11-777 Spring 2021 Class Project
ALFRED Dataset Analysis

Christian Deverall[*] **Jingyuan Li** [*] **Artidoro Pagnoni** [*]
{cdeveral, jingyua4, apagnoni}@andrew.cmu.edu

1 Introduction and Problem Definition

Understanding and following task-oriented instructions are natural and straightforward skills for human beings. This remains to be a great challenge for robotic systems due to the difficulties in both visual and linguistic understanding. In this project, we are interested in allowing the robots to understand the instructions and take actions based on the instructions in a manner similar to the humans. As a result, the ALFRED dataset (Shridhar et al., 2020) was chosen for this project.

The high-level task for the ALFRED dataset is to predict the correct sequence of household actions given textual instructions and visual feedback. The ALFRED dataset is unique in that the actions are long, compositional, and non-reversible. An example of this is that the agent must pick up a knife in order to slice a potato. Moreover, once a potato is sliced, it cannot be put back together. Actions consist of navigation within a room or interaction with an object. For interaction, a pixelwise mask must be predicted whereby the object with the highest intersection-over-union score with the interaction mask is acted upon. The textual instructions come in the form of both high-level task descriptions and step by step commands. The visual feedback is ego-centric and the next image is provided after each action is performed. Compared to previous datasets, the ALFRED dataset contains a diverse range of tasks. Specifically there are 7 high-level task types parameterized by a combination of 84 object classes and 120 household scenes.

In this project, the focus will be on the navigation and interaction of the agent in the simulator. The input space would be a set of instructions (represented in natural language) and an image (represented in ResNet-18 (He et al., 2016) features) reflecting the current environment perceived by the agent. The output space is the prediction of the next action that helps the agent achieve the ultimate goal. The predicted actions can be described in three dimensions, including 1) a classification outcome indicating the type of action, 2) the magnitude of action when the action is for navigation, and 3) an object mask indicating the bounding box of an object acted upon when the object is involved in the action. The input image is updated for the next prediction after an action is performed. To evaluate the success of the prediction sequence, the number of desired state changes in the target object completed by the agent is measured.

split	mean	std	min	max
low-level train	49.78	24.78	5	234
low-level val seen	46.98	20.22	13	137
low-level val unseen	50.12	25.61	12	165
high-level train	6.51	2.49	2	19
high-level val seen	6.47	2.62	3	16
high-level val unseen	6.16	2.33	3	14

Table 1: Statistics on the number of low-level and high-level actions for the splits of the data.

2 Data Analysis

We present an analysis of the data. Our objective is to get a sense of the types of demonstrations that are present in the dataset. The sizes of the seen validation set, unseen validation set and training set are 251, 255 and 6574 respectively.

2.1 Action Space Analysis

In the ALFRED dataset, tasks are described by low-level and high-level actions. In this section, we examine the patterns behind these actions and the objects which are acted upon. In order to gauge the size of tasks and the granularity of actions, we provide statistics about the number of low-level

[*] Everyone Contributed Equally – Alphabetical order

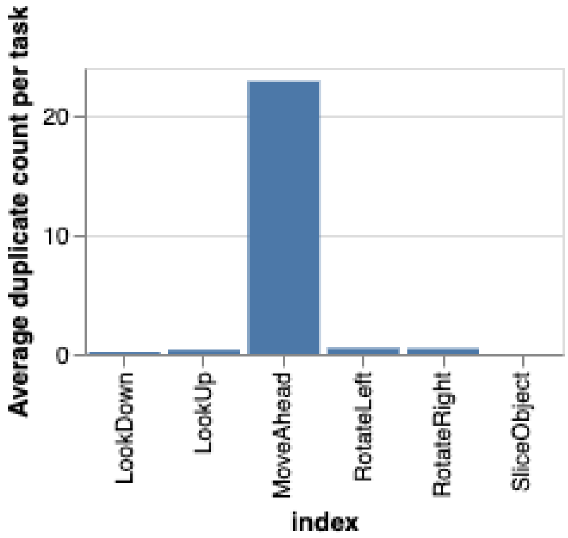


Figure 1: Average number of duplicate actions per task

and high-level actions per task in Table 1.

