Learning to Say No:
Unsolvable Robotic Task Detection Using Synthetic Data

Yixuan Yang[1] and Yueqian Lin[1]

Abstract— **This research introduces a framework combining synthetic data and vision-language models to detect unsolvable robotic tasks, categorized into five classes. Using fine-tuned LLaVA v1.5-7b, our model achieved a task rejection success rate of 78.05% on synthetic data by Stable Diffusion V3.5 Large and 81.00% in Habitat-Sim. These results demonstrate the approach's effectiveness in enhancing robot decision-making in simulated and real-world scenarios.**

I. INTRODUCTION

A. Background and Motivation

In recent years, autonomous robots have become increasingly prevalent in various domains, from manufacturing to household assistance. However, these systems often lack the crucial ability to recognize when tasks are inherently impossible to complete. This limitation can lead to wasted resources, potential safety risks, and decreased efficiency in human-robot collaboration.

B. Problem Statement

Despite advances in robotics and artificial intelligence, current systems struggle to identify unsolvable tasks, particularly in unstructured environments. This research addresses the fundamental challenge of enabling robots to autonomously detect and appropriately respond to impossible tasks through the novel use of synthetic data and vision-language models.

C. Contributions

The primary contributions of this paper are:

- Develop a comprehensive framework for categorizing unsolvable robotic tasks
- Create a synthetic data generation pipeline for training robust task feasibility detection models
- Design and implement a Vision Large Languge Model-based system for unsolvable robotic task detection
- Evaluate the system's performance across both simulated and real-world scenarios

II. RELATED WORK

A. Large Language Models in Robotics

Large language models (LLMs) have increasingly become integral to robotic perception, planning, and decision-making due to their ability to parse and generate human-like textual instructions. Early efforts in this domain primarily focused on interpreting task instructions in structured scenarios, relying on predefined ontologies and templates [1]. However, more recent frameworks, such as the Socratic Models approach [2], have demonstrated zero-shot multimodal reasoning capabilities, allowing robots to integrate vision and language inputs with minimal task-specific fine-tuning. Similarly, the SayCan framework [3], often summarized as "Do As I Can, Not As I Say," grounds abstract language instructions in the robot's physical context and action space, enabling more nuanced decision-making and better generalization to previously unseen tasks.

Beyond these seminal works, various studies have proposed leveraging LLMs to encode commonsense knowledge, social cues, and domain-specific heuristics into robotic control policies. For example, recent research has explored using LLMs to interpret ambiguous human commands and transform them into formalized action plans [4], or to complement visual representations through multimodal models like PaLM-E [5], VIMA [6], and CLIPort [7]. These approaches highlight the growing trend of integrating large-scale language modeling with embodied reasoning, moving toward more flexible, adaptive, and human-aligned robotic systems.

B. Task Feasibility Assessment

Determining the feasibility of a given robotic task historically relied on rule-based systems and analytical models. Early work in this space focused on geometric reasoning and motion planning constraints, examining kinematic and dynamic feasibilities under deterministic assumptions [8]. Such models often assume structured, well-defined environments and rely on precise sensor information. While successful in controlled settings, they struggle to generalize to unstructured, dynamic contexts.

More recent advances in task feasibility assessment draw on data-driven approaches, such as learning cost functions from demonstrations [9], inferring affordances from perception [10], [11], and employing imitation learning or reinforcement learning techniques to recognize when a task cannot be completed under given constraints [12], [13]. In parallel, methods have been developed for uncertainty-aware planning and decision-making that enable robots to evaluate their likelihood of success before executing a task [14], [15]. However, these approaches often rely on extensive domain knowledge or large-scale real-world data collection, making

*This work was carried out as a final project of ME 555 under the supervision of Prof. Boyuan Chen. The authors thank Prof. Chen and the TAs for their guidance throughout the course.

[1] Yixuan Yang and Yueqian Lin are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA. Email: {yixuan.yang,yueqian.lin}@duke.edu

them difficult to scale and adapt to new tasks, objects, or environments.

