Attention-based Fusion for Multi-source Human Image Generation

Stéphane Lathuilière[1], Enver Sangineto[2], Aliaksandr Siarohin[2] and Nicu Sebe[2,3]
[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, France,
[2] DISI, University of Trento, Italy, [3]Huawei Technologies Ireland, Dublin, Ireland

In this Supplementary Material we present some additional qualitative experiments. First, we show how a model, trained with 2 source images, can be tested with a varying number of sources (Sec. 1). Then, in Sec. 2, we compare the *Avg* with the *Full* model presented in Sec. 4.4 of the main paper. In Sec. 3, we show other qualitative results for both the datasets used in the main paper. Finally, in Sec. 4 we show some *failure* cases, i.e. examples of images wrongly generated by our method.

1. Non-fine tuned model: varying the testing source number

In this section we use a model *trained* with only 2 source images for each training sample ($M_n = 2$ for each $n \in \{1, ..., N\}$) but then *tested* by varying M_n over $\{2, 3, 4\}$ source images, *without* fine-tuning the networks with the specific M_n value used for testing. This is possible because, as explained in the main paper, the attention-based decoder A can aggregate a variable number of source image representations ($\mathcal{E}_n$) provided by replicas of E (see Sec. 3 of the main paper).

We qualitatively evaluate the ability of the non-fine tuned model to handle a varying number of source images (M_n) and the advantage of using $M_n > 1$ at testing time. In Fig. 1 we show a few images generated using the DeepFashion dataset. In the first row we see that the model combines the images and reduces the artefacts. For instance, when the models combines 4 source images, the artefacts due to the hood disappear and the collar is generated with more details. This is probably due to the fact that the collar is clearly visible only in the fourth source image. In the second row the model cannot generate correctly the frontal view using only the first two images. When using more images, the motif on the shirt is better rendered. In rows 3, 4 and 5, the model does not correctly generate the collar shape from the first two images because the collar is clearly visible only in the fourth source image. Generally speaking, using more images improves the generated image quality. It the last row the generated images have some artefacts on the shoulders because the hair in the source images hide a part of the jacket. Although the shoulders of the *target* pose are never clearly visible in the 4 source images, with increasing M_n, the artefacts on the generated shoulders gradually disappear and the jacket colour becomes the dominant pattern.

2. Comparison between the attention-based and the average-based decoder

In Fig. 2 we show other qualitative results in order to further compare the *Avg* model with our *Full* model (see Sec. 4.4 of the main paper). These images confirm the results reported in the main paper since *Full* generates better details of the clothes. For instance, in the first two rows, the images generated by *Avg* contain more artefacts. In the third row, although not perfect, the edges between the cardigan and the dress are sharper and the texture of the two clothes are less blended when using *Full*. The forth-row example is particularly challenging because of the complex texture. First, we observe that *Avg* slightly changes the color of the yellow texture. Moreover, the multicoloured motif next to the collar is better generated when using the proposed attention-based model. Finally, *Avg* generates large artefacts on the left arm, probably due to the tatoo in the source images. In the last two rows, the shirts are tucked in the pants in the source frontal views but not in the source rear views. *Avg* predicts semi-transparent shirts by simply averaging the corresponding two source images. Conversely, in both cases *Full* correctly generates the shirt tucked in the pants by exploiting the frontal views, which are the closest to the pose view. Note that in the last row both methods fail when generating the bottom part of the shirt, by mistake blending its color with the right arm which partially overlaps the shirt in the frontal source image.

3. Additional qualitative results

In this section we use networks *fine-tuned* using the same cardinality (M_n) of the source image set then used for testing (see Sec. 4 of the main paper).

In Fig. 3 and 4 we show images generated by using the Market-1501 and DeepFashion dataset, respectively, and we adopt the same visualization technique used in Fig. 4 of the main paper. Similarly to the results reported in the main paper, also Figs 3 and 4 show that, when the number of source images increases, our model generates better images with more realistic details.

In Fig 3, we use the Market-1501 dataset and $M_n \in \{2, 3, 5\}$. In the first row, when $M_n = 2$ or $M_n = 3$, the body of the person in the corresponding generated images is blended with the color of the occluding umbrella depicted in the second source image. However, when using $M_n = 5$, the model generates a better image by exploiting the fourth source image. In rows 2, 3, 4, 7 and 8, exploiting more source images leads to sharper images with more realistic details. In row 6, the backpack is correctly generated only when using five source images since the backpack is visible only in the fifth image. This example clearly illustrates the benefit of using several source image.

