Abstract

The study of object representations in computer vision has primarily focused on developing representations that are useful for image classification, object detection, or semantic segmentation as downstream tasks. In this work we aim to learn object representations that are useful for control and reinforcement learning (RL). To this end, we introduce Transporter, a neural network architecture for discovering concise geometric object representations in terms of keypoints or image-space coordinates. Our method learns from raw video frames in a fully unsupervised manner, by transporting learnt image features between video frames using a keypoint bottleneck. The discovered keypoints track objects and object parts across long time-horizons more accurately than recent similar methods. Furthermore, consistent long-term tracking enables two notable results in control domains – (1) using the keypoint co-ordinates and corresponding image features as inputs enables highly sample-efficient reinforcement learning; (2) learning to explore by controlling keypoint locations drastically reduces the search space, enabling deep exploration (leading to states unreachable through random action exploration) without any extrinsic rewards. Code for the model is available at: https://github.com/deepmind/deepmind-research/tree/master/transporter.

1 Introduction

End-to-end learning of feature representations has led to advances in image classification [18], generative modeling of images [7] and agents which outperform expert humans at game play [23, 30]. However, this training procedure induces task-specific representations, especially in the case of reinforcement learning, making it difficult to re-purpose the learned knowledge for future unseen tasks. On the other hand, humans explicitly learn notions of objects, relations, geometry and cardinality in a task-agnostic manner [31] and re-purpose this knowledge to future tasks. There has been extensive research inspired by psychology and cognitive science on explicitly learning object-centric representations from pixels. Both instance and semantic segmentation has been approached using supervised [22, 25] and unsupervised learning [2, 9, 14, 10, 16, 21, 6] methods. However, the representations learned by these methods do not explicitly encode fine-grained locations and orientations of object parts, and thus they have not been extensively used in the control and reinforcement learning literature. We argue that being able to precisely control objects and object parts is at the root of many complex sensory motor behaviours.

In recent work, object keypoint or landmark discovery methods [39, 15] have been proposed to learn representations that precisely represent locations of object parts. These methods predict a set of Cartesian co-ordinates of keypoints denoting the salient locations of objects given image frame(s).
Figure 1: **Transporter.** Our model leverages object motion to discover keypoints by learning to transform a source video frame \((x_s)\) into another target frame \((x_t)\) by *transporting* image features at the discovered object locations. During training, spatial feature maps \(\Phi (x)\) and keypoints co-ordinates \(\Psi (x)\) are predicted for both the frames using a ConvNet and the fully-differentiable *KeyNet* [15] respectively. The keypoint co-ordinates are transformed into Gaussian heatmaps (same spatial dimensions as feature maps) \(H(\Psi (x))\). We perform two operations in the transport phase: (1) the features of the source frame are set to zero at both locations \(H(\Psi (x_s))\) and \(H(\Psi (x_t))\); (2) the features in the source image \(\Phi (x_s)\) at the target positions \(\Psi (x_t)\) are replaced with the features from the target image \(H(\Psi (x_t)) \cdot \Phi (x_t)\). The final refinement ConvNet (which maps from the transported feature map to an image) then has two tasks: (i) to inpaint the missing features at the source position; and (ii) to clean up the image around the target positions. During inference, keypoints can be extracted for a single frame via a feed-forward pass through the *KeyNet* (\(\Psi\)).

However, as we will show, the existing methods struggle to accurately track keypoints under the variability in number, size, and motion of objects present in common RL domains.

We propose **Transporter**, a novel architecture to explicitly discover spatially, temporally and geometrically aligned keypoints given only videos. After training, each keypoint represents and tracks the co-ordinate of an object or object part even as it undergoes deformations (see fig. 1 for illustrations). As we will show, **Transporter** learns more accurate and more consistent keypoints on standard RL domains than existing methods. We will then showcase two ways in which the learned keypoints can be used for control and reinforcement learning. First, we show that using keypoints as inputs to policies instead of RGB observations leads to drastically more data efficient reinforcement learning on Atari games. Second, we show that by learning to control the Cartesian coordinates of the keypoints in the image plane we are able to learn skills or options [32] grounded in pixel observations, which is an important problem in reinforcement learning. We evaluate the learned skills by using them for exploration and show that they lead to much better exploration than primitive actions, especially on sparse reward tasks. Crucially, the learned skills are task-agnostic because they are learned without access to any rewards.

In summary, our key contributions are:

- **Transporter** learns state of the art object keypoints across a variety of commonly used RL environments. Our proposed architecture is robust to varying number, size and motion of objects.