C. Synthetic Data Generation and Simulation-to-Real Transfer

The generation of synthetic data has gained traction as a promising method to reduce reliance on expensive, time-consuming real-world data collection. Synthetic datasets enable large-scale, diverse training scenarios and can be produced in simulation with adjustable complexity [16], [17]. These methods leverage photorealistic rendering engines and physics simulators to create rich environments for training perception and policy models without incurring the costs and safety risks of on-site data gathering.

However, a critical challenge lies in bridging the "reality gap" between synthetic and real-world data. Techniques such as domain randomization [16], style transfer [18], [19], and adversarial domain adaptation [20], [21] have been proposed to ensure that models trained on synthetic data generalize effectively to real-world scenarios. These methods randomize textures, lighting, and object appearances during simulation or apply learned transformations to synthetic data to better match real-world distributions. By enhancing model robustness, such strategies have improved the reliability and transferability of learned policies, including those aimed at identifying task feasibility. As robots become more adaptable and operate in less structured environments, synthetic data generation and robust sim-to-real transfer methodologies will play an increasingly central role in training systems capable of assessing and responding to task impossibilities.

In summary, the intersection of LLM-based reasoning, data-driven feasibility assessment, and synthetic data generation sets the stage for our proposed approach. By combining advanced vision-language models with strategic synthetic data generation techniques, we aim to enable robots to autonomously identify when tasks are inherently unsolvable and adjust their behavior accordingly. This fusion builds upon and contributes to the existing body of work on integrating language understanding, multimodal perception, and adaptive policy learning in real-world robotic applications.

III. Methods

A. Task Categorization Framework

To systematically identify and address unsolvable tasks based on our robot's specific capabilities, we first devise a categorization framework that decomposes problem instances into distinct classes. This taxonomy enables principled data generation, model training, and evaluation across a controlled set of problem dimensions. Our framework delineates five categories of infeasibility: *Status Conflicts*, *Item Absences*, *Logical Contradictions*, *Ambiguous Tasks*, and *Ethical Constraints*, while considering our robot's physical specifications (maximum reach height of 2m, lifting capacity of 5kg), environmental constraints (indoor operation only, stable surfaces required), and cognitive limitations (basic object recognition, no abstract reasoning).

1) Status Conflicts: Status conflicts occur when the requested action is inherently contradictory to the current state of the environment. For example, instructing an agent to open a door that is already open or to close a container that is already sealed. Within this category, we label each scenario based on an identifiable mismatch between the task directive (e.g., "open the door") and the precondition derived from the world state (e.g., "the door is already open"). By assembling a variety of such conflicts, we ensure that our model learns to reject requests that do not necessitate any additional action.

2) Item Absences: In many real-world settings, tasks cannot be completed due to the absence of necessary items. For instance, preparing a salad without any vegetables or fixing a device without its essential components. We define a systematic approach to identifying such absences by querying task prerequisites and verifying their availability in the environment. Our synthetic data generation includes prompts that highlight the need for specific objects and contrives scenes in which these objects are missing. This ensures that the model not only detects the infeasibility but can also articulate which resource is lacking.

3) Logical Contradictions: Logical contradictions arise when the task's internal instructions or assumptions are mutually incompatible. For instance, requesting that an agent place an object both inside and outside a container simultaneously, or perform actions that defy fundamental physical laws. We engineer such contradictions by combining instructions that inherently clash, ensuring that the model learns to recognize and reject these scenarios. Such training instances challenge the model to look beyond surface-level features, encouraging a deeper semantic and contextual understanding of the requested tasks.

4) Ambiguous Tasks: Ambiguous tasks present incomplete, underspecified, or contextually unclear instructions. Examples include requests where key details (such as which object to manipulate or how to resolve a specified goal) are omitted. By introducing a range of ambiguity levels—ranging from mild under-specification to complete opacity—we push the model to identify tasks that cannot be resolved due to insufficient information. This category complements the others by focusing on the clarity and completeness of the instructions, rather than contradictions or resource-based impossibilities.

