# Self-Supervised Attention-Aware Reinforcement Learning

**Haiping Wu,**[1, 2] **Khimya Khetarpal,**[1, 2] **Doina Precup** [1, 2, 3]
[1] McGill University
[2] Mila
[3] Google DeepMind, Montreal.
{haiping.wu2, khimya.khetarpal}@mail.mcgill.ca, dprecup@cs.mcgill.ca

## Abstract

Visual saliency has emerged as a major visualization tool for interpreting deep reinforcement learning (RL) agents. However, much of the existing research uses it as an analyzing tool rather than an inductive bias for policy learning. In this work, we use visual attention as an inductive bias for RL agents. We propose a novel self-supervised attention learning approach which can 1. learn to select regions of interest without explicit annotations, and 2. act as a plug for existing deep RL methods to improve the learning performance. We empirically show that the self-supervised attention-aware deep RL methods outperform the baselines in the context of both the rate of convergence and performance. Furthermore, the proposed self-supervised attention is not tied with specific policies, nor restricted to a specific scene. We posit that the proposed approach is a general self-supervised attention module for multi-task learning and transfer learning, and empirically validate the generalization ability of the proposed method. Finally, we show that our method learns meaningful object keypoints highlighting improvements both qualitatively and quantitatively.

## Introduction

In recent years, deep reinforcement learning methods (Mnih et al. 2013, 2015, 2016) have achieved great success in large part driven by the revolution in convolution neural networks (CNN) and feed-forward networks as function approximators. Most methods directly use the CNN extracted features of entire images as state representation and then perform reasoning over this representation. Humans, on the other hand, tend to focus on salient areas of interest such as objects (Borji, Sihite, and Itti 2012) and faces (Judd et al. 2009) for understanding a scene, allowing them to quickly process most relevant parts of the observations during decision making (Wyart and Tallon-Baudry 2009).

Researchers (He et al. 2015; Wang et al. 2015; Zhao et al. 2015; Hou et al. 2017) have made significant efforts in scene understanding by performing saliency detection via image segmentation. However, most of these methods depend on human-annotated training datasets (Liu et al. 2010; Alpert et al. 2011). Collecting such datasets can be infeasible and come at the expense of time and manual labor. Moreover, these methods are highly incumbent on the data distribution seen during training and generalize poorly to unseen tasks. Instead, we resort to *unsupervised* learning methods that do not need extra information and explicit supervision in the form of labelled datasets. Specifically, we are interested in building reinforcement learning (RL) agents which learn representations guided by an understanding of what is important in a scene for sequential decision making.

One approach to learning such meaningful representations is via attention masks; as they create a bottleneck where the gradients are driven by the final RL task objective (specified via a reward signal). We refer to this approach as *top-down attention*. However, the meaning and quality of the learned masks via the top-down approach are typically task-specific and therefore hard to generalize across unseen scenarios. Moreover, top-down attention masks are often used as interpretation tools for understanding the learned policies (Yang et al. 2018; Shi et al. 2020). Contrary to the top-down attention methods, we propose an approach for generic understanding of the scene regardless of the tasks or policies, and use it to guide the learning of policies as opposed to explaining them (Greydanus et al. 2018).

Object-oriented representation is a long-standing approach to understanding and simplifying a scene (Eslami et al. 2016; Greff, Van Steenkiste, and Schmidhuber 2017; Kosiorek et al. 2018; Greff et al. 2019; Lin et al. 2020). Recent works (Zhang et al. 2018; Jakab et al. 2018; Kulkarni et al. 2019; Minderer et al. 2019; Gopalakrishnan, van Steenkiste, and Schmidhuber 2021) try to obtain object keypoints in an unsupervised manner. However, current unsupervised keypoints detection methods including the Transporter (Kulkarni et al. 2019) are limited in that they do not deal with variable number of objects, scale, and classes of objects. Furthermore, the use of object-oriented representation for deep RL has not been extensively explored. For instance, how to obtain a meaningful state representation given these objects is not immediately clear and remains an open question.

We here argue that attention masks are a better technique to help the learning of policies given current tools. Attention masks aim to find salient areas in a scene and account for any number of regions in that they are not restricted to specific object categories or count. More importantly, since it has the same form as that of the original image (i.e. a map of

