inspired by the successful utilization in NLP, Transformer has been gradually applied to computer vision tasks [34], [33], [5], which largely improved conventional CNN and LSTM based networks [35], [7]. For instance, Ranftl et al. presented a novel architecture called dense prediction transformers, which employed Vision Transformer instead of convolutional networks as the foundational structure for tasks requiring dense predictions [29]. In [27], Vision Transformer was employed alongside Convolutional Neural Networks (CNN) to forecast urban traffic congestion. TransDepth is also a novel transformer-based network, aiming to make pixel-wise predictions in various computer vision tasks, such as depth estimation, surface normal estimation. Lu et al. introduced TransFlow which used a pure transformer for optical flow

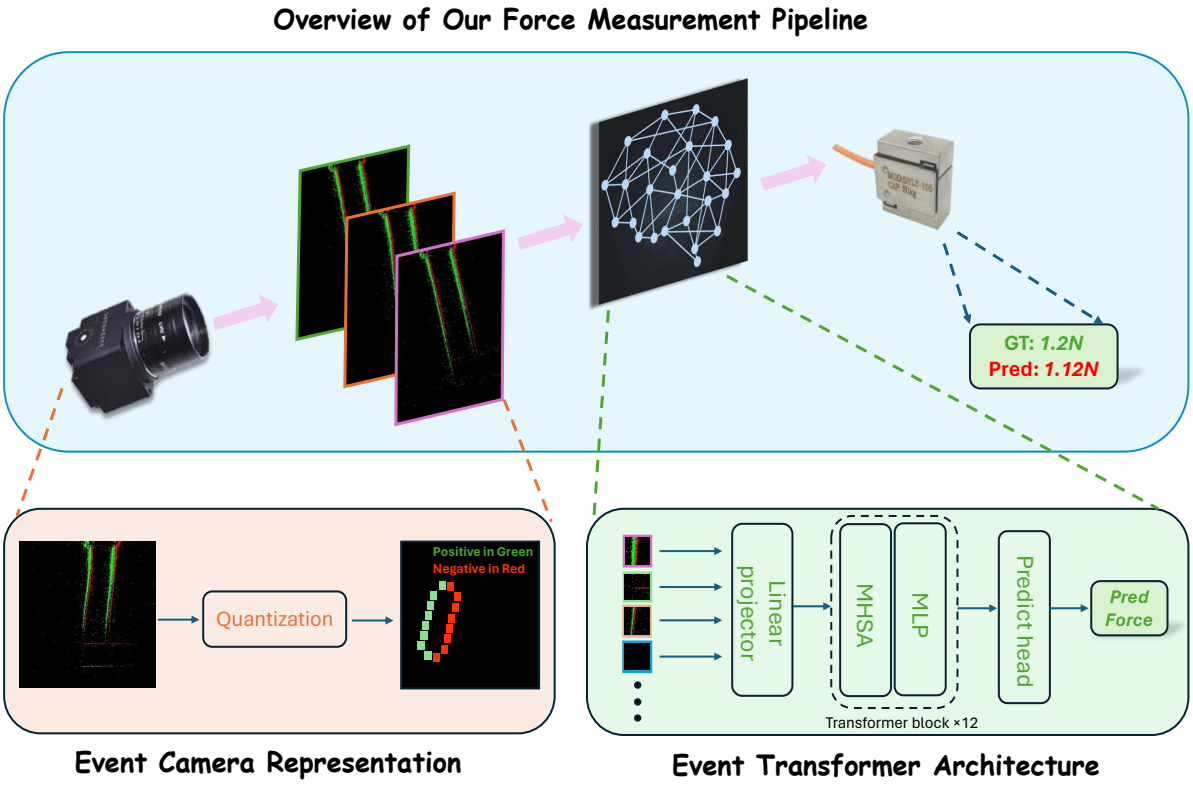


Fig. 2. Overview of the proposed method Force-EvT. The output of the event camera is converted into event frames over a certain time interval T. The events are colored according to polarity (positive in green, negative in red). Then, the event frames are processed by Vision Transformer (ViT) network, which makes force prediction of the robotic gripper.

estimation, demonstrating the effectiveness of incorporating spatial self-attention and cross-attention mechanisms [21].

