Part I

## Unsupervised Visual Representation Learning by Context Prediction

1. This paper is trying to have a CNN be able to determine relative positions of two patches of a larger image. The data used for training this CNN were unlabeled and the training was done in an unsupervised fashion. Wants to show that if the CNN does well, it has gained an understanding of scenes and objects and has a good visual representation.
2. They take a sample of random pairs that are in one of the eight spatial configurations. The present those to a machine learner with no prior information of the original position for the pairs. The algorithm will then make a prediction in of where it things the pairs are relative to each other.

## The Curious Robot: Learning Visual Representations via Physical Interactions

1. The paper is trying to prove that using a more biological approach to Visual Representations (i.e. similar to how animals and humans approach it) is a more effective way than a passive observation approach. They want to see if having datapoints from not only visual observation but also grabbing, pushing, and poking with a haptic sensor will improve the system's Visual Representation.
2. The team built a robot that observes an object from a static point. That robot has a moveable arm with different ends, a grabbing head, a pushing head, and a poking head. The system will then utilize datapoints gathered from all four of these interactions in order to better broadly classify objects.

## Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

1. They want to create a system that has learned the mapping between input and output images using a set of aligned image pairs. They need to do this in such a way that there is no paired training data necessary as that is often non-existent. The system would be able to translate some input image into the form or style of another image, i.e. photo -> painting and vis-versa.
2. They utilized an algorithm that can learn to translate between domains without paired input-output examples. They trained the mapping such that the input and output would be deemed indistinguishable by an adversary trained to distinguish them apart in order to find close enough pairs.

Part II

## Generative Adversarial Nets

1. The paper is trying to propose a new framework for estimating generative networks. The way it hopes to do this is through the use of adversarial networks. Other work on deep generative networks had issues as they would require numerous approximations to the likelihood gradient. Other work has been done with variational autoencoders (VAEs) which are similar to GANs but the discriminative model in a VAE is a recognition model that performs approximate interference. GANs require differentiation through visible units while VAEs require differentiation through hidden units.
2. This proposed framework of utilizing adversarial networks would work as follows: two models would be trained simultaneously. The first model would be a generative model which is trained to generate things similar in form to the training data. The second model would be a discriminator model with the goal of predicting if a given sample is part of the training data utilized for the generative model or if it was generated by the first model. Both models would follow a two-player minimax game where each "compete" against each other in order to either generate data nearly indistinguishable from training data (at least to the discriminator model) or where the discriminative model becomes too advances for the generative model to be successful, thus making it a cycle of improvement for future iterations of training.
3. The adversarial nets were trained on a range of datasets including MNIST, the Toronto Face Database (TFD), and CIFAR-10. The generative model utilized rectifier linear activations and sigmoid activations. The discriminator model utilized maxout activations and dropout was applied. The use of noise was the input only to the bottommost layer of the generator network. The results of the experimented method of estimating the likelihood has somewhat high variance and does not perform well in high dimensionality, but it is the best method available at the time of the paper being written (2014).
4. All of the listed advantages were primarily computational based. Some examples being Markov chains are never needed, only backprop is used to obtain gradients, no inference is needed during learning, and a wide variety of functions can be incorporated into the model.
5. A drawback to the approach is that the two models, generative and discriminative, must be synchronized well during the training as if the generative model is trained too much, "the Helvetia scenario" may occur where it collapses too many values.
6. I do wonder if utilizing increasing the "game" from a two-player minimax game to a, say 4 player, 2 on each team, would provide benefits to this type of model. I could see the use of two separate models working to make data to fool the discriminators and using the result to both become better.

## Image-to-Image Translation with Conditional Adversarial Networks

1. The paper is trying to investigate and apply conditional adversarial networks as a general-purpose solution for image-to-image translation problems. Prior approaches have used CNNs which work well but require a lot of manual intervention and attention to detail in that intervention as it is possible to mess up the network after the intervention. cGANs (conditional Generative Adversarial Networks) provide a more hands-off solution to the issue CNNs provide. Current work with Image-to-image translation treat the output space as "unstructured" in the

sense that each output pixel is considered conditionally independent from all others given the input image. cGANs instead learn a structured loss, which penalize the joint configuration of the output. Not the first people to utilize cGANs as others had used them to produce normal maps, future frame prediction, product photo generation, and image generation from sparse annotations. Some other papers have utilized GANs for image-to-image mappings but the GANs were applied unconditionally. The approach of the paper differs in that nothing is application specific making it simpler than the other setups.

2. As stated, the paper utilizes cGANs for their general use applications. Within the cGANs, the generative network utilizes a “U-Net” architecture with a skip element to it and the discriminator utilizes a developed “PatchGAN” architecture which checks NxN patches within the image to determine if they are real or fake. Both generator and discriminator will use modules of the form convolution-BatchNorm-ReLu.
3. They want to test the generality of cGANs with photo generation and semantic segmentation. Their results were shown in a paper and on a GitHub link which shows their results compared to other methods. Link. They had two method of evaluating the results. First they would run “real vs fake” perceptual studies using the Amazon Mechanical Turk. Testers were presented with a series of trials that pitted a “real” image against a “fake” image and then were asked to determine which one was real and which one was fake. They also measured if their results were realistic enough that off-the-shelf recognition systems could recognize the generated objects within the output. The idea being, that if off-the-shelf classifiers, which are trained on real images, can detect objects within the generated images, the generated images are realistic. Compared to other methods, their results show that their method was significantly better. Compared to encoder-decoder method against their U-Net skip method, they found that encoder-decoder was unable to learn to generate realistic images in their experiments whereas the U-Net skip method was able to be generalized and achieves the superior results. Results in this paper suggest that conditional adversarial networks are a promising approach for many image-to-image translation tasks
4. As stated in the previous answer, the U-Net utilized for generation appear not to be specific to cGANs, or in other words, is able to be generalized whereas other methods were not.
5. They state that while cGANs achieve some success, that they are far from the best available method for solving some problems they tacked, specifically photo->labels. A simpler method of using a L1 regression gets better scores making it more sufficient.
6. If I had to expand some part of this paper, I would see if it could work for text-in-image (i.e. photo of text).