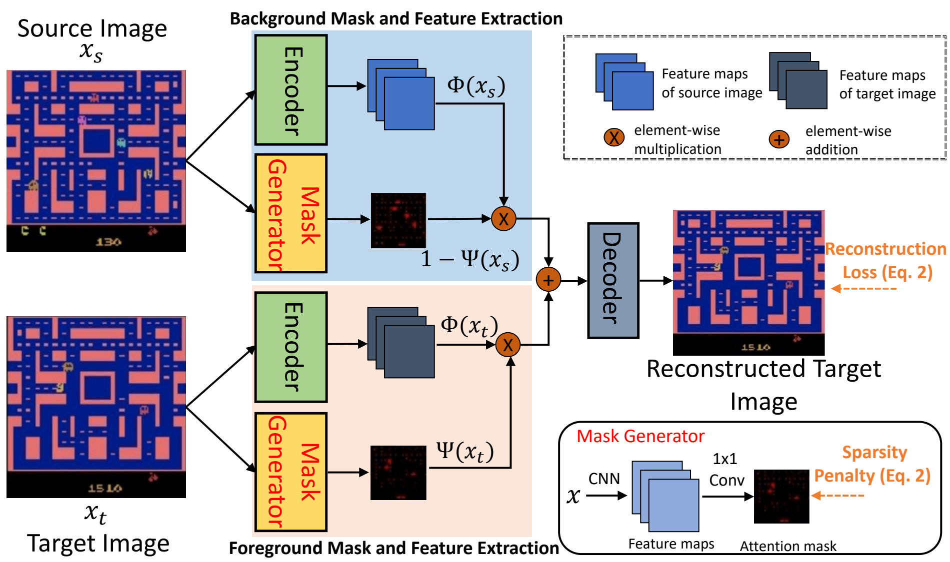


Figure 1: Proposed self-supervised attention module pipeline. The core idea is to employ a self-supervised loss through an auto-encoder architecture with a bottleneck. The module tries to reconstruct the target image $x_t$ by using minimal information (features of foreground regions) from the target image $x_t$, and other needed information from source image $x_s$. The mask generator outputs the foreground attention masks for the input images as $\Psi(x_s)$ and $\Psi(x_t)$. We have $1-\Psi(x_s)$ as the background regions of the source image. The decoder reconstructs the target image using the foreground features of the target image and background features of the source image. Image reconstruction losses and $\ell_1$ sparsity over attention masks are used.

the attention values v.s. a map of pixel values), it is straightforward to plug in any existing deep RL methods for decision making once we have constructed such a representation. Inspired by Transporter (Kulkarni et al. 2019), we propose a self-supervised attention module which is designed for learning attention masks instead of object keypoints. The module is an auto-encoder like architecture with a bottleneck in attention masks that it needs to correctly identify the regions of interest to perform the image reconstruction. The learning is performed in an unsupervised manner, where the self-supervised attention module is not related to a specific task as the top-down attention methods. This view would potentially lead to a better generalization ability. The proposed attention module is shown in Figure 1.

The learned attention masks are class-agnostic, in that they are not limited in response to certain object categories. Moreover, they account for various shapes, number of objects, as opposed to pre-defined number of objects during training (Kulkarni et al. 2019). Furthermore, we show that one could easily extract object keypoints as well from our learned masks, demonstrating the flexibility of our method. Our main contributions and findings are as follows:

1. We design a self-supervised attention mask module that learns general-purpose attention masks through a novel self-supervised loss.
2. We incorporate the self-supervised attention module in deep RL methods, and empirically show gains in both the convergence speed and final scores in single-task setting.
3. We empirically demonstrate that the attention learned via our self-supervised approach results in generalization capabilities in both transfer and multi-task settings.
4. We extract object keypoints from our masks and show that they are qualitatively and quantitatively better than the Transporter (Kulkarni et al. 2019), further highlighting the efficacy of our method.

## Self-Supervised Attention for Reinforcement Learning

In this section, we first describe the proposed method for learning attention masks in a self-supervised manner. We then show how the self-supervised attention module can be plugged in existing RL methods to improve policy learning.

### Method: Self-Supervised Attention Module

Our aim is to learn a mask that indicates the potential of each location in the visual input being the region of interest. Hereafter, we refer to the region of interest as the *foreground*, and *background otherwise*. Inspired by the Transporter (Kulkarni et al. 2019) model, we design a bottleneck architecture to reconstruct images, which could ideally differentiate between the interested foreground and background, in a self-supervised manner. Contrary to Transporter, our attention module learns the foreground **attention mask** rather than a **pre-defined number of keypoints**. The overall architecture is shown in Figure 1.

Given a source frame $x_s$ and a target frame $x_t$, randomly sampled from one game-play, we design the self-supervised learning task as reconstructing the target frame $x_t$ from the source frame $x_s$. We use auto-encoder with bottleneck to

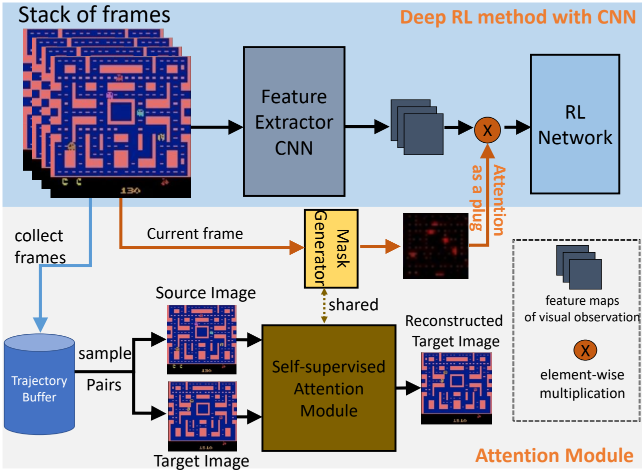


Figure 2: Attention-aware Reinforcement Learning. Pipeline demonstrating the proposed self-supervised attention module as a plug for existing deep RL methods. The shaded blue area shows the original deep RL pipeline with a CNN. In addition, the mask generator outputs the attention mask for the current frame. We then use this attention mask as a plug by changing the original feature maps of visual observation to a multiplication of the attention mask and feature maps. The RL method then reasons upon this modified feature maps. The mask generator is identical to the one in the self-supervised attention module, which is detailed in Figure 1. The Self-supervised Attention Module is trained using pairs of images from a trajectory buffer.

construct $x_t$. First, the encoder extracts features of $x_s$ and $x_t$ as $\Phi(x_s), \Phi(x_t) \in \mathbb{R}^{H' \times W' \times D}$ respectively.

