IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

Multimodal Detection and Classification of Robot Manipulation Failures

Arda Inceoglu, *Student Member, IEEE*, Eren Erdal Aksoy, and Sanem Sariel, *Member, IEEE*

***Abstract*—An autonomous service robot should be able to interact with its environment safely and robustly without requiring human assistance. Unstructured environments are challenging for robots since the exact prediction of outcomes is not always possible. Even when the robot behaviors are well-designed, the unpredictable nature of the physical robot-object interaction may lead to failures in object manipulation. In this letter, we focus on detecting and classifying both manipulation and post-manipulation phase failures using the same exteroception setup. We cover a diverse set of failure types for primary tabletop manipulation actions. In order to detect these failures, we propose FINO-Net (Inceoglu et al., 2021), a deep multimodal sensor fusion-based classifier network architecture. FINO-Net accurately detects and classifies failures from raw sensory data without any additional information on task description and scene state. In this work, we use our extended FAILURE dataset (Inceoglu et al., 2021) with 99 new multimodal manipulation recordings and annotate them with their corresponding failure types. FINO-Net achieves 0.87 failure detection and 0.80 failure classification F1 scores. Experimental results show that FINO-Net is also appropriate for real-time use.**

***Index Terms*—Deep learning methods, data sets for robot learning, failure detection and recovery, sensor fusion.**

I. Introduction

CURRENTLY, robots are not robust enough to safely handle all house chores on their own, especially while executing manipulation tasks in unstructured environments where they continuously interact with humans and everyday objects [2]. It is hard to predict the actual outcomes of actions in these settings, where unintended or harmful consequences may arise. This is due to the fact that the estimations on manipulation/interaction parameters leading to success may be wrong due to incomplete or incorrect internal representation of the world. Furthermore, environmental factors (e.g., external events) may also prevent proper action performance, leading to undesirable outcomes. In such circumstances, ensuring the reliability and safety of the robot manipulations is crucial [3].

Manuscript received 23 July 2023; accepted 29 November 2023. Date of publication 22 December 2023; date of current version 2 January 2024. This letter was recommended for publication by Associate Editor M. Li and Editor M. Vincze upon evaluation of the reviewers' comments. This work was supported by the Scientific and Technological Research Council of Türkiye under Grant 119E-436. *(Corresponding author: Arda Inceoglu.)*

Arda Inceoglu and Sanem Sariel are with the Artificial Intelligence and Robotics Laboratory, Faculty of Computer and Informatics Engineering, Istanbul Technical University, Maslak 34718, Türkiye (e-mail: inceoglua@itu.edu.tr; sariel@itu.edu.tr).

Eren Erdal Aksoy is with the School of Information Technology, Center for Applied Intelligent Systems Research, Halmstad University, 30118 Halmstad, Sweden (e-mail: eren.aksoy@hh.se).

Digital Object Identifier 10.1109/LRA.2023.3346270

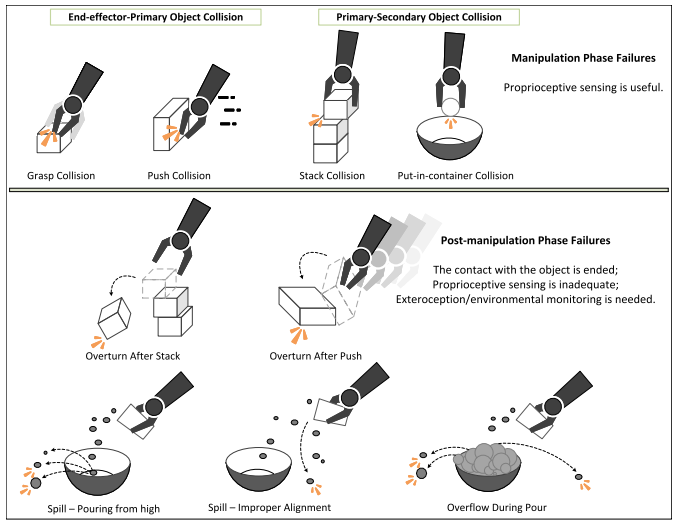


Fig. 1. Manipulation failures can be categorized along two dimensions: (i) one for the collision time: either at the manipulation phase or at the post-manipulation phase; and (ii) one for the object types that are in collision: either end-effector-primary object collisions or primary-secondary object collisions. Failure types on a set of primary manipulation actions (e.g., *pick*, *push*, *stack*, *put-in-container*, *pour*) that we focus on in this study are categorized along these dimensions.

In industrial settings, safety measures are taken by external safeguarding devices, and certain regulations and standards are applied by operators [4], [5]. However, for autonomous operations in human workspaces, the robot itself is responsible for the safety of its own actions. Situation awareness is, therefore, required to make robots able to effectively operate in unstructured environments without assistance [6], [7], [8]. In this study, we focus on these onboard situation awareness capabilities for service robots.

If the robot can detect and classify a failure effectively, it can react appropriately when confronted with it [9]. This can be achieved through real-time monitoring of the robot workspace during manipulation. In this letter, we present Failure Is Not an Option (FINO)-Net to sense the presence of failures by multimodal exteroceptive sensors without integrating any domain/task-specific knowledge. Previously, failure detection [10], [11] and failure type identification [12] were addressed at the symbolic level [13]. In the context of scene analysis, egocentric vision and audition were separately processed in various relevant works, [14], [15]. Our works differ in that we present FINO-Net as a data-driven framework for monitoring, while we further integrate egocentric vision with audition to complement each other.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

Our main interest is in failures that may occur during real-time execution of primitive object manipulation actions, where the robot end-effector is in close contact (i.e., without tools) with objects on the tabletop. Failure types that we consider involve collisions with objects (either manipulated or secondary objects), missing objects being manipulated, overflowing or spilling the contents of a container, and overturning of objects (see Fig. 1). The causes of these failures are as follows:

- improper alignment of the robot arm with respect to the manipulated object due to unspecified or misidentified parameters (e.g., grasping or pushing from a misaligned orientation, pouring from high, etc.) based on the internal representations of the world,
- incorrect estimations or perceptual errors on objects (e.g., underestimating the contents of a container),
- wrong assumptions about the skills of the self.

