Paper

- **Title:** ST-GAN: Spatial Transformer Generative Adversarial Networks for Image Compositing
- **Authors:** Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, Simon Lucey
- **arXiv link:** https://arxiv.org/abs/1803.01837

TL;DR

The paper presents a technique to find geometric corrections to a foreground object such that it appears natural when composited into a background image using a Spatial Transformer Network which is trained on adversarial loss. It is used to place indoor furniture in rooms and placing glasses on real facial portraits.

Problem statement

- Given a background image $\mathcal{I}_{bg}$ and a foreground image $\mathcal{I}_{fg}$, find a composition of them that looks realistic.
- Note that the background has to remain the same and the transformation has to be applied to the foreground only.
- Appearance differences are not taken into account here i.e. it will not affect the lighting, white balance, shading, contrast and such things because Poisson Blending solves such problems.

Proposed method

- Predicting large displacement warp parameters from image pixels is extremely challenging, so they predict small geometric transformations in an iterative fashion.
- At the i^{th} iteration, given the input image $\mathcal{I}$ and the previous warp state p_{i-1} and a warp update Δp_i, the new warp parameter is:

$$\Delta p_i = \mathcal{G}_i(\mathcal{I}_{FG}(p_{i-1}), \mathcal{I}_{BG})$$
$$p_i = p_{i-1} \circ \Delta p_i$$

Sequential Adversarial Training

- STNs are embedded into a WGAN (Arjovsky et al 2017), where the iterative STN is a generator and a the discriminator is a Fully Convolutional Network.
- $\mathcal{G}$ generates a set of low-dimensional warp parameter updates.
- $\mathcal{D}$ gets as input the warped foreground image composited with the background image.
- Training is also iterative. They start by training a single $\mathcal{G}_1$ and each subsequent new generator $\mathcal{G}_i$ is added and trained by fixing the weights of all previous generators $\{\mathcal{G}_j\}_{j=1\ldots i-1}$
- The WGAN objective is

$$\underset{\mathcal{G}_i}{min}\ \underset{\mathcal{D}}{max} \underset{x\sim P_{fake}, p_i\sim P_{p_i|p_{i-1}}}{\mathbb{E}}[\mathcal{D}(x(p_i))] - \underset{y\sim P_{real}}{\mathbb{E}}[\mathcal{D}(y)]$$

- The loss for Generator $\mathcal{G}$ and Discriminator $\mathcal{D}$ are:

$$\mathcal{L}_{\mathcal{G}} = -\mathbb{E}_{x,p_i}[\mathcal{D}(x(p_i))] + \lambda_{update} \cdot \mathcal{L}_{update}$$
$$\mathcal{L}_{\mathcal{G}} = \mathbb{E}_{x,p_i}[\mathcal{D}(x(p_i))] - \mathbb{E}_y[\mathcal{D}(y)] + \lambda_{grad} \cdot \mathcal{L}_{grad}$$

  Here, λ_{update} is the penalty weight for the warp update Δp_i to ensure that warp updates are small. and λ_{grad} is the penalty weight for the gradient of Discriminator as suggested in Gulrajani et. al 2017