Novel View Synthesis by Geometric Transformation and Image Completion Neural Network

Miao Zhang, Xiaobai Ma (CS 236) {miaoz2, maxiaoba}@stanford.edu

Overview

The objective of synthesizing novel views is to build a network that can generate images of an object viewed from different angles, given only a single view of that object. Its has practical applications in the field of robotics, virtual reality and computer graphics. This task is generally challenging due to the ambiguity of 3D object shape given only one single view, especially inferring the unseen parts of the object.

We inherited the idea of synthesizing novel views by combining appearance flow network (DOAFN) and completion network (TVSN) presented in [2], and added generated object contour as an additional input to the completion network, using various ways of incorporating this new information, as shown in Fig.1.

Dataset

We used the ShapeNet 3D objects dataset [3] car category as in [1] and [2], with an open source image render engine [4] to render 2D views of the 3D object from different angle.

SHAPENET

The render engine also can generate the surface normal and object coordinates which are later used to generate visibility maps, which is an intermediate value in the network generating process. The dataset contains 7900 different cars, and each car object has 18×3 images sweeping through azimuth angles from 0 to 340° (20° interval), and elevation angles from 0 to 20° (10° interval). The figure in the banner gives an example of rendered images of a single vehicle in all 18 azimuth angle views with 0° elevation angle.

Model

The full generation network is composed of two parts. The first part is the appearance flow network noted as dis-occlusion-aware appearance flow network (DOAFN). The second part is the completion network noted as transformation-grounded view synthesis network (TVSN). The two networks are trained sequentially.

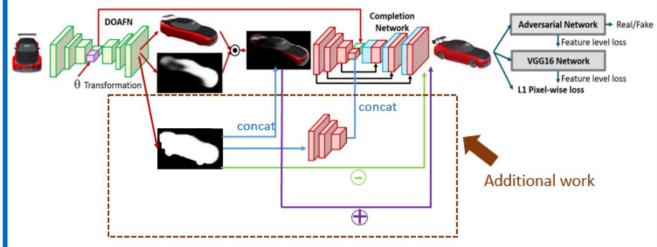


Fig. 1 Network architecture

Training loss:

$$L_{doafn} = L_1(I_t, I_{afn}) + BCE(M_{vis}, M_{vis,groundtruth}) + BCE(M_c, M_{c,groundtruth})$$

$$L_g = -logD(G(I_s)) + \alpha L_2(F_D(G(I_s)), F_D(I_t)) + \beta L_2(F_{vgg}(G(I_s)), F_{vgg}(I_t)) + \gamma L_1(G(I_s), I_t) + \lambda L_{TV}(G(I_s))$$

$$L_D = -log(I_s) - log(1 - D(G(I_s)))$$

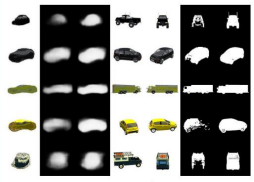
Discussion/Future Work

Our model shows comparable performance to the baseline model presented in the paper, but no significant improvement either; however, applying the generated contour mask to the final network output does show a more clearly defined object outline. Our GAN suffers from generator instability and diminished generator gradient due to an overly powerful discriminator. For future work, we will further carefully tune the weight for each term in the generator loss, and perhaps further reduce the discriminator complexity. With a satisfying first 100 epoch performance, we'll train the network for a full 300 epochs, and reset the discriminator parameters every 100 epochs.

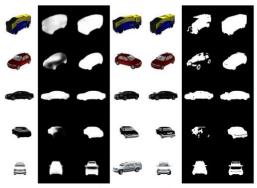
1. Zhou, T., et al.: "View Synthesis by Appearance Flow", arXiv:1605.03557 [cs.CV], Feb. 2017
2. Park, E., et al.: "Transformation-Grounded Image Generation Network for Novel 3D View Synthesis", arXiv:1703.02921 [cs.CV], Mar. 2017
3. A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015
4. https://github.com/sunweilun/ObjRenderer

Results

DOAFN:



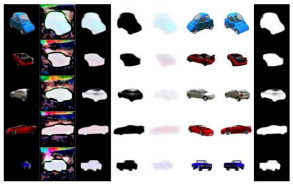
1 epoch

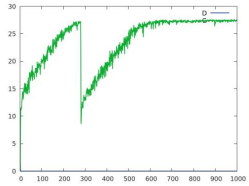


200 epoch

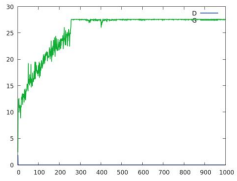
TVSN:

Concat & output mask after 100 epoch





Concat only after 100 epoch



Concat & residual after 21 epoch