Overview

**Goal:**
Learning disentangled representations that allow us to control only a specific factor in the image for unsupervised image manipulation.

**Proposal:**
We propose a novel generative model that learns to disentangle the appearance, the x-axis, the y-axis factors, assemble the factors, and then synthesize images.

---

**Method:** Disentangle-Assemble-Synthesize

- **Latent-specific Network**: Given the appearance, the x-axis, and the y-axis noises, each network outputs a vector of the corresponding factor.
- **Assemble module**: The module assembles these vectors into a structurally constrained map by tiling and concatenating operations.
- **Upscale network**: It performs two-stage upscaling until the output size. In the first step, it upscales the feature map while maintaining the structural property. Then, it upscales by a vanilla deconvolution that ignores the property.

**Key idea: Structural Constraint**
Packing the appearance and location in each position of the feature map

→ Since these vectors have the same values for each axis, the generator must assign the corresponding properties to these vectors and assemble them.

---

**Visual Results on Translated MNIST**

![Interpolation of appearance (app.) and location (lo.)](image1)

![Comparison between initial frame and interpolated frame](image2)