OCCLUSION DETECTION WITH FULLY CONVOLUTIONAL NETWORKS

YOUNGSUN KIM, GYUHAK KIM @ NYU

Introduction

In this project, we have constructed fully convolutional networks which enable us to learn the properties of occlusion by supervised learning and reconstruct an occlusion mask in full resolution for a pair of unknown frames.

During the learning phase, a unit input will be a pair of 2D images that are successive video frames, and there will be a corresponding ground truth mask. After finishing learning, it will take the same format of a unit input and produce a same size of make which consists of pixels with 1 for occlusion area and 0 for otherwise.

Algorithm

In this project, we have constructed fully convolutional networks which enable us to learn the properties of occlusion by supervised learning and reconstruct an occlusion mask in full resolution.

Two phases of learning:

- Convolutions
  - 5 levels (by 2x2 pooling, 3x3 kernels)
  - 3 sublayers per level (1x1 kernels)
- Deconvolutions
  - 5 levels and 3 sublayers per level
  - In each level, concatenated with a corresponding convolution layer

Methodology

The model was trained to minimize the square loss function. Each predicted pixel was matched to the corresponding pixel of ground truth. The predicted value was

$$\hat{y} = \arg\min_{f} \sum \left(f(x_i) - y_i\right)^2$$

For qualitative evaluation, we recovered the detected occlusion into visual images so a direct comparison with the provided occlusion frames is possible. For quantitative evaluation, we estimated the proportion of correct prediction to the number of pixels.