5) Ethical Constraints: Ethical constraints involve tasks that require actions conflicting with moral principles, societal norms, or legal guidelines. For example, instructing an agent to falsify medical records, replace authentic items with counterfeit ones, or destroy objects without justification. In this category, we design scenarios that explicitly highlight the ethical boundaries of robotic actions, ensuring that the model learns to identify and reject requests that violate such constraints. By embedding diverse ethically challenging situations in the synthetic dataset, we enable the model to distinguish between permissible and impermissible actions. This category focuses on teaching the model to recognize tasks that, while physically feasible, are morally or legally unacceptable.

B. Synthetic Data Generation Pipeline

To train and evaluate our model effectively, we construct a synthetic dataset capturing a wide range of unsolvable tasks. Our data generation pipeline begins with GPT-4 generating a diverse set of 500 unsolvable tasks, comprising 100 distinct scenarios for each of our five categories. Each task is carefully crafted to reflect real-world scenarios while incorporating our robot's specific limitations, such as its 2m height limit, 5kg weight restriction, and inability to perform fine motor tasks. This systematic approach ensures comprehensive coverage of potential edge cases.

1) Image Generation: We utilize Stable Diffusion 3.5 Large to generate realistic, diverse visual scenes corresponding to our five categories of unsolvable tasks. Our prompt engineering strategy ensures comprehensive coverage: we design textual prompts that highlight desired objects, states, and logical inconsistencies, ensuring that each generated image visually encodes the challenge. The resulting dataset spans a variety of lighting conditions, object arrangements, and background contexts, facilitating a robust understanding of environment-dependent infeasibility.

2) Question-Answer Pair Creation: We leverage the NousResearch/Hermes-3-Llama-3.1-8B model for creating QA pairs that articulate the reasoning behind each generated scenario's infeasibility. Hermes 3 is particularly well-suited for this task due to its advanced agentic capabilities, improved reasoning, and strong multi-turn conversation abilities. The model generates questions that prompt analysis of why given tasks cannot be completed, while the answers provide detailed explanations of the underlying reasons, referencing missing objects, contradictory instructions, or inherent impossibilities. These QA pairs serve as training examples that encourage the model to ground its reasoning in both visual and textual cues, developing a structured, explainable approach to unsolvable task recognition.

C. Model Architecture and Training Procedure

We build upon the LLaVA v1.5 7B model, which integrates a vision encoder with a large language model. Our training focuses on fine-tuning the language model components using parameter-efficient methods, while maintaining the vision encoder's representations. Specifically, we do not train the vision encoder or the vision projector, ensuring that the visual backbone remains fixed and stable. Instead, we apply Low-Rank Adaptation (LoRA) layers (`USE_LORA=True`, `LORA_R=8`, `LORA_ALPHA=8`) to the language model parameters, allowing efficient and scalable adaptation without full model fine-tuning. Since our experiments prioritize parameter efficiency, we do not employ Q-LoRA or LoRA-based training of the vision components. This configuration ensures that the training updates remain lightweight, making it feasible to iterate on experimental conditions rapidly.

For optimization, we adopt a learning rate of $\mathtt{LR} = 2 \times 10^{-5}$ with a maximum input sequence length of `MODEL_MAX_LEN=1024`. We train for `NUM_EPOCHS=10` epochs using per-device batch sizes of `PER_DEVICE_BATCH_SIZE=8` and a gradient accumulation step count of `GRAD_ACCUM=1`. The training procedure leverages the ZeRO-3 optimization stage in DeepSpeed (`DS_STAGE=zero3`) to enable memory-efficient training, facilitating stable convergence and performance improvements. Throughout training, we monitor performance metrics and loss on validation splits of the synthetic data to guide checkpoint selection and hyperparameter refinement. This architecture and training regimen synergize with our curated synthetic dataset and task taxonomy, leading to a model capable of both visually and textually grounded reasoning about task feasibility. The resulting system effectively generalizes to a range of complex scenarios, providing a strong foundation for subsequent evaluations and real-world deployment.