The mask generator outputs the mask maps of $x_s$ and $x_t$ as $\Psi(x_s), \Psi(x_t) \in [0,1]^{H' \times W'}$, indicating the probability of being interested for the corresponding feature map location. The features used for reconstructing $x_t$ are then calculated as follows:

$$\hat{\Phi}\left(\boldsymbol{x}_s, \boldsymbol{x}_t\right) \triangleq \overbrace{\left(1-\Psi\left(\boldsymbol{x}_s\right)\right) \cdot \left(1-\Psi\left(\boldsymbol{x}_t\right)\right) \cdot \Phi\left(x_s\right)}^{\text{background features}} + \underbrace{\Psi\left(\boldsymbol{x}_t\right) \cdot \Phi\left(x_t\right)}_{\text{foreground features}}. \tag{1}$$

Finally, besides the original auto-encoder pipeline that the decoder reconstructs $\hat{x_t}^{\text{auto}}$ from features $\Psi(x_t)$, the decoder also takes in the features $\hat{\Phi}\left(\boldsymbol{x}_s, \boldsymbol{x}_t\right)$ and outputs the reconstructed $\hat{x_t}$.

Ideally, we want the decoder to use the features that combine the background features from $x_s$ and foreground features from $x_t$ to reconstruct $x_t$ as in Eq. 1. However, directly optimizing the reconstruction loss between $x_t$ and $\hat{x_t}$ would give a trivial solution for masks that $\Psi(x_t) = 1$, which is not in our interest. Therefore, we propose to add a penalty term for the masks that leads to minimize the locations that are identified as regions of interest. We can also interpret this penalty term acts as a sparsity regularizer. The overall loss for training the self-supervised attention mask is defined as a combination of the reconstruction losses and sparsity penalty, as follows:

$$\mathcal{L}_{\text{attention}} = \underbrace{\|\hat{x_t} - x_t\|_{2*}^2 + \|\hat{x_t}^{\texttt{auto}} - x_t\|_2^2}_{\text{reconstruction loss}} + \underbrace{\lambda_m \|\Psi(x_t)\|_1}_{\text{sparsity penalty}}. \tag{2}$$

where $\|\cdot\|_{2*}^2$ is squared-$\ell_2$ norm with threshold $\delta$, that ignores the terms that have a squared value less than $\delta$. It is defined as follows:

$$\|\boldsymbol{y}\|_{2*}^2 = \sum_k y_k^2, \quad \forall k \text{ if } y_k^2 \geqslant \delta. \tag{3}$$

We ignore the error when the squared-$\ell_2$ distance of a pixel location between the reconstruct $\hat{x_t}$ and target $x_t$ is below $\delta$. This allows the model to ignore small changes that might occur in the background, focusing on salient parts of the reconstruction. $\delta$ is a hyper-parameter. The second term in the reconstruction loss is the original auto-encoder loss, which is used for regulating the feature space to be meaningful.

$\lambda_m$ is a hyper-parameter that balances the total number of regions of interest. Since there is a penalty for positions that are identified as regions of interest, the loss would force the model to select relatively more important (necessary) parts from $x_t$ and ignoring the background in $x_s$ with less penalty.

## Attention-Aware Reinforcement Learning

We now discuss the utilization of the self-supervised attention module as a plug for existing deep RL methods. For any deep RL methods that uses a convolutional neural network (CNN), the idea is to exploit the intermediate features extracted by the CNN. Specifically, we multiply the features learned via the attention mask, and leave everything else unchanged for the policy learning. The deep RL methods with CNN is abstracted in the top blue area of Figure 2 demonstrating the attention-aware RL pipeline in Figure 2. The *attention module* is highlighted in bottom gray area and can be used as a plug for *any* deep RL method.

For the baseline RL algorithm, we use Advantage Actor Critic (A2C) (Mnih et al. 2016). For a visual observation $x$, a convolutional neural network (CNN) extracts intermediate feature maps as $f(x) \in \mathbb{R}^{H' \times W' \times C}$. The policy and state value function is then predicted using the feed-forward networks $\pi(a_t|f(x))$, $V(f(x))$ as function approximators. Policy gradient is used to train the networks, and we refer to the loss as $\mathcal{L}_{\text{RL}}$, as used in Mnih et al. (2016).

For the attention-aware RL learning shown in Figure 2, in addition to the original CNN extracting intermediate feature maps as $f(x)$, an additional self-supervised attention module is used, which takes the visual state $x$ and produces the attention mask $\Psi(x) \in \mathbb{R}^{H' \times W'}$ through the mask generator. The original feature $f(x)$ is multiplied by the attention mask, obtaining the new feature as $\Psi(x)f(x)$. Thus, the policy and state value functions for A2C method are predicted as $\pi(a_t|\Psi(x)f(x))$, $V(\Psi(x)f(x))$.

