1 IMPLEMENTATION DETAILS OF CONDITIONING INPUT IMAGES.

We clip the depth values in \( D \) to a depth range of two meters starting from the minimum depth value, and normalize them to the interval \([0,255] \). The size of \( X \) is \( W \times H \times (P \cdot C + B) \), where \( W = 256 \) and \( H = 256 \) denote the image width and height, respectively. \( P = 6 \) is the number of body parts, \( C = 4 \) denotes the number of channels of the rendered RGBD images, and \( B = 3 \) denotes the number of channels of the background image (RGB).

2 ADDITIONAL COMPARISONS

In this section, we provide comparisons to the concurrent unpublished work of \cite{Aberman2018}, \cite{Chan2018} and \cite{Wang2018}. Note that, neither the code nor the dataset of their work is available. Therefore, we can only provide qualitative comparisons of the results on similar poses. As can be seen from Fig. 2, our method yields clearly much sharper imagery with less artifacts than the methods of \cite{Aberman2018} and \cite{Wang2018}. In general, the quality of our results is comparable with or better than that of \cite{Chan2018}. The results of \cite{Chan2018} occasionally exhibit strong artifacts such as broken and unnaturally extended arms (see Fig. 1). While our method exhibits occasional failures too, such strong errors are less likely to occur in our method.

Fig. 1. The results of \cite{Chan2018} exhibit occasional strong artifacts such as broken and unnaturally extended arms.

which is based on a more sophisticated shape model as conditioning input.

REFERENCES


Fig. 2. Qualitative comparison to the concurrent unpublished work of [Aberman et al. 2018], [Chan et al. 2018] and [Wang et al. 2018]: Our method yields clearly much sharper imagery with less artifacts than the methods of [Aberman et al. 2018] and [Wang et al. 2018], and comparable or better results than [Chan et al. 2018].