Synthesizing photo-realistic images from text descriptions is a challenging problem. Previous studies have shown remarkable progresses on visual quality of the