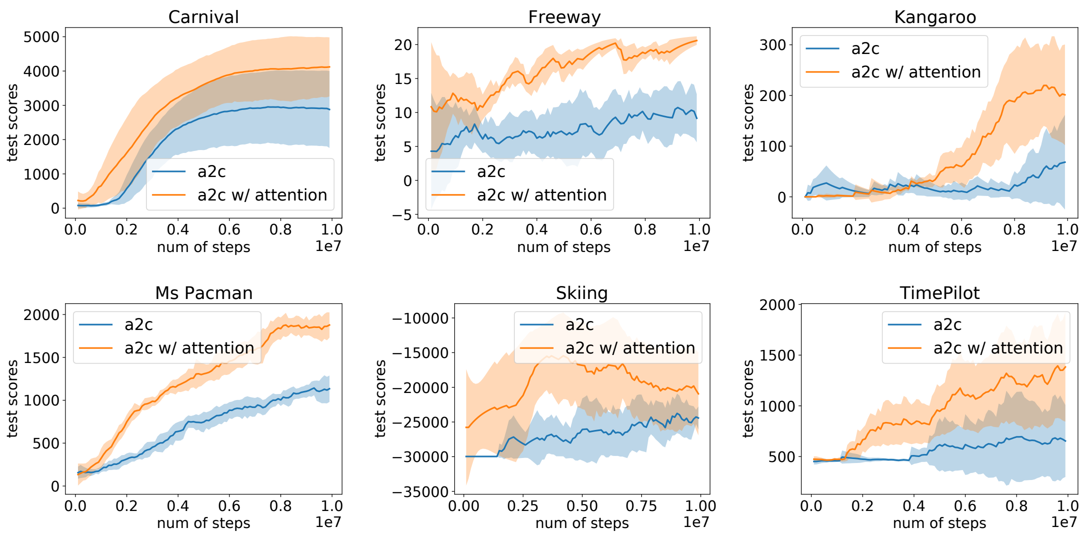


Figure 3: Single-task Learning. Average (over 5 random seeds) test scores during learning of A2C with/without the our self-supervised attention mask. Our method consistently performs better than the baseline in both convergence speed and test scores.

The self-supervised attention module could be trained offline or jointly trained in an online fashion with the RL agent. For offline training, we sample source and target image pairs $\{(x_s, x_t)\}$ from a pre-collected image set or offline trajectories, and minimize the loss $\mathcal{L}_{\text{attention}}$ in Eq. 2. For joint training with RL agent, the source and target image pairs $\{(x_s, x_t)\}$ are sampled from the online trajectory of the current agent (as in single task learning experiments). The total training loss is:

$$\mathcal{L} = \mathcal{L}_{\text{RL}} + \mathcal{L}_{\text{attention}}. \tag{4}$$

The plugged attention module tries to simplify the original features by suppressing the response of background regions, which helps the abstraction of the observation, and thus improves policy learning.

## Experiments

We now evaluate the proposed method in different settings to demonstrate the efficacy of the self-supervised attention mask module. We evaluate our method on Atari ALE (Bellemare et al. 2013; Brockman et al. 2016) games. First, we show that RL agents equipped with the self-supervised attention masks perform better in both convergence speed and the scores obtained in a single-task setting. Then, we demonstrate that one universal attention mask could be applied across different tasks, showcasing the generalization ability. Finally, we show that the learned attention mask could enable transfer to unseen tasks. The implementation details including experiment setups, network architectures and hyperparameters are provided in the appendix. The source code is available here. [1]

### Single-task Learning

In the single-task setting, the self-supervised attention module and the RL agent are jointly trained in an online fashion for each game using the loss defined in Eq 4. Pairs of source and target frames are randomly sampled from the agent's trajectory to train the attention module. The results are shown in Figure 3. We observe that by masking the features using the attention, the agents learn faster and perform better than the baseline A2C method on Atari games shown in Figure 3. The qualitative performance indicates that the attention mask learned is indeed helpful for the understanding of the scene. By only seeing regions of interest, the scene is potentially simplified for understanding and reasoning.

We show additional results using another baseline method i.e. ACKTR (Wu et al. 2017) shown in Figure 4. It is observed that the self-supervised attention-aware agents perform consistently better on the games shown. We report the average performance averaged over 5 independent runs.

**Comparison with Top-Down Attention.** To compare with the top-down (goal-driven) attention, we simply utilize the attention mask as well without the use of $\mathcal{L}_{\text{attention}}$, where the supervision signals come from the RL objectives. More specifically, the intermediate feature $f(x)$ is multiplied by the masks $\Psi(x)$ generated from the mask generator, obtaining $\Psi(x)f(x)$, the same way as the self-supervised attention-aware RL. However, the parameters of $\Psi(x)$ are learned by back-propagating gradients from $\pi(a_t|\Psi(x)f(x))$, $V(\Psi(x)f(x))$ using chain-rule, with the only loss $\mathcal{L}_{\text{RL}}$. We compare this top-down attention guided RL with our self-supervised attention-aware RL. The results are shown in Figure 5. We find that the top-down attention guided RL could perform better or worse than the baseline method without attention masks. The top-down attention is guided only by the final objective. Thus the quality and meaning of it highly depend on the task-specific RL objective. On the other hand, the self-supervised attention-aware RL agents perform better than both the top-down attention guided RL as well as the baseline.

[1] https://github.com/happywu/Self-Sup-Attention-RL

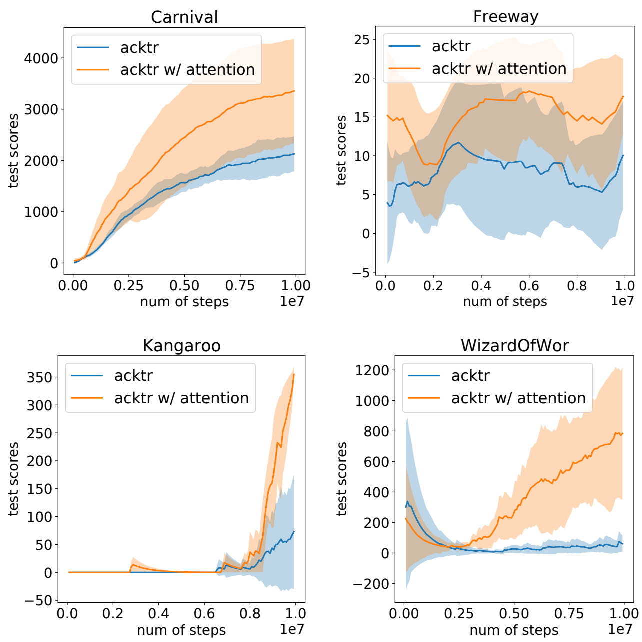


Figure 4: Single-task Learning. Average (over 5 random seeds) test scores during learning of ACKTR algorithm with/without our self-supervised attention mask.Our method consistently performs better than the baseline in both convergence speed and test scores.

