# Supplementary Material for Structure-Aware Human Body Reshaping with Adaptive Affinity-Graph Network

Qiwen Deng*
University of Electronic Science and Technology of China
don2889632705@gmail.com

Yangcen Liu*
Georgia Institute of Technology
yliu3735@gatech.edu

## 1. Global Affinity

For a specific body part, ensuring consistency with other parts involves extracting affinity across the portraits in the Adaptive Affinity-Graph Block (AAG). As illustrated in Figure 1, we visualize the affinity map $W$ for two single points, separately in arms and legs. In (a), we visualize the activated attention map for legs, torso, and arms, with a query point in the legs, and we mark the top 20 activated points. In (b) we visualize the results activated by a query point from the arms. The body parts features are $F_i \in \mathbb{R}^{H\times W\times C}, i \in \{torso, legs, arms\}$.

In (b) we let $q$ represent the query point in $F_{arms}$ and $k$ represent a key point in $F_{torso}$. The attention map of point $q$ to $F_{torso}$ is computed as follow:

$$W_{q,k} = \frac{1}{C}\sum_{c=1}^{C} Q_q^c * K_k^c, \tag{1}$$

where $Q_q^c$ represents query value ($Q = \Phi_Q^{arms,torso}(F_{arms})$) of point $q$ in channel $c$ and $K_k^c$ represents the key value ($K = \Phi_K^{arms,torso}(F_{torso})$) of point $k$. The higher value of $W_{i,j}$ indicates the stronger affinity between point $q$ in the arms and point $k$ in the torso.

Our observations indicate that our model effectively directs attention to primary areas with affinity, thereby preserving photo aesthetics. For example, in Figure 1(a), the attention map to torso, points with high activation scores accumulated on the back (left) side of the torso, indicative of a strong association with the leg region. Leveraging the Adaptive Affinity-Graph Block (AAG), which extracts affinity between each pair of body parts to obtain global affinity, our model achieves consistency across the entire body of the generated image.

## 2. Ablation Study

**Ablation Study on Convolutional Block Attention Module** The Adaptive Affinity-Graph Block (AAG) is composed of two primary components: a fully connected graph network, wherein each node corresponds to a distinct body part for affinity extraction, and the Convolutional Block Attention Module (CBAM). In our method, we organize affinities into the channel of $A$ and use a Convolutional Block Attention Module (CBAM) following (4) to re-assign and refine the weights to these affinities utilizing the inductive bias of both average pooling and max pooling.

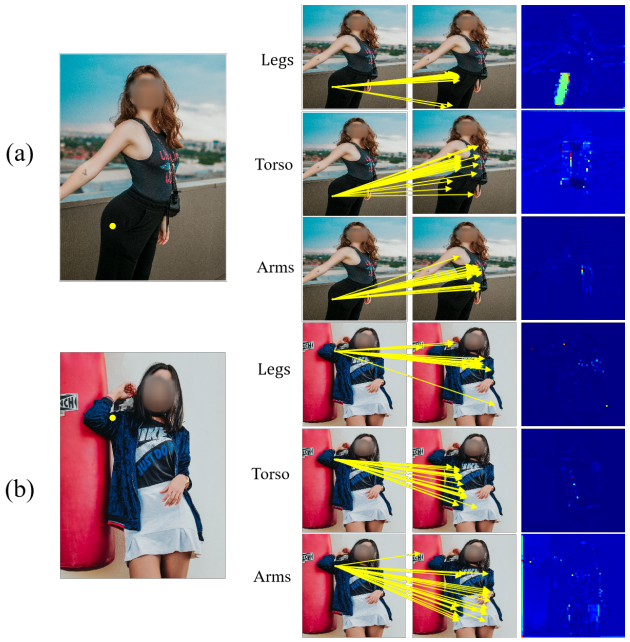


Figure 1. The visualization of $W$ queried by a single point. (a) shows the 20 highest activated points of a sample point in the legs, and (b) a sample point in the arms.

The ablation study with or without the CBAM is shown in Table 1. An improvement is evident across all three metrics. The results prove that while acquiring the global affinity, re-assigning and reigning the attention score to different body parts with CBAM is necessary for the output affinity map $A$.

**Reliance on pose estimation.** Our method requires skele-

*Equal contribution.