C. Force Measurement

The existing force measurement methods applied to robot hands usually fall into two categories, namely traditional force sensors and non-sensor methods. Researchers have proposed various force sensors integrated into gripper structures. In [14], the Gifu hand II is introduced, featuring the capability to be equipped with a six-axis force sensor at each fingertip, showcasing high integration levels. Dai et al. developed a contact force transducer based on a six-component Stewart platform to enhance reliability and precision [6]. In [15], Kuang et al. introduce a novel hinged-joint cantilever beam sensor structure designed to reduce sensor nonlinearity. The aforementioned traditional contact-type force sensors typically exhibit high accuracy and reliability but come with drawbacks, such as occupying substantial structural space and increasing the complexity of the structure. Therefore, with the advancements in technologies like machine learning and computer vision, researchers have interdisciplinarily proposed new sensorless methods for force sensing. For instance, [38] captures deformations in nodes located on the framework of the fin ray gripper, [39] observes the movement of markers on the soft layer, [41] measures changes in the angle of the fingers. Furthermore, Baghaei Naeini et al. proposed a dynamic-vision based approach to measure contact force on silicone membrane, using Convolutional Recurrent Neural Networks [1]. These sensorless methods simplify the structure of robotic hands. However, they also face the challenge of insufficient precision. Additionally, the RGB cameras used to capture deformations operate continuously, potentially consuming plenty of system resources and resulting in high energy consumption.

III. Methodology

As shown in Figure 2, our methodology is designed to precisely quantify the forces applied to a robotic gripper, utilizing data captured by an event camera. In this section, we first address the conversion of raw event data into a structured frame format, according to a certain time interval. Then, we employ a regression algorithm based on the Vision Transformer architecture to estimate the forces applied to the robotic gripper. Finally, we introduce the loss function that guides training towards precise predictions.

A. Event Frame based Representation

Deep learning algorithms, which stand at the forefront of recent advancements in machine learning, have been developed with a focus on processing conventional frame-based data [18]. To bridge the gap between the unique data structure produced by event cameras and the requirements of these advanced algorithms, we first perform an event-to-frame conversion. In this case, the asynchronous events will

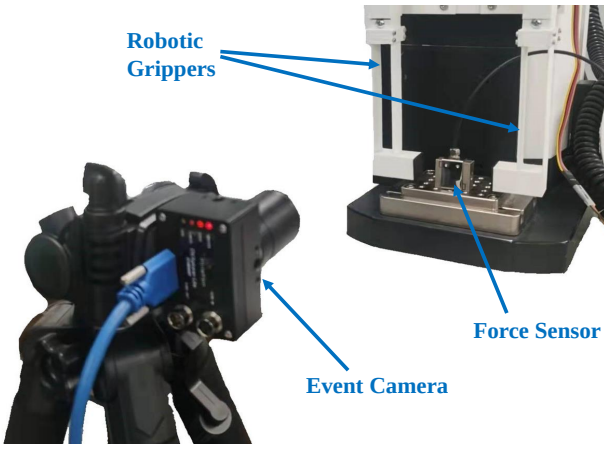


Fig. 3. Experimental setup to conduct data collection using an event camera, a force sensor, and robotic grippers.

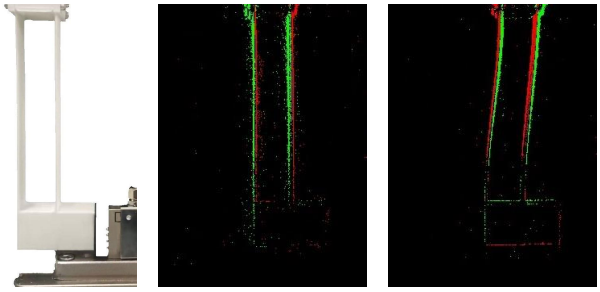


Fig. 4. The gripper is captured by an RGB camera and an event camera. (a) displays our designed soft robotic gripper captured by a standard RGB camera in a state without any applied force, (b) and (c) show the gripper under different deformation states, as captured by an event camera.

be converted into synchronous frames. As mentioned before, an event, as captured by an event camera, consists of four key attributes: the x-coordinate (x_i), the y-coordinate (y_i), the timestamp (t_i), and the polarity (p_i) of the change in brightness. Therefore, we use $e_i = (x_i, y_i, t_i, p_i)$ to represent the attributes.

