ViSaRL: Visual Reinforcement Learning Guided by Human Saliency

Anthony Liang, Jesse Thomason and Erdem Bıyık

***Abstract*— Training robots to perform complex control tasks from high-dimensional pixel input using reinforcement learning (RL) is sample-inefficient, because image observations are comprised primarily of task-irrelevant information. By contrast, humans are able to visually attend to task-relevant objects and areas. Based on this insight, we introduce Visual Saliency-Guided Reinforcement Learning (ViSaRL). Using ViSaRL to learn visual representations significantly improves the success rate, sample efficiency, and generalization of an RL agent on diverse tasks including DeepMind Control benchmark, robot manipulation in simulation and on a real robot. We present approaches for incorporating saliency into both CNN and Transformer-based encoders. We show that visual representations learned using ViSaRL are robust to various sources of visual perturbations including perceptual noise and scene variations. ViSaRL nearly doubles success rate on the real-robot tasks compared to the baseline which does not use saliency.**

I. INTRODUCTION

Studies in neuroscience [1] show that humans utilize selective attention to focus on task-relevant information for efficiently processing and understanding complex visual scenes [2]. We employ selective attention when performing everyday pick-and-place tasks to identify the target objects, focus on the grasp points, and execute precise hand-eye coordination. We hypothesize that saliency maps capturing human visual attention is a useful signal to process visual observations for AI agents. In this paper, we investigate whether *human* visual attention helps *agents* perform tasks.

A key ingredient in solving visual control tasks is to learn visual representations that capture useful features of the sensory input to simplify the decision-making process. Many works in the deep reinforcement learning (RL) community have proposed to learn such representations through various self-supervised objectives including contrastive learning [3] and data augmentation [4]. By contrast, we focus on self-supervision using *saliency* as additional human domain knowledge to inform the representation of task-relevant features in the visual input while filtering out perceptual noise.

We present **Vi**sual **Sa**liency **R**einforcement **L**earning (ViSaRL), a general approach for incorporating human-annotated saliency maps as an inductive bias for learned visual representations. The key idea of ViSaRL is to train a visual encoder using both RGB and saliency inputs and an RL policy that operates over lower dimensional image representations as shown in Figure 1. By using a multimodal autoencoder trained using a self-supervised objective, our learned representations attend to the most salient parts of an image for downstream task learning making them robust to visual distractors. To circumvent the expensive process of manually annotating saliency maps, we train a state-of-the-art saliency predictor using only a few human-annotated examples to pseudo-label RGB observations with saliency.

All authors are with the Thomas Lord Department of Computer Science, University of Southern California, Los Angeles, CA 90089 USA. (Correspondence to: anthony.liang@usc.edu).

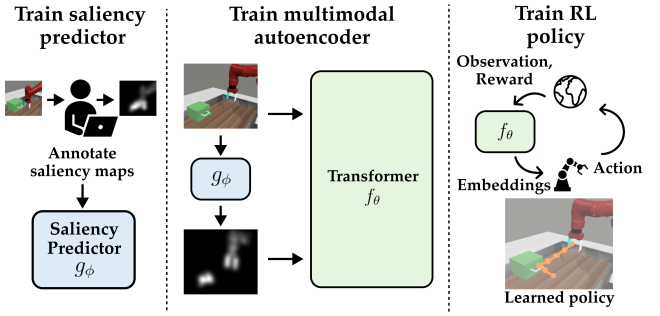

Fig. 1: ViSaRL trains a saliency prediction model from a few human-annotated saliency maps. This model is used to augment an offline image dataset with saliency. A visual encoder is pretrained with the dataset and used during downstream policy learning to generate latent representations of the agent's observations.

We evaluate ViSaRL on a diverse set of challenging continuous control tasks in the DeepMind Control (DMC) suite [5] and robot manipulation tasks in Meta-World [6] and a real robot. Our method improves in sample-efficiency and robustness over state-of-the-art vision-based RL methods across all environments. Remarkably, ViSaRL nearly doubles the task success rate on a real-robot.

Our contributions can be summarized as follows:

1) We propose ViSaRL, a framework for incorporating human-annotated saliency maps to learn robust representations for visual control tasks;
2) We present approaches for utilizing saliency information in both CNN and Transformer encoders; and
3) We conduct extensive experiments that demonstrate ViSaRL consistently outperform prior state-of-the-art methods for various visual control tasks both in simulation and on a real robot.

