

DiffBody: Human Body Image Restoration with Generative Diffusion Prior

Yiming Zhang, Lionel Z. Wang*, Sizhuo Ma*, Xinjie Li, Jian Ren, Zhihang Zhong[†], and Jian Wang[†]

Abstract—Human body image restoration is crucial for various applications but remains challenging due to the limitations of generative models: General image restoration methods built on generative models may generate unnatural textures, noticeable structural misalignments, and significant loss of fine details. To address these shortcomings, we present DiffBody, a novel human body-aware diffusion model that incorporates domain-specific knowledge to significantly enhance restoration quality. Our approach adopts a two-stage framework: (1) a multi-branch joint diffusion model generates preliminary priors, including normal and depth maps supported by a robust reconstruction pre-processing step; (2) a restoration stage refines the output using a body-prior ControlNet and a color adapter, ensuring structural accuracy and color consistency. Extensive quantitative evaluations, qualitative evaluations, and user studies validate the superior performance of DiffBody in producing perceptually high-quality human body restoration results. Code is available at <https://github.com/yimingz1218/DiffBody>.

Index Terms—Image Restoration, Diffusion Model



1 INTRODUCTION

BLIND image restoration (BIR) aims to enhance the quality of degraded images through processes like denoising [1], sharpening [2], deblurring [3], super-resolution [4], *etc.*, a domain that has been significantly advanced by neural networks. Although general BIR has made substantial strides, users often exhibit a greater interest in restoration of specific subjects such as human faces [5] and bodies. Restoration of human body images can have a profound impact on various human-centric applications, such as improving portrait quality in social media and aiding downstream tasks like person re-identification [6], 3D reconstruction [7], *etc.*

While the end-to-end reconstruction paradigm [8], [9] has made great progress for BIR, it struggles to handle complicated combinatorial and severe degradations. Recently, a generative paradigm has emerged, which harnesses the power of generative models such as generative adversarial networks (GANs) [10] and diffusion models [11]. Generative models possess comprehensive *prior knowledge* of how a natural, high-quality image looks like learned from large amounts of data, which can be used to fill in reasonable details to the degraded images. Recent diffusion models have enhanced the perceptual quality and versatility of image restoration [12], [13], [14], thereby expanding the applicability of image restoration in practical contexts.

Despite these advancements, the specific domain of human body image restoration remains underdeveloped. Current diffusion-based general restoration models [13], [14],

[15] often produce artifacts in low-quality human images, such as unnatural and overly smoothed textures, incorrect anatomy, and loss of body details, as shown in Fig. 1. This problem can be examined through the perception-distortion tradeoff [16]: image restoration models inherently favor either perceptual quality or low distortion but cannot excel at both. While the preference for quality or distortion varies by task, perceptual quality is crucial in *human body restoration* due to our higher sensitivity to distortions in limbs and skin. For example, while minor distortion in background objects may go unnoticed, artifacts in human bodies such as plastic-like skin and missing fingers are more noticeable and can make users immediately reject the result.

In this work, we aim to push the performance of human body image restoration by prioritizing perceptual quality to enhance *viewer comfort* [17] and ensure a more natural and pleasant user experience. We present DiffBody, a novel diffusion model tailored to human body restoration. The key idea is to effectively guide a pretrained diffusion model to restore clear and realistic human bodies using extracted human priors [18], implemented through a two-stage process. In Stage 1, we use SwinIR [8] to preprocess degraded images and produce a preliminary restoration, from which we extract essential priors: pose, text, depth, and normal maps. Depth maps ensure structural alignment, while normal maps preserve surface details and correct unnatural textures. Pose information maintains anatomical coherence, ensuring a visually consistent human body structure. Stage 2 integrates these priors for detailed restoration. To address inconsistent colors that could undermine structural corrections, we introduce a *emphcolor* adapter for accurate color alignment. Although no existing metric quantifies viewer comfort, we conduct the *viewer comfort test* in user studies to confirm that our method delivers the most visually coherent and comfortable human body restoration compared to existing approaches.

Our main contributions are as follows: (1) Rather than

- Y. Zhang is with Cornell University. L. Wang is with The Hong Kong Polytechnic University. S. Ma, J. Ren and J. Wang are with Snap Research, Snap Inc. X. Li is with The Pennsylvania State University. Z. Zhong is with Shanghai AI laboratory.
- E-mail: yz2926@cornell.edu, zhe-leo.wang@connect.polyu.hk, sma@snap.com, xql5497@psu.edu, renjianustc@gmail.com, zhongzhihang@pjlab.org.cn, jwang4@snap.com
- * Contribute equally in this paper, [†] are both the corresponding authors of this paper.