Low-level actions are often repeated in order to fully complete a high-level goal. For example, numerous fixed-distance movement actions are required to travel from one end of the kitchen to the other. This repetition pattern is relevant to action prediction because for highly repeated actions, it is likely that the next action is the exact same as the current one. In Figure 1, it can be seen that by far the most duplicated action is the ”MoveAhead” operation encompassing approximately 44% of all actions.

To further understand the sequential order of actions, we display the distribution of next actions given current actions in Figure 2. Only data from the training split is shown here. Current actions are shown on the left axis while next actions are shown on the bottom axis. The top three most common sequences of actions are “MoveAhead” then “MoveAhead”, “RotateLeft” then “MoveAhead” and “RotateRight” then “MoveAhead”.

Objects are utilized in two manners. They are acted upon or they are decorative in that they exist in the room but are not used. To get a sense of the distribution of the acted upon objects, we plot the top 20 most used objects in Figure 3. The microwave and fridge are used far more frequently than other objects. This is most likely because they are highly common household items that essential for “OpenObject” and “CloseObject” actions. The distribution of the top 20 most common objects regardless of whether they are used or not is shown in Figure 4.

split	mean	std	min	max
train	6.72	2.49	3	19
val seen	6.79	2.73	4	16
val unseen	6.26	2.33	4	14

Table 2: Statistics on the number of sentence instructions (sub-goals) for the splits of the data.

split	mean	std	min	max
train	3.20	0.68	3	6
val seen	3.27	0.80	3	6
val unseen	3.22	0.72	3	6

Table 3: Statistics on the number sets of instructions for each trajectory by split.

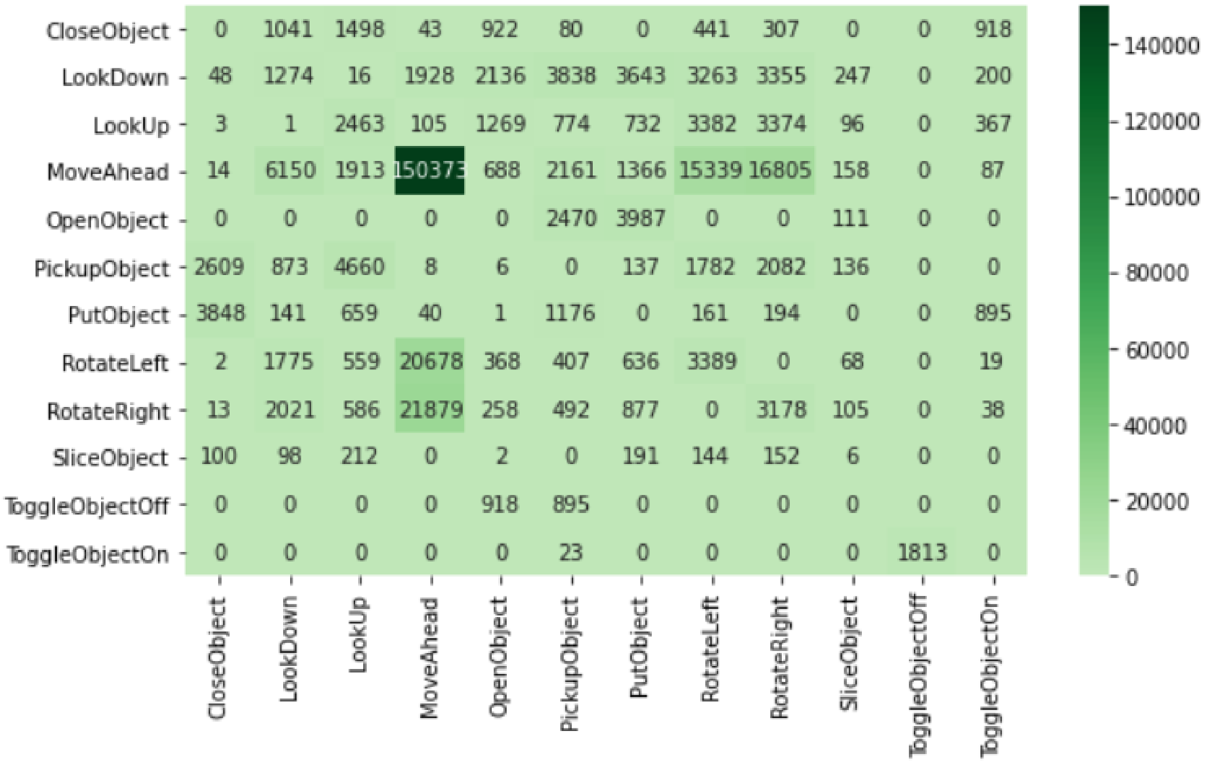
Certain objects are more frequently used for certain actions. Intuitively, a potato is more likely to be sliced than a microwave. Thus, to understand which objects are more frequently used for each action, we plot an action-object heatmap in Figure 5.