II. RELATED WORK

Different forms of human data can be leveraged when solving control tasks. Researchers have created various interfaces to collect different data modalities from humans such as reward sketches [7], feature traces [8], scaled comparisons [9], and abstract trajectories [10]. Attention saliency maps, in contrast, do not require humans to work with abstract concepts like rewards and task features, and do not require watching and comparing lengthy trajectories.

Saliency Maps. Saliency maps approximate which parts of an image tend to attract human visual attention, corresponding to where the human eye would likely fixate when viewing

an image [11]. Saliency maps have been used in both computer vision and machine learning for various applications including activity recognition [12], question answering [13], and object segmentation [14]. The explainable AI community uses saliency maps to understand how a model is making its predictions and to identify the most informative regions of an image for a particular task [15], [16], [17]. Most existing works explore using saliency maps only as tools for interpretation [18], [19]. For example, Atrey et al. [18] and Rosynski et al. [19] use saliency maps to rationalize and explain the actions of RL agents in Atari games. Boyd et al. [20] show saliency maps encoding prior human knowledge enable better generalization of deep learning models.

Bertoin et al. [21] uses neural network saliency in a self-supervised regularization objective to encourage better visual representations. We do not use a model's saliency, but rather human saliency to identify salient regions of the input image and distill this knowledge into the visual representation.

User Interfaces for Human Saliency. ViSaRL needs a small number of human-annotated saliency maps to bootstrap the saliency prediction network. Prior work used superpixel segmentation [22] to first divide each image into segments, and then asked humans to click on the segments that are salient [17]. However, that method requires manually checking and combining the segments that belong to the same object before showing the images to annotators, burdening system designers. As an alternative, Boyd et al. [23] used interfaces where the annotators created binary masks by simply clicking on images. We employ a similar but simpler interface: an annotator clicks on the salient parts of the image, and a Gaussian kernel is applied around selected pixels to achieve smooth saliency maps shown in Figure 2.

Representation Learning for RL. Saliency maps are representations of the environment that carry domain knowledge about which regions of the visual input are important for the downstream task. Such representations are crucial in RL because they enable agents to tractably deal with high-dimensional image observation spaces.

Prior works have shown self-supervised learning with data augmentation helps achieve good performance in image-based RL. Contrastive Unsupervised RL (CURL) [3] employs a contrastive learning objective as an auxiliary loss to learn representations for off-policy RL. RL with Augmented Data (RAD) [4] use simple image augmentations such as random cropping and color jittering as regularization to learn representations invariant to visual perturbations. ViSaRL does not use data augmentation directly in the value function or policy update. Instead, saliency augmentation is introduced during the visual encoder pretraining phase.

Nair et al. [24] and Karamcheti et al. [25] propose to combine internet scale language and vision datasets to learn visual representations applicable across all robot tasks. While they focus on learning general visual representations, ViSaRL augments small task-specific datasets with saliency information to improve pretrained visual representations.

Sax et al. [26] demonstrated that mid-level visual representations such as surface normals or depth predictions

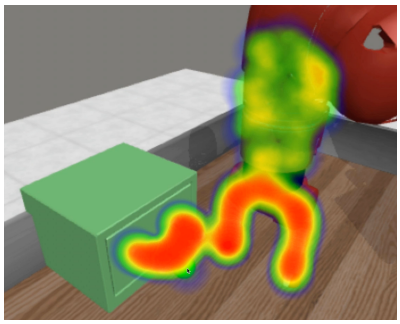

Fig. 2: **Annotation Interface**. Custom click-based saliency annotation interface. Each click generates a Gaussian centered at the clicked coordinate with some variance. Warmer colors denote more salient regions such the drawer handle and the robot's end-effector.

from RGB images removes unimportant information and captures useful invariances about the visual world leading to better success on downstream RL tasks. Similar to Sax et al. [26], ViSaRL utilizes saliency maps as a mid-level feature. However, we empirically show that our approach for incorporating the saliency information into the visual representation improves task performance over other mid-level features including depth and surface normals.

III. VISUAL SALIENCY-GUIDED RL

We propose ViSaRL, a simple approach for incorporating human-annotated saliency to learn representations for visual control tasks. ViSaRL can be implemented on top of any standard RL algorithm for learning a policy. It aims to learn representations that encode useful task-specific inductive biases from human saliency maps. ViSaRL consists of three learned components: a saliency predictor g_ϕ, an image encoder f_θ, and a policy network π_ψ shown in Figure 1. We will elaborate on each component in the following sections.