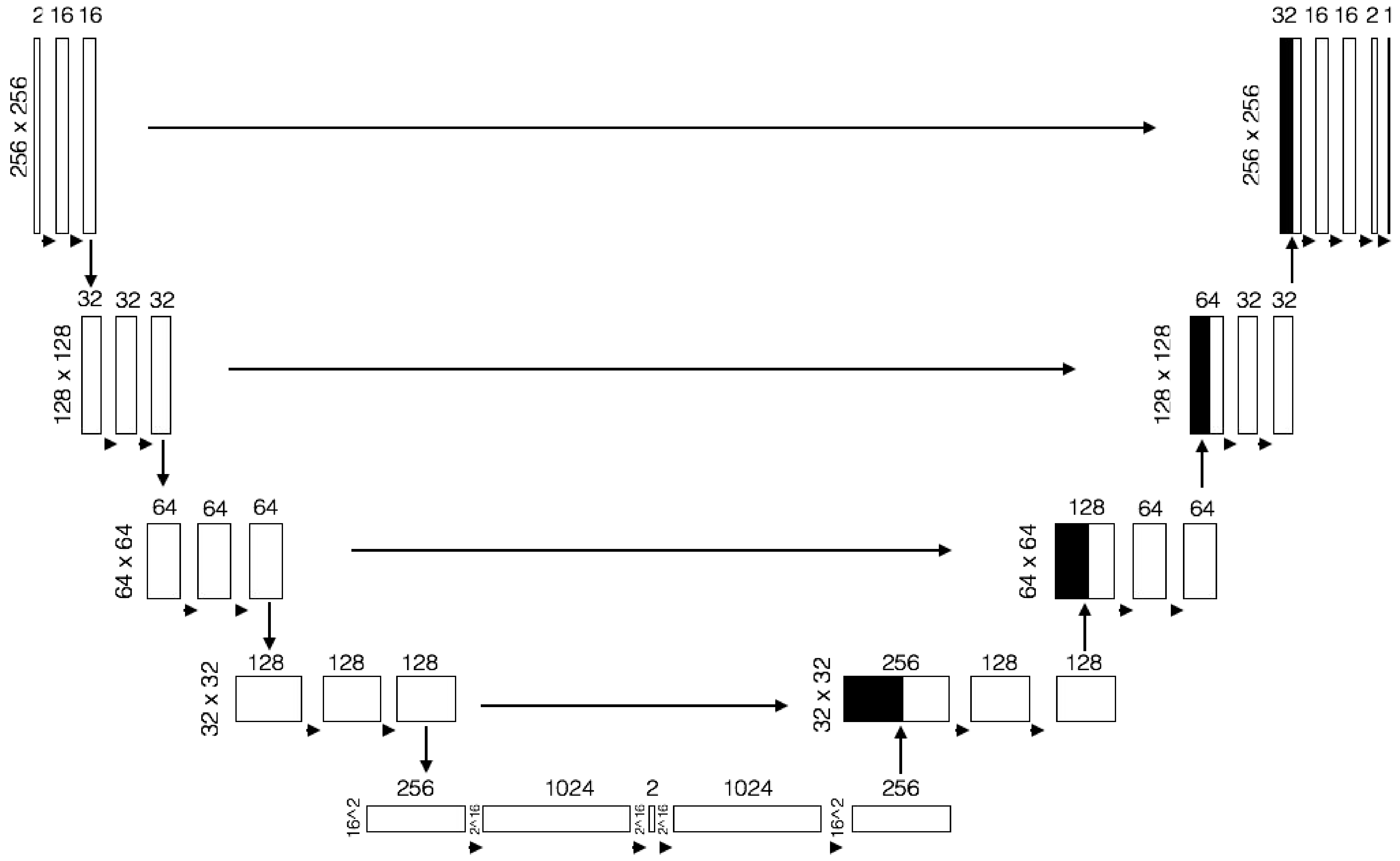

Figure 1. Graphical illustration of the network

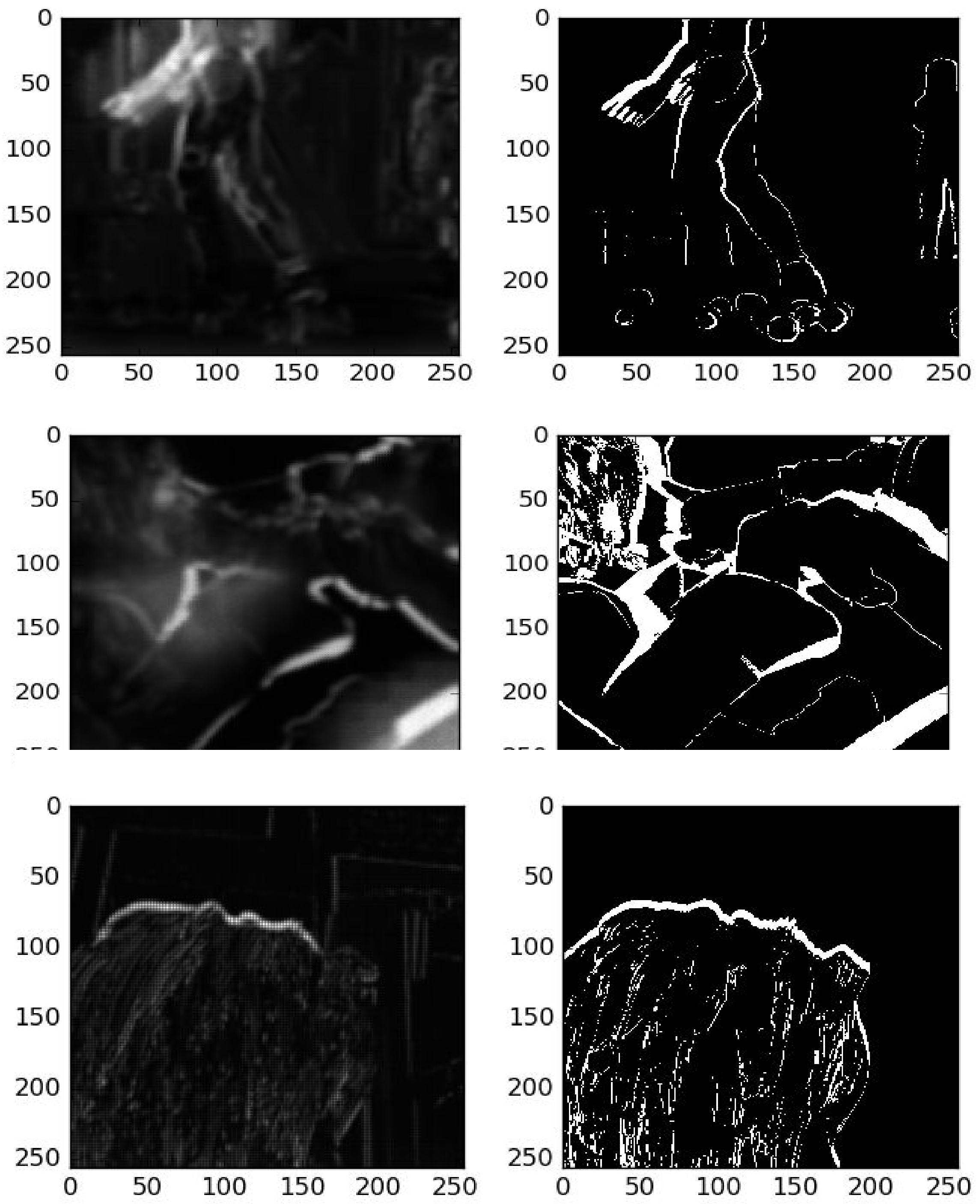

Figure 3. Model result. The pictures on the left are model predicted and ones on the right are ground truth