There is a large body of research addressing collision detection both in physical human-robot interaction and manipulation tasks; most of them are, however, referring to collision detection using only proprioceptive sensors [16], [14], [17], [15], [18]. For instance, torque/force sensors, microphones, and inertial measurement units are used to detect physical collisions with objects in manipulation settings [17]. Joint/end-effector and object/human collisions are of particular interest in these works. On the contrary, we focus on classifying both contact phase and post-manipulation phase failures [19] affecting either the primary objects being manipulated or the secondary objects that are not in direct contact with the end-effector (Fig. 1).

In our earlier work [10], we presented a preliminary analysis on contributions of different modalities, specifically proprioception, audition, and vision in the detection of manipulation failures. Our analysis reveals that proprioceptive sensors constitute the primary source of information for detecting failures during the manipulation phase, such as those encountered during grasp and push actions. Nevertheless, proprioceptive sensing does not yield valuable insights when addressing post-manipulation phases (i.e. when the contact between the end-effector and manipulated object ends). Effective post-manipulation failure detection requires environmental monitoring and the use of exteroceptive sensors since post-impact phase failures are further expected. We, therefore, particularly address how to detect and identify a robot manipulation failure by using exteroception and provide an analysis of how exteroception enhances the capabilities of proprioception.

When it comes to sensory data, which has a crucial role in detecting and identifying failures, multisensory integration is more desirable to take advantage of each sensor's perceptual contribution [1]. Recently, there have been great efforts to collect large-scale multimodal robot manipulation data [20], [21], [22], [23], [24]. These efforts, however, center around manipulation skill learning tasks and ignore failures emerging during manipulation executions. To the best of our knowledge, there are no publicly available multimodal datasets on robot manipulation failures.

In this context, the FAILURE [1] dataset is a collection of the failed attempts of various robot manipulation actions (e.g., push, pour, pick&place, etc.) coming with multimodal sensor readings such as RGB images, depth data, and audio waves. Note that manipulation actions in our dataset are defined based on the ontology introduced in [25]. At least one manipulation type from each goal-oriented manipulation category is selected from this ontology. FAILURE covers a variety of manipulation failures that emerge from different affected objects or at different manipulation phases. Fig. 1 depicts a categorization of failure types. A failure can occur either during the manipulation phase or in the post-manipulation phase. On the other hand, the end-effector may get involved in a collision with the primary object, or even the primary and secondary objects may collide with each other. In this work, we address all these interactions during the manipulation and post-manipulation phases.

To address the above-mentioned failures, we present FINO-Net [1], a deep sensor fusion-based multimodal classifier network for detecting and identifying manipulation failures. FINO-Net can accurately detect manipulation failures by incorporating visual (RGB and depth) and auditory modalities. In addition, our network employs early fusion to combine RGB and depth frames, while late fusion is further integrated to merge vision and audio data. To represent spatio-temporal features in sensory observations, modalities are processed separately with a series of convolutional and convolutional-LSTM layers, and the latent space representations are finally combined to detect potential failures. In this work, we extend FINO-Net to further classify manipulation failure types. Furthermore, we deeply analyze real-time on-demand failure detection and classification capabilities of FINO-Net on the extended FAILURE dataset.

Contributions of this letter are as follows: i) We introduce the multimodal sensor fusion-based FINO-Net architecture. ii) We analyze the FINO-Net performance for real-time on-demand failure detection and classification. iii) We release the extended FAILURE dataset as a multimodal (RGB, depth, and audio) real-world dataset involving, in total, 324 manipulation scenarios from 5 different manipulation types. This dataset also contains annotations of failure types.

We here note that existing execution monitoring and failure detection studies cover only a single manipulation action (e.g., grasping), where the main focus is on robot-primary object interactions. In comparison, we cover a set of primary manipulation actions (Fig. 1). Furthermore, we also address failures involving both primary (i.e., manipulated main object) and secondary objects (i.e., other objects with which the robot is not in direct contact) in the workspace.

II. Related Work

There is a large body of literature on failure detection and execution monitoring. In some of these works, failure, fault or anomaly keywords are used interchangeably. The work in [26] summarizes deep learning-based anomaly detection for various application domains. From the robotics perspective, there are model-based and model-free approaches proposed for execution monitoring [27]. The former approach compares the already known models with observations, whereas the latter uses sensory observation to make predictions [28], [29] [30].

Among recent works, [31] extends planning with a vision-based execution monitoring system, [32] analyzes different preprocessing techniques for introspective data to detect gearbox failures. Non-parametric Bayesian models [33] and Non-parametric Hidden Markov Models (HMMs) [34] are investigated to detect and classify anomalies.

A multimodal execution monitoring system for assistive feeding tasks is proposed in [35], [36], where the authors adopt LSTM-based variational autoencoders to process multimodal input from a sensor set including a camera, a microphone, a

joint encoder, and a force sensor. In another work [37], multimodal cues are used to detect book manipulation failures on shelves. The work in [38] fuses visuo-tactile cues for grasp failure detection. Furthermore, [12] extracts predicates from multimodal inputs, which are then combined for anomaly-cause identification. In a later work [39], end-to-end approaches are also investigated.

Our work differs from these studies in that it addresses the problem of detecting and identifying manipulation failures, which are mainly caused by uncertainties in perception and execution. Unlike the works in [35], [36], where the focus is on the failures emerging during human-robot interaction, our work investigates the robot-object interaction failures observed over the course of object manipulation.