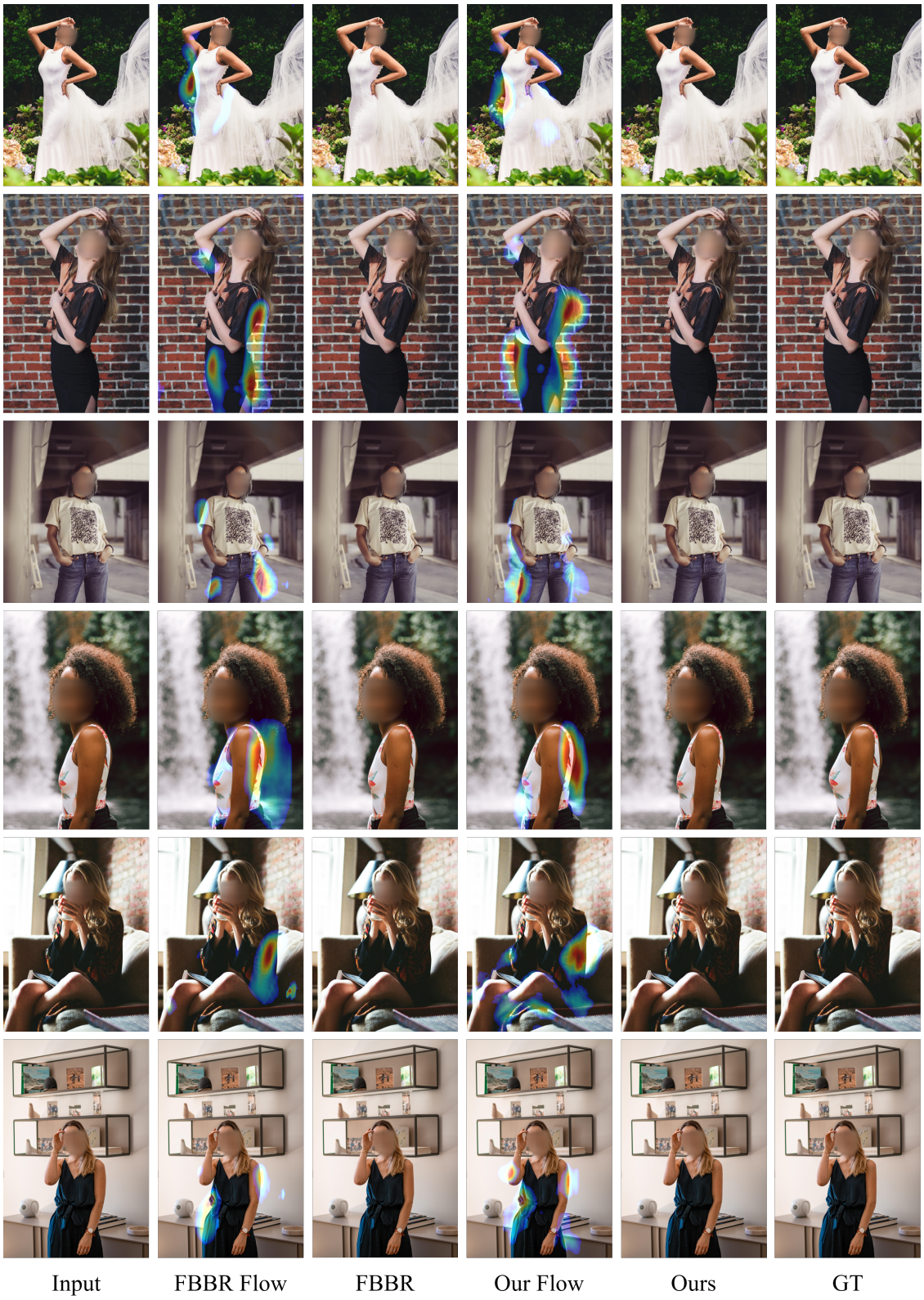


Figure 2. The qualitative results of our method and FBBR.

| Method | SSIM ↑ | PSNR ↑ | LPIPS ↓ |
|---|---|---|---|
| without CBAM | 0.8398 | 25.8853 | 0.0679 |
| with CBAM | **0.8427** | **26.4100** | **0.0643** |

Table 1. Ablation study of Convolutional Block Attention Module (CBAM). The results are evaluated when Adaptive Affinity-Graph Block (AAG) is with or without CBAM design.

ton maps as input. However, this reliance is acceptable for the following reasons: (1) A lower-bound analysis in Table 3 indicates performance degradation in the extreme case where the input pose is masked. Despite this, the lower-bound performance still surpasses that of RGB-based methods. (2) Pose estimation is not the key point of our paper. We use the pose estimator aligned with FBBR for fair comparison, though more advanced pose estimation methods (1; 2) could further improve the results.