Dataset

Our dataset is MPI-Sintel Flow Dataset. It is an open source dataset consist of ~3000 color frames of animated short film produced for learning optical flows with different environments, actions, and settings. We partitioned each color frame into 6 equal sized smaller frames (each frame became 256 x 256 pixels). Then we did the same work for ground truth. Thus, we finally get ~18,000 pairs of frames and corresponding ground truth bitmaps. We believe this enables us to have much diverse images to train the model with.

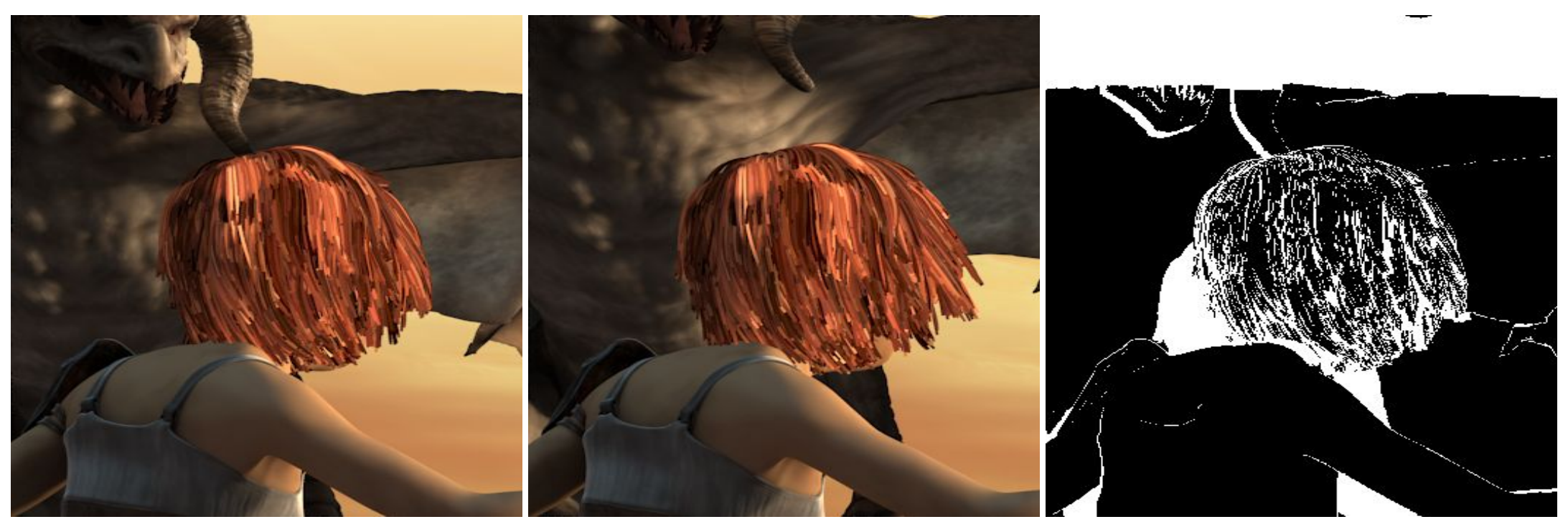

Figure 2. Two consecutive images and their ground truth image. The white area represents occlusion

Results

We have restored the occluded area from the pixel-wise prediction for qualitative evaluation. The examples of the result are shown on Figure 3. The model captured the shape of occlusion. It shows high accuracy on the area where the occlusion is relatively large such as the first and second images of Figure 3. However, it does not capture the subtle change between frames. The third image of Figure 3 shows a vague shape of human's head, but lacks the detail since it does not show the occluded area generated from hair movement.

We also did quantitative evaluation on the result by matching each pixels of the restored image and the ground truth. The matching rate was approximately 90% on average. However, we do not believe this average accuracy indicates the quality of result since most pixels are 0.

Future Work

For further study, one could modify the model and test on dataset which contains images with blur and environmental effects. FlowNet also uses fully convolutional networks, but it constructed two identical processing streams and combine them at a later stage. This method is suitable for more complicated dataset.

For quantitative evaluation, one could measure the proportion of the matching pixels of occluded area of the ground truth and of the predicted result to the number of the occlusion pixels.