Instead of hand-crafted features such as gripper status, audio events, and object displacements used in [10], [40] or sound energy, spoon position, and mouth position employed in [35], [36], our proposed perception framework learns feature representations directly from the raw multimodal sensory data in an end-to-end fashion.

III. Real-Time Detection and Classification of Manipulation Failures

In the following subsections, the problem description, data, and network details are presented.

A. Problem Description

Manipulation failures are inevitable in unstructured environments due to limitations in perception, estimation, and the physical capabilities of the robot. For safety reasons, robots need to be situationally aware of their actions' outcomes. Situation awareness can be formally defined as the spatio-temporal perception of the elements in the environment, their interpretation, and the projection of their state in the near future [41].

In order to build situationally aware robots, their execution should be continuously monitored, as an unexpected event may occur in any phase of the execution. An execution monitoring system should, therefore, incorporate both spatial and temporal information obtained from the sensors to interpret changes in the scene during manipulation. Furthermore, monitoring the execution with multiple sensors is helpful, as each sensing modality provides complementary information [1], [10], [40].

In this work, we focus on real-time and on-board detection and classification of tabletop manipulation failures caused by the robot itself (i.e., egocentric perception with sensors mounted on the robot). We define manipulation failure detection as the process of detecting unexpected outcomes during robot execution, whereas failure classification refers to explaining the types of failures.

For addressing the aforementioned problems, failure detection and failure classification problems are modeled as classification tasks, where the inputs are multisensory observation sequences and the targets are $y \in \{success, fail\}$ for detection, and $y \in \{success, collision, miss, overflow, spill, overturn\}$ for classification.

Let $M = \bigcup_m (m \in \{RGB, Depth, Audio, etc.\})$ be the set of sensing modalities. x_i^m is a complete observation sequence obtained from modality m (e.g., all RGB images of a pouring action), and i is the manipulation recording index of the multimodal observation sequence. $x_i^{m_{t_m}}$ represents an observation

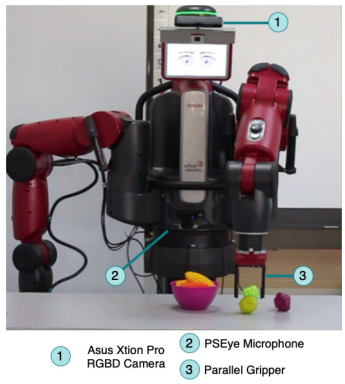


Fig. 2. Experimental environment.

TABLE I
Distribution of the Manipulation Data

	#Successes			#Failures			*Total*
Manipulation	**Existing**	**New**	**Total Success**	**Existing**	**New**	**Total Failure**	
Push	12	18	30	19	12	31	61
Pick&place	13	2	15	30	2	32	47
Pour	25	24	49	42	4	46	95
Put-in-container	23	21	44	31	9	40	84
Stack-a-tower	9	4	13	21	3	24	37
Total	82	69	151	143	30	173	324

(e.g., a single RGB frame) at time step t_m where t_m is the sampled time index for modality m. We construct a dataset D ($|D| = N$) containing multimodal observation sequences as:

$$D = \{\{(x_i^{m_{t_1}}, \ldots, x_i^{m_{t_m}})\}_{m=1}^{M}, y_i\}_{i=1}^{N} . \quad (1)$$

The goal is to learn a function $\Phi(\cdot)$ that classifies multimodal sensory data to a label y as:

$$y = \Phi(\phi_1(x^{1_1}, \ldots, x^{1_{t_1}}), \ldots, \phi_m(x^{m_1}, \ldots, x^{m_{t_m}})) . \quad (2)$$

B. Data

In our previous work, we introduced the FAILURE dataset [1]. To the best of our knowledge, there currently exist no public datasets with multimodal robot execution traces for both successful and failed cases, covering a diverse set of manipulation actions.

Our FAILURE dataset is constructed by using a Baxter humanoid robot equipped with the following equipment: a parallel gripper, an Asus Xtion Pro RGB-D camera mounted on the head, and a PSEye microphone mounted on the lower torso (See Fig. 2). During the data collection, the robot is tasked to execute a manipulation action, and all synchronized sensor readings (i.e., RGB/RGB-D image streams and audio waves) are then simultaneously recorded.

In this work, we extend the dataset with 99 new manipulation execution recordings. Table I presents the distribution of the existing and newly collected data.

Furthermore, we have annotated failure cases in each recording in the FAILURE dataset. The resulting execution status types are as follows:

- *Success:* Execution is completed as expected. Minor deviations from the expected target location or orientation are tolerated.
- *Collision:* A collision failure occurs during all action executions except pour. In this failure type, both the manipulated object and secondary objects are affected.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