Fig. 1. While the preference for quality or distortion varies by task, perceptual quality is crucial in human body image restoration due to our higher sensitivity to distortions in limbs and skin. Our DiffBody model shows superior performance compared to other methods, particularly passing the viewer comfort test as demonstrated in our user study.

forcing the model to strictly fit the distribution of low-quality images, we adopt a perception-prioritized approach that guides generation to pass the viewer comfort test. (2) We propose a novel two-stage framework: in Stage 1, we generate priors from low-quality images to guide restoration, and in Stage 2, we use these priors to refine human body restoration, assessing the impact of different priors on output quality; (3) We introduce an adapter module to address color inconsistencies, ensuring accurate and realistic color reproduction in restored images.

2 RELATED WORK

Perception-distortion tradeoff and evaluation methods: [16], [19] shows a tradeoff between perception and distortion: As the mean distortion (the dissimilarity to the ground truth image) decreases, the perceptual quality (the consistency with natural image statistics) must decrease as well. This tradeoff can be visualized as a distortion-quality curve as shown in Fig. 1. Restoration results that pass the viewer comfort test are unattainable using conventional approaches. Our goal is to improve perceptual quality, delivering visually appealing images that better align with human perception. This is reflected in improved perceptual metrics such as LPIPS [20], ManIQA [21], ClipIQA [22], and MUSIQ [23]. However, achieving this may come at the expense of potential visual distortions and lower scores on traditional objective metrics like PSNR and SSIM [14], which are less important for human body restoration. To assess viewer comfort, which cannot be measured by existing methods, we introduce the comfort pass test and comfort scoring in our user study.

Blind image restoration: Blind Image Restoration aims to restore images without prior knowledge of the specific

degradation model. Rather than relying on a known corruption process, BIR algorithms generalize across different types of degradation, making it a more challenging task. Predominantly, existing literature [5], [24], [25], [26], [27] has concentrated on discerning a latent code situated in the latent space of a pre-trained GAN. Recent advancements in this domain [28], [29], [30], [31] have transitioned towards the utilization of DDPMs [28], marking a notable shift from conventional approaches. Other novel approaches such as DDRM [32] utilizes SVD to address linear image restoration challenges, presenting an innovative and simplified approach. DDNM [33] delves into vector range-null space decomposition to develop a novel sampling strategy, enhancing image restoration efficiency. DiffBIR [13] and SUPIR [14] aims to exploit a pretrained powerful generative prior to solve the BIR problem. In the realm of domain-specific image restoration models, a predominant emphasis has been placed on blind face restoration, as evidenced by works such as [34], [35], [36], [37], [38]. In contrast, the equally critical domain of human body restoration has not seen comparable development, a gap that our DiffBody model seeks to address.

Controllable Human Image Generation: Traditional methods for generating controllable human images mainly fall into two categories: those based on Generative Adversarial Networks (GANs) [39], [40] and those using Variational Autoencoders (VAEs) [41], [42], both leveraging reference images and specific conditions for input. Recent studies have ventured into enabling the generation process through textual instructions, though these tend to limit user input to basic pose or style adjustments [43], [44]. State-of-the-art methods enable detailed control over vocabulary and pose including ControlNet [45], T2I-Adapter [46], HumanSD [47], HyperHuman [48]. These works have shown that diffusion models are capable to generate human images that contain

rich detail and natural texture, which give us confidence that they can be utilized for high-quality human body image restoration.

3 PROPOSED METHOD

3.1 Preliminary: Latent Diffusion Model

Our method leverages the exceptional generative capabilities of Latent Diffusion Models (LDM) [30]. By compressing images into a lower-dimensional latent space before performing the diffusion process, LDMs achieve remarkable efficiency and detail in image synthesis. The model initiates a reverse diffusion process starting from a distribution of latent noise, gradually denoising this representation to reconstruct the image. This process is facilitated by a U-Net architecture, which iteratively refines the latent features under the guidance of textual conditions embedded by a pre-trained text encoder such as CLIP. The primary objective in training these models involves minimizing the error in the predicted noise.

3.2 Degraded Image-driven Joint Diffusion for Human-centric Prior

In Stage 1, the framework takes the degraded image as input to generate the human-centric priors, as shown in Fig. 2 (top). As illustrated in the model structure, degraded image I_{LQ} is preprocessed by a robust image restoration model SwinIR [8] to produce preliminary restoration $I_{IR} = \text{SwinIR}(I_{LQ})$. I_{IR} is subsequently passed to MMPose [49] and LLaVA [50] to extract the human pose I_{pose} and the corresponding textual prompt T , respectively: $I_{pose} = \text{MMPose}(I_{IR})$, $T = \text{LLaVA}(I_{IR})$. The description T is then input into CLIP to extract the textual features $c_t = \text{CLIP}(T)$ as text prompts to the network. With these foundational elements in place, we encode the latents of I_{IR} and I_{pose} using a VAE, producing $c_R = E(I_{IR})$ for the restored image and $c_p = E(I_{pose})$ for the pose. c_p is then concatenated with z_t (noisy latent at timestep t) to form \hat{z}_t .

