We appreciate all reviewers' valuable comments. We shall address the concerns raised point by point, as follows.

**To R1 & R2: 1. The qualitative results are not very satisfying:** In fact, there is very little supervision used in our method and so it is very challenging. Firstly, we do not use ground-truth 3D texture supervision. Second, the human parsing, body shape and pose are directly obtained from the HMR [17] model pre-trained on other datasets, which may cause inaccurate estimations and affect the performance of texture generation. The method would be much stronger with an end-to-end trained human body model, as suggested by R3.

**2. Comparison from multiple views:** Fig. 1 illustrates multiple views of our rendered images on Market-1501 and DeepFashion. It can be seen that the details are still preserved well even for unseen parts of input images. We will show more qualitative results and comparison with RSTG [33] from multiple views in the future version.

**3. Comparison to PIFu:** To be general, our method does not use any ground-truth 3D textures during training. This is more applicable, however, limited by this it is more challenging. Therefore, they are along different developments and it is unfair to compare our method with PIFu, which requires ground-truth 3D texture supervision.

**To R1 & R3: Novelty and differences w.r.t. RSTG [33]:** Different from RSTG [33], we explicitly encourage the cross-view consistency during training. Besides, we also propose to use semantic parsing of human image as the model input to reduce the appearance variation and preserve pose information of human body. To our knowledge, these have not been explored before in the task of texture generation.

**To R1: 1. It's unclear if there is significant improvement over RSTG [33] from Fig. 5:** Compared to RSTG [33], our method can keep more details of input images, e.g., T-shirts in Row 1-2 and shorts in Row 3 in Fig. 5 (a). Our method also achieves much better quantitative results in terms of SSIM and mask-SSIM, as reported in Table 2.

**2. Using LPIPS for evaluation:** We evaluate our method with LPIPS on Market-1501. For our method, the average distance computed by the LPIPS metric is 2.440, and for RSTG [33], the one is 2.685. Thus, our method still outperforms RSTG in terms of LPIPS.

**3. CMR has much higher mask-SSIM score than RSTG:** Although CMR looks worse from a global perspective, it does better than RSTG in some local details. This may cause that CMR obtains higher mask-SSIM score which focuses on the area of human body.

**4. Using the disjoint UV mapping:** Good suggestion. Actually, although we only use the simple spherical UV mapping, our method still achieves better performance than RSTG [33] which already uses the disjoint UV mapping, especially in detail preserving. This further demonstrates the effectiveness of the proposed components.

**5. Missing relevant citations:** Thanks for your suggestion. We will include them in the updated version.

**To R2: 1. The texture copying is a little "randomly" when the texture is complicated (Fig.1 (b)):** Since we learn to predict the texture flow without ground-truth 3D texture supervision, it is impossible to perfectly match the pixel locations between the mesh surface and input image. When the texture pattern is complicated, such dislocation will be magnified visually and the texture appears to be copied "randomly".

**2. The texture tends to have front-back symmetry in Fig.1:** You are right. Since only one image is used as input and ground-truth 3D textures are unavailable during training, it is very hard to infer the accurate texture of the back (front) view from the front (back) view, even for humans. Thus, the model tends to simply copy the predicted texture flow. Further improvements could be possible with prior knowledge about front-back texture relations learned from data.

**To R3: 1. Relation with [A]:** Both the task and method are different. [A] is a nice work aiming at learning a powerful representation for 3D pose estimation, where a spatial transform based bidirectional novel view synthesis is further proposed to exploit view consistency. In contrast, our task is to generate a physical 3D model for computer graphics system, with a parameterized 3D mesh along with detailed textures on it. This explicit 3D modeling is different from spatial transform. Thanks for your suggestion. We will add related discussion in the updated version.

**2. Could the existing model be trained end-to-end, including pose and shape estimation:** Yes, it could be end-to-end trained if annotations of human joints and masks are given. In such way, more accurate 3D poses and shapes can be estimated to further improve the performance of texture generation, and at the same time the computation can be reduced by sharing the same backbone network. Thanks for the suggestion as a nice future work.

**3. What happens when the pose or shape estimation is inaccurate:** The perceptual loss can mitigate the impact of inaccurate pose or shape estimation to a certain extent, but inaccurate human parsing may affect the performance of texture flow prediction.

**4. Why is the head excluded in DeepFashion:** We exclude the head to make visualization focus on the body texture generation, which is mainly considered in this paper.