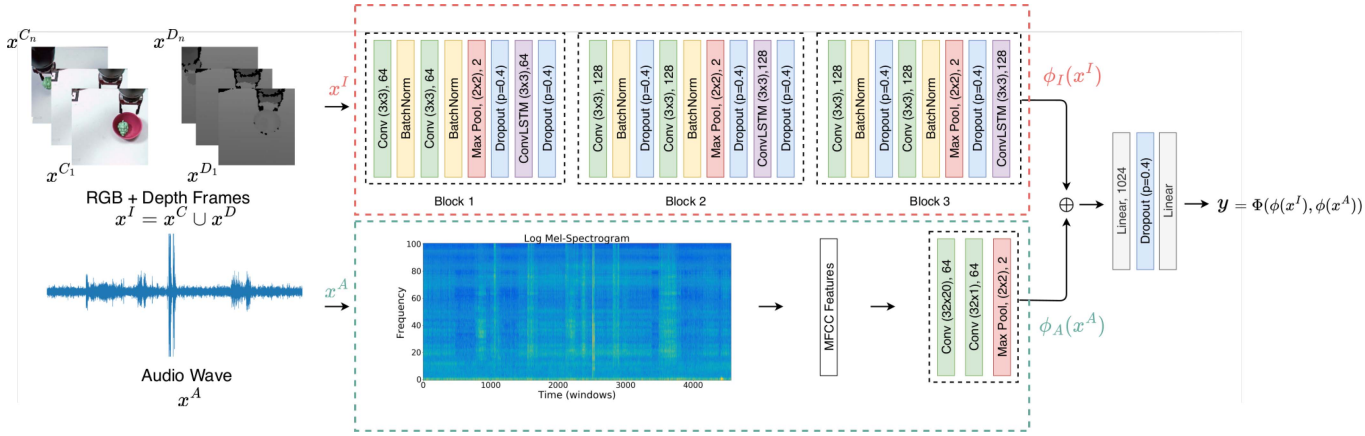


Fig. 3. FINO-Net Architecture.

TABLE II
DISTRIBUTION OF FAILURE TYPES BY MANIPULATION ACTION

Type \Action	**put-in-container**	**pour**	**push**	**place**	**stack-a-tower**	*Total*
Success	44	50	30	15	13	152
Collision	33		2	23	19	77
Miss	7		6	7	2	22
Overflow		25				25
Spill		20				20
Overturn			23	2	3	28
Total	84	95	61	47	37	324

- *Miss:* A miss failure occurs during all action executions except pour. The robot either completely fails to interact with the object or fails to grasp it due to alignment errors. This failure type emerges when the location or size of the target object is miscalculated.
- *Overflow:* An overflow failure is particularly obtained during the execution of a pouring action. The poured content completely fills the target container and keeps on overflowing onto the table.
- *Spill:* A spill failure occurs during a pouring action. The poured content is spilled onto the table due to alignment errors, the pouring velocity, the size mismatch between source and target containers, etc. In comparison to overflow failure, in this failure type, the target container is not full at the end of the execution.
- *Overturn:* An overturn failure occurs during either pushing, placing, or stacking-a-tower action. In this failure type, only the manipulated object is affected and overturned during the manipulation phase.

During failure annotation, we grouped the interactions between the robot, the target object, and the other objects in the workspace. Table II presents the distribution of failure types over manipulation actions. Note that the presented failure classes are challenging, as i) a single failure type can occur during the execution of different actions. For instance, an overturn failure can be observed during either a put-in-container, place, or stack-a-tower action; ii) a manipulation action can result in different types of failures. For instance, during the execution of the pour action, overflow or spill failures can be observed.

C. FINO-Net

We present FINO-Net, a deep sensor fusion-based multimodal classifier network to detect manipulation failures using onboard sensory data in real-time. An earlier version has appeared in [1]. This work extends FINO-Net to further classify manipulation failure types. We further present a detailed analysis of failure detection with the extended dataset.

The inputs of the network are composed of RGB (x^C) and depth (x^D) frames captured from the head camera and audio waves (x^A) recorded over the course of a robot manipulation action. FINO-Net adopts early fusion ($x^I = x^C \cup x^D$) to combine RGB and depth frames, while applying late fusion (Φ) to combine visual (ϕ_I) and auditory (ϕ_A) features. The architecture processes visual and auditory inputs individually with a series of convolutional and convolutional-LSTM (convLSTM) layers. Finally, in the fusion step, the latent space representations are concatenated into a feature vector, and fed to the fully connected layers. The overall FINO-Net architecture is depicted in Fig. 3. In the following subsections, we elaborate more on the network architecture.

1) Vision (ϕ_I): In the preprocessing step, 8 RGB and depth image pairs are sampled, representing the complete execution sequence. Prior to the sampling step, self-occluded frames are eliminated with a depth-based thresholding approach. The remaining samples are roughly segmented into *approach*, *manipulate*, and *retreat* phases. Then, from each *approach* and *retreat* phase, 4 frames are sampled.

To process spatio-temporal features in RGB and depth frames, convLSTM cells are employed. A typical LSTM cell is implemented using fully connected layers, while a convLSTM replaces these with convolution operators.

The visual branch consists of three main blocks (see the top branch in Fig. 3). Each block is composed of two convolutional layers and a convLSTM layer. RGB and depth frames are early fused by stacking on top of each other before feeding into the first visual block. Inside each block, the filter numbers remain the same for all convolutional and convLSTM layers. Each convolutional layer has 3x3 filters. Before applying convLSTM layers, max pooling is applied to cut the number of features by half. Each block also has batch normalization and dropout layers.

We note that the total number of convolutional-LSTM blocks in the visual branch is determined empirically. Reducing the number of these blocks results in lower representation capacity. Although increasing the number of blocks improves accuracy slightly, the number of parameters increases drastically. Thus, having three blocks is the most ideal choice when

considering the trade-off between accuracy and computational complexity.

2) Audition (ϕ_A): In our earlier works [10], [40] we have shown that the use of audition data helps in detecting drop and collision types of failures for which RGB and depth data do not provide sufficient clues due to occlusions, etc. Therefore, audition data was used as a complementary modality to the others. Our earlier work on audition data processing involved Mel Frequency Cepstral Coefficients (MFCCs) as a representation for auditory monitoring, where Support Vector Machines were employed to classify audio features into symbolic states: *drop*, *collision*, *idle* and *ego-noise* [10], [40]. In comparison, FINO-Net directly classifies MFCC features into execution states: *success* or *failure*.