To summarize our action-space analysis, we have provided statistical insights into the number of high-level and low-level actions for each task in the three pre-test splits of our data. We showed that the probability distribution of the next action greatly depends on the current action being performed. Finally, we displayed the probability distribution of used items and showed that each one has varying levels of association with each action.

2.2 Instructions Analysis

There are two types of instructions: the task description and the detailed instructions. The task description is generally a single sentence outlining the goal of the task while the detailed instructions break the task down into sub-goals and provide a sentence instruction for each sub-goal. We measure statistics on the number of sub-goals for each task in Table 2. We observe that there are up to 19 distinct sub-goals in each trajectory and that on average there are 6.7 sub-goals. The tasks are not simple as they require a minimum of three sub-goals.

There are different sets of each instruction for each trajectory. We examine the statistics of the number of separate sets of instructions for each trajectory in Table 3. We observe that there are generally three sets for each trajectory, but they can go up to six.



	CloseObject	LookDown	LookUp	MoveAhead	OpenObject	PickupObject	PutObject	RotateLeft	RotateRight	SliceObject	ToggleObjectOff	ToggleObjectOn
CloseObject	0	1041	1498	43	922	80	0	441	307	0	0	918
LookDown	48	1274	16	1928	2136	3838	3643	3263	3355	247	0	200
LookUp	3	1	2463	105	1269	774	732	3382	3374	96	0	367
MoveAhead	14	6150	1913	150373	688	2161	1366	15339	16805	158	0	87
OpenObject	0	0	0	0	0	2470	3987	0	0	111	0	0
PickupObject	2609	873	4660	8	6	0	137	1782	2082	136	0	0
PutObject	3848	141	659	40	1	1176	0	161	194	0	0	895
RotateLeft	2	1775	559	20678	368	407	636	3389	0	68	0	19
RotateRight	13	2021	586	21879	258	492	877	0	3178	105	0	38
SliceObject	100	98	212	0	2	0	191	144	152	6	0	0
ToggleObjectOff	0	0	0	0	918	895	0	0	0	0	0	0
ToggleObjectOn	0	0	0	0	0	23	0	0	0	0	1813	0

Figure 2: Heatmap to show the frequency of low-level action bigrams in the training set.

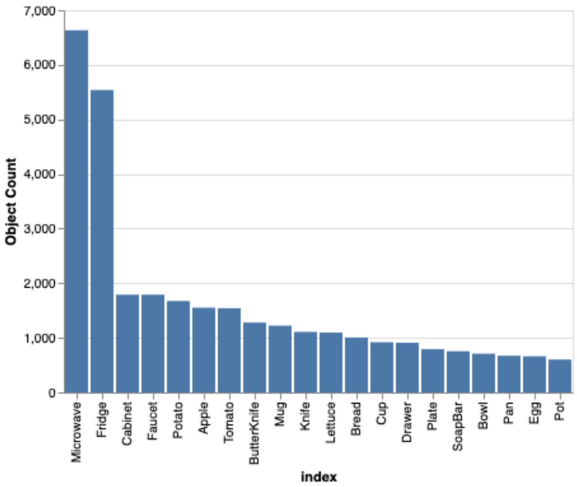


Figure 3: The frequency distribution of objects used within actions.