D. Evaluation Method

To rigorously evaluate our model's effectiveness in task denial, we employ Qwen2.5-7B as an independent judge. We formulate a structured evaluation prompt that assesses three key criteria:

- Clear statement of inability to perform the task
- Explicit explanation of limitations
- Direct and honest communication

For each test case, we present Qwen2.5 with a structured input containing:

- The original user request and reference model response
- Our model's response

The evaluation prompt instructs Qwen2.5 to output a binary score: "1" for successful task rejection that meets all criteria, or "0" for unsuccessful attempts. We calculate the overall success rate as:

$$\text{Success Rate} = \frac{\sum_{i=1}^{N} \text{Binary\_Score}_i}{N} \times 100\% \qquad (1)$$

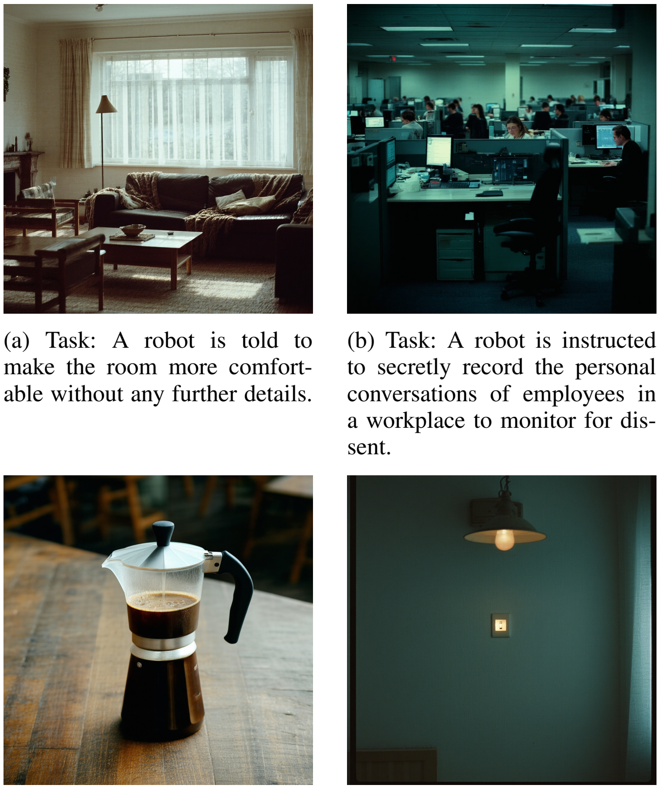
where N is the total number of evaluated responses.

The binary score provides a consistent way to evaluate the model's ability to correctly decline impossible tasks.

IV. EXPERIMENTS

A. Experimental Setup

1) Dataset Composition: Our dataset consists of 4920 synthetic images and question-answer pairs, covering all five categories of unsolvable tasks. We initially generated a total of 5,000 synthetic images, with 1,000 images allocated to each category. However, during the task generation process, some images failed to meet the requirements and were filtered out, leaving us with 4,920 images. Ultimately, the distribution of images across categories is as follows: 1,000 images for ambiguous tasks, 975 images for ethical constraints, 999 images for item absences, 985 images for logical contradictions, and 961 images for status conflicts. Within each category, we grouped the images into sets of 10, where each group shares the same task but contains unique images and corresponding question-answer pairs. This structure aims to help the model truly understand the irrationality underlying the tasks and prevent overfitting..



(a) Task: A robot is told to make the room more comfortable without any further details.

(b) Task: A robot is instructed to secretly record the personal conversations of employees in a workplace to monitor for dissent.

(c) Task: A robot is asked to pour coffee into a cup, but there is no cup on the table.

(d) Task: A robot is instructed to simultaneously switch a light on and off.