In general, event-to-frame conversion can be approached through various methodologies, primarily based on timestamps, the number of events, or the polarity of events [40]. In this work, we adopt a timestamp-based approach for constructing event frames. Given a certain time interval *T*, a flow of events would be divided into numerous event-based frames. Since the event camera is highly sensitive to changes of brightness, the slight deformation of the robotic gripper could be clearly captured, and the edges of the gripper are clearly depicted in each event frame.

B. Vision Transformer based Force Measurement

Vision Transformer (ViT) is a powerful deep learning architecture that can be used in computer vision tasks. In this work, we leverage the ViT as a foundational architecture to estimate the forces applied to a robotic gripper, an application where precision and contextual understanding of spatial relationships are important.

Unlike traditional Convolutional Neural Networks (CNNs) which analyze images through a hierarchical sequence of localized filters, ViT approaches the task by dividing the input event frames into fixed-size patches (8×8 pixels in our implementation). Each patch is then transformed into a high-dimensional vector through a linear embedding process. This transformation not only preserves the spatial hierarchy of the original image but also allows for a more granular analysis of the visual content. Once embedded, these patches are fed into a series of transformer encoder, allowing the model to capture both local and global features within patches and learn the relationships across the entire image. For force estimation on a robotic gripper, this means the ViT can intelligently focus on critical regions of the input frames that are most indicative of the applied forces, such as areas of significant deformation or contact points between the gripper and objects. In addition, the adaptability and efficiency of ViT are further enhanced by its self-attention mechanism, which allows for selective focus on salient features within the image patches, effectively ignoring irrelevant information [9].

C. Loss Function

In the development of our force measurement model, an essential component is the choice of an appropriate loss function to guide the training process towards accurate predictions. For this purpose, we employ the Mean Squared Error (MSE) as loss function in our experiments [31], [13]. MSE is widely recognized for its efficacy in regression tasks and its ability to quantify the variance between predicted values and ground truth. The selection of MSE is motivated by its sensitivity to large errors, making it particularly suitable for ensuring precision in force measurement.

The MSE is mathematically defined as the average of the squared differences between the predicted forces ($\hat{y}_i$) and the actual measured forces (y_i), overall N samples in the dataset. The formula for MSE is given by:

$$MSE = \frac{1}{N}\sum_{N}^{i=1}(\hat{y}_i - y_i)^2 \tag{1}$$

During the model training phase, the minimization of MSE facilitates the adjustment of model parameters that incrementally improves the accuracy of force predictions.

IV. Experiments

In this section, we first present the experimental settings of the proposed approach. Following this, we introduce the process of data collection and data preprocessing steps taken to prepare the data for analysis. Finally, force measurement evaluations are provided.

A. Experimental Setup

As shown in Figure 3, in this work, the experimental setup contains a tension and compression sensor (commonly referred to as a force sensor), an Arduino microcontroller, an event camera (DVS sensor), and a custom 3D-printed robotic

gripper. The event camera, known for its high-speed and low-latency imaging capabilities, provided asynchronous visual information crucial for dynamic scenes. The force sensor, grasped by the robotic gripper, enabled precise measurement of the forces exerted during object manipulation tasks. An Arduino microcontroller served as the central processing unit, orchestrating data acquisition and communication between the force sensor and robotic gripper. In the experiments, the force sensor is positioned at the center of the two grippers to ensure optimal force measurement. Through the action of grasping the robotic grippers, the force sensor receives the applied force data at a rate of 10 samples per second.