- Using learned keypoints as state input leads to policies that perform better than state-of-the-art model-free and model-based reinforcement learning methods on several Atari environments, while using only up to 100k environment interactions.

- Learning skills to manipulate the most controllable keypoints provides an efficient action space for exploration. We demonstrate drastic reductions in the search complexity for exploring challenging Atari environments. Surprisingly, our action space enables random agents to play several Atari games without rewards and any task-dependent learning.
Figure 2: **Keypoint visualisation.** Visualisations from our and state-of-the-art unsupervised object keypoint discovery methods: Jakab et al. [15] and Zhang et al. [39] on Atari ALE [1] and Manipulator [34] domains. Our method learns more spatially aligned keypoints, e.g. froebite and stack_4 (see section 4.1). Quantitative evaluations are given in fig. 4 and further visualisations in the supplementary material.

2 Related Work

Our work is related to the recently proposed literature on unsupervised object keypoint discovery [39, 15]. Most notably, Jakab and Gupta et al. [15] proposed an encoder-decoder architecture with a differentiable keypoint bottleneck. We reuse their bottleneck architecture but add a crucial new inductive bias – the feature transport mechanism – to constrain the representation to be more spatially aligned compared to all baselines. The approach in Zhang et al. [39] discovers keypoints using single images and requires privileged information about temporal transformations between frames in form of optical flow. This approach also requires multiple loss and regularization terms to converge. In contrast, our approach does not require access to these transformations and learns keypoints with a simple pixel-wise L2 loss function. Other works similarly either require known transformations or output dense correspondences instead of discrete landmarks [35, 29, 33, 37]. Deep generative models with structured bottlenecks have recently seen a lot of advances [3, 19, 36, 38, 12] but they do not explicitly reason about geometry.

Unsupervised learning of object keypoints has not been widely explored in the control literature, with the notable exception of [5]. However, this model uses a full-connected layer for reconstruction and therefore can learn non-spatial latent embeddings similar to a baseline we consider [15]. Moreover, similar to [39] their auto-encoder reconstructs single frames and hence does not learn to factorize geometry. Object-centric representations have also been studied in the context of intrinsic motivation, hierarchical reinforcement learning and exploration. However, existing approaches either require hand-crafted object representations [20] or have not been shown to capture fine-grained representations over long temporal horizons [14].
Figure 3: **Temporal consistency of keypoints.** Our learned keypoints are temporally consistent across hundreds of environment steps, as demonstrated in this classical hard exploration game called Montezuma’s revenge [1]. Additionally, we also predict the most controllable keypoint denoted by the triangular markers, without using any environment rewards. This prediction often corresponds to the avatar in the game and it is consistently tracked across different parts of the state space. See section 4.2.2 for further discussion.

3 Method

In section 3.1 we first detail our model for unsupervised discovery of object keypoints from videos. Next, we describe the application of the learned object keypoints to control for – (1) data-efficient reinforcement learning (section 3.2.1) and (2) learning keypoint based options for efficient exploration (section 3.2.2).

3.1 Feature Transport for learning Object Keypoints

Given an image \( x \), our objective is to extract \( K \) 2-dimensional image locations or keypoints, \( \Psi(\mathbf{x}) \in \mathbb{R}^{K \times 2} \), which correspond to locations of objects or object-parts without any manual labels for locations. We follow the formulation of [15] and assume access to frame pairs \( \mathbf{x}_s, \mathbf{x}_t \) collected from some trajectories such that the frames differ only in objects’ pose / geometry or appearance. The learning objective is to reconstruct the second frame \( \mathbf{x}_t \) from the first \( \mathbf{x}_s \). This is achieved by computing ConvNet (CNN) feature maps \( \Phi(\mathbf{x}_s), \Phi(\mathbf{x}_t) \in \mathbb{R}^{H^' \times W' \times D} \) and extracting \( K \) 2D locations \( \Psi(\mathbf{x}_s), \Psi(\mathbf{x}_t) \in \mathbb{R}^{K \times 2} \) by marginalising the keypoint-detector feature-maps along the image dimensions (as proposed in [15]). A transported feature map \( \hat{\Phi}(\mathbf{x}_s, \mathbf{x}_t) \) is generated by suppressing both sets of keypoint locations in \( \Phi(\mathbf{x}_s) \) and compositing in the featuremaps around the keypoints from \( \mathbf{x}_t \):