To generate the priors, we use a multi-branch U-Net with copies of input and output layers, which has been shown to be able to simultaneously generate high-quality, spatially-aligned images of different domains [48]. This model is trained in two steps, with the first step focusing on generating the depth, normal, and RGB components based on the pose and textual conditions c_p and c_t :

$$L_U = E_{z_t; t; c_t; c_p} \sum_{d, n, i} k_{d, n, i}(\hat{z}_t; t; c_t) k_2^2; \quad (1)$$

where \hat{z}_t and t represents the noise added during diffusion and predicted during denoising, respectively. The subscripts $d; n; i$ denotes the depth, normal and RGB branches.

Once the U-Net has been trained, we introduce the latent c_{IR} from the restored image and shift to training ControlNet [45] with the following objective:

$$L_{C_1} = E_{z_t; t; c_t; c_r; c_p} \sum_{d, n, i} k_{d, n, i}(\hat{z}_t; t; c_t; c_r) k_2^2 + k_{n, i}(\hat{z}_t; t; c_t; c_r) k_2^2; \quad (2)$$

In this phase, only the ControlNet is trained such that the whole network is retargeted to an image restoration process conditioned on the restored image latent c_{IR} . Stage 1 outputs three separate channels: $I_{res}; I_{depth}; I_{normal}$, which are then used in Stage 2 to further enhance the overall performance of human image restoration. The textual prompt is also updated in this stage, where $T^0 = \text{LLaVA}(I_{res})$ is generated from the restored image I_{res} .

3.3 Enhancing Human Image Restoration through Human-centric Prior and Color Adapter

In Stage 2, the human priors I_{IR} , I_{pose} , I_{depth} , and I_{normal} obtained from Stage 1 are processed by convolutional layers F_i , followed by a linear fusion layer $c_g = {}_1F_1(I_{IR}) + {}_2F_2(I_{pose}) + {}_3F_3(I_{depth}) + {}_4F_4(I_{normal})$, as shown in Fig. 2 (bottom). The fused features are then fed to a ControlNet to guide a pretrained latent diffusion model. In parallel, we use a color adapter where the restored image I_{res} is initially encoded by CLIP and subsequently aligned through a dedicated projection module [51]. After a cross-attention module, the text prompt and I_{res} are encoded as c_t^0 and c_r^0 , respectively.

The Stage 2 model is again trained in two steps: In the first step, we train the ControlNet with the fused prior features and the text features. The loss can be defined as:

$$L_{C_2} = E_{z_t; t; c_t^0; c_r; c_p} \sum_{d, n, i} k_{d, n, i}(z_t; t; c_t^0; c_g) k_2^2; \quad (3)$$

Empirically, we find that providing I_{IR} (the initial restoration) to the model, rather than I_{res} (the further restored image), helps mitigate potential artifacts that may be introduced during restoration process in Stage 1. Our rationale is that diffusion models, particularly those based on ControlNet, are inherently more effective at adding plausible details to under-specified regions than at altering or correcting erroneous details that were not present in the original input, while I_{res} may carry forward artifacts introduced in the first stage, potentially leading the model to reinforce or amplify them.

Once the ControlNet has been trained, we train the color adapter with the image CLIP features c_r^0 using the full loss:

$$L_A = E_{z_t; t; c_t^0; c_r^0; c_g} \sum_{d, n, i} k_{d, n, i}(z_t; t; c_t^0; c_g) k_2^2; \quad (4)$$

This step is crucial because providing the depth and normal maps, rather than only the degraded image, gives the model greater flexibility in generating outputs. Although I_{res} may contain distorted image details, it generally retains accurate color information, as it is guided solely by I_{IR} and is not affected by the structural priors (depth and normal maps). Therefore, an additional color adapter is needed to enhance and preserve the original color characteristics during generation. Fusing the I_{res} information with the CLIP embedding broadens the model's learning paradigm to better capture color information. This fusion enables the model to handle color inconsistencies more effectively, resulting in more robust and higher-fidelity restoration. The synergy of complementary information from the degraded images, text prompts, poses, depth maps and normal maps allows the model to restore images with greater accuracy, especially when critical information, such as color and fine details, has been obscured due to image degradation.

Fig. 2. Method overview. In the first stage, we employ a multi-channel joint diffusion model with robust reconstruction to generate priors from degraded images. In the second stage, these priors are used to guide image reconstruction, enhancing quality through a body-prior ControlNet and a color adapter for improved structural and color consistency. Since each stage involves two separate training processes, we use the terms "Trainable" and "Frozen After Training" to distinguish their training order and status.