Saliency Predictor. Saliency maps highlight regions in an image likely to capture human attention or are considered crucial for a given task. Having a human expert annotate saliency maps for every image observation is impractical and not scalable to complex domains. To alleviate the burden of manual annotations, we propose to learn a saliency network using only a few hand-annotated examples of saliency maps collected using a custom user interface.

Formally, given an input RGB image observation, $I \in \mathbb{R}^{H\times W\times C}$, a saliency predictor g_ϕ maps an input image I to a continuous saliency map $M = g_\phi(I) \in [0,1]^{H\times W}$ highlighting important parts of the image for the downstream task. We use a state-of-the-art saliency model, Pixel-wise Contextual Attention network (PiCANet) [28]. PiCANet uses global and local pixel-wise attention modules to selectively attend to informative context. Global attention can attend to backgrounds for foreground objects while local attention can attend to regions that have similar appearance. The mixture of attention at different scales allows for more homogeneous and consistent saliency predictions. We emphasize that our method is agnostic to the choice of saliency model.

Pretraining Visual Representation. We use our trained g_ϕ to pseudo-label an offline image dataset collected using any behavior policy (random, replay buffer, expert demon-

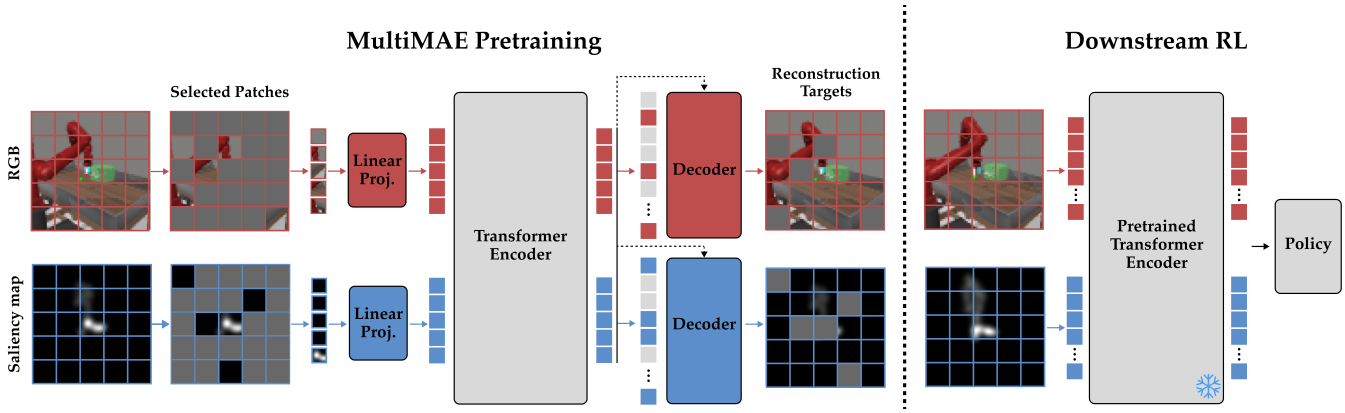

Fig. 3: **ViSaRL**. We pretrain a MultiMAE [27] Transformer on a dataset of paired images and saliency maps. MultiMAE employs a self-supervised objective in which masked patches for both input modalities are reconstructed given only the visible patches. The pretrained model is frozen and used for extracting representations during task learning. There is no input masking during downstream RL.

Algorithm 1 Visual Saliency-Guided RL

Input: env, ϕ, ψ, θ randomly initialized parameters
Collect image dataset $\mathscr{D}$ with any behavioral policy π_B
Annotate N random frames from $\mathscr{D}$ with saliency
Train g_ϕ on $\{(I,M)\}_{i=1}^{N}$ using PiCANet loss
Annotate the full dataset $\mathscr{D} = \{(I, g_\phi(I))\}_{i=1}^{N}$
Train f_θ using masked reconstruction
for every environment step **do** ▷ RL Training
    Select action $a \sim \pi_\psi(f_\theta(o, g_\phi(o)))$
    Optimize $\mathscr{L}_{RL}$ with respect to ψ

strations, etc.) with saliency maps. We then use the paired image and saliency dataset to pretrain an image encoder, f_θ. We experiment with two models for our backbone visual encoder, CNN and Transformer, and investigate different techniques for augmenting each with saliency input. To add saliency to a CNN, we can use saliency as a continuous mask or simply add it as an additional channel per pixel. For a Transformer encoder, we pretrain the model with saliency as an additional input using a masked reconstruction objective.