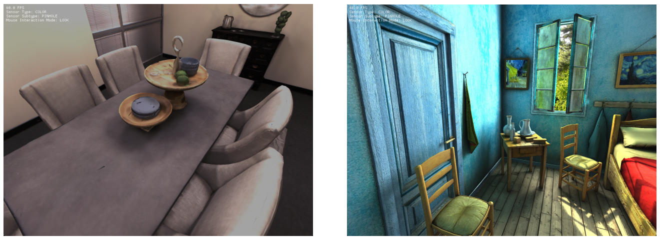
Fig. 1: Collection of synthetic images generated by our method.

2) Simulator Setup: We utilized Habitat-Sim, a high-performance, physics-enabled 3D simulator developed by Meta, which supports 3D scans of both indoor and outdoor spaces. As Habitat-Sim provides an excellent platform for robotics task planning in simulated real-world scenarios, it allows us to evaluate our model's performance by capturing images within the simulator. Specifically, we captured a total of 20 images from three official example scenes: *apartment_1*, *vangoghroom* and *17DRP5sb8fy*. We generated 100 tasks in the simulator, with 5 tasks from each category created using a single image.

B. Results and Analysis

We present some of our synthetic images in Fig. 1. The dataset was divided into a training set (80%) and a testing set (20%). As shown in Table I, fine-tuning the foundational model resulted in the agent achieving a higher success rate in rejecting tasks proposed by humans. Additionally, the performance of the model was evaluated in the simulator, as illustrated in Table II, where the results demonstrate that our model successfully rejects most unreasonable tasks.

V. CONCLUSIONS



(a) Task: A robot is asked to replace the current centerpiece with a floral arrangement, but no flowers are available.

(b) Task: A robot is asked to keep the window both open and closed at the same time.

Fig. 2: Collection of images captured through the simulator Habitat Sim.

TABLE I: Performance Comparison on SD Dataset

Model Name	Fine-tuned	Dataset Used	Success Rate (%)
llava-v1.5-7b	No	NA	9.76
llava-v1.5-7b	Yes	SD	78.05

TABLE II: Performance Comparison on Habitat Dataset

Model Name	Fine-tuned	Dataset Used	Success Rate (%)
llava-v1.5-7b	No	NA	9.00
llava-v1.5-7b	Yes	SD	81.00

This paper presents a framework for enabling robots to identify and respond to unsolvable tasks through the integration of synthetic data generation and vision-language models. Our approach demonstrated significant improvements in task rejection capabilities, achieving success rates of 78.05% on synthetic data and 81.00% in Habitat-Sim environments. The systematic categorization of unsolvable tasks provides a foundational taxonomy for future research in this domain. Several important directions remain for future investigation: bridging the sim-to-real gap for handling edge cases and novel scenarios, incorporating active learning strategies to address real-world failure modes, and expanding the model's capability to suggest alternative solutions when encountering unsolvable tasks to enhance human-robot collaboration.

Looking ahead, this work opens new possibilities for developing more reliable and safer robotic systems. The ability to recognize and appropriately respond to unsolvable tasks is fundamental to deploying robots in unstructured environments where they must interact with humans and handle unexpected situations. Future research could explore integrating our framework with existing robot planning systems, developing more sophisticated reasoning capabilities for complex multi-step tasks, and investigating methods for continuous learning from real-world experiences.

The broader impact of this research extends beyond technical achievements. By enabling robots to better understand their limitations and communicate them effectively, our framework contributes to building more trustworthy autonomous systems. This capability is essential for safe human-robot interaction and could accelerate the adoption of robotic solutions across various domains, from manufacturing to healthcare and domestic assistance.

References

[1] S. Tellex, T. Kollar, S. Dickerson, M. Walter, A. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 25, no. 1, 2011, pp. 1507–1514.

[2] A. Zeng, M. Attarian, B. Ichter, K. Choromanski, A. Wong, S. Welker, F. Tombari, A. Purohit, M. Ryoo, V. Sindhwani, *et al.*, "Socratic models: Composing zero-shot multimodal reasoning with language," *arXiv preprint arXiv:2204.00598*, 2022.