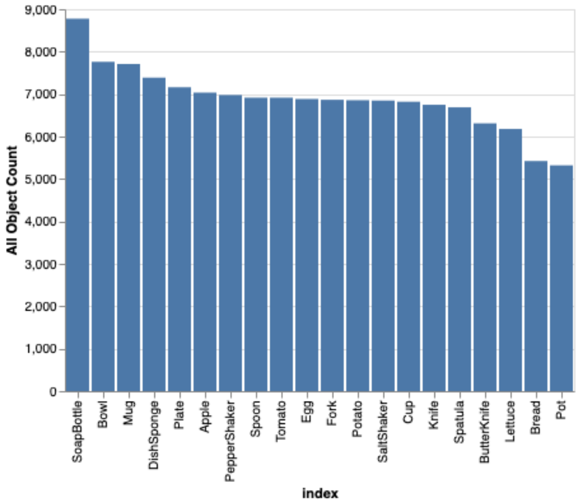


Figure 4: The frequency distribution of all objects.

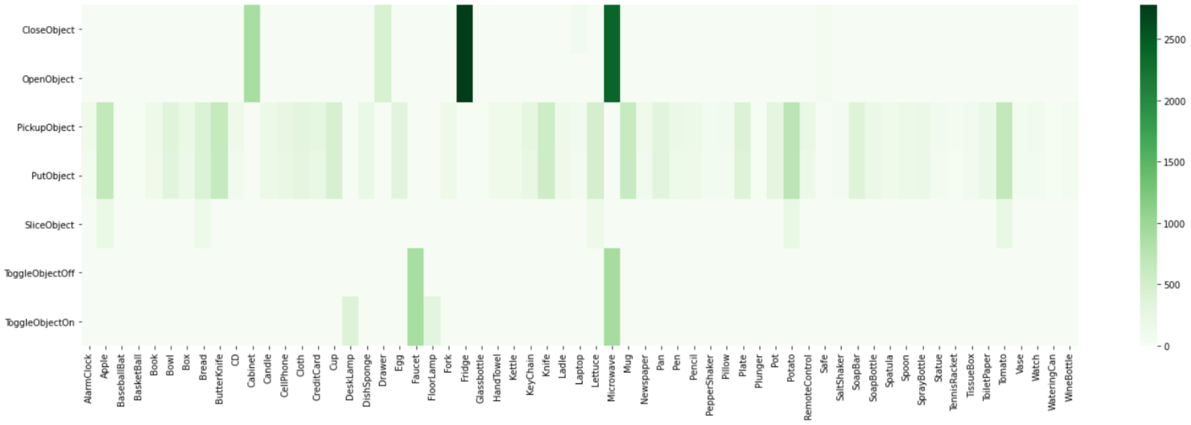


Figure 5: A heatmap to show which object is utilized more frequently for each action.

When inspecting the instructions we notice that verbs indicate actions, while nouns indicate objects. We measure statistics on the number of verbs and nouns in both types of instructions. In both cases, we use the lemma of the noun and verb. We report the results in Table 4. The number of verbs used can reach as high as 68 verbs for one set of detailed instructions. On average the number of verbs is around 13 in the detailed instructions and slightly over one in the task instructions.

Besides the actual values, we observe that the ratio of nouns to verbs is higher in the task instructions. This indicates that in the task instructions there are more complements used to specify the actions described by the verbs. By manual analysis, the task instructions have specifications on the location where an action should be completed. This is information that comes in addition to the object that needs to be acted on and makes the sentence structure more complex.

It is surprising that some task instructions do not contain any nouns. We inspected some of the sentences without nouns and they seem to be truncated sentences for example: “The ro”. This indicates that there is some noise in the data.

In order to get a sense of the type of objects and verbs that appear in the instructions, we plot their distributions. In both cases they are long tailed distributions with most of the elements occurring very rarely. The distribution over verbs is steeper than that over nouns. This partly relates to the fact that there are fewer action types than objects in the ALFRED world. We do also notice that there are a lot of synonyms in both cases. For example, in the verbs, both “turn” and “veer” are present. These two verbs are very close in meaning.