In Fig. 4 we use the DeepFashion dataset and $M_n \in \{1, 2, 3\}$. In the first two rows, the single-source model suffers from the self-occlusion in the source image. As a consequence, in the first row, the left shoulder is not correctly generated while in the second row, a dark line appears in the abdomen. When the model disposes of more source images, these artefacts disappear. In row 3, the single-source model is not able to separate the sleeve from the pants and, consequently, wrongly generates the pants. Conversely, the multi-source model corrects this issue. In row 4, the single-source model is not able to handle the large pose difference whereas the multi-source model can exploit the additional images having a pose closer to the target pose. In row 5, the single-source model cannot handle the unusual texture of the dress and generates non-existing shorts. Finally, in the last two rows, the single-source model generates many artefacts which disappear in the images generated by the multi-source model.

4. Failure cases

We show in Fig 5 some *failure* examples. Similarly to the previous section, we use *fine-tuned* networks. In rows 2 and 4 our model fails to correctly transfer the corresponding textures even when it disposes of more images. This is likely due to the fact that, in both rows, none of the source images has a pose similar to the target pose. In the other rows, the model is not able to generate correctly the image most likely because of the unusual target pose.

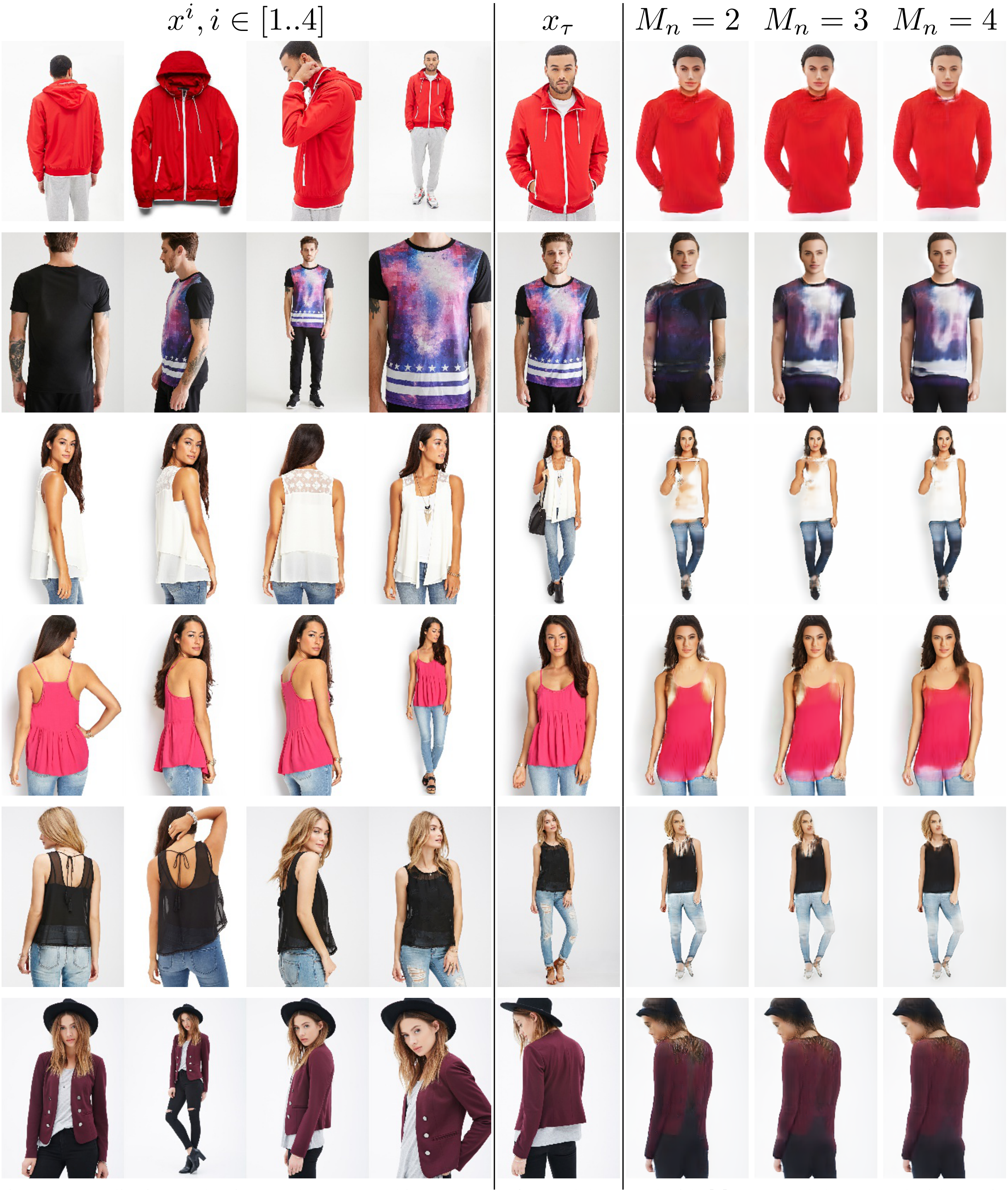

Figure 1: Qualitative evaluation on the DeepFashion Dataset of a model *trained* with $M_n = 2$ and then *tested* with varying M_n values (from 2 to 4). The first 4 columns show the *testing* source images which correspond to the last 3 columns showing the generated images. For instance, the first 2 columns correspond to testing with $M_n = 2$ sources, whose result is shown in the corresponding column; while the first 3 columns correspond to testing with $M_n = 3$ sources, etc.

