SKETCHHMR: GENERATIVE PROJECTION GUIDANCE FOR HUMAN MESH RECOVERY FROM FREEHAND SKETCHES

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing a 3D human mesh from a freehand sketch is complicated by the inherent sparsity and style variance of hand-drawn strokes, which offer insufficient cues regarding body surface, joint ordering, occlusion, and depth. To bypass the ambiguity of single-stage regressors that map raw strokes directly to parametric body dimensions, we propose SketchHMR, a two-stage paradigm that decouples sketch interpretation from mesh estimation via an intermediate, camera-aligned generated SMPL projection. The first stage, the core component, GeoInk, combines a spatially aligned sketch latent with DINOv2 visual features and input-anchored guidance to translate the sketch into a structured body observation. The second stage iteratively estimates SMPL pose, shape, and weak-perspective camera from the generated projection. To reduce the domain gap between clean rendered projections and generated projections, we pretrain the regressor on clean inputs and subsequently adapt it using paired clean and generated projections. On the Sketch3D dataset, SketchHMR achieves 96.60 mm MPJPE, 66.10 mm PA-MPJPE, and 115.90 mm PVE. Reducing flow sampling from 25 to 10 steps also decreases translation latency by 45.2% without degrading reconstruction accuracy. SketchHMR provides a modular and interpretable approach to sketch-based human mesh recovery by making sketch interpretation an explicit intermediate stage.

1 INTRODUCTION

Freehand sketches offer an intuitive way to specify human pose, but recovering a poseable 3D mesh from them is difficult. Unlike RGB photographs, sketches provide little texture or shading and may omit limb boundaries or occluded body parts. Their sparse, stylized strokes also leave depth and body-part ordering ambiguous.

Most HMR methods directly map image features to the pose and shape parameters of a body model such as SMPL (Loper et al., 2015; Kanazawa et al., 2018). SketchBodyNet extends this direct-regression paradigm to freehand drawings through separate decoders for pose, shape, and camera (Wang et al., 2023). Although direct regression offers a compact pipeline, it requires one network to solve two distinct problems simultaneously: interpreting sparse and ambiguous strokes as a coherent body observation, and recovering valid parametric body geometry from that observation. The former is dominated by cross-domain ambiguity and missing visual evidence, whereas the latter is a structured regression problem constrained by the human-body model. Coupling these mappings makes it difficult to determine whether a reconstruction error originates from sketch interpretation or SMPL parameter estimation. This motivates a central question: can an explicit intermediate body observation bridge the representational gap between freehand strokes and parametric human mesh recovery?

We address this question with SketchHMR, a two-stage framework that introduces a generated, camera-aligned SMPL projection between the input sketch and the final mesh. The first stage, termed GeoInk, translates the freehand sketch into a grayscale projection of the corresponding SMPL body. Compared with sparse keypoints or a binary silhouette, this representation retains a richer combination of body scale, silhouette, limb relationships, self-occlusion, and coarse surface

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

structure. At the same time, it suppresses nuisance appearance factors such as clothing texture, illumination, and background. GeoInk combines a spatially aligned sketch latent with visual semantic features to construct this structured body observation. The second stage then applies an iterative regressor to estimate SMPL pose, shape, and weak-perspective camera from the generated projection.

The generated projection creates a mismatch between clean projections used to pretrain the regressor and GeoInk outputs received at inference. We address this gap by adapting the clean-pretrained regressor with paired clean and generated projections that share the same SMPL labels. This retains accurate clean observations during training while exposing the regressor to generation artifacts.

Experiments on Sketch3D show that SketchHMR achieves 96.60 mm MPJPE under the ten-step configuration and outperforms the matched direct-regression baselines. The results also demonstrate that the explicit projection bridge makes translation failures observable and enables efficient sketch-to-projection sampling.