FINO-Net adopts a convolutional network composed of two convolutional layers followed by a max pooling layer (see the bottom branch in Fig. 3). There are 64 filters in each layer, with a filter size of 32. As input, we use single-channel audio recordings with a 16 kHz sampling rate. The raw audio signal is divided into 32-millisecond windows. For each window, Short Time Fourier transform is applied to convert the signal into the frequency domain. A Mel filter bank is then employed, and 20 MFCCs are obtained. The number of audio windows is fixed by either applying padding or clipping.

3) Fusion (Φ): FINO-Net adopts a late fusion approach to combine visual and auditory modalities. We introduce the following model:

$$y = \Phi(\phi_I(x^{I_1}, \ldots, x^{I_{t_I}}) \oplus \phi_A(x^{A_1}, \ldots, x^{A_{t_A}})), \qquad (3)$$

where ϕ_m is a unimodal convolutional neural network that acts as a feature extractor, $\oplus$ is the concatenation operator, and Φ is the late fusion-based classifier network. In the fusion step, vision and audition features, obtained from the final output of each modality, are concatenated into a single feature vector. The fusion layer is composed of two fully connected layers as shown in Fig. 3.

4) Regularization: Batch normalization is applied after each convolutional layer for regularization. To boost the roles of basic features (e.g., edges and curves), a central dropout approach is adopted with a probability rate of 0.4. After each convolutional, convLSTM, and fully connected layers, a dropout layer is inserted except for the first convolutional layer in Block 1 and the last convLSTM layer in Block 3.

IV. Experiments

In this section, we quantitatively assess the prediction performance of the FINO-Net architecture on both failure detection and classification tasks using our extended FAILURE dataset. Additionally, we explore the real-time monitoring performance of FINO-Net. For this purpose, we first look at the effect of sampled frames from a robot execution. Next, we investigate the prediction accuracy within different temporal segments of the execution.

A. Quantitative Evaluation

For the quantitative assessment, the extended FAILURE dataset is split into training (70%), validation (10%), and test sets (20%). All network weights are initialized randomly and trained for 250 epochs using the Adam optimizer with a learning rate of $1e-5$. An early stopping strategy is adopted to select the best model based on validation set scores. The test scores obtained with the selected best models are reported in the following subsections.

TABLE III
Quantitative Evaluation

	Failure Detection			Failure Classification		
	Pr	Re	F1	Pr	Re	F1
FINO-Net-RGB	0.7667	0.7567	0.7617	0.6924	0.6216	0.6551
FINO-Net-D	0.6898	0.6756	0.6826	0.7302	0.6486	0.6869
FINO-Net-RGB-D	0.7817	0.7567	0.7690	0.6846	0.6756	0.6801
FINO-Net-A	0.6801	0.6486	0.6640	0.7475	0.6216	0.6788
FINO-Net-RGB-D-A	**0.8665**	**0.8648**	**0.8656**	**0.8085**	**0.7837**	**0.7959**

To prevent overfitting, we augment the data by applying color augmentation and random flipping. For instance, the brightness, contrast, saturation, and hue values of all images in a sequence are randomly changed with a probability of 0.2. In a similar fashion, each image sequence is flipped vertically with a probability of 0.5.

FINO-Net is trained with the following inputs for both detection and classification tasks:

- *FINO-Net-RGB:* The network is trained with RGB frames.
- *FINO-Net-D:* The network is only trained with the depth (D) frames.
- *FINO-Net-A:* Only the audio (A) branch is trained. After the convolutional layers, a single fully connected layer with 64 neurons is appended.
- *FINO-Net-RGB-D:* The visual branch of FINO-Net is trained by stacking the RGB and depth frames as input to the network.
- *FINO-Net-RGB-D-A:* The entire network is trained with all the given modalities.

Table III presents obtained quantitative results. FINO-Net-RGB-D-A is able to achieve the highest F1 scores for failure detection and classification (0.8656 and 0.7959, respectively). The same trend is observed in precision and recall scores in both tasks. These results verify that multimodal sensor fusion yields higher scores as visual and auditory modalities provide complementary information.

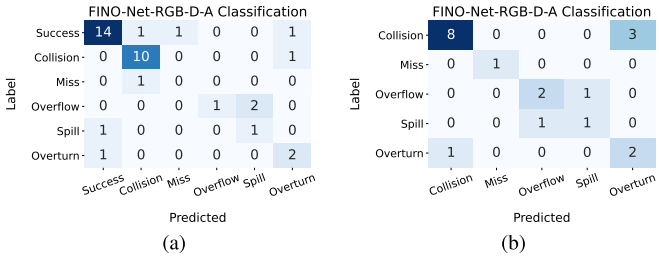
We further investigate two different approaches for the failure-type classification task. First, a stand-alone approach is conducted to simultaneously detect and classify failures. In this approach, the network is trained with both successful and failed executions, thus, there are in total 6 class labels (i.e., success and failure types). Next, a cascaded approach is considered, where the failure classification network is only triggered once a failure is detected. For this purpose, we excluded successful executions from the dataset and trained the network only with failed samples.

Fig. 4 presents confusion matrices for stand-alone and cascaded approaches, respectively. Looking deeper into the results, we observe that models confuse classes due to the scene similarities in the post-manipulation phases. For instance, particles are laid on the table due to overflow or spill failures. Primary object locations and orientations change whenever a collision or overturn failure occurs. Even though the sources of the failures are distinct, and different failure indicators are identified in the manipulation phase, similar scene states are very likely to be observed in the post-manipulation phase. All these reasons form

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

TABLE IV
EFFECT OF FRAME SAMPLING. AVERAGES OF 50 PREDICTIONS

	Failure Detection			Failure Classification		
	Pr	Re	F1	Pr	Re	F1
FINO-Net-RGBD	0.7331 ± 0.045	0.7276 ± 0.041	0.7274 ± 0.041	0.6987 ± 0.045	0.6805 ± 0.049	0.6820 ± 0.046
FINO-Net-RGBDA	$\mathbf{0.8201 \pm 0.044}$	$\mathbf{0.8092 \pm 0.043}$	$\mathbf{0.8054 \pm 0.045}$	$\mathbf{0.8154 \pm 0.039}$	$\mathbf{0.7795 \pm 0.047}$	$\mathbf{0.7884 \pm 0.043}$



(a) (b)

Fig. 4. (a) Confusion matrix for stand-alone FINO-Net-RGB-D-A failure classification. Success class is included. (b) Confusion Matrix for cascaded FINO-Net-RGB-D-A failure classification. Success class is omitted.