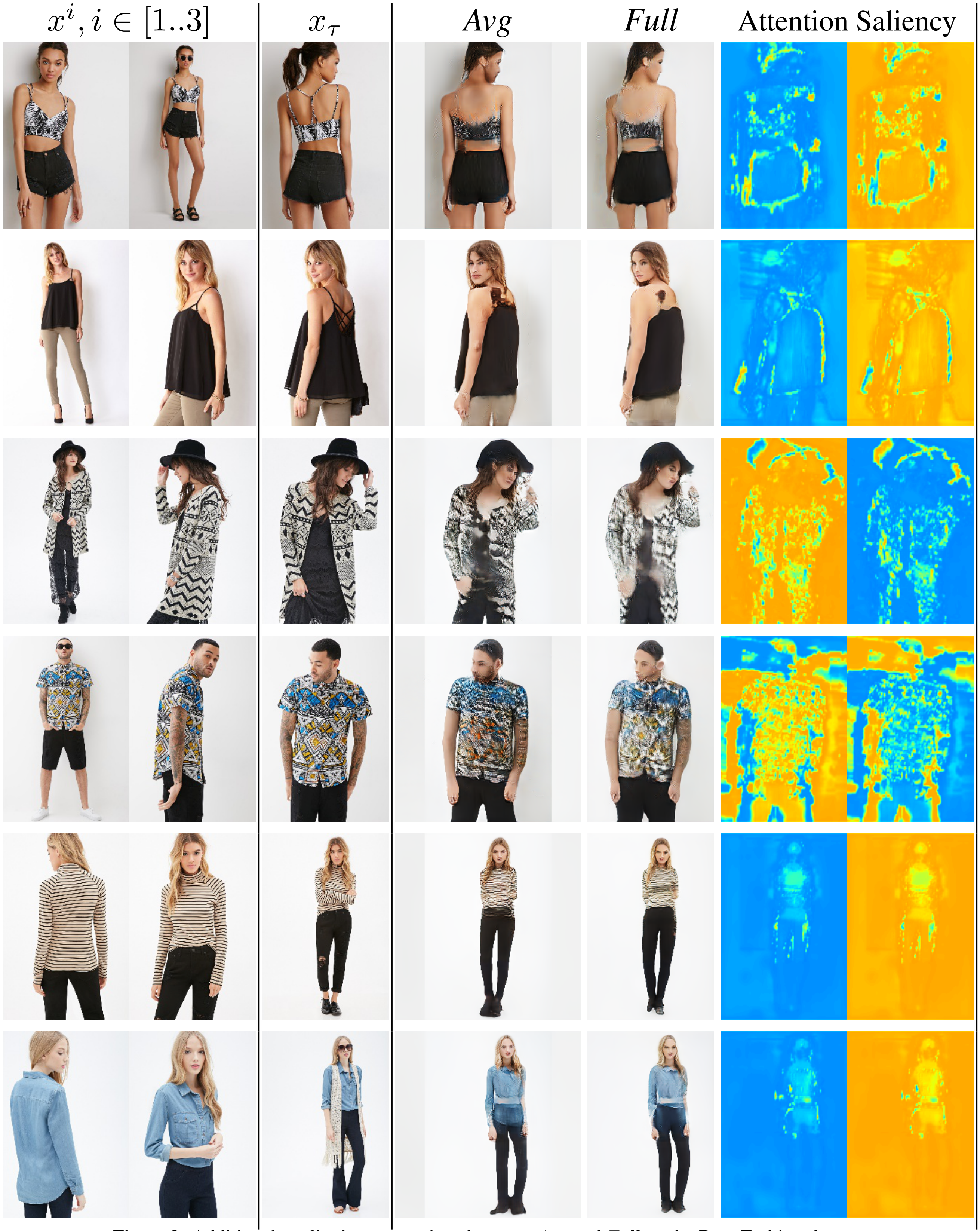

Figure 2: Additional qualitative comparison between *Avg* and *Full* on the DeepFashion dataset.