\[
\hat{\Phi}(\mathbf{x}_s, \mathbf{x}_t) = (1 - H_{\Psi(\mathbf{x}_s)}) \cdot (1 - H_{\Psi(\mathbf{x}_t)}) \cdot \Phi(\mathbf{x}_s) + H_{\Psi(\mathbf{x}_t)} \cdot \Phi(\mathbf{x}_t)
\]

where \( H_{\Psi} \) is a heatmap image containing fixed-variance isotropic Gaussians around each of the \( K \) points specified by \( \Psi \). A final CNN with small-receptive field refines the transported reconstruction \( \hat{\Phi}(\mathbf{x}_s, \mathbf{x}_t) \) to regress the target frame \( \hat{x}_t \). We use pixel-wise squared-\( \ell_2 \) reconstruction error \( ||x_t - \hat{x}_t||_2^2 \) for end-to-end learning. The keypoint network \( \Psi \) learns to track moving entities between frames to enable successful reconstruction.

In words, (i) the features in the source image \( \Phi(\mathbf{x}_s) \) at the target positions \( \Psi(\mathbf{x}_t) \) are replaced with the features from the target image \( H_{\Psi(\mathbf{x}_t)} \cdot \Phi(\mathbf{x}_t) \) — this is the transportation; and (ii) the features at the source position \( \Psi(\mathbf{x}_s) \) are set to zero. The refine net (which maps from the transported feature map to an image) then has two tasks: (i) to inpaint the missing features at the source position; and (ii) to clean up the image around the target positions. Refer to fig. 1 for a concise description of our method.

Note, unlike [15] who regress the target frame from stacked target keypoint heatmaps \( H_{\Psi(\mathbf{x}_t)} \) and source image features \( \Phi(\mathbf{x}_s) \), we enforce explicit spatial transport for stronger correlation with image locations leading to more robust long-term tracking (section 4.1).
3.2 Object Keypoints for Control

Given learned keypoints, we want to use them within the context of control and exploration. Consider a Markov Decision Process (MDP) with visual observations \( x \in \mathcal{X} \) as states, actions \( a \in \mathcal{A} \) and a transition function \( T^1 : (x_t, a) \rightarrow (x_{t+1}, r_t) \). We use a Transporter model \( \{\Psi, \Phi\} \) which is pre-trained in an unsupervised fashion without extrinsic rewards. The agents output actions \( a_t \) and receive rewards \( r_t \) as normal.

3.2.1 Data-efficient reinforcement learning

Our first hypothesis is that task-agnostic learning of object keypoints can enable fast learning of goal-directed policies. This is because once we learn keypoints, the control policy can be much simpler and does not have to relearn all visual features using temporal difference learning. In order to test this hypothesis, we use a variant of the neural fitted Q-learning framework [27] with learned keypoints as input and a recurrent neural network Q function to output behaviors. The agent observes \( \Psi(x_t) \) and \( \Phi(x_t) \) only at the corresponding masked keypoint locations. We encode one hot vectors to denote positions of keypoints and their corresponding (keypoint mask averaged) feature vectors at that location. To get these averaged features, the (Gaussian) heat-map (\( \mathcal{H}_q \)) is multiplied with the feature tensor (\( \Phi \)) and then spatially averaged to obtained this feature vector. The Transporter is trained by collecting data using a random policy and without any reward functions (see supplementary material for details). The Transporter network weights are fixed during behavior learning given environment rewards.

3.2.2 Keypoint-based options for efficient exploration

Our second hypothesis is that learned keypoints can enable significantly better task-independent exploration. Typically, raw actions are randomly sampled to bootstrap goal-directed policy learning. This exploration strategy is notoriously inefficient. We leverage the Transporter representation to learn a new action space. The actions are now skills grounded in the control of co-ordinate values of each keypoint. This idea has been explored in the reinforcement learning community [20, 14] but it has been hard to learn spatial features with long temporal consistency. Here we show that Transporter is particularly amenable to this task. We show that randomly exploring in this space leads to significantly more rewards compared to raw actions. Our learned action space is agnostic to the control algorithm and hence other exploration algorithms [24, 4, 26] can also benefit from using it.