4 EXPERIMENTAL RESULTS

4.1 Datasets

To address challenges like fragmented human figures and varying image quality, we developed a dataset of five million high-quality human images, each annotated with MMPose, MiDaS depth [52], OmniNormal [53], and LLaVA captions. We used a bucket-based resizing strategy, similar to SDXL [54], categorizing images into five resolution buckets (512 512, 512 768, 512 1024, 768 512, and 1024 512) to accommodate different resolutions. To ensure consistent quality, we applied Real-ESRGAN's [9] degradation settings to simulate realistic image degradation.

The final training set consists of around four million human images from the CosmicMan dataset [55], refined through cropping and annotation, and one million web-sourced images, offering broader diversity in poses and environments. For evaluation, we used the high-quality SHHQ [56] dataset, which provides consistent resolution and quality, making it an ideal benchmark for testing our diffusion model's performance.

4.2 Experimental Details

For prior generation in Stage 1, we employ Stable Diffusion 2.1-base [57] as the base model. An SDXL-based multi-channel prior generator is not used due to its large number of parameters, which surpasses the memory limitations of our available GPU resources. The three-branch architecture is initialized using the HumanSD [47] framework, with re-tuning applied only to the U-Net for 100,000 steps and a batch size of 64. The model is optimized with the Adam [58] optimizer at a learning rate of 10^{-5} . After this phase, we

freeze the U-Net and re-tune the ControlNet for another 100,000 steps with the same batch size, optimization settings and hardware. For image restoration in Stage 2, we use Stable Diffusion XL-1.0-base (SDXL) as the backbone. We re-tune the ControlNet over 100,000 steps, with a batch size of 32 and gradient accumulation of 2. This phase is optimized using Adam with a learning rate of 10^{-5} . Following this, we initialize the color adapter with IP-AdapterXL Plus parameters and re-tune it for an additional 200,000 steps with a batch size of 64. This final phase uses Adam with a learning rate of 10^{-4} and is trained under the same conditions and duration. All training sessions are conducted using 8 NVIDIA A100 GPUs. For inference, we utilize DDPM sampler [28] with 200 steps for both stages.

4.3 Comparisons with State-of-the-Art Methods

Quantitative Comparison: We use PSNR, SSIM, and LPIPS [20] for full-reference evaluation. To better evaluate the perceptual image quality, we further incorporate non-reference image quality assessment (IQA) metrics: MANIQA [21], CLIPIQA [22], and MUSIQ [23]. Since no human body-specific BIR methods have been developed to our knowledge, we compare DiffBody with leading general image restoration methods: BSRGAN [59], Real-ESRGAN [9], DiffBIR [13], PASD [15], and SUPIR [14]. As shown in Table 1, DiffBody achieves strong performance on non-reference IQA metrics such as MANIQA, CLIPIQA, and MUSIQ. However, we observe relatively lower results on PSNR and SSIM, which can be attributed to the limitations of traditional metrics like PSNR and SSIM in accurately reflecting true image quality in restoration tasks as reported

TABLE 1

Quantitative comparison across different degradation scenarios. Bold and underline represent the best and second-best performance, respectively. For metrics marked with $_$, lower values are better, while for the other metrics, higher means better.