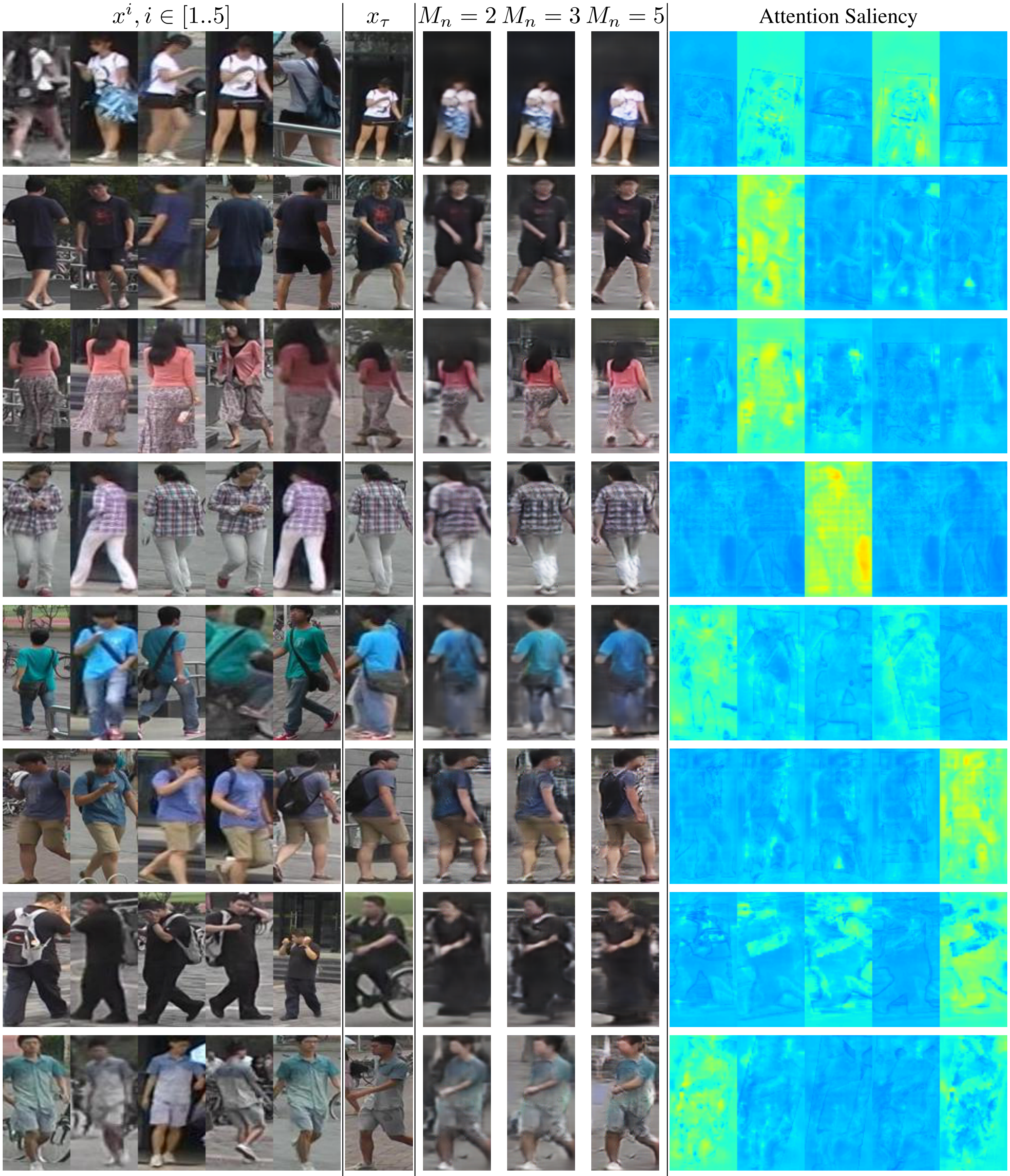

Figure 3: Additional qualitative results on the Market-1501 dataset.