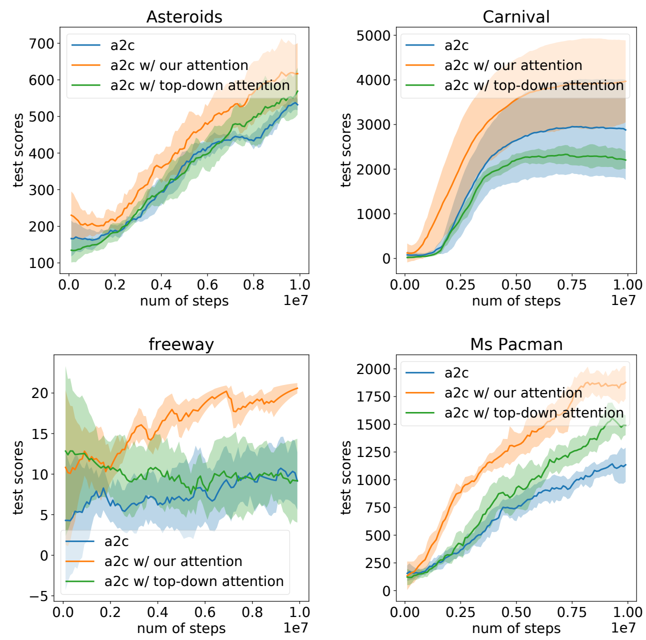


Figure 5: Single-task Learning - Comparison with top-down attention. Comparison between our self-supervised attention module, and top-down mask module as a plug for A2C. Average (over 5 random seeds) test scores during training are reported. Our self-supervised attention consistently outperforms the top-down attention method.

## Multi-task Learning

**One mask module across different tasks.** Unlike the keypoints representation in Transporter (Kulkarni et al. 2019), where the keypoints are linked to specific objects, or the top-down attention masks that are related to specific RL objectives, the self-supervised attention masks are not semantically restricted in specific scenes or RL objectives. Consequentially, we could intuitively train the mask module across a range of tasks, potentially resulting in a *universal* (of multiple games) attention mask for many tasks.

To show the generalization ability of the self-supervised attention module, we train the attention module in a multi-task setting. More specifically, we train the self-supervised mask module on frames jointly collected from three different games (Asteroids, Assault, Ms.Pacman) using a random policy. For each training iteration, image pairs $(x_s, x_t)$ are randomly sampled from the three games, and the networks are trained using $\mathcal{L}_{\text{attention}}$. We then apply the *universal* trained self-supervised mask module to RL learning of these different games by multiplying the intermediate features $f(x)$ to get $\Psi(x)f(x)$. The agents learn to play each game separately from scratch using $\mathcal{L}_{\text{RL}}$ and the attention module parameters are fixed. The results are shown in Figure 6. We see that this *universal* attention module facilitates learning policies on different games, achieving nearly the same performance compared to using the self-supervised attention module specifically trained on one game as in single-task setting (as shown in Figure 3). We conjure that the training data covering a variety of tasks is the potential cause for the attention module to have better generalization abilities.

## Transfer Learning

**Transfer mask across tasks.** To further validate the transfer ability of the self-supervised attention module, we design a pipeline that shows the learned attention mask can generalize to related scenes which it has never seen during training. First, the self-supervised mask module is trained on frames from the source domain Atari game, *JourneyEscape* or *Assault* using the loss $\mathcal{L}_{\text{attention}}$. We then fix the parameters of the attention module, and apply it to the RL learning of the target domain game *Asteroids* and *Carnival* in another instance, using the self-supervised attention-aware RL. The results are shown in Figure 7. Notably, even when the attention module has never seen any frames from the target games, the attention masks are still beneficial for the learning of agents as it provides significant gains in the performance. This further highlights that the proposed self-supervised module can generalize to unseen scenarios that have similar visual components, indicating the transfer ability.

# Bottom-up Object Extraction

While our approach does not rely on a predefined number of keypoints and results in task-agnostic attentive representation learning, the ability to extract object keypoints could be potentially useful as it provides means to represent the knowledge in the form of objects which is akin to human understanding of the world. In this section, we show prelim-

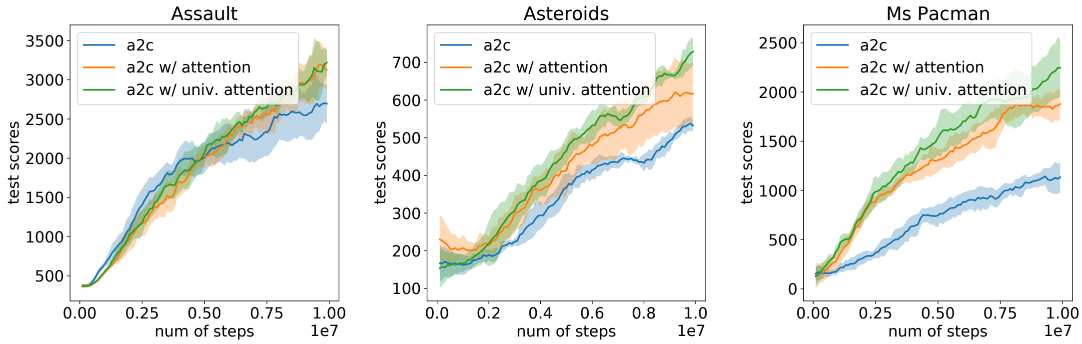


Figure 6: Multi-task Learning. Comparison between the baseline method A2C, A2C with the self-supervised attention module, A2C with the the universal attention module jointly trained on three games. Average (over 5 random seeds) test scores during training are reported. The universal attention guided policies perform comparable to those with the self-supervised attention trained on single-task.

