1 Main Motivations

1.1 Learning from images

The benefits of learning from images listed are

1. **Natural representation for scenes with multiple objects**. If we want to encode the joint state of a variable number of objects into a fixed-size vector, we are forced to choose one representation out of a combinatorially large number of valid representation. Representing the states of the world with low-level pixel observation also removes the need to incorporate a pre-processing step to extract the state of the objects.
2. **Removing the need for a pre-processing system to design reward function**. If given two images, representing the current state of the world and the desired state, we can learn a meaningful reward function that can be used to train a policy, that removes the need to manually design a reward function.
3. **Remove the need for access to GT state of the world**

1.2 Self-supervised goal discovery and practice

1. If the goal the agent needs to accomplish at test time is unknown during training time, the agent should learn to accomplish a wide variety of goals during training time.
2. For this to be very scalable, this would also require a scalable method of goal generation.

1.3 Sample efficient learning from images with structured representation

1. What is the right representation of images to enable sample efficient learning?

2 Main Components

2.1 A goal-condition policies

1. In standard RL, the policy is trained to accomplish a single goal, expressed implicitly through the reward function. However, if we want the trained agent to be able to perform a wide variety of goal, then we can reformulate the problem as a goal-conditioned RL problem.
2. A goal-condition Q function $Q(s, a, g)$ computes the expected return starting in state s, a when the policy needs to accomplish a goal g.
3. The Bellman error now includes the goal g and can be used to train a parameterized Q_w function.

$$\mathcal{E}(w) = \frac{1}{2}\left\| Q_w(s,a,g) - \left(r + \gamma \max_{a'} Q_{\overline{w}}\left(s', a', g\right)\right)\right\|^2$$

2.2 VAE

The VAE has many different purposes:

1. provide structured representation of pixel input. The policy receives the encoded image instead of the raw pixel. They argue that this leads to more sample efficient learning.
2. relabelling the transitions collected so far with new goals to generate synthetic training data to train the value function. There are two strategies for relabelling the transition: sampling from the VAE prior and the Future strategy in Hindsight Experience Replay. They demonstrated that a mix of both works best. For sampling from the VAE prior, they argue that the distribution of the latents do not match the prior exactly. They thus use a fitted prior: the distribution of the latent encoding of the VAE training data is fitted with a diagonal Gaussian.

3. reward function: given two images, the distance in the latent space of the encoding of the two images are used as the reward signal to train the policy. They argue that this has a nice probabilistic interpretation: minimizing this distance is equivalent to reaching states that maximize the probability of the latent goal.
4. custom goal generation: to generate goal for the robot to practice, the goal states are sampled from the VAE prior. It is nice they chose to do this since this is a scalable and simple method. There are probably better ways to do it but this is not what the paper is about.

3 Questions

1. Why does a mix of sampling from the VAE prior and the Future strategy of relabelling transition work best?
2. The goal in this paper is only a picture. This can be ill-defined for more complex goals (such as manufacturing an object where the internal components are not visually visible from the outside).
3. The paper argues that the distance in the latent space between the latent encoding of the current and target observation provide more well-shaped reward function. However, they never explain what well-shaped means in their context or demonstrate why this distance is well-shaped.
4. What does instrumentation means in this sentence "In the absence of domain knowledge and instrumentation"?
5. It is interesting that they chose to use TD3 instead of SAC. Any particular reason why?
6. In the experiment section, pick and place is the most complex task. It is unclear from Figure 2 if the task actually requires the robot to pick up an object, can it simply push it?
7. In Figure 3, what is the scale of the y-axis? It is unclear if the performance is good without knowing the answer to this question.
8. When using the distance in the latent space as the reward function, they did not use the distance with the probabilistic interpretation, instead opt for a simpler distance. They mention that this works better, but it is unclear why.
9. In section 5.1, for the visual pusher task on real robot, they mention that they do not have the grouth truth reward function to evaluate the performance of the trained policy, they thus use the distance in the VAE latent as a measure of performance. Is this measure highly correlated with the true distance? This distance is changing during training as well, is this a confounding factor?