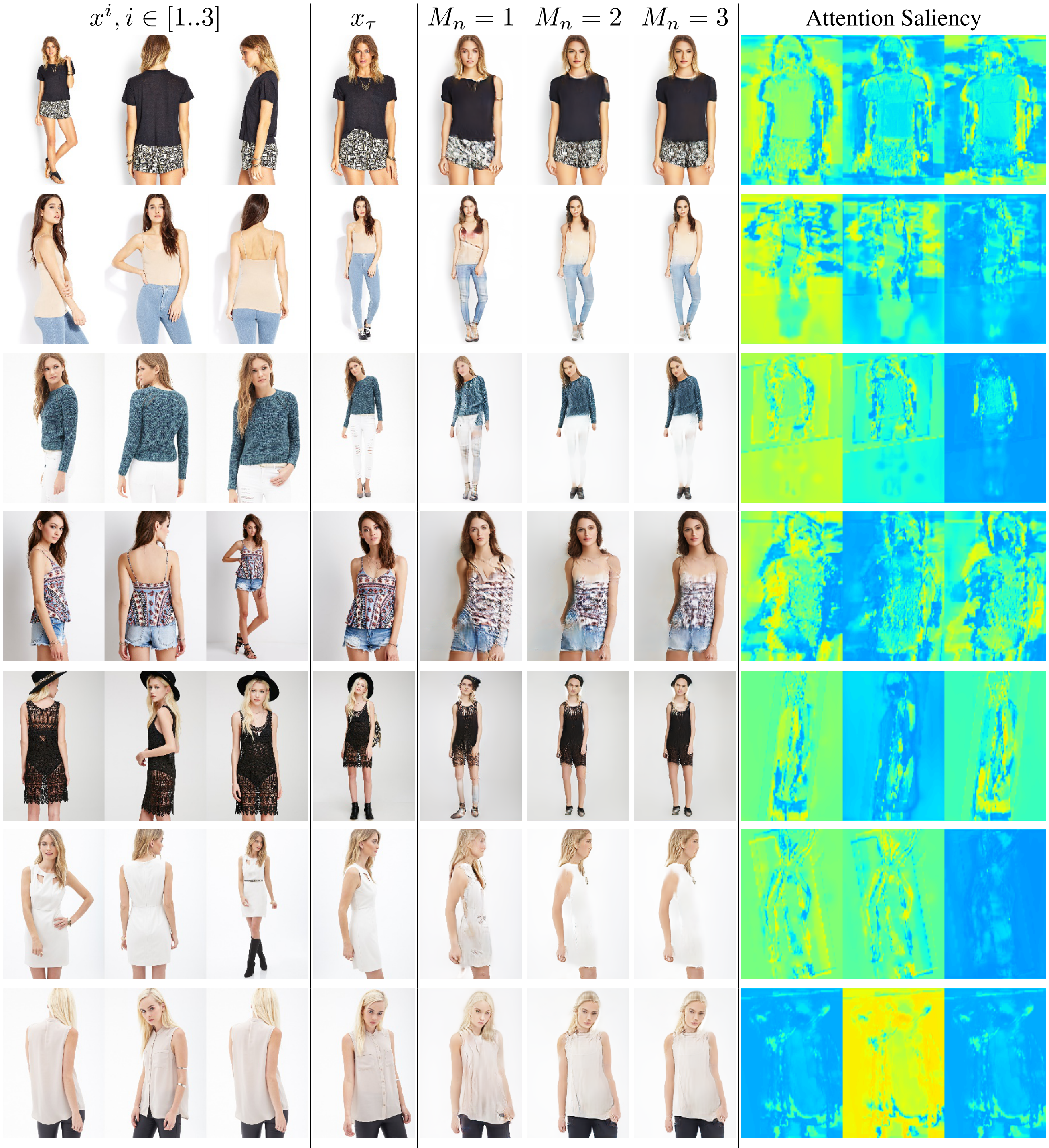

Figure 4: Additional qualitative results on the DeepFashion dataset.