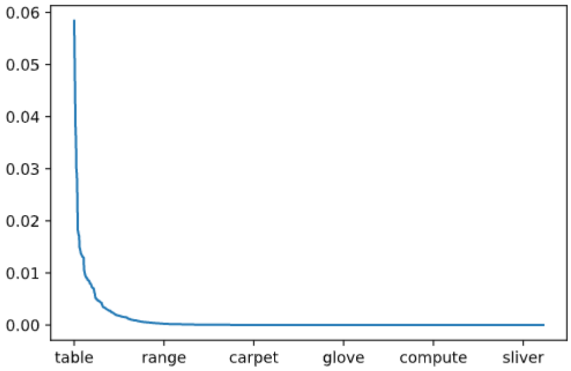


Figure 6: Frequency of occurrences of nouns in the detailed instructions in the training set.

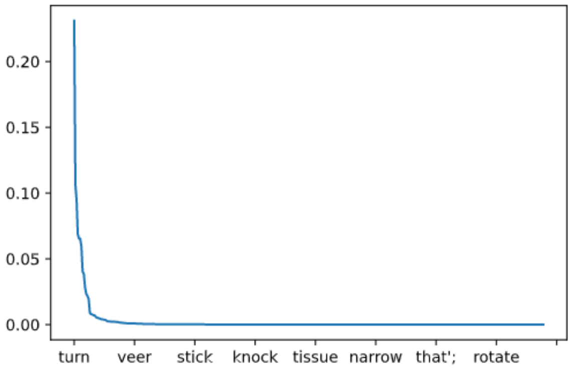


Figure 7: Frequency of occurrences of verbs in the detailed instructions in the training set.

		Nouns				Verbs			
Instruction Type	split	mean	std	min	max	mean	std	min	max
Task Instructions	train	3.08	1.10	0	10	1.26	0.53	0	8
	val seen	3.09	1.16	0	11	1.28	0.58	0	5
	val unseen	3.09	1.16	0	11	1.28	0.58	0	5
Detailed Instructions	train	19.89	9.15	2	98	13.13	6.81	1	68
	val seen	19.89	10.16	5	82	12.98	7.31	2	53
	val unseen	19.89	10.16	5	82	12.98	7.31	2	53

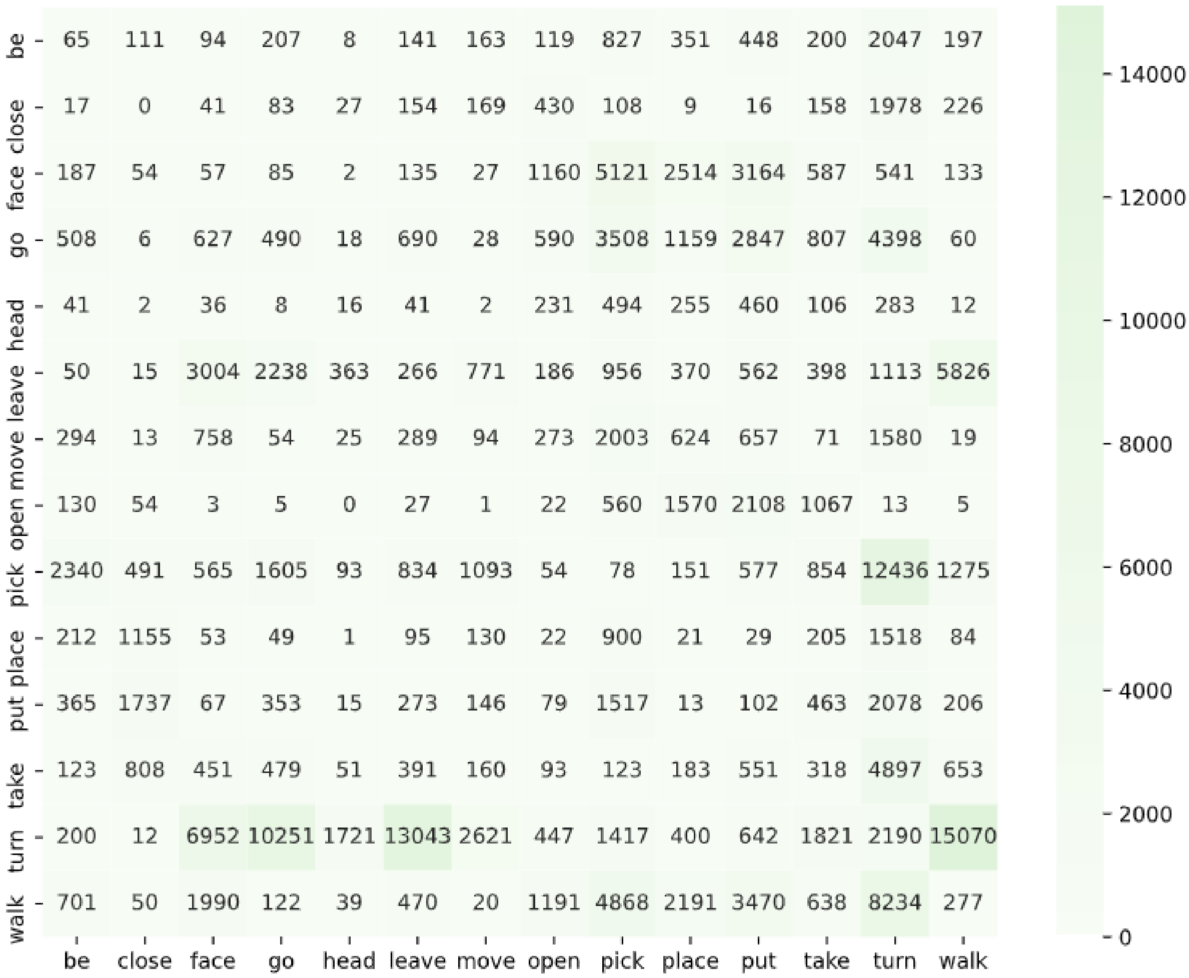
Table 4: Statistics on the number of verbs and nouns in the instructions for each split.