Masked autoencoders (MAE) [29] are an effective and scalable approach for learning visual representations. MAE masks out random patches of an image and reconstructs the masked patches using a Vision Transformer (ViT) [30]. An image $I \in \mathbb{R}^{H\times W\times C}$ is processed into a sequence of 2D patches $h \in \mathbb{R}^{K\times(P^2C)}$ where P is the patch size and $K = HW/P^2$ is the number of patches. A subset of these patches are randomly masked out with a masking ratio of m. Only the *visible, unmasked patches* are used as input to the ViT encoder. Masking reduces the input sequence length and encourages learning global, contextualized representations.

The image patches are embedded via a linear projection and added to positional embeddings. The resulting tokens are processed via a series of Transformers. Finally, a ViT decoder reconstructs the original input by processing all of the tokens including the encoded visible patches and placeholder mask tokens. Following He et al. [29], we set a high masking ratio $m = 0.75$ and a heavy-encoder, light-decoder architecture.

MultiMAE for Encoding Saliency. The standard MAE architecture is limited to processing just RGB modality. We propose to incorporate saliency using the MultiMAE [27] architecture shown in Figure 3. MultiMAE extends MAE to encode multiple input modalities in a way that these modalities are contributing synergistically to the resulting representation. Specifically, MultiMAE uses a different linear projection and decoder for each input modality. A cross attention layer is used in each decoder to incorporate information from the encoded tokens of other modalities. Crucially, MultiMAE's pretraining objective requires the model to perform well in both the original MAE objective of RGB in-painting and cross-modal reconstruction, resulting in a stronger cross-modal visual representation.

Downstream Policy Learning. After pretraining the MultiMAE model, we freeze the encoder and use it to compute latent representations of environment observations for policy training. ViSaRL is not only compatible with online RL algorithms such as Soft-Actor Critic (SAC) [31] in which the agent learns through environment interactions but also imitation learning from expert demonstrations. Image inputs are not masked during policy learning. We average the patch embeddings to generate a global image representation. The full procedure for ViSaRL is summarized in Algorithm 1.

IV. EXPERIMENT SETUP

To demonstrate the effectiveness of using human-annotated saliency information to enhance visual representations for task learning, we show quantitative results of our approach with two different encoder backbones, CNN and Transformer, across multiple simulated environments including the Meta-World manipulation [6] and DMC benchmarks [21] and real-robot manipulation with a Kinova Jaco 2 arm. We train the downstream policy using SAC [31] for the simulation experiments and behavioral cloning with expert demonstrations for the real robot experiments. [1]

Saliency Map Annotation. We created a simple user interface to collect saliency annotations shown in Figure 2. An annotator clicks on the pixels in the image that they think are relevant for performing the given downstream task. The

[1]The code implementation for reproducing the results and additional analysis can be found on: https://aliang8.github.io/visarl_site.