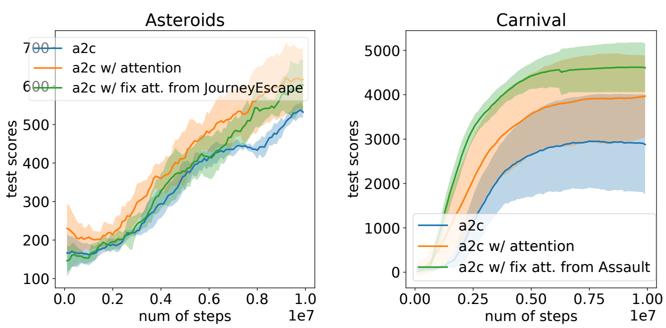


Figure 7: Transfer Learning. Comparison between the baseline method A2C and A2C with the fix attention module trained on JourneyEscape or Assault. Average (over 5 random seeds) test scores during training are reported. We note that the attention module has the ability to transfer across games.

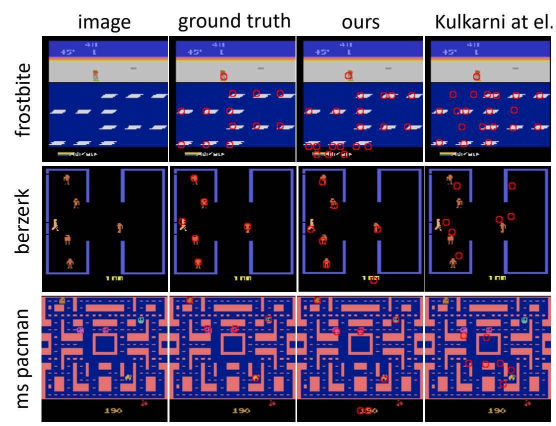


Figure 8: Qualitative Analysis. Comparison of object keypoints extracted form the self-supervised attention masks, Transporter (Kulkarni et al. 2019) and the ground truth. The number of object keypoints $k$ are set to the same as Transporter. Our method successfully focuses on important objects and is visually better than Transporter.

inary results on using our self-supervised attention module to extract object keypoints, with the potential to facilitate object-centric RL. We extract object locations from the self-supervised attention masks. Specifically, each cell in the attention mask map is considered as a candidate for the center of one object. Non-maximum suppression (NMS) (Rosenfeld and Thurston 1971) is applied upon the learned attention mask to get the object center proposals. We end up with $k$ object keypoints by taking the $k$ max object proposals with the attention mask value. The pseudocode is provided in the appendix.

We compare with Transporter (Kulkarni et al. 2019) as shown in Figure 8. We find that the objects extracted from the self-supervised attention masks are reasonably focused on salient objects, as compared to both the ground truth objects extracted from (Anand et al. 2019) and the Transporter (Kulkarni et al. 2019) method. Further, we could easily adjust the number of object keypoints by using different $k$, in contrast to Transporter (Kulkarni et al. 2019), in which the number of keypoints has to be predefined during the training. We notice that the object keypoints extracted for Ms. Pacman in Figure 8 focused mainly on Ms. Pacman and monsters, when the number of keypoints $k$ was set to 7. The resulting keypoints also focused on the remaining pellets, when increasing $k$ from 7 to 10 as shown in Figure 9. This further demonstrates the flexibility of our method.

We further quantify the improvements in comparison with the baseline i.e. Transporter (Kulkarni et al. 2019) through recall and precision metrics. We compute these metrics using the predicted object locations and the ground truth locations from Anand et al. (2019). Two keypoints with a distance less than a threshold $\epsilon$ are considered as a successful detection. $\epsilon$ is determined as in the baseline. The number of keypoints $k$ is the same as in Transporter. We report in Table 1 that our method performs better than (Kulkarni et al. 2019) in both recall and precision. Both quantitative and qualitative measures highlight the soundness of the self-supervised attention mask and extracted object keypoints. Given the flexibility of our self-supervised attention

| | Frostbite | | Berzerk | | Ms. Pacman | |
|---|---|---|---|---|---|---|
| | Recall | Prec | Recall | Prec | Recall | Prec |
| Transporter | 0.74 | 0.56 | 0.32 | 0.27 | 0.33 | 0.23 |
| Ours | **0.76** | **0.57** | **0.54** | **0.46** | **0.72** | **0.52** |

Table 1: Recall and Precision (Prec) comparison on different games with Transporter (Kulkarni et al. 2019). Our method performs better than Transporter in both recall and precision metrics in all 3 games tested here.

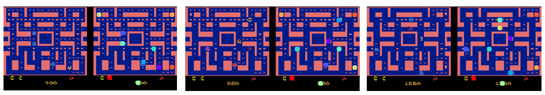

Figure 9: Variable Number of Object Keypoints Extraction using the self-supervised attention masks for Ms.Pacman. The number of keypoints are easily adjustable.

masks, the extracted objects could potentially be used to form object-centric representation for RL agents, which is scope for future work.