Degradation	Visual Example	Method	PSNR	SSIM	LPIPS	#	ManIQA	ClipIQA	MUSIQ
Mixture: Blur ($\sigma = 2$) SR ($\times 4$)		BSRGAN	<u>32.42</u>	0.7522	0.3604		0.3203	0.7329	58.06
		Real-ESRGAN	31.08	0.7741	0.4944		0.1364	0.6234	15.03
		DiffBIR	32.30	0.7368	0.3302		0.2918	0.7067	54.35
		PASD	32.52	<u>0.7637</u>	<u>0.2793</u>		0.4029	0.7142	72.16
		SUPIR	31.90	0.7143	0.2871		<u>0.4475</u>	<u>0.7251</u>	74.04
		DiffBody (ours)	28.69	0.6423	0.1986		0.4532	0.7621	<u>73.20</u>
Mixture: Noise ($\sigma = 40$) SR ($\times 4$)		BSRGAN	<u>33.78</u>	<u>0.8400</u>	0.1734		<u>0.4548</u>	<u>0.7306</u>	71.01
		Real-ESRGAN	32.99	0.8428	<u>0.1624</u>		0.4235	0.5836	72.29
		DiffBIR	34.15	0.8369	<u>0.1610</u>		0.3427	0.7156	69.66
		PASD	33.31	0.7897	0.1733		0.4513	0.7224	<u>75.63</u>
		SUPIR	33.55	0.7977	0.1633		0.4741	0.7250	75.06
		DiffBody (ours)	29.36	0.6973	0.1973		0.4521	0.7421	76.34
Mixture: Blur ($\sigma = 2$) Noise ($\sigma = 40$)		BSRGAN	31.04	<u>0.7488</u>	0.5071		0.2422	0.7120	18.73
		Real-ESRGAN	30.87	<u>0.7633</u>	0.5341		0.2094	0.5984	14.35
		DiffBIR	30.94	0.7104	0.4996		0.1794	0.6903	48.55
		PASD	<u>31.23</u>	0.6897	0.5171		0.2607	0.6737	34.23
		SUPIR	31.44	0.7028	<u>0.3489</u>		0.5103	<u>0.7182</u>	<u>69.72</u>
		DiffBody (ours)	29.48	0.6327	0.1598		<u>0.4494</u>	0.7366	70.01
Mixture: Blur ($\sigma = 2$) Noise ($\sigma = 40$) SR ($\times 4$)		BSRGAN	32.93	0.7997	<u>0.2832</u>		0.2355	0.7111	24.44
		Real-ESRGAN	30.88	<u>0.7665</u>	0.5162		0.1707	0.5436	14.33
		DiffBIR	31.65	0.7211	0.4493		0.2197	0.6960	60.25
		PASD	<u>31.85</u>	0.7544	0.3470		0.4001	0.7022	56.89
		SUPIR	31.50	0.7102	0.3474		<u>0.4609</u>	<u>0.7131</u>	<u>66.02</u>
		DiffBody (ours)	29.86	0.6360	0.1360		0.4690	0.7405	68.82
Mixture: Blur ($\sigma = 2$) Noise ($\sigma = 20$) SR ($\times 4$) JPEG ($q = 50$)		BSRGAN	<u>32.93</u>	0.7997	0.4800		0.3331	0.7150	58.91
		Real-ESRGAN	31.55	0.7790	0.2719		0.3541	0.6011	61.02
		DiffBIR	33.03	<u>0.7879</u>	0.2622		0.3427	0.7043	62.44
		PASD	32.79	0.7854	<u>0.2117</u>		0.4019	0.7208	74.18
		SUPIR	32.37	0.7533	0.2334		<u>0.4780</u>	<u>0.7231</u>	<u>74.45</u>
		DiffBody (ours)	30.11	0.7202	0.1402		0.4861	0.7561	75.71

in previous work [14]. Furthermore, it also demonstrates the inherent tradeoff between distortion and perception. As human viewers are especially sensitive to artifacts on human bodies, we primarily focus on perceptual quality to prioritize viewer comfort.

Qualitative Comparison: Fig. 3 shows visual comparisons on the SHHQ dataset using the highest degradation setting from Table 1 Row 5. Fig. 4 highlights finer details on human bodies, in which our model performs the best in skin texture and limb details. Additionally, Fig. 5 presents comparisons on real-world images from the Market1501 dataset [60] (no synthetic degradation). Our model achieves the best on fidelity and visual quality.

4.4 Ablation Studies

Effectiveness of LQ Image and Pose Conditioning in Stage 1: We evaluate the effectiveness of three different conditioning mechanisms on the low-quality (LQ) image and pose in our Stage 1 model, as shown in Table 2. Specifically, LQ Only uses only the LQ image as input to ControlNet without

TABLE 2

L_2 loss comparison of normal map and depth map. L_2^n represents the loss between normal map and ground truth, L_2^d represents the loss between depth map and ground truth.

Mode	L_2^n	L_2^d
LQ Only	151.9	531.2
LQ+Pose	180.2	561.8
LQ+Pose2U	106.8	488.7

the pose information. LQ+Pose (ours) feeds both the pose and the LQ image into the ControlNet. LQ+Pose2U sends the LQ image to the ControlNet while provide the pose for the U-Net. Both quantitative results and visual examples in Fig 6 show that the depth and normal maps generated by LQ+Pose2U are the closest to the ground truth.

Effectiveness of Human Body Priors: To evaluate the individual contributions of different human body priors in the restoration, we trained three separate models, each excluding one of the priors (without pose, without depth, and

Fig. 3. Qualitative comparison. Our model excels at generating fine body details, natural textures, and preserving the overall visual quality of the human body. Zoom in for details .

TABLE 3

Quantitative comparisons demonstrating the effectiveness of incorporating multiple priors. Notations follow those in Table 1. The model utilizing all priors achieves the overall best results.