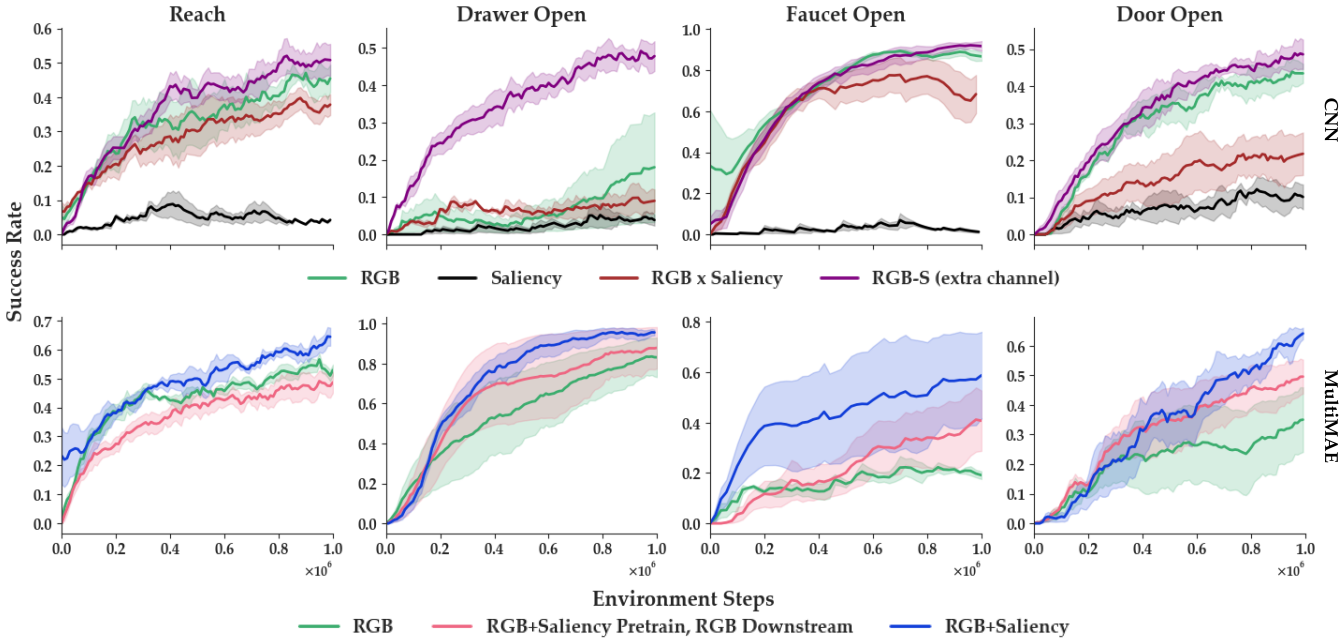

Fig. 4: **Learning curves** for four robot manipulation tasks in Meta-World evaluated by task success rate. **(Top)** CNN encoder methods. **(Bottom)** Transformer encoder methods. We select tasks that require manipulating small objects with different motions such as a pushing, pulling, and reaching. The solid lines represent the mean and shaded region the standard error across three seeds.

interface creates a Gaussian centered around the clicked pixel with $\sigma = 10$ on an input image of resolution $256 \times 256 \times 3$.

V. SIMULATION EXPERIMENTS

Figure 4 and Table I summarize our main findings on 4 Meta-World robot manipulation tasks and 5 DMC tasks using CNN and Transformer backbones. We compare against two state-of-the-art methods for visual representation learning: CURL [3], a contrastive representation learning method and RAD [4], a method to combine various image augmentations to induce visual invariances in the learned representations.

Saliency input improves downstream task success rates. Incorporating saliency improves the task success rate in Meta-World using CNN and Transformer encoders by 13% and 18% respectively over the next best baseline. For DMC environments, we observe a 256% relative improvement in average return when using saliency input. Our Transformer encoder results in an average 4% relative improvement in environment returns across all tasks over the next best baseline with a 7.5% improvement in `Cartpole Swing`.

A. CNN Encoder

We follow the CNN implementation used in prior work [4], [21] and compare several methods for incorporating saliency. In each approach, the CNN encoder and policy are trained jointly but take different inputs.

A saliency channel achieves the best task success rate for CNN encoder. In Table I, we find that naive ways of utilizing saliency, such as using saliency directly as input the policy (**Saliency**), are unable to achieve good performance on the task. We hypothesize that the saliency map alone is not sufficient to infer the exact orientation of the end-effector position critical for fine control. Supporting this hypothesis, we find that using saliency to mask the RGB observation (**RGB $\times$ Saliency**) achieves higher task success rate than **Saliency**, but is still worse than providing the raw RGB input (**RGB**). Although masking should help the encoder identify the important image features, it may still be nontrivial for the encoder to differentiate between similarly masked observations. Lastly, we find that incorporating saliency as an additional channel to the RGB input (**RGB(S)**) improves task success rate by more than 10% across all tasks. We hypothesize that the CNN encoder is able to utilize the saliency information to more effectively associate the observed rewards to the relevant features in the image.

B. MultiMAE Transformer

We compare MultiMAE representations pretrained with RGB only (**RGB**) and both RGB and saliency (**RGB+Saliency (PO)**, **RGB+Saliency (Ours)**). **RGB+Saliency (PO)** uses saliency only during pretraining while **RGB+Saliency (Ours)** uses saliency in both pretraining and downstream RL.

Training encoder with saliency improves RGB-only success rates at inference time. Even without saliency input during downstream RL, using saliency as an additional input modality during pretraining still improves downstream performance on 3 of the 4 tasks. Except for the `Reach` task, where performances are similar, **RGB+Saliency (PO)** achieves better success rate than **RGB**, with an average absolute gain of 10% across tasks.

