1 Key Ideas

1. The goal of the paper is to generate landmarks of objects without manual supervision.
2. Source image to target image generation is hard due to ambiguity in predicting motion of objects.
3. Source image and encoded target image to target image generation is much easier.

1. The argument is: landmark encodes geometry of the object
2. Thus, they need a training procedure that distills the target image into the geometry of the object in the target image in an unsupervised way.
3. They pick generating the target image from the source image and encoded target image as the training procedure.
4. To generate the target image, we need: appearance + geometry.
5. The appearance is taken from the source image.
6. The geometry is supposed to be taken from the encoding of the target image.
7. To encourage the encoding of the target image to distill the geometry of the target image, a heatmap bottleneck is enforced on the encoding.
8. When the source image and encoded target image are used as input into a network which is trained to minimize construction loss, the encoded target image extracts keypoint structure from the target image, which can be interpreted as landmarks.

2 Other interesting bits in the paper

1. Gaussian heatmap and encoding.
2. Perceptual loss. They argue that they do not use GAN because GAN is good for aligning distribution, but what they are trying to do is to predict only a single target image. The perceptual loss compares the activation at various layers of a NN when the input is the target image and the predicted target image.
3. Finetuning with a few labelled examples by learning a linear regressor between the predicted landmarks and the grouth truth landmarks. If this finetuning stage helps, that means that the geometrical relationship between the predicted landmarks is correct, but their positions within the target maybe off by a linear transformation.