This work makes the following contributions:

- We formulate freehand-sketch-based HMR as a two-stage problem and introduce a generated, camera-aligned SMPL projection as an explicit intermediate observation between sparse strokes and parametric human-body regression.
- We develop a paired mixed-domain adaptation strategy that combines geometrically accurate clean projections with deployment-matched generated projections, reducing the projection-domain gap while preserving a stable geometric anchor.
- We provide a diagnostic analysis of the resulting pipeline, separating projection-generation error from downstream regression error and identifying an efficient sampling configuration for sketch-to-projection translation.

2 Related Work

2.1 Sketch-Based Human Reconstruction

Sketch2Pose predicts 2D bone tangents, self-contacts, and foreshortening before optimizing a 3D character pose (Brodt & Bessmeltsev, 2022). Interactive systems infer and refine mannequin poses from stylized drawings (Ünlü et al., 2022). More recently, Sketch2PoseNet combines synthesized training sketches with generative visual priors to estimate human pose and shape across sketch styles (Wang et al., 2025). SketchBodyNet uses a shared encoder and separate attention decoders to estimate SMPL pose, shape, and camera directly from freehand sketches (Wang et al., 2023). SketchHMR instead generates a camera-aligned SMPL projection before parameter regression, exposing the inferred body layout as an intermediate result.

2.2 Human Mesh Recovery

SMPL maps pose and shape parameters to a 3D human mesh (Loper et al., 2015). Image-based HMR methods have developed direct regression and optimization-based training (Kanazawa et al., 2018; Kolotouros et al., 2019a), as well as mesh convolution (Kolotouros et al., 2019b), part-aware features (Kocabas et al., 2021), inverse kinematics (Li et al., 2021), transformer decoding (Goel et al., 2023), dense surface cues (Lê et al., 2024), and discrete pose representations (Dwivedi et al., 2024). These methods improve the extraction or regression of body geometry from visual inputs. SketchHMR retains an iterative SMPL regressor and instead changes the observation supplied to it.

2.3 Intermediate Representations and Visual Translation

Neural Body Fitting uses semantic part segmentations (Omran et al., 2018); STRAPS uses silhouettes and 2D joints with corrupted synthetic inputs (Sengupta et al., 2020); DenseRaC combines estimated and rendered IUV maps during training (Xu et al., 2019). SketchHMR instead adapts its regressor using paired clean and generated SMPL projections. GeoInk generates camera-aligned grayscale projections using a pretrained VAE latent space (Rombach et al., 2022), flow matching (Lipman et al., 2023), and DINOv2 features (Oquab et al., 2024).