Depth	Normal	Pose	PSNR	SSIM	LPIPS	#	ManIQA	ClipIQA	MUSIQ
X	X		28.72	0.7265	0.1907	0.4394	0.7603	73.7625	
X		X	30.25	0.7243	0.1986	0.4436	0.7498	71.0442	
	X	X	28.11	0.6924	0.2105	0.4332	0.7492	70.8363	
X	X	X	<u>30.11</u>	0.7402	0.1402	0.4861	<u>0.7561</u>	75.7115	

without normal), and compared their performance against our full model. Results are shown in Table 3. Our full model, leveraging all three priors, achieves the best overall performance. Visual comparisons in Fig. 7 demonstrate that incorporating pose enhances limb details, the normal map reveals skin textures, and the depth map improves the 3D spatial relationships between body parts.

Effectiveness of the Color Adapter: Finally, we evaluate the impact of incorporating the color adapter (color-Ada). Instead of PSNR and SSIM, we use CPSNR and CD-SSIM

TABLE 4

Qualitative comparison with and without the color adapter. The results show that incorporating the color adapter significantly enhances fidelity and overall visual quality.

Method	CPSNR	CD-SSIM	LPIPS	#	ManIQA	ClipIQA	MUSIQ
w/o color-Ada	24.31	0.6423	0.1872	0.5160	0.7410	72.9950	
w/ color-Ada	29.12	0.6821	0.1402	0.5380	0.7561	75.7115	

instead as they evaluate the distortion on all three color channels. Table 4 shows that the color adapter improves the performance on all quality metrics. Fig. 7 further demonstrates the color adapter's capability to achieve more accurate and faithful color recovery.

4.5 User Study

We conducted a user study to assess whether a method passes the viewer comfort test, as metrics like PSNR, LPIPS, or ManIQA cannot evaluate this aspect [61]. We processed 50 low-quality human body images using four diffusion-

Fig. 4. Qualitative comparison on limb details and skin textures. Our DiffBody model outperforms other state-of-the-art diffusion-based methods on human body images, particularly in limb details and skin textures. Zoom in for details.

Fig. 5. Qualitative comparison on real-world LQ images. DiffBody effectively restores human body details, enhancing real-world LQ images from 64×128 to 512×1024 high-quality (HQ) resolutions.

Fig. 6. Visual comparison of the generated normal and depth maps demonstrates that our LQ+Pose2U method achieves results that are most consistent with the ground truth, closely preserving the structural and geometric details evident in the original maps.

Fig. 7. Ablation studies on body priors: (First Row): Ablating the pose: Incorporating the pose leads to improved limb details. (Second Row): Ablating the normal map. Incorporating the normal map improves human skin textures. (Third Row): Ablating the depth map. Incorporating depth improves 3D spatial relationships in the generated images. (Fourth Row): Ablating in color adapter. Incorporating the color adapter significantly enhances detail and overall visual quality.

based methods, including ours, and presented them to 10 volunteers, who were given two questions: (1) "Rank the image quality from best (1st) to worst (4th)." (ranking-based comparison for best-performing method), and (2) "Select the best output from the four diffusion-based methods by evaluating each one based on its fidelity to the input image, overall quality, and viewer's comfort level." (choice-based comparison for best-performing method). The results are

shown in Fig. 8. Since GAN artifacts (e.g., poor quality, lack of detail) differ from diffusion models, we only present results from four diffusion-based methods for these two questions. Our method achieved the highest comfort pass rate (81.25%) and proportion of highest restoring quality (59.18%), outperforming other models.

body structure reconstruction and ensuring the preservation of personal identity throughout restoration. Future work will focus on handling more challenging scenarios, including complex poses, multi-human images, and cases where subjects are partially occluded. These extensions will further enhance the robustness and applicability of human image restoration models.

IMPACT STATEMENT

This paper presents work whose goal is to advance the field of Machine Learning. The ability to restore human images could lead to unwanted alterations of an individual's likeness, potentially infringing on personal rights. It is essential that this model is applied responsibly, with explicit consent, and that strong safeguards are in place to prevent misuse.

