Spatial Transformer
Generative Adversarial Networks
for
Image Compositing

Burak Kahveci
Outline

- Spatial Transformer Networks
- Generative Adversarial Networks
- Image Compositing
- Solution / Approach
- Results
Spatial Transformer Networks

- Jaderberg et al., 2015.
- Can be inserted into existing convolutional architectures.
- Localisation network + grid generator + sampler
Spatial Transformer Networks

- Localisation network regresses the transformation parameters $\theta = f_{loc}(U)$.
- The regular spatial grid $G$ over $V$ is transformed to the sampling grid $T_\theta(G)$, which is applied to input feature map $U$. 
Spatial Transformer Networks
Generative Adversarial Networks

- Goodfellow et al., 2014.
- Realistic image generation
- Generative and Discriminative networks
- Better than direct image generation
Image Compositing

- Foreground object usually comes from a different scene than the background.

- Appearance differences; Poisson blending or more recent deep learning approaches.

- Geometric differences!
Approach

\[ I_{\text{comp}} = I_{\text{FG}} \odot M_{\text{FG}} + I_{\text{BG}} \odot (1 - M_{\text{FG}}) \]
\[ = I_{\text{FG}} \oplus I_{\text{BG}}. \]

- Realistic geometric correction for image compositing
- Background image: \( I_{\text{BG}} \)
- Foreground object \( I_{\text{FG}} \)
- A corresponding mask \( M_{\text{FG}}. \)
Iterative Geometric Correction

- G, from GAN predicts a correcting update $\Delta p_1$.
  
  
  $\Delta p_i = G_i(I_{FG}(p_{i-1}), I_{BG})$
  
  
  $p_i = p_{i-1} \circ \Delta p_i$,

- STNs, integrated into GANs, preserve the original images from loss of information due to multiple warping operation.
  1. G generates a set of low-dimensional warp parameter updates instead of images (the whole set of pixel values).
  2. D gets as input the warped foreground image composited with the background.
Full Pipeline
Objective Function

- Minimax objective

\[
\min_{g_i} \max_{D \in D} \mathbb{E}_{x \sim P_{\text{fake}}} [D(x(p_i))] - \mathbb{E}_{y \sim P_{\text{real}}} [D(y)]
\]

- Loss functions of discriminative and generative networks

\[
L_D = \mathbb{E}_{x, p_i} [D(x(p_i))] - \mathbb{E}_y [D(y)] + \lambda_{\text{grad}} \cdot L_{\text{grad}}
\]

\[
L_{g_i} = -\mathbb{E}_{x, p_i} [D(x(p_i))] + \lambda_{\text{update}} \cdot L_{\text{update}},
\]
Model Architecture

- **C(k)**: 2D convolutional layer with k filters of size 4 x 4 and stride 2
- **L(k)**: Fully-connected layer with k output nodes.
- **The input of Gi**: 7 x 120 x 160, RGBA for foreground and RGB for background.
- **G**: C(32)-C(64)-C(128)-C(256)- C(512)-L(256)-L
- **Input to the discriminator D**: The composite image with dimensions 3 x 120 x 160 channels (RGB).
- **D**: C(32)-C(64)-C(128)-C(256)- C(512)-C(1).
- ReLU for G and LeakyReLU with slope 0.2 for D.
- Parameterized warp can be applied to full-resolution images at test.
Experiments - 3D Cubes

- Simple start
- Right perspective, translational offset from GT
Experiments - Indoor objects

- Remove occluding objects
- Remove the object itself for training
- Perturb the 6-DoF camera pose
Experiments - Indoor objects
Experiments
- Glasses
Conclusion

- A class of methods to model geometric realism.

- Will open up new revenues to the research community to continue to explore in this direction

- Suffers at rare examples
Questions & Discussion