To do this, we define \( K \times 4 \) intrinsic reward functions using the keypoint locations, similar to [14]. Each reward function corresponds to how much each keypoint moves in the 4 cardinal directions (up, down, left, right) between consecutive observations. We learn a set of \( K \times 4 \) Q functions \( \{Q_{i,j}\} \in \{1, ... , K\} \times \{1, 2, 3, 4\} \), to maximise each of the following reward functions:

\[
\begin{align*}
    r_{i,1} &= \Psi_x^i(x_{t+1}) - \Psi_x^i(x_t), \\
    r_{i,2} &= \Psi_x^i(x_{t+1}) - \Psi_y^i(x_t), \\
    r_{i,3} &= \Psi_y^i(x_{t+1}) - \Psi_x^i(x_t), \\
    r_{i,4} &= \Psi_y^i(x_{t+1}) - \Psi_y^i(x_t).
\end{align*}
\]

These functions correspond to increasing/decreasing the \( x \) and \( y \) coordinates respectively. The Q functions are trained using n-step Q(\( \lambda \)).

During training, we randomly sample a particular Q function to act with and commit to this choice for \( T \) timesteps before resampling. All Q functions are trained using experiences generated from all policies via a shared replay buffer. Randomly exploring in this Q space can already reduce the search space as compared to raw actions. We further reduce this search space by learning to predict the most controllable keypoint. For instance, in many Atari games there is an avatar that is directly controllable on the screen. We infer this abstraction via a fixed controllability policy to select the single "most controllable" keypoint: \( \pi_{Q_{\text{top}}}(s) = \arg\max_{i, j} \frac{1}{4} \sum_{j=1}^{4} \max_{a} Q_{i,j}(s;a) - \min_{a} Q_{i,j}(s;a) \).

This procedure picks keypoints where one action leads to more prospective change in all spatial directions than all other keypoints. Given this keypoint, we randomly sample a \( j \) with a fixed temporal commitment \( T \) as the random exploration policy. Consider a sequence of 100 actions with 18 choices before receiving rewards, which is typically the case in hard exploration Atari games (e.g. montezuma’s revenge). A random action agent would need to search in the space of \( 18^{100} \) raw actions. However, observing 5 keypoints and \( T = 20 \) only has \( (5 \times 4)^{100/20} \), giving a search space reduction of \( 10^{100} \). The search space reduces further when we explore with the most controllable keypoints. Since our learned action space is agnostic to the control mechanism, we evaluate them by randomly searching in this space versus raw actions. We measure extrinsically defined game score as the metric to evaluate the effectiveness of both search procedures.
Figure 4: Long-term tracking evaluation. We compare long-term tracking ability of our keypoint detector against Jakab et al. [15] and Zhang et al. [39] (visualisations in fig. 2 and supplementary material). We report precision and recall for trajectories of varying lengths (lengths $= 1 – 200$ frames; each frame corresponds to 4 action repeats) against ground-truth keypoints on Atari ALE [1] and Manipulator [34] domains. Our method significantly outperforms the baselines on all games (100% on pong), except for ms_pacman where we perform similarly especially for long trajectories (length $= 200$). See section 4.1 for further discussion.

4 Experiments

In section 4.1 we first evaluate the long-term tracking ability of our object keypoint detector. Next, in section 4.2 we evaluate the application of the keypoint detector on two control tasks — comparison against state-of-the-art model-based and model-free methods for data-efficient learning on Atari ALE games [1] in section 4.2.1, and in section 4.2.2 examine efficient exploration by learning to control the discovered keypoints; we demonstrate reaching states otherwise unreachable through random explorations on raw-actions, and also recover the agent self as the most-controllable keypoint. For implementation details, please refer to the supplementary material.

Datasets. We evaluate our method on Atari ALE [1] and Manipulator [34] domains. We chose representative levels with large variations in the type and number of objects. (1) For evaluating long-term tracking of object keypoints section 4.1 we use — pong, frostbite, ms_pacman, and stack_4 (manipulator with blocks). (2) For data-efficient reinforcement learning (section 4.2.1) we train on diverse data collected using random exploration on the Atari games indicated in fig. 6. (3) For keypoints based efficient-exploration (section 4.2.2) we evaluate on one of the most difficult exploration game — montezuma revenge, along with ms_pacman and seaquest.

A random policy executes actions and we collect a trajectory of images before the environment resets; details for data generation are presented in the supplementary material. We sample the source and target frames $x_s, x_t$ randomly within a temporal offset of 1 to 20 frames, corresponding to small or significant changes in the configuration between these two frames respectively. For Atari ground-truth object locations are extracted from the emulated RAM using hand crafted per-game rules and for Manipulator it is extracted from the simulator geoms. The number of keypoints $K$ is set to the maximum number of moving entities in each environment.