## Related Work

Unsupervised bottom-up salient object detection and segmentation (Itti, Koch, and Niebur 1998) methods have immense potential to simplify the scene for decision making in the RL paradigm (Sutton and Barto 2018). A naive approach is to use an off-the-shelf saliency method to foveate regions of interest in an input image for policy learning (Khetarpal and Precup 2018). However this would heavily rely on the training dataset used for the pre-trained saliency model and therefore has limited performance guarantees. Seeking self-supervision in the form of non-explicit labels is more appealing instead. Greydanus et al. (2018) adapts saliency methods to visualize and interpreting agents. Goel, Weng, and Poupart (2018) use optical flow as a label to supervise the learning of the segmentation, and the features for segmentation are augmented for policy learning. Yuezhang, Zhang, and Ballard (2018) also adapt optical flow between two frames to serve as an attention map, and then incorporate the attention by multiplying the intermediate features of agents. Optical flow captures motion information between frames and thus identifying moving objects. Unlike Goel, Weng, and Poupart (2018), our self-supervised attention mask does not aim to find moving objects between two frames, but minimal regions of interest that could reconstruct the scene. Our supervision signal does not come from optical flows, but through reconstruction via a bottleneck architecture. Optical flow captures local temporal information and is therefore reliable only for nearby frames. However, our attention module is able to capture information across a relatively larger temporal window.

Mott et al. (2019) uses a soft, recurrent, top-down attention by creating a bottleneck for the learning of agents, leading to the attention maps which focus on task-relevant information. They manage to achieve comparable performance to the baseline while being interpretable. Shi et al. (2020); Yang et al. (2018) also utilize a top-down attention map, where they use an attention map to interpret the behavior of the policies. In contrast to top-down attention methods, our self-supervised attention masks are not task-specific, and could be used across different tasks as shown in the universal mask experiments in Figure 6.

More recently, Zhang et al. (2019) introduced a human action and gaze dataset for Atari games. They utilize the annotated gaze to predict human action labels, showing that the gaze information is useful for imitation learning. An interesting approach then is to design auxiliary gaze loss (Saran et al. 2020) that uses AtariHead dataset to help the inverse RL and behavioral cloning problems. Unlike these methods, our self-supervised attention masks do not require any annotation of the human gaze or actions. Moreover, we can directly apply the learned attention masks to model-free RL methods instead of imitation learning.

Closely related to our work, Zhang et al. (2018); Jakab et al. (2018) design an auto-encoder architecture with keypoints bottleneck to perform unsupervised object keypoints detection. Transporter (Kulkarni et al. 2019) extends the pipeline with a feature transporter mechanism to extract object keypoints without the use of temporal transformations in the form of optical flow. Our self-supervised attention mask module also utilizes a bottleneck architecture similar to Transporter. However, our approach has two key differences. First, our attention module does not directly generate object keypoints, and instead learns to produce foreground/background focused attention masks. As a result, we do not have to predefine the number of keypoints to be detected and could obtain variable number of keypoints from the learned attention masks. Second, due to the large temporal window and the ability to capture most relevant regions of interest, our attention masks could be used across multiple scenarios (such as tasks or scenes), showcasing its better generalization ability.

## Discussion

We designed a self-supervised attention module which can identify salient regions of interest without explicit hand labelled annotations. Our approach is flexible in that the attention mask is not related to particular object semantics or restricted to specific downstream tasks. It is straightforward to plug-and-play the proposed method in existing deep RL approaches with CNNs as feature extractor since the attention mask has the same form as the CNN extracted feature maps. Extensive experiments show that the self-supervised attention module not only improves policy learning in the single-task setting, but also, in transfer and multi-task settings.

Additionally, we show preliminary results for extracting object keypoints from the self-supervised attention mask. The extracted keypoints reasonably focus on interested objects and are comparable to baseline specially designed for object keypoints detection. Our approach allows to change the number of extracted keypoints at inference time without re-training as required. In the future, this ability to extract task-agnostic object keypoints could be potentially used to build symbolic high level representations.

## References


Alpert, S.; Galun, M.; Brandt, A.; and Basri, R. 2011. Image segmentation by probabilistic bottom-up aggregation and cue integration. *IEEE transactions on pattern analysis and machine intelligence* 34(2): 315–327.

Anand, A.; Racah, E.; Ozair, S.; Bengio, Y.; Côté, M.-A.; and Hjelm, R. D. 2019. Unsupervised state representation learning in atari. In *Advances in Neural Information Processing Systems*, 8766–8779.

Bellemare, M. G.; Naddaf, Y.; Veness, J.; and Bowling, M. 2013. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research* 47: 253–279.

Borji, A.; Sihite, D.; and Itti, L. 2012. Salient Object Detection: A Benchmark. Computer Vision—ECCV 2012: the 12th European Conference on Computer Vision; 2012 Oct 7-13; Florence, Italy.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. Openai gym. *arXiv preprint arXiv:1606.01540* .

Eslami, S. A.; Heess, N.; Weber, T.; Tassa, Y.; Szepesvari, D.; Hinton, G. E.; et al. 2016. Attend, infer, repeat: Fast scene understanding with generative models. In *Advances in Neural Information Processing Systems*, 3225–3233.

Goel, V.; Weng, J.; and Poupart, P. 2018. Unsupervised video object segmentation for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, 5683–5694.

Gopalakrishnan, A.; van Steenkiste, S.; and Schmidhuber, J. 2021. Unsupervised Object Keypoint Learning using Local Spatial Predictability. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=GJwMHetHc73.