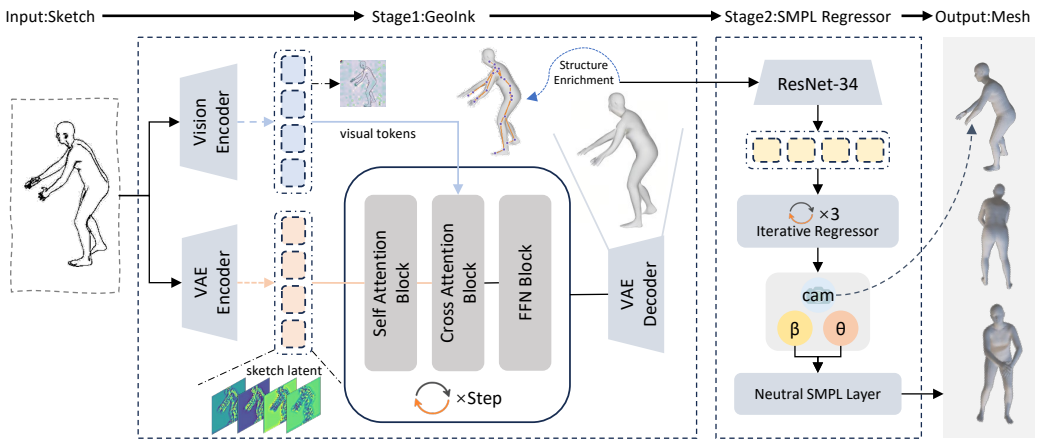

Figure 1: Overview of SketchHMR. GeoInk translates the input sketch into a camera-aligned SMPL projection using frozen VAE and vision encoders with a latent-flow Transformer. A ResNet-34-based iterative regressor estimates SMPL pose $\boldsymbol{\theta}$, shape $\boldsymbol{\beta}$, and weak-perspective camera $\boldsymbol{\pi}$ through three residual updates, after which a frozen neutral SMPL layer produces the final mesh from the predicted pose and shape.

3 Methodology

3.1 Framework Overview

Given a freehand human sketch x, our goal is to recover the SMPL pose $\boldsymbol{\theta}$, shape $\boldsymbol{\beta}$, and weak-perspective camera $\boldsymbol{\pi} = [s, t_x, t_y]$. Rather than directly regressing these parameters from sparse and style-dependent strokes, SketchHMR decomposes the task into structured body-observation generation and parametric body regression.

As illustrated in Fig. 1, Stage 1 uses GeoInk to translate the sketch into a generated, camera-aligned SMPL projection $\hat{y}$. Stage 2 then estimates the SMPL parameters from this structured observation:

$$\hat{y} = G_\phi(x), \qquad (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}) = R_\psi(\hat{y}), \tag{1}$$

where G_ϕ denotes the projection generator and R_ψ denotes the projection-to-SMPL regressor. The predicted pose and shape are subsequently passed through a frozen neutral SMPL layer to obtain the final mesh vertices and body joints.

The clean target projection is rendered from the annotated neutral SMPL body under a scaled-orthographic camera with a white background and uniform gray material. It preserves the body geometry and image-plane configuration required for regression while suppressing appearance-related nuisance factors. Here, camera-aligned indicates that the generated projection retains the location, scale, and viewing configuration expressed by the sketch without using ground-truth camera parameters at inference. The two stages are trained sequentially, and only the generated-projection pathway is retained during inference.

3.2 GeoInk: Structured SMPL Projection Generation

GeoInk translates sketch latents and visual tokens into camera-aligned SMPL projections. Figure 2 shows the decoded latent states after successive Euler updates.

Structural and semantic conditioning. A frozen Stable-Diffusion-2.1 variational autoencoder (Rombach et al., 2022) encodes the input sketch x and clean target projection y into latent tensors z_x and z_y. Each 256×256 image is represented by a $4 \times 32 \times 32$ latent. The sketch latent z_x acts as a spatial structural condition and retains image-plane position, body scale, stroke layout, and coarse body configuration.



Figure 2: GeoInk sampling trajectories from input sketches to camera-aligned SMPL projections at Euler steps 4, 6, 8, and 10.

In parallel, the input sketch is resized to 448×448 and processed by a frozen DINOv2-Base encoder (Oquab et al., 2024). We extract local visual tokens h_x from Transformer block index 8, with a feature dimension of 768. These tokens provide complementary information about global pose and human-part relationships and are injected into the latent Transformer through cross-attention. The VAE latent therefore preserves spatial layout, while the DINOv2 tokens provide higher-level semantic guidance.

Latent-flow generation. Following flow matching (Lipman et al., 2023), we sample a timestep $t \sim \mathcal{U}(0,1)$ and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$. The interpolation path and its target velocity are defined as

$$z_t = (1-t)z_y + t\epsilon, \qquad v^\star = \epsilon - z_y. \tag{2}$$

GeoInk is trained with a fully conditioned branch and a weakly conditioned branch. Their structural and semantic conditions are

$$(c^\kappa_{\text{str}}, c^\kappa_{\text{sem}}) = \begin{cases} (z_x, h_x), & \kappa = c, \\ (\alpha z_x, \mathbf{0}), & \kappa = w, \end{cases} \qquad \