REFERENCES

- [1] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C.-W. Lin, "Deep learning on image denoising: An overview," *Neural Networks* vol. 131, pp. 251–275, 2020.
- [2] W. Wang, X. Wu, X. Yuan, and Z. Gao, "An experiment-based review of low-light image enhancement methods," *IEEE Access* vol. 8, pp. 87 884–87 917, 2020.
- [3] K. Zhang, W. Ren, W. Luo, W.-S. Lai, B. Stenger, M.-H. Yang, and H. Li, "Deep image deblurring: A survey," *International Journal of Computer Vision* vol. 130, no. 9, pp. 2103–2130, 2022.
- [4] A. Liu, Y. Liu, J. Gu, Y. Qiao, and C. Dong, "Blind image super-resolution: A survey and beyond," *IEEE transactions on pattern analysis and machine intelligence* vol. 45, no. 5, pp. 5461–5480, 2022.
- [5] X. Wang, Y. Li, H. Zhang, and Y. Shan, "Towards real-world blind face restoration with generative facial prior," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9168–9178.
- [6] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE transactions on pattern analysis and machine intelligence* vol. 44, no. 6, pp. 2872–2893, 2021.
- [7] J. Wang, S. Tan, X. Zhen, S. Xu, F. Zheng, Z. He, and L. Shao, "Deep 3d human pose estimation: A review," *Computer Vision and Image Understanding* vol. 210, p. 103225, 2021.
- [8] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1833–1844.
- [9] X. Wang, L. Xie, C. Dong, and Y. Shan, "Real-esrgan: Training real-world blind super-resolution with pure synthetic data," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1905–1914.
- [10] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *CoRR* vol. abs/1812.04948, 2018. [Online]. Available: <http://arxiv.org/abs/1812.04948>
- [11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *CoRR* vol. abs/2112.10752, 2021. [Online]. Available: <https://arxiv.org/abs/2112.10752>
- [12] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Refusion: Enabling large-size realistic image restoration with latent-space diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1680–1691.
- [13] X. Lin, J. He, Z. Chen, Z. Lyu, B. Fei, B. Dai, W. Ouyang, Y. Qiao, and C. Dong, "Diffbir: Towards blind image restoration with generative diffusion prior," *arXiv preprint arXiv:2308.15070*, 2023.
- [14] F. Yu, J. Gu, Z. Li, J. Hu, X. Kong, X. Wang, J. He, Y. Qiao, and C. Dong, "Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2024, pp. 25 669–25 680.
- [15] T. Yang, R. Wu, P. Ren, X. Xie, and L. Zhang, "Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization," *arXiv preprint arXiv:2308.14469*, 2023.

Fig. 8. User study. Questions include the viewer comfort pass test and image quality ranking. Results clearly demonstrate that our method significantly outperforms the others.

5 CONCLUSION

DiffBody introduces a novel framework for human body restoration, achieving realistic results by incorporating human body priors into the pre-trained Stable Diffusion model, surpassing the capabilities of existing general image restoration models in addressing artifacts. A key aspect of our approach is balancing different priors, such as pose, depth, and normal maps, to strike a balance between the viewer comfort and fidelity to the low-quality (LQ) image. However, there are still areas for improvement, such as exploring advanced techniques like mesh modeling for precise