Greff, K.; Kaufman, R. L.; Kabra, R.; Watters, N.; Burgess, C.; Zoran, D.; Matthey, L.; Botvinick, M.; and Lerchner, A. 2019. Multi-object representation learning with iterative variational inference. *arXiv preprint arXiv:1903.00450* .

Greff, K.; Van Steenkiste, S.; and Schmidhuber, J. 2017. Neural expectation maximization. In *Advances in Neural Information Processing Systems*, 6691–6701.

Greydanus, S.; Koul, A.; Dodge, J.; and Fern, A. 2018. Visualizing and understanding atari agents. In *International Conference on Machine Learning*, 1792–1801.

He, S.; Lau, R. W.; Liu, W.; Huang, Z.; and Yang, Q. 2015. Supercnn: A superpixelwise convolutional neural network for salient object detection. *International journal of computer vision* 115(3): 330–344.

Hou, Q.; Cheng, M.-M.; Hu, X.; Borji, A.; Tu, Z.; and Torr, P. H. 2017. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3203–3212.

Itti, L.; Koch, C.; and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence* 20(11): 1254–1259.

Jakab, T.; Gupta, A.; Bilen, H.; and Vedaldi, A. 2018. Unsupervised learning of object landmarks through conditional image generation. In *Advances in neural information processing systems*, 4016–4027.

Judd, T.; Ehinger, K.; Durand, F.; and Torralba, A. 2009. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, 2106–2113. IEEE.

Khetarpal, K.; and Precup, D. 2018. Attend before you act: Leveraging human visual attention for continual learning. *arXiv preprint arXiv:1807.09664* .

Kosiorek, A.; Kim, H.; Teh, Y. W.; and Posner, I. 2018. Sequential attend, infer, repeat: Generative modelling of moving objects. In *Advances in Neural Information Processing Systems*, 8606–8616.

Kulkarni, T. D.; Gupta, A.; Ionescu, C.; Borgeaud, S.; Reynolds, M.; Zisserman, A.; and Mnih, V. 2019. Unsupervised learning of object keypoints for perception and control. In *Advances in Neural Information Processing Systems*, 10723–10733.

Lin, Z.; Wu, Y.-F.; Peri, S. V.; Sun, W.; Singh, G.; Deng, F.; Jiang, J.; and Ahn, S. 2020. Space: Unsupervised object-oriented scene representation via spatial attention and decomposition. *arXiv preprint arXiv:2001.02407* .

Liu, T.; Yuan, Z.; Sun, J.; Wang, J.; Zheng, N.; Tang, X.; and Shum, H.-Y. 2010. Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence* 33(2): 353–367.

Minderer, M.; Sun, C.; Villegas, R.; Cole, F.; Murphy, K. P.; and Lee, H. 2019. Unsupervised learning of object structure and dynamics from videos. In *Advances in Neural Information Processing Systems*, 92–102.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, 1928–1937.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* .

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; Petersen, S.; Beattie, C.; Sadik, A.; Antonoglou, I.; King, H.; Kumaran, D.; Wierstra, D.; Legg, S.; and Hassabis, D. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540): 529–533. URL http://dx.doi.org/10.1038/nature14236.

Mott, A.; Zoran, D.; Chrzanowski, M.; Wierstra, D.; and Rezende, D. J. 2019. Towards interpretable reinforcement learning using attention augmented agents. In *Advances in Neural Information Processing Systems*, 12350–12359.

Rosenfeld, A.; and Thurston, M. 1971. Edge and curve detection for visual scene analysis. *IEEE Transactions on computers* 100(5): 562–569.

Saran, A.; Zhang, R.; Short, E. S.; and Niekum, S. 2020. Efficiently Guiding Imitation Learning Algorithms with Human Gaze. *arXiv preprint arXiv:2002.12500* .

Shi, W.; Wang, Z.; Song, S.; and Huang, G. 2020. Self-Supervised Discovering of Causal Features: Towards Interpretable Reinforcement Learning. *arXiv preprint arXiv:2003.07069* .

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Wang, L.; Lu, H.; Ruan, X.; and Yang, M.-H. 2015. Deep networks for saliency detection via local estimation and global search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3183–3192.

Wu, Y.; Mansimov, E.; Grosse, R. B.; Liao, S.; and Ba, J. 2017. Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation. In *Advances in neural information processing systems*, 5279–5288.

Wyart, V.; and Tallon-Baudry, C. 2009. How ongoing fluctuations in human visual cortex predict perceptual awareness: baseline shift versus decision bias. *Journal of Neuroscience* 29(27): 8715–8725.

Yang, Z.; Bai, S.; Zhang, L.; and Torr, P. H. 2018. Learn to interpret atari agents. *arXiv preprint arXiv:1812.11276* .

Yuezhang, L.; Zhang, R.; and Ballard, D. H. 2018. An initial attempt of combining visual selective attention with deep reinforcement learning. *arXiv preprint arXiv:1811.04407* .

Zhang, R.; Liu, Z.; Guan, L.; Zhang, L.; Hayhoe, M. M.; and Ballard, D. H. 2019. Atari-HEAD: Atari Human Eye-Tracking and Demonstration Dataset. *arXiv preprint arXiv:1903.06754* .

Zhang, Y.; Guo, Y.; Jin, Y.; Luo, Y.; He, Z.; and Lee, H. 2018. Unsupervised discovery of object landmarks as structural representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2694–2703.

Zhao, R.; Ouyang, W.; Li, H.; and Wang, X. 2015. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1265–1274.