**Ablation study of flow generator.** The effectiveness of VGG loss is demonstrated in Table 2. We observed that without $L_{VGG}$, the performance in LPIPS is even better. We believe that $L_{VGG}$ helps with high-frequency information extraction, but might lead to extra flow blur in those irrelevant areas.

| Exp | $L_{VGG}$ | $L_{img}$ | SSIM ↑ | PSNR ↑ | LPIPS ↓ |
|---|---|---|---|---|---|
| 1 | ✗ | ✓ | 0.8351 | 26.27 | **0.0662** |
| 2 | ✓ | ✗ | 0.8346 | 25.97 | 0.0719 |
| 3 | ✓ | ✓ | **0.8365** | **26.29** | 0.0669 |

Table 2. Ablation study on Flow Generator.

## 3. More Visualization Results

**More Visualization** We show more qualitative comparison results in Figure 2. Here we only compare our method with FBBR (3) as shown in Figure 2. Our approach can produce more consistent and visually pleasing body-reshaping results. We can observe that FBBR tends to edit a particular part individually rather than considering all parts collectively.

**Dynamic Visualization** In this section, we present a package of 8 sample GIFs included in the supplementary material. We exclusively compare our model with the previous state-of-the-art method, FBBR. Our aim is to provide a clearer demonstration of the application process. The results highlight the advantages of our method and emphasize that optical flow-based methods excel at preserving the background of the original image. Optical flow primarily transforms specific areas, making it more suitable for our task. Conversely, recent trends in image-to-image transformation or relevant downstream tasks, such as motion transfer, have leaned towards employing diffusion models.

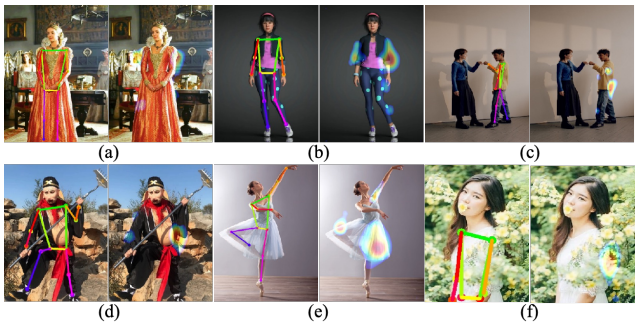


Figure 3. Results of multiple types of open-domain images. (a) Complex clothes. (b) Digital human. (c) Multiple persons. (d) Fantasy characters. (e) Complicated pose. (f) Occlusions.

| Exp | SSIM ↑ | PSNR ↑ | LPIPS ↓ |
|---|---|---|---|
| w/o pose | 0.8132 | 23.9434 | 0.0892 |
| w/ pose | **0.8427** | **26.4100** | **0.0643** |

Table 3. Lower-bound analysis without pose input.

**Calculation and structure differences compared to FBRR.**

The results of the calculation and parameter scale comparison are shown in Table 4. Our AAGN improves performance while remaining lightweight compared to previous work, with only a 0.11ms FPS decrease and an additional 0.2M parameters. Our Flow Generator (FG) retains the same structure as in FBBR but omits the Structure Affinity Self-attention module.

| Exp | FBBR | FG | AAGN |
|---|---|---|---|
| FPS (ms) | 3.75 | 3.65 | 3.54 |
| Param (M) | 6.7 | 6.6 | 6.9 |

Table 4. Quantitative comparison of FPS (Frame Per Second) and Parameter Scale of processing images (including pose detector) on a single RTX 3090.

**Visualization on open-world datasets.** Testing the methods on different datasets is essential. However, similar to previous work, the BR-5K dataset is the only publicly available dataset. Despite concerns about generalizability, our method, trained on the limited BR-5K dataset, still delivers remarkable results on open-domain images. We visualize six types of open-domain images downloaded from the web, as shown in Figure 3. The visualization results still maintain high quality.

## References

[1] Yuanzheng Ci, Yizhou Wang, Meilin Chen, Shixiang Tang, Lei Bai, Feng Zhu, Rui Zhao, Fengwei Yu, Donglian Qi, and

Wanli Ouyang. Unihcp: A unified model for human-centric perceptions, 2023. 3

[2] Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. Human pose as compositional tokens, 2023. 3

[3] Jianqiang Ren, Yuan Yao, Biwen Lei, Miaomiao Cui, and Xuansong Xie. Structure-aware flow generation for human body reshaping. In 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 7744–7753, 2022. 3

[4] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, Computer Vision – ECCV 2018, pages 3–19, Cham, 2018. Springer International Publishing. 1