In the implementation, we train and test the specific vit_base_patch8_224 model on our collected dataset. In the experiment, the dataset is randomly divided into a training set, a validation set, and a testing set in the ratio of 7: 1.5: 1.5. Our force prediction training uses an Adam optimizer with a learning rate set to 0.001 and a batch size of 16. This entire pipeline is deployed on two GeForce RTX 3090s GPU platforms using the PyTorch framework.

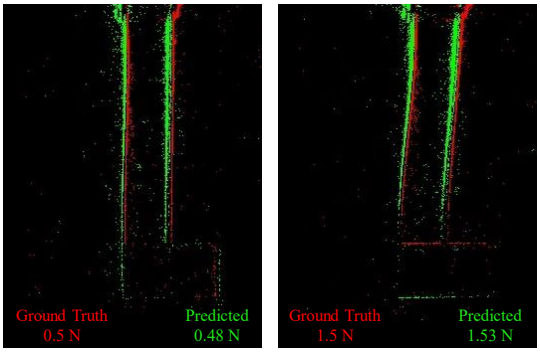
B. Dataset

To estimate the force exerted on our designed soft robotic gripper, we conduct a comprehensive data collection using the event camera. Employing the experimental setup described previously, we repeat the process of closing, grasping, and opening the gripper for 25 times. Given our interest in capturing the critical moments when the gripper undergoes deformation, our dataset exclusively encompasses event frames corresponding to the grasping phase.

Finally, a total of 1000 event frames were gathered using the DVXplorer Lite camera, and we call the dataset as **RG-Event**. During each experimental iteration, the force applied to the gripper varied within a range from 0 N to 1.6 N throughout the grasping stage. We synchronize the data collection windows for both the force sensor and the event camera to a duration of T = 100 ms, ensuring precise temporal alignment between the sensory inputs and the visual data. Illustrated in Figure 4, there are two representative images extracted from our dataset, showcasing the progressive deformation of the gripper from its initial state to the point of maximum deformation.

C. Force Measurement Evaluations

In order to evaluate the performance of using ViT in the force prediction task, we use the Root Mean Squared Error (RMSE) [4], [36] and R-squared (R^2) [3] as evaluation metrics. RMSE is a measure of the average deviation of the predictions made by a model from the actual observed values. Lower RMSE values indicate better performance of the model, as it means the model's predictions are closer to the actual values. R^2 is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variables in a regression model. It ranges from 0 to 1, where R^2 value closer to 1 indicates a better fit of the model to the data. In our testing stage, we get RMSE as 0.13 N and R^2 as 0.93. As shown in Figure 5, the deformed grippers with 0.5 N and 1.5 N are predicted as 0.48 N and 1.53 N respectively. The prediction results with high accuracy demonstrate the effectiveness of the proposed method. Furthermore, as shown in Table I, we provide the performance comparisons between the novel event-based approach and our previous marker-based approach [11].



(a) **(b)**

Fig. 5. The prediction results demonstrate the efficacy and accuracy of employing Force-EvT in force measurement task.

	$RMSE$	R^2	$Error$
Marker-based Approach [11]	-	-	19.5%
Force-EvT	0.13	0.93	13%

TABLE I

OUR NOVEL FORCE-EVT MODEL IS ABLE TO ACHIEVE BETTER PERFORMANCE IN FORCE MEASUREMENT COMPARED WITH OUR PREVIOUS MARKER-BASED METHOD.

V. CONCLUSION AND FUTURE WORK

In this paper, we introduce a novel approach named Force-EvT for predicting forces applied to soft robotic grippers using event-based vision. Leveraging a Dynamic Vision Sensor, particularly the DVXplorer Lite event camera, we capture and record gripper deformation processes. By employing the Vision Transformer (ViT) algorithm, our proposed method demonstrates promising results and potential for force estimation in robotic applications. Experimental evaluations validate the effectiveness of the approach, highlighting its suitability for measuring forces applied to soft robotic grippers.

For future works, we intend to expand our experiments to encompass different illumination conditions, including both very bright and dark environments, to demonstrate the superiority of using event cameras in force measurement projects. Moreover, we plan to incorporate more complex designs of robotic grippers into our training data. By diversifying our dataset, we can enhance the robustness and adaptability of our approach to a wider range of gripper configurations and applications.