Fig. 5. Analysis on real-time demands for failure detection and classification. Different rates of the observation sequence of the execution are provided to FINO-Net-RGB-D-A model.

the main source of misclassifications captured in the confusion matrices.

FINO-Net can also make predictions on demand. In order to assess the suitability of FINO-Net for real-time deployment on the robot, two analyses are conducted. First, each manipulation is temporarily segmented by execution's completion rate. Next, 8 RGB and depth frame pairs are sampled from the beginning to the end of each segment for the detection and identification tasks. Fig. 5 reports the results of this analysis. Typically, indicators of failures are observed during the second half of the execution. Results verify that the model makes more reliable predictions towards the end of manipulations as more descriptive information becomes available. Note that the models are only trained with frames sampled from the complete manipulation sequence. Yet, FINO-Net is still capable of detecting and identifying failures in the absence of a complete manipulation sequence.

The performance of our FINO-Net model is also affected by the sampled images, as these frames may not necessarily contain enough information about the failure phase or may fail to capture the effects of the failure. In order to measure the effect of the quality of the sample images, 50 different image sequences are randomly sampled for each manipulation. Average results are given in Table IV.

B. Qualitative Evaluation

We also demonstrate how FINO-Net behaves in a longer compound scenario where multiple manipulations are chained. For instance, in the scenario illustrated in Fig. 6, the robot first executes a successful pouring followed by a successful put-in-container and a failed pushing action. In this long manipulation sequence, the scene contains not only task-related objects but also novel secondary objects, which are not presented in the training phase.

As depicted in Fig. 6, FINO-Net-RGB-D detection and classification models make correct predictions for all three manipulation actions. For the second manipulation action, FINO-Net-RGB-D-A detection and classification models make incorrect predictions as, during the put-in-container phase, a sound event is observed (i.e., the transparent container collides with the purple container). For the final manipulation action, all models except the FINO-Net-RGB-D-A classification model make correct predictions. Even though the model catches the failure, it confuses overturn with collision.

The audio modality is challenging to work with since environmental background noise and the robot's ego noise affect the model's performance. Furthermore, most of the object interactions generate a sound event regardless of the manipulation success (e.g., pouring pasta into the pan). Various types of sounds are generated from object-object interactions [15]. Even though the FAILURE dataset includes a variety of object materials, the FINO-Net architecture naturally faces difficulties in discriminating successful and failed sound events on new unseen object materials.

C. Generalization Performance on Novel Actions

Supplementary experiments are carried out to assess generalization performance of FINO-Net on novel actions. A wiping scenario is selected as a case study where 11 new out-of-distribution test samples are collected in the real world. In this particular scenario, a cooking scene that is more cluttered, as compared to the training dataset, is arranged featuring novel objects. There is spilled pasta on the table, and the goal of the robot is to wipe pasta into a predefined location (i.e., to the left side of the workspace) (See Fig. 7). Occasionally, a human participant also intervenes the task execution by placing a new object on the table, removing or displacing the object from the table. The manipulation is considered failed if any collisions occur within the scene.

The obtained F1 scores for the wiping scenario are 0.64 and 0.36 for detection and classification, respectively. Note that we applied no additional network training. This test setup imposes quite a few challenges (e.g., novel objects and unseen scene setup; dynamic intervention of the human, etc.). Despite all these challenges, we still obtain a relatively high detection score (0.64). The same performance is indeed not visible when it comes to the fine-grained classification task, as the network starts

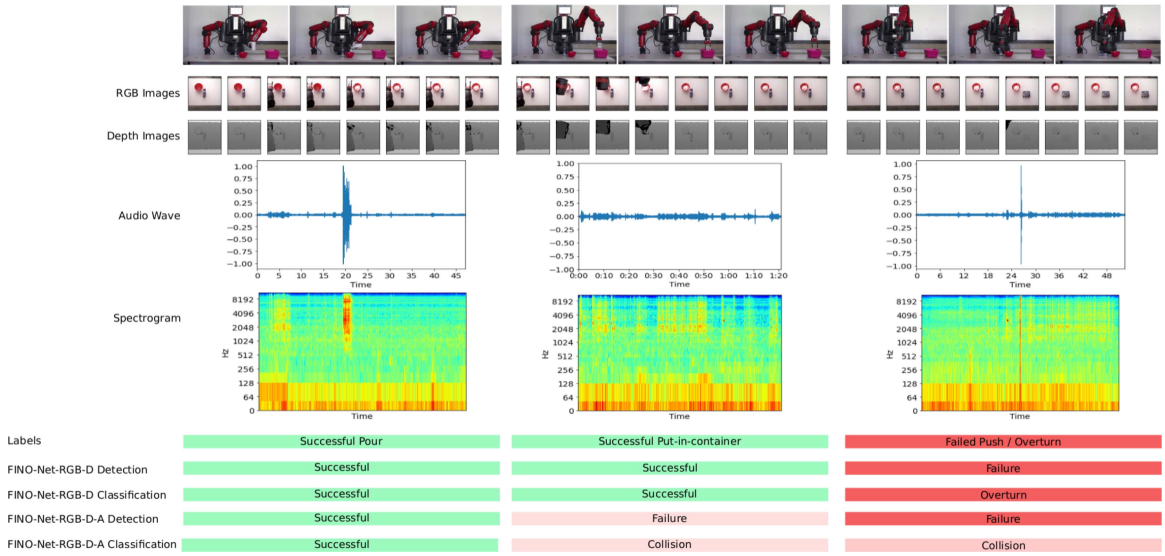


Fig. 6. FINO-Net evaluation on a compound scenario. The robot starts withs a successful pouring, followed by a successful put-in-container and a failed pushing manipulation. From top to bottom, rows present the following: third-person camera view, 8 RGB and depth frame pairs sampled from the execution (i.e., input to FINO-Net), audio wave, and corresponding spectrogram images, labels, and predictions.