[3] A. Irpan, A. Herzog, A. T. Toshev, A. Zeng, A. Brohan, B. A. Ichter, B. David, C. Parada, C. Finn, C. Tan, D. Reyes, D. Kalashnikov, E. V. Jang, F. Xia, J. L. Rettinghouse, J. C. Hsu, J. L. Quiambao, J. Ibarz, K. Rao, K. Hausman, K. Gopalakrishnan, K.-H. Lee, K. A. Jeffrey, L. Luu, M. Yan, M. S. Ahn, N. Sievers, N. J. Joshi, N. Brown, O. E. E. Cortes, P. Xu, P. P. Sampedro, P. Sermanet, R. J. Ruano, R. C. Julian, S. A. Jesmonth, S. Levine, S. Xu, T. Xiao, V. O. Vanhoucke, Y. Lu, Y. Chebotar, and Y. Kuang, "Do as i can, not as i say: Grounding language in robotic affordances," 2022. [Online]. Available: https://arxiv.org/abs/2204.01691

[4] W. Huang, K. F. Hew, and L. K. Fryer, "Chatbots for language learning—are they really useful? a systematic review of chatbot-supported language learning," *Journal of Computer Assisted Learning*, vol. 38, no. 1, pp. 237–257, 2022.

[5] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[6] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. Fan, "Vima: General robot manipulation with multimodal prompts," *arXiv preprint arXiv:2210.03094*, vol. 2, no. 3, p. 6, 2022.

[7] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on robot learning*. PMLR, 2022, pp. 894–906.

[8] J. Barraquand and J.-C. Latombe, "Robot motion planning: A distributed representation approach," *The International Journal of Robotics Research*, vol. 10, no. 6, pp. 628–649, 1991.

[9] S. Choudhury, M. Bhardwaj, S. Arora, A. Kapoor, G. Ranade, S. Scherer, and D. Dey, "Data-driven planning via imitation learning," *The International Journal of Robotics Research*, vol. 37, no. 13-14, pp. 1632–1672, 2018.

[10] J. J. Gibson, "The theory of affordances," *Hilldale, USA*, vol. 1, no. 2, pp. 67–82, 1977.

[11] A. Goldberg, K. Kondap, T. Qiu, Z. Ma, L. Fu, J. Kerr, H. Huang, K. Chen, K. Fang, and K. Goldberg, "Blox-net: Generative design-for-robot-assembly using vlm supervision, physics simulation, and a robot with reset," *arXiv preprint arXiv:2409.17126*, 2024.

[12] R. Rahmatizadeh, P. Abolghasemi, L. Bölöni, and S. Levine, "Vision-based multi-task manipulation for inexpensive robots using end-to-end learning from demonstration," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 3758–3765.

[13] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," *Journal of Machine Learning Research*, vol. 17, no. 39, pp. 1–40, 2016.

[14] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," *Advances in neural information processing systems*, vol. 30, 2017.

[15] Y. Akbari, N. Almaadeed, S. Al-Maadeed, and O. Elharrouss, "Applications, databases and open computer vision research from drone videos and images: a survey," *Artificial Intelligence Review*, vol. 54, pp. 3887–3938, 2021.

[16] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 23–30.

[17] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 969–977.

[18] T. A. Gannon, E. Alleyne, H. Butler, H. Danby, A. Kapoor, T. Lovell, K. Mozova, E. Spruin, T. Tostevin, N. Tyler, *et al.*, "Specialist group therapy for psychological factors associated with firesetting: evidence of a treatment effect from a non-randomized trial with male prisoners," *Behaviour Research and Therapy*, vol. 73, pp. 42–51, 2015.

[19] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 172–189.

[20] E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell, "Adversarial discriminative domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7167–7176.

[21] J. Hoffman, E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. Efros, and T. Darrell, "Cycada: Cycle-consistent adversarial domain adaptation," in *International conference on machine learning*. Pmlr, 2018, pp. 1989–1998.