Using saliency in both pretraining and inference yields the best performance. We compare the full ViSaRL method (**RGB+Saliency (Ours)**) to pretraining using only the RGB images (**RGB**) in Tables I demonstrating that multimodal pretraining with saliency information significantly outperforms single modality pretraining by at least a 10% margin across all tasks. Notably, **RGB** achieves only 19% success on `Faucet Open`, while our approach solves the task with 62% success rate. Using saliency as an input for both

Meta-World	CNN				MMAE		
	RGB	Saliency	RGB × Saliency	RGB(S)	RGB	RGB+Saliency (PO)	RGB+Saliency (Ours)
Reach	0.40±0.12	0.04±0.02	0.38±0.05	0.52±0.08	0.50±0.02	0.48±0.06	**0.62**±0.06
Drawer Open	0.18±0.25	0.04±0.02	0.10±0.04	0.48±0.06	0.84±0.02	0.88±0.04	**0.94**±0.04
Faucet Open	0.82±0.02	0.02±0.02	0.72±0.16	**0.86**±0.02	0.18±0.05	0.40±0.20	0.62±0.16
Door Open	0.42±0.04	0.10±0.06	0.22±0.10	0.48±0.06	0.36±0.18	0.52±0.08	**0.64**±0.02
Average	0.46±0.11	0.05±0.03	0.36±0.08	0.58±0.05	0.48±0.12	0.57±0.10	**0.65**±0.07

TABLE I: **Success rate** on four Meta-World manipulation tasks averaged across 50 rollouts and 3 seeds for the CNN and MultiMAE (MMAE) visual encoder backbones. Text in **maroon** indicates the best performing method per task.

	DMC-GB	CURL	RAD	RGB +Saliency (Ours)
color	Walker Walk	645±55	636±33	**823**±55
	Cartpole Swing	668±74	763±29	**870**±21
	Ball Catch	565±160	727±87	**962**±14
	Finger Spin	781±139	789±160	**823**±102
video	Walker Walk	572±121	595±85	**756**±42
	Cartpole Swing	418±72	434±58	**730**±32
	Ball Catch	402±169	520±44	**802**±78
	Finger Spin	612±55	588±82	**702**±83

TABLE II: **Average return** of ViSaRL and baseline methods on the color and video environments from DMC-GB.

	Saliency	RGB	RGB + Depth	RGB + SN
Reach	✗	0.50±0.02	0.43±0.07	0.46±0.04
	✓	**0.62**±0.06	**0.58**±0.04	**0.64**±0.06
Drawer Open	✗	0.82±0.02	0.76±0.06	0.80±0.04
	✓	**0.94**±0.04	**0.90**±0.04	**0.92**±0.04
Faucet Open	✗	0.18±0.04	0.22±0.04	0.24±0.04
	✓	**0.62**±0.16	**0.54**±0.06	**0.58**±0.10
Door Open	✗	0.36±0.18	0.28±0.14	0.34±0.10
	✓	**0.64**±0.02	**0.62**±0.04	**0.58**±0.04

TABLE III: Human-annotated saliency versus depth and surface normals (SN) as input modalities to MultiMAE model.

pretraining and downstream RL (**RGB+Saliency (Ours)**) improves task success rate over **RGB+Saliency (PO)** because there are new observations during online training that were not in the pretraining dataset.

ViSaRL representations generalize to unseen environments. We evaluate the generalizability of our learned representations on the challenging *random colors* and *video backgrounds* benchmark from DMControl-GB [32]. In DMControl-GB, agents trained in the original environment are evaluated on their generalization to the same environment with visually perturbed backgrounds using randomized color and video overlays. ViSaRL significantly outperforms the baselines across all tasks as shown in Table II, with an average 19% and 35% relative improvement respectively for the color and video settings.

Human-annotated saliency improves performance compared to depth and surface normals. We conduct ablation experiments to compare saliency versus other mid-level input modalities such as depth and surface normals proposed by Sax et al. [26]. We substitute saliency with these other modalities as input to the MultiMAE. In Table III, we observe that neither depth nor surface normal features alone improves task success over just using RGB image input. By contrast, adding saliency as an additional modality consistently improves task success suggesting that human-annotated saliency information can help learn better visual representations compared to other input modalities.

VI. REAL ROBOT EXPERIMENTS

We use a Kinova Jaco 2 (6-DoF) robot arm with a 1-DoF gripper. The observation space consists of a front-view image ($224 \times 224 \times 3$) from a Logitech webcam and proprioceptive information. We consider four tabletop manipulation tasks shown in Figure 5. In two of these tasks, we purposefully include distractor objects to evaluate the robustness of our learned representations to scene variations. We collected 10 demonstrations per task, resulting in an offline imitation learning dataset of around 10,000 transitions.