4.1 Evaluating Object Keypoint Predictions

Baselines. We compare our method against state-of-the-art methods for unsupervised discovery of object landmarks — (1) Jakab et al. [15] and (2) Zhang et al. [39]. For (1) we use exactly the same architecture for $\Phi$ and $\Psi$ as ours; for (2) we use the implementation released online by the authors where the image-size is set to $80 \times 80$ pixels. We train all the methods for $10^6$ optimization steps and pick the best model checkpoint based on a validation set.
ε-greedy policy
Transporter pre-trained offline
CNN
LSTM
Recurrent agent trained with a variant of Neural Fitted Q-iteration
Recurrent state
Observation
CNN
PNet

Figure 5: Agent architecture for data-efficient reinforcement learning. Transporter is trained off-line with data collected using a random policy. A recurrent variant of the neural-fitted Q-learning algorithm [27] rapidly learns control policies using keypoint co-ordinates and features at the corresponding locations given game rewards.

Game | KeyQN (ours) | SimPLe | Rainbow | PPO (100k) | Human | Random
--- | --- | --- | --- | --- | --- | ---
breakout | 19.3 (4.5) | 12.7 (3.8) | 3.3 (0.1) | 5.9 (3.3) | 31.8 | 1.7
frostbite | 388.3 (142.1) | 254.7 (4.9) | 140.1 (2.7) | 174.0 (40.7) | 4334.7 | 65.2
ms_pacman | 999.4 (145.4) | 762.8 (331.5) | 364.3 (20.4) | 496.0 (379.8) | 15693.0 | 307.3
pong | 10.8 (5.7) | 5.2 (9.7) | -19.5 (0.2) | -20.5 (0.6) | 9.3 | -20.7
seaquest | 236.7 (22.2) | 370.9 (128.2) | 206.3 (17.1) | 370.0 (103.3) | 20182.0 | -20.7

Figure 6: Atari Mean Scores. Mean scores (and std-dev in parentheses) obtained by our method (three random seeds) in comparison with Rainbow [11], SimPLe [17] and PPO [28] trained on 100K steps (400K frames). See section 4.2.1 for details. Numbers (except for KeyQN) taken from [17].

Metrics. We measure the precision and recall of the detected keypoint trajectories, varying their lengths from 1 to 200 frames (200 frames ≈ 13 seconds @ 15-fps with action-repeat of 4) to evaluate long-term consistency of the keypoint detections crucial for control. The average Euclidean distance between each detected and ground-truth trajectory is computed. The time-steps where a ground-truth object is absent are ignored in the distance computation. Distances above a threshold (ϵ) are excluded as potential matches. One-to-one assignments between the trajectories are then computed using min-cost linear sum assignment, and the matches are used for reporting precision and recall.

Results. Figure 2 visualises the detections while fig. 4 presents precision and recall for varying trajectory lengths. Transporter consistently tracks the salient object keypoints over long time horizons and outperforms the baseline methods on all environments, with the notable exception of [15] on pacman where our method is slightly worse but achieves similar performance for long-trajectories.

4.2 Using Keypoints for Control

4.2.1 Data-efficient Reinforcement Learning on Atari

We demonstrate that using the learned keypoints and corresponding features within a reinforcement learning context can lead to data-efficient learning in Atari games. Following [17], we trained our Keypoint Q-Network (KeyQN) architecture for 100,000 interactions, which corresponds to 400,000 frames. As shown in fig. 6, our approach is better than the state-of-the-art model-based SimPLe architecture [17] and model-free Rainbow architecture [11] on four out of five games. Applying this approach to all Atari games will require training Transporter inside the reinforcement learning loop because pre-training keypoints on data from a random policy is insufficient for games where new objects or screens can appear. However, these experiments provide evidence that the right visual abstractions and simple control algorithms can produce highly data efficient reinforcement learning algorithms.

1The threshold value (ϵ) for evaluation is set to the average ground-truth spatial extent of entities for each environment.
We demonstrate that it is possible to learn stable object keypoints across thousands of environment steps, without having access to task-specific reward functions. Therefore, object keypoints could provide a flexible and re-purposable representation for efficient control and reinforcement learning. Scaling keypoints to work reliably on richer datasets and environments is an important future area of research. Further, tracking objects over long temporal sequences can enable learning object dynamics.
and affordances which could be used to inform learning policies. A limitation of our model is that we do not currently handle moving backgrounds. Recent work [8] that explicitly reasons about camera / ego motion could be integrated to globally transport features between source and target frames. In summary, our experiments provide clear evidence that it is possible to learn visual abstractions and use simple algorithms to produce highly data efficient control policies and exploration procedures.

Acknowledgements. We thank Loic Matthey and Relja Arandjelović for valuable discussions and comments.

References