(a) Successful wipe execution.

(b) Failed wipe execution.

Fig. 7. Sample images for wipe action execution. (a) The robot wipes the spilled pasta on the tabe. (b) The robot collides with the pasta box during manipulation. FINO-Net correctly detects failure and success cases.

getting confused with the causes of failures. This deficiency is due to the fact that the scene involves many out-of-distribution objects, such as spilled pasta, frying pan, etc. This naturally leads to confusion, even though FINO-Net accurately detects the failure. These findings convey the fact that the scalability of the proposed network needs to be improved with more data, particularly for fine-grained identification tasks. Supplementary results and materials are available at the project page.[1]

V. Conclusion

Previous studies investigated model-based and data-driven execution monitoring systems. In our previous works [10], [40], we have also represented auditory and visual inputs as symbolic predicates. However, data-driven manipulation failure detection and classification approaches have been underexplored due to the lack of open-source robot execution datasets. This work bridges the gap by providing FAILURE, an open-source, multimodal real-robot execution dataset. In this work, we extend the FAILURE dataset with 99 new multimodal manipulation executions and annotate each with the corresponding failure type. We present the FINO-Net architecture as an end-to-end framework that directly learns the relationship between inputs and targets for both failure detection and failure classification tasks. Quantitative results indicate that FINO-Net is capable of successfully detecting and classifying failures.

Working with the audio modality is challenging as different object material types have different sound characteristics. We have included materials such as letter, plastic, and wooden objects in our dataset. However, we may expect performance drops on novel object-object & robot-object interactions, in particular, when the object is from out-of-distribution. Environmental noises (sound events generated outside the robot's workspace) are also left as future work. The current version of the data is recorded in a semi-controlled environment where the room was near silent, although some recordings still contain robot fan and air conditioner noise in the background.

Our findings also verify that visual and auditory modalities are complementary to each other and performance is boosted via fusion. In future work, we plan to focus on anticipation of failures even before they occur. Thus, any potential damage can be minimized and safety can be assured.

[1][Online]. Available: https://air.cs.itu.edu.tr/projects/finonet.html

Acknowledgment

The authors would like to thank A. Cihan Ak for his efforts during data collection.

References

[1] A. Inceoglu, E. E. Aksoy, A. C. Ak, and S. Sariel, "Fino-Net: A deep multimodal sensor fusion framework for manipulation failure detection," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 6841–6847.

[2] M. Ersen, E. Oztop, and S. Sariel, "Cognition-enabled robot manipulation in human environments: Requirements, recent work, and open problems," *IEEE Robot. Automat. Mag.*, vol. 24, no. 3, pp. 108–122, Sep. 2017.

[3] G. Veruggio, F. Operto, and G. Bekey, "Roboethics: Social and ethical implications," in *Springer Handbook of Robotics*, Berlin, Germany: Springer, 2016, pp. 2135–2160.

[4] *Robots and Robotic Devices–Safety Requirements for Industrial Robots – Part 1: Robots*, ISO Standard 10218:2011, ISO, Geneva, Switzerland, 2011.

[5] *Robots and Robotic Devices–Safety Requirements for Industrial Robots – Part 2: Robot Systems and Integration*, ISO Standard 10218-2:2011, ISO, Geneva, Switzerland, 2011.

[6] A. Morin, "Self-awareness part 1: Definition, measures, effects, functions, and antecedents," *Social Pers. Psychol. Compass*, vol. 5, no. 10, pp. 807–823, 2011.

[7] V. Nitsch, "Situation awareness in autonomous service robots situation awareness in autonomous service robots," 2013. [Online]. Available: https://www.researchgate.net/publication/257958102_Situation_Awareness_in_Autonomous Service_Robots

[8] C. W. Dos Santos, L. Nelson Filho, D. B. Espíndola, and S. S. Botelho, "Situational awareness oriented interfaces on human–robot interaction for industrial welding processes," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 10168–10173, 2020.

[9] A. C. Ak, A. Inceoglu, and S. Sariel, "When to stop for safe manipulation in unstructured environments?," in *Proc. 18th Int. Conf. Auton. Agents MultiAgent Syst.*, 2019, pp. 1767–1769.

[10] A. Inceoglu, G. Ince, Y. Yaslan, and S. Sariel, "Comparative assessment of sensing modalities on manipulation failure detection," in *Proc. IEEE ICRA Workshop Percep., Inference, Learn. Joint Semantic, Geometric, Phys. Understanding*, 2018. [Online]. Available: https://natanaso.github.io/rcw-icra18/assets/ref/ICRA-MRP18_paper_19.pdf

[11] M. Diehl and K. Ramirez-Amaro, "Why did i fail? A causal-based method to find explanations for robot failures," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 8925–8932, Oct. 2022.

[12] D. Altan and S. Sariel, "What went wrong? Identification of everyday object manipulation anomalies," *Intell. Serv. Robot.*, vol. 14, no. 2, pp. 215–234, 2021.