ACKNOWLEDGMENT

This work is supported by the National Science Foundation (NSF) grant under CMMI-2131711.

REFERENCES

[1] F. Baghaei Naeini, D. Makris, D. Gan, and Y. Zweiri. Dynamic-vision-based force measurements using convolutional recurrent neural networks. *Sensors*, 20(16):4469, 2020.

[2] P. Bardow, A. J. Davison, and S. Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 884–892, 2016.

[3] A. C. Cameron and F. A. Windmeijer. An r-squared measure of goodness of fit for some common nonlinear regression models. *Journal of econometrics*, 77(2):329–342, 1997.

[4] T. Chai and R. R. Draxler. Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. *Geoscientific model development*, 7(3):1247–1250, 2014.

[5] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

[6] J. Dai and D. Kerr. A six-component contact force measurement device based on the stewart platform. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 214(5):687–697, 2000.

[7] W. Dai, J. Tao, X. Yan, Z. Feng, and J. Chen. Addressing unintended bias in toxicity detection: An lstm and attention-based approach. In *2023 5th International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, pp. 375–379. IEEE, 2023.

[8] M. D. A. Dhanawade and M. N. V. Sabnis. A review: State of the art of robotic grippers. *Int Res J Eng Technol*, 5(5):371–375, 2018.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] J. Fu, Q. Guo, and D. Gan. Machine learning based deflection prediction and inverse design for discrete variable stiffness units. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, vol. 87363, p. V008T08A043. American Society of Mechanical Engineers, 2023.

[11] J. Fu, Z. Yu, Q. Guo, L. Zheng, and D. Gan. A variable stiffness robotic gripper based on parallel beam with vision-based force sensing for flexible grasping. *Robotica*, pp. 1–19, 2023.

[12] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.

[13] Q. Guo and J. Wen. Multi-level fusion based deep convolutional network for image quality assessment. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part VI*, pp. 670–678. Springer, 2021.

[14] H. Kawasaki, T. Komatsu, and K. Uchiyama. Dexterous anthropomorphic robot hand with distributed tactile sensor: Gifu hand ii. *IEEE/ASME transactions on mechatronics*, 7(3):296–303, 2002.

[15] L. Kuang, Y. Lou, and S. Song. Design and fabrication of a novel force sensor for robot grippers. *IEEE Sensors Journal*, 18(4):1410–1418, 2017.

[16] M. Kyrarini, F. Lygerakis, A. Rajavenkatanarayanan, C. Sevastopoulos, H. R. Nambiappan, K. K. Chaitanya, A. R. Babu, J. Mathew, and F. Makedon. A survey of robots in healthcare. *Technologies*, 9(1):8, 2021.

[17] T. Lien. Gripper technologies for food industry robots. In *Robotics and automation in the food industry*, pp. 143–170. Elsevier, 2013.

[18] Y. Lu, Q. Guo, and G. Lu. A geometric convolutional neural network for 3d object detection. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1–5. IEEE, 2019.

[19] Y. Lu and G. Lu. Superthermal: Matching thermal as visible through thermal feature exploration. *IEEE Robotics and Automation Letters*, 6(2):2690–2697, 2021.

[20] Y. Lu, Z. Sun, J. Shao, Q. Guo, Y. Huang, S. Fei, and V. Chen. Lidar-forest dataset: Lidar point cloud simulation dataset for forestry application. *arXiv preprint arXiv:2402.04546*, 2024.

[21] Y. Lu, Q. Wang, S. Ma, T. Geng, Y. V. Chen, H. Chen, and D. Liu. Transflow: Transformer as flow learner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18063–18073, 2023.

[22] A. I. Maqueda, A. Loquercio, G. Gallego, N. García, and D. Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5419–5427, 2018.

[23] A. Mondal, J. H. Giraldo, T. Bouwmans, A. S. Chowdhury, et al. Moving object detection for event-based vision using graph spectral clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 876–884, 2021.

