Deep Scene Interpolation

Brendan Corcoran, John Clow, Matt Volk

Introduction

- Novel view synthesis
- Automate a rotation/translation of an urban scene
- Potential applications
  - Google Street View
  - Video compression
  - Photo editing

Project Objective

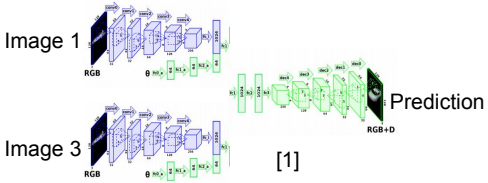
- Input: 1st and 3rd frame (perspectives on a scene)
- Output: a middle frame that would have been captured half way between the 1st and 3rd frame

Dataset

- About 40,000 image triples, generated from the KITTI Vision Benchmark dataset
- Two classes of generated data:
  - **Simple**: Images separated by 1 frame
  - **Hard**: Images separated by 3-7 frames
- Cropped to 224x224

Models

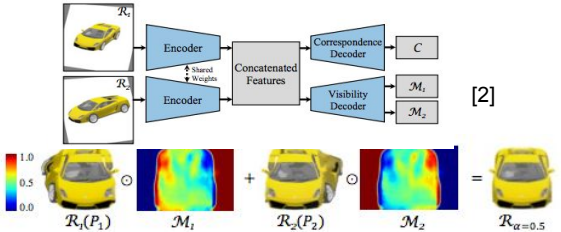
Direct Pixel Generation (Baseline)



[1]

- Encoder/decoder network (5 layers each)
- Encoder (shared weights) captures high-level representation of input images
- Representations concatenated and input to decoder, which directly predicts the output (intermediate image)

Appearance Flow



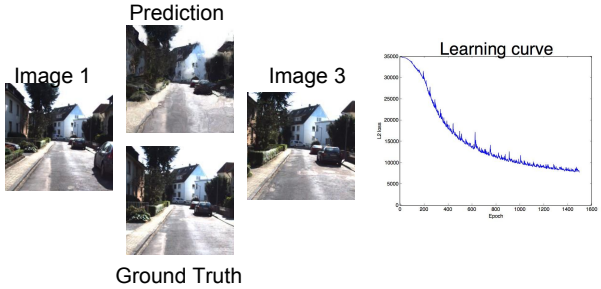
[2]

- Rather than learn what pixels to generate, this approach learns a transformation to apply to the original images
- One decoder recovers dense correspondences, which are used to create a "best guess" for each input image
- The other decoder recovers masks which are used to blend the 2 guesses into a final prediction

Results

Direct Pixel Generation (Baseline)

Appearance Flow



References

[1] M. Tatarchenko, A. Dosovitskiy: "Multi-view 3D Models from Single Images with a Convolutional Network", 2015;
[2] D. Ji, J. Kwon, M. McFarland, S. Savarese: "Deep View Morphing", 2017