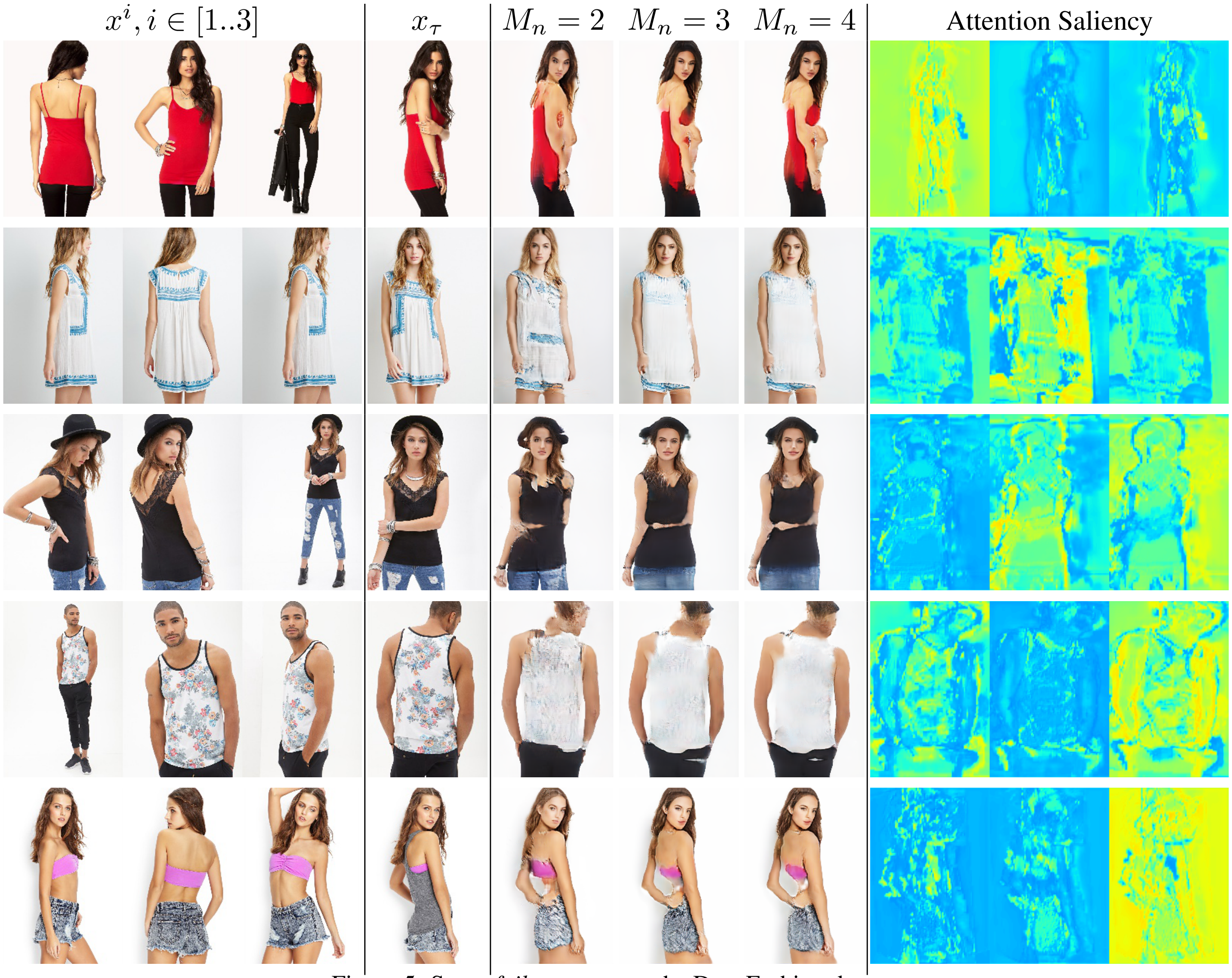

Figure 5: Some *failure* cases on the DeepFashion dataset.