- [16] Y. Blau and T. Michaeli, "The perception-distortion tradeoff," in *Proceedings of the IEEE conference on computer vision and pattern recognition* 2018, pp. 6228–6237.
- [17] Y. Liang, J. He, G. Li, P. Li, A. Klimovskiy, N. Carolan, J. Sun, J. Pont-Tuset, S. Young, F. Yang et al., "Rich human feedback for text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2024, pp. 19401–19411.
- [18] R. Khirodkar, T. Bagautdinov, J. Martinez, S. Zhaoen, A. James, P. Selednik, S. Anderson, and S. Saito, "Sapiens: Foundation for human vision models," in *European Conference on Computer Vision* Springer, 2024, pp. 206–228.
- [19] H. Wang, I. Katsavounidis, J. Zhou, J. Park, S. Lei, X. Zhou, M.-O. Pun, X. Jin, R. Wang, X. Wang et al., "Videoset: A large-scale compressed video quality dataset based on jnd measurement," *Journal of Visual Communication and Image Representation* vol. 46, pp. 292–302, 2017.
- [20] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition* 2018, pp. 586–595.
- [21] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang, "Maniqa: Multi-dimension attention network for no-reference image quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2022, pp. 1191–1200.
- [22] J. Wang, K. C. Chan, and C. C. Loy, "Exploring clip for assessing the look and feel of images," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 2555–2563.
- [23] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "Musiq: Multi-scale image quality transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* 2021, pp. 5148–5157.
- [24] A. Bora, A. Jalal, E. Price, and A. G. Dimakis, "Compressed sensing using generative models," in *International conference on machine learning* PMLR, 2017, pp. 537–546.
- [25] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, "Pulse: Self-supervised photo upsampling via latent space exploration of generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2020, pp. 2437–2445.
- [26] X. Pan, X. Zhan, B. Dai, D. Lin, C. C. Loy, and P. Luo, "Exploiting deep generative prior for versatile image restoration and manipulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* vol. 44, no. 11, pp. 7474–7489, 2021.
- [27] T. Yang, P. Ren, X. Xie, and L. Zhang, "Gan prior embedded network for blind face restoration in the wild," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 2021, pp. 672–681.
- [28] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems* vol. 33, pp. 6840–6851, 2020.
- [29] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems* vol. 32, 2019.
- [30] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 2022, pp. 10684–10695.
- [31] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125* vol. 1, no. 2, p. 3, 2022.
- [32] B. Kawar, M. Elad, S. Ermon, and J. Song, "Denoising diffusion restoration models," *Advances in Neural Information Processing Systems* vol. 35, pp. 23593–23606, 2022.
- [33] Y. Wang, J. Yu, and J. Zhang, "Zero-shot image restoration using denoising diffusion null-space model," *arXiv preprint arXiv:2212.00490* 2022.
- [34] Z. Yue and C. C. Loy, "Difface: Blind face restoration with diffused error contraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2024.
- [35] Z. Wang, Z. Zhang, X. Zhang, H. Zheng, M. Zhou, Y. Zhang, and Y. Wang, "Dr2: Diffusion-based robust degradation remover for blind face restoration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2023, pp. 1704–1713.
- [36] P. Yang, S. Zhou, Q. Tao, and C. C. Loy, "Pgdiff: Guiding diffusion models for versatile face restoration via partial guidance," *Advances in Neural Information Processing Systems* vol. 36, pp. 32194–32214, 2023.
- [37] Y. Miao, J. Deng, and J. Han, "Waveface: Authentic face restoration with efficient frequency recovery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2024, pp. 6583–6592.
- [38] M. Suin, N. G. Nair, C. Pong Lau, V. M. Patel, and R. Chellappa, "Diffuse and Restore: A Region-Adaptive Diffusion Model for Identity-Preserving Blind Face Restoration," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. Los Alamitos, CA, USA: IEEE Computer Society, Jan. 2024, pp. 6331–6340. [Online]. Available: <https://doi.ieeecomputersociety.org/10.1109/WACV57701.2024.00622>
- [39] S. Zhu, R. Urtasun, S. Fidler, D. Lin, and C. Change Loy, "Be your own prada: Fashion synthesis with structural coherence," in *Proceedings of the IEEE international conference on computer vision* 2017, pp. 1680–1688.
- [40] A. Siarohin, S. Lathuilliere, E. Sangineto, and N. Sebe, "Appearance and pose-conditioned human image generation using deformable gans," *IEEE transactions on pattern analysis and machine intelligence* vol. 43, no. 4, pp. 1156–1171, 2019.
- [41] Y. Ren, X. Yu, J. Chen, T. H. Li, and G. Li, "Deep image spatial transformation for person image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2020, pp. 7690–7699.
- [42] L. Yang, P. Wang, C. Liu, Z. Gao, P. Ren, X. Zhang, S. Wang, S. Ma, X. Hua, and W. Gao, "Towards fine-grained human pose transfer with detail replenishing network," *IEEE Transactions on Image Processing* vol. 30, pp. 2422–2435, 2021.
- [43] P. Roy, S. Ghosh, S. Bhattacharya, U. Pal, and M. Blumenstein, "Tips: Text-induced pose synthesis," in *European Conference on Computer Vision* Springer, 2022, pp. 161–178.
- [44] Y. Jiang, S. Yang, H. Qiu, W. Wu, C. C. Loy, and Z. Liu, "Text2human: Text-driven controllable human image generation," *ACM Transactions on Graphics (TOG)* vol. 41, no. 4, pp. 1–11, 2022.
- [45] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* 2023, pp. 3836–3847.
- [46] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," *arXiv preprint arXiv:2302.08453* 2023.
- [47] X. Ju, A. Zeng, C. Zhao, J. Wang, L. Zhang, and Q. Xu, "Humansd: A native skeleton-guided diffusion model for human image generation," *arXiv preprint arXiv:2304.04269* 2023.
- [48] X. Liu, J. Ren, A. Siarohin, I. Skorokhodov, Y. Li, D. Lin, X. Liu, Z. Liu, and S. Tulyakov, "Hyperhuman: Hyper-realistic human generation with latent structural diffusion," *arXiv preprint arXiv:2310.08579* 2023.
- [49] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns," *IEEE Sensors Journal* vol. 20, no. 17, pp. 10032–10044, 2020.
- [50] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems* vol. 36, 2024.
- [51] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721* 2023.
- [52] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE transactions on pattern analysis and machine intelligence* vol. 44, no. 3, pp. 1623–1637, 2020.
- [53] A. Eftekhari, A. Sax, J. Malik, and A. Zamir, "Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* 2021, pp. 10786–10796.
- [54] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952* 2023.
- [55] S. Li, J. Fu, K. Liu, W. Wang, K.-Y. Lin, and W. Wu, "Cosmicman: A text-to-image foundation model for humans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2024, pp. 6955–6965.
- [56] J. Fu, S. Li, Y. Jiang, K.-Y. Lin, C. Qian, C. C. Loy, W. Wu, and Z. Liu, "Stylegan-human: A data-centric odyssey of human generation," in *European Conference on Computer Vision* Springer, 2022, pp. 1–19.
- [57] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models,"