[24] R. Muthusamy, X. Huang, Y. Zweiri, L. Seneviratne, and D. Gan. Neuromorphic event-based slip detection and suppression in robotic grasping and manipulation. *IEEE Access*, 8:153364–153384, 2020.

[25] F. B. Naeini, A. M. AlAli, R. Al-Husari, A. Rigi, M. K. Al-Sharman, D. Makris, and Y. Zweiri. A novel dynamic-vision-based approach for tactile sensing applications. *IEEE Transactions on Instrumentation and Measurement*, 69(5):1881–1893, 2019.

[26] E. Perot, P. De Tournemire, D. Nitti, J. Masci, and A. Sironi. Learning to detect objects with a 1 megapixel event camera. *Advances in Neural Information Processing Systems*, 33:16639–16652, 2020.

[27] K. Ramana, G. Srivastava, M. R. Kumar, T. R. Gadekallu, J. C.-W. Lin, M. Alazab, and C. Iwendi. A vision transformer approach for traffic congestion prediction in urban areas. *IEEE Transactions on Intelligent Transportation Systems*, 24(4):3922–3934, 2023.

[28] B. Ramesh, S. Zhang, Z. W. Lee, Z. Gao, G. Orchard, and C. Xiang. Long-term object tracking with a moving event camera. In *Bmvc*, p. 241, 2018.

[29] R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12179–12188, 2021.

[30] H. Rebecq, G. Gallego, E. Mueggler, and D. Scaramuzza. Emvs: Event-based multi-view stereo—3d reconstruction with an event camera in real-time. *International Journal of Computer Vision*, 126(12):1394–1414, 2018.

[31] U. Sara, M. Akter, and M. S. Uddin. Image quality assessment through fsim, ssim, mse and psnr—a comparative study. *Journal of Computer and Communications*, 7(3):8–18, 2019.

[32] C. Scheerlinck, H. Rebecq, D. Gehrig, N. Barnes, R. Mahony, and D. Scaramuzza. Fast image reconstruction with an event camera. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 156–163, 2020.

[33] Z. Sun, S. Cao, Y. Yang, and K. M. Kitani. Rethinking transformer-based set prediction for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3611–3620, 2021.

[34] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 568–578, 2021.

[35] W. Weimin, L. Yufeng, Y. Xu, X. Mingxuan, and G. Min. Enhancing liver segmentation: A deep learning approach with eas feature extraction and multi-scale fusion. *International Journal of Innovative Research in Computer Science & Technology*, 12(1):26–34, 2024.

[36] J. Wen and Q. Guo. When distortion meets perceptual quality: A multi-task learning pipeline. In *PRICAI 2021: Trends in Artificial Intelligence: 18th Pacific Rim International Conference on Artificial Intelligence, PRICAI 2021, Hanoi, Vietnam, November 8–12, 2021, Proceedings, Part III 18*, pp. 353–365. Springer, 2021.

[37] W. Weng, Y. Zhang, and Z. Xiong. Event-based video reconstruction using transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2563–2572, 2021.

[38] W. Xu, H. Zhang, H. Yuan, and B. Liang. A compliant adaptive gripper and its intrinsic force sensing method. *IEEE Transactions on Robotics*, 37(5):1584–1603, 2021.

[39] T. Zhang, Y. Cong, X. Li, and Y. Peng. Robot tactile sensing: Vision based tactile sensor for force perception. In *2018 IEEE 8th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*, pp. 1360–1365. IEEE, 2018.

[40] X. Zheng, Y. Liu, Y. Lu, T. Hua, T. Pan, W. Zhang, D. Tao, and L. Wang. Deep learning for event-based vision: A comprehensive survey and benchmarks. *arXiv preprint arXiv:2302.08890*, 2023.

[41] H. Zhu, X. Li, W. Chen, and C. Zhang. Flexure-based variable stiffness gripper for large-scale grasping force regulation with vision. In *Intelligent Robotics and Applications: 12th International Conference, ICIRA 2019, Shenyang, China, August 8–11, 2019, Proceedings, Part I 12*, pp. 346–357. Springer, 2019.