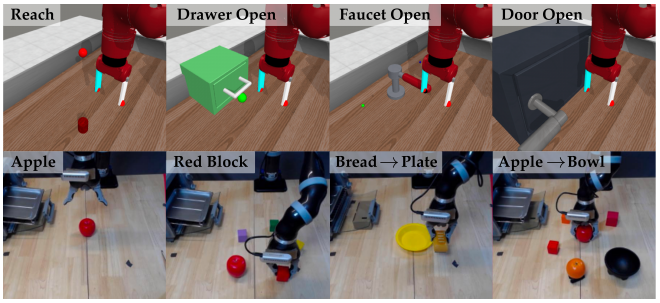

Fig. 5: **Evaluation Tasks**. Four Meta-World (top) simulation tasks and four real-robot tabletop manipulation tasks (bottom).

For each task, 10 randomly sampled frames are hand-annotated with saliency. Even with real-world images, only a small number of annotated frames are required to learn a good saliency predictor. We train an imitation learning policy by minimizing the mean-squared error between predicted end-effector pose and expert actions. We use a recurrent policy to encode history information and a 2-layer MLP to predict continuous actions.

ViSaRL scales to real-robot tasks and is robust to distractor objects. Videos of evaluation trajectories for each task can be found on the project website. Table IV reports the task success rates on 10 evaluation rollouts. Even on the easier Pick Apple task, using saliency augmented representations, **RGB+Saliency**, improves the success rate over **RGB**. On tasks with distractor objects and longer-horizon tasks such as Put Apple in Bowl, saliency-augmented representations nearly double the success rate.

VII. CONCLUSION

We proposed to use human-annotated saliency as an additional input modality for solving challenging visual robot control tasks. We present a simple approach, ViSaRL, to utilize saliency to learn robust image representations enabling more sample-efficient and generalizable policy learning.

Limitations and Future Work. One potential limitation of our user interface is that it could be tedious to collect

MultiMAE	Apple	Red Block	Bread → Plate	Apple → Bowl	Cumulative
RGB	6/10	4/10	3/10	1/10	14/40
+Saliency	8/10	7/10	6/10	6/10	27/40

TABLE IV: Task success rates in real-world tabletop manipulation tasks for **RGB** and **RGB+Saliency** with MultiMAE.

saliency annotations when scaling to more complex real world applications or video saliency [33]. Future work could investigate alternative interfaces that will enable collecting more saliency data, e.g., area-based methods or by tracking the eye gaze of the user [34].

One can further evaluate the generalizability of ViSaRL on the recent benchmark, The Colosseum [35], a suite of manipulation tasks design to measure the robustness of trained robot policies against visual perturbations.

In this paper, we only considered static frame saliency maps for single-object manipulation tasks. We plan to extend our approach to handle longer-horizon multi-object tasks using video saliency models [36] which can learn to encode more flexible temporal saliency representations across a sequence of frames. This extension could be implemented by asking the human users to watch video clips of the trajectories and annotate saliency over these clips.

REFERENCES

[1] K. P. Darby, S. W. Deng, D. B. Walther, and V. M. Sloutsky, "The development of attention to objects and scenes: From object-biased to unbiased," *Child development*, 2021.

[2] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998.

[3] M. Laskin, A. Srinivas, and P. Abbeel, "CURL: Contrastive Unsupervised Representations for Reinforcement Learning," in *International Conference on Machine Learning (ICML)*, 2020.

[4] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas, "Reinforcement Learning with Augmented Data," *Neural Information Processing Systems (NeurIPS)*, 2020.

[5] S. Tunyasuvunakool, A. Muldal, Y. Doron, S. Liu, S. Bohez, J. Merel, T. Erez, T. Lillicrap, N. Heess, and Y. Tassa, "dm_control: Software and Tasks for Continuous Control," *Software Impacts*, 2020.

[6] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, "Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning," in *Conference on Robot Learning (CoRL)*, 2020.

[7] S. Cabi, S. G. Colmenarejo, A. Novikov, K. Konyushkova, S. Reed, R. Jeong, K. Zolna, Y. Aytar, D. Budden, M. Vecerik, *et al.*, "Scaling data-driven robotics with reward sketching and batch reinforcement learning," *Robotics: Science and Systems (RSS)*, 2020.

[8] A. Bobu, M. Wiggert, C. Tomlin, and A. D. Dragan, "Feature Expansive Reward Learning: Rethinking Human Input," in *Human-Robot Interaction (HRI)*, 2021.

[9] N. Wilde, E. Biyik, D. Sadigh, and S. L. Smith, "Learning Reward Functions from Scale Feedback," in *Conference on Robot Learning (CoRL)*, 2022.