We compare the occurrence of objects in the instructions between the training and validation set to estimate how often there will be unseen objects in the test set. We observe that there are only 22 unobserved objects in the validation set that did not appear in the training set. When looking at verbs, we only observe 6 verbs that appeared in the validation sets but not in the training set. This is relatively small and not surprising since the set of actions remains the same across the two splits. In both cases, the proportion of unseen objects and verbs is lower than 0.1% so it should not be of major concern.

We also look at the actions that generally follow each other. To this end we build a confusion matrix of pairs of verbs in Figure 8. We study how verbs follow each other in the instructions. In general, we observe that the most frequent pairs of verbs involve navigation. The interactive actions are less frequent, and they appear after other navigation actions (like “face” or “walk”).

3 Conclusion

In this report, we analyze the ALFRED (Shridhar et al., 2020) dataset. We explore both the actions and the instructions in the training and validation data splits. We observed a highly imbalanced distribution of actions, and of action pairs, which indicates potential biases in the dataset. Strong symmetry is observed in the interactive actions that change the states of object, meaning potential rule-of-thumb that could be exploited in designing our own method. With respect to the instructions, we observe that almost all objects and verbs are common to the training and validation data splits. Furthermore, we see that there are long tailed distributions of nouns and verbs. Finally, the high-level task description includes more mentions of objects to specify the actions beyond the objects acted on.



	be	close	face	go	head	leave	move	open	pick	place	put	take	turn	walk
be	65	111	94	207	8	141	163	119	827	351	448	200	2047	197
close	17	0	41	83	27	154	169	430	108	9	16	158	1978	226
face	187	54	57	85	2	135	27	1160	5121	2514	3164	587	541	133
go	508	6	627	490	18	690	28	590	3508	1159	2847	807	4398	60
head	41	2	36	8	16	41	2	231	494	255	460	106	283	12
leave	50	15	3004	2238	363	266	771	186	956	370	562	398	1113	5826
move	294	13	758	54	25	289	94	273	2003	624	657	71	1580	19
open	130	54	3	5	0	27	1	22	560	1570	2108	1067	13	5
pick	2340	491	565	1605	93	834	1093	54	78	151	577	854	12436	1275
place	212	1155	53	49	1	95	130	22	900	21	29	205	1518	84
put	365	1737	67	353	15	273	146	79	1517	13	102	463	2078	206
take	123	808	451	479	51	391	160	93	123	183	551	318	4897	653
turn	200	12	6952	10251	1721	13043	2621	447	1417	400	642	1821	2190	15070
walk	701	50	1990	122	39	470	20	1191	4868	2191	3470	638	8234	277

Figure 8: Occurrence of verbs following each other. Only displaying the most frequent verbs.

References

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.