[13] A. Inceoglu, C. Koc, B. O. Kanat, M. Ersen, and S. Sariel, "Continuous visual world modeling for autonomous robot manipulation," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 49, no. 1, pp. 192–205, Jan. 2019.

[14] F. Xiaoran et al., "Acoustic collision detection and localization for robotic devices," U.S. Patent App. 17/084,257, Sep. 23, 2021.

[15] I. Saltali, S. Sariel, and G. Ince, "Scene analysis through auditory event monitoring," in *Proc. Int. Workshop Social Learn. Multimodal Interact. Designing Artif. Agents*, 2016, Art. no. 5.

[16] F. Min, G. Wang, and N. Liu, "Collision detection and identification on robot manipulators based on vibration analysis," *Sensors*, vol. 19, no. 5, 2019, Art. no. 1080.

[17] C. M. C. Valle, A. Kurdas, E. P. Fortunić, S. Abdolshah, and S. Haddadin, "Real-time imu-based learning: A classification of contact materials," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 1965–1971.

[18] S. Haddadin, A. De Luca, and A. Albu-Schäffer, "Robot collisions: A survey on detection, isolation, and identification," *IEEE Trans. Robot.*, vol. 33, no. 6, pp. 1292–1312, Dec. 2017.

[19] B. Proper, A. Kurdas, S. Abdolshah, S. Haddadin, and A. Saccon, "Aim-aware collision monitoring: Discriminating between expected and unexpected post-impact behaviors," *IEEE Robot. Automat. Lett.*, vol. 8, no. 8, pp. 4609–4616, Aug. 2023.

[20] A. Mandlekar et al., "Roboturk: A crowdsourcing platform for robotic skill learning through imitation," in *Proc. Conf. Robot Learn.*, 2018, pp. 879–893.

[21] A. Mandlekar et al., "Scaling robot supervision to hundreds of hours with roboturk: Robotic manipulation dataset through human reasoning and dexterity," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 1048–1055.

[22] S. Dasari et al., "Robonet: Large-scale multi-robot learning," 2019, *arXiv:1910.11215*.

[23] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *Int. J. Robot. Res.*, vol. 37, no. 4/5, pp. 421–436, 2018.

[24] C. Finn, I. Goodfellow, and S. Levine, "Unsupervised learning for physical interaction through video prediction," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 64–72.

[25] F. Worgotter, E. E. Aksoy, N. Kruger, J. Piater, A. Ude, and M. Tamosiunaite, "A simple ontology of manipulation actions based on hand-object relations," *IEEE Trans. Auton. Ment. Develop.*, vol. 5, no. 2, pp. 117–134, Jun. 2013.

[26] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," 2019, *arXiv:1901.03407*.

[27] J. Gertler, *Fault Detection and Diagnosis in Engineering Systems*. Boca Raton, FL, USA: CRC Press, 1998.

[28] C. Fritz, "Execution monitoring–A survey," University of Toronto,Toronto, ON, Canada, Tech. Rep., 2005. [Online]. Available: https://bibbase.org/network/publication/fritz-executionmonitoringasurvey-2005

[29] O. Pettersson, "Execution monitoring in robotics: A survey," *Robot. Auton. Syst.*, vol. 53, no. 2, pp. 73–88, 2005.

[30] O. Pettersson, L. Karlsson, and A. Saffiotti, "Model-free execution monitoring in behavior-based robotics," *IEEE Trans. Syst., Man, Cybern., Part B: Cybern.*, vol. 37, no. 4, pp. 890–901, Aug. 2007.

[31] L. Mauro et al., "Visual search and recognition for robot task execution and monitoring," 2019, *arXiv:1902.02870*.

[32] V. Sathish, M. Orkisz, M. Norrlof, and S. Butail, "Data-driven gearbox failure detection in industrial robots," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 193–201, Jan. 2020.

[33] X. Zhou, H. Wu, J. Rojas, Z. Xu, and S. Li, "Nonparametric Bayesian Method for Robot Anomaly Monitoring," in *Nonparametric Bayesian Learning for Collaborative Robot Multimodal Introspection*. Singapore: Springer, 2020, pp. 51–93.

[34] H. Wu, Y. Guan, and J. Rojas, "A latent state-based multimodal execution monitor with anomaly detection and classification for robot introspection," *Appl. Sci.*, vol. 9, no. 6, 2019, Art. no. 1072.

[35] D. Park, Y. Hoshi, and C. C. Kemp, "A multimodal anomaly detector for robot-assisted feeding using an LSTM-based variational autoencoder," *Robot. Autom. Lett.*, vol. 3, no. 3, pp. 1544–1551, 2018.

[36] D. Park, H. Kim, and C. C. Kemp, "Multimodal anomaly detection for assistive robots," *Auton. Robots*, vol. 43, no. 3, pp. 611–629, 2019.

[37] S. Thoduka, J. Gall, and P. G. Plöger, "Using visual anomaly detection for task execution monitoring," in *Proc. IEEE Int. Conf. Intell. Robots Syst.*, 2021, pp. 4604–4610.

[38] P. Gohil, S. Thoduka, and P. G. Plöger, "Sensor fusion and multimodal learning for robotic grasp verification using neural networks," in *Proc. IEEE 26th Int. Conf. Pattern Recognit.*, 2022, pp. 5111–5117.

[39] D. Altan and S. Sariel, "Clue-AI: A convolutional three-stream anomaly identification framework for robot manipulation," *IEEE Access*, vol. 11, pp. 48347–48357, 2023.

[40] A. Inceoglu, G. Ince, Y. Yaslan, and S. Sariel, "Failure detection using proprioceptive, auditory and visual modalities," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 2491–2496.

[41] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Hum. Factors*, vol. 37, no. 1, pp. 32–64, 1995.