[10] S. Tao, X. Li, T. Mu, Z. Huang, Y. Qin, and H. Su, "Abstract-to-Executable Trajectory Translation for One-Shot Task Generalization," in *Neural Information Processing Systems (NeurIPS) Deep Reinforcement Learning Workshop*, 2022.

[11] Y. Tong, H. Konik, F. Cheikh, and A. Tremeau, "Full Reference Image Quality Assessment Based on Saliency Map Analysis," *Journal of Imaging Science and Technology*, 2010.

[12] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond Frame-level CNN: Saliency-Aware 3-D CNN With LSTM for Video Action Recognition," *IEEE Signal Processing Letters*, 2016.

[13] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra, "Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?" *Computer Vision and Image Understanding*, 2017.

[14] Q. Li, Y. Zhou, and J. Yang, "Saliency Based Image Segmentation," in *International Conference on Information and Multimedia Technology (ICIMT)*, 2011.

[15] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *arXiv preprint arXiv:1312.6034*, 2013.

[16] T. N. Mundhenk, B. Y. Chen, and G. Friedland, "Efficient Saliency Maps for Explainable AI," *arXiv preprint arXiv:1911.11293*, 2019.

[17] R. Zhao, W. Oyang, and X. Wang, "Person Re-Identification by Saliency Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[18] A. Atrey, K. Clary, and D. Jensen, "Exploratory Not Explanatory: Counterfactual Analysis of Saliency Maps for Deep Reinforcement Learning," *International Conference on Learning Representations (ICLR)*, 2020.

[19] M. Rosynski, F. Kirchner, and M. Valdenegro-Toro, "Are Gradient-based Saliency Maps Useful in Deep Reinforcement Learning?" *arXiv preprint arXiv:2012.01281*, 2020.

[20] A. Boyd, P. Tinsley, K. W. Bowyer, and A. Czajka, "CYBORG: Blending Human Saliency Into the Loss Improves Deep Learning," in *Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[21] D. Bertoin, A. Zouitine, M. Zouitine, and E. Rachelson, "Look where you look! Saliency-guided Q-networks for generalization in visual Reinforcement Learning," *Neural Information Processing Systems (NeurIPS)*, 2022.

[22] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC Superpixels Compared to State-of-the-Art Superpixel Methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2012.

[23] A. Boyd, K. W. Bowyer, and A. Czajka, "Human-Aided Saliency Maps Improve Generalization of Deep Learning," in *Winter Conference on Applications of Computer Vision (WACV)*, 2022.

[24] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3M: A Universal Visual Representation for Robot Manipulation," *Conference on Robot Learning (CoRL)*, 2022.

[25] S. Karamcheti, S. Nair, A. S. Chen, T. Kollar, C. Finn, D. Sadigh, and P. Liang, "Language-Driven Representation Learning for Robotics," *arXiv preprint arXiv:2302.12766*, 2023.

[26] A. Sax, B. Emi, A. R. Zamir, L. Guibas, S. Savarese, and J. Malik, "Mid-Level Visual Representations Improve Generalization and Sample Efficiency for Learning Visuomotor Policies," *Conference on Robot Learning (CoRL)*, 2019.

[27] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "MultiMAE: Multi-modal Multi-task Masked Autoencoders," *European Conference on Computer Vision (ECCV)*, 2022.

[28] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning Pixel-wise Contextual Attention for Saliency Detection," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[29] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," in *Computer Vision and Pattern Recognition (CVPR)*, 2022.

[30] 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," *International Conference on Learning Representations (ICLR)*, 2021.

[31] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, *et al.*, "Soft Actor-Critic Algorithms and Applications," *International Conference on Machine Learning (ICML)*, 2018.

[32] N. Hansen and X. Wang, "Generalization in reinforcement learning by soft data augmentation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 611–13 617.

[33] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting Video Saliency: A Large-scale Benchmark and a New Model," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[34] D. P. Papadopoulos, A. D. Clarke, F. Keller, and V. Ferrari, "Training Object Class Detectors from Eye Tracking Data," in *European Conference on Computer Vision (ECCV)*, 2014.

[35] W. Pumacay, I. Singh, J. Duan, R. Krishna, J. Thomason, and D. Fox, "The colosseum: A benchmark for evaluating generalization for robotic manipulation," *arXiv preprint arXiv:2402.08191*, 2024.

[36] D. Rudoy, D. B. Goldman, E. Shechtman, and L. Zelnik-Manor, "Learning video saliency from human gaze using candidate selection," in *Computer Vision and Pattern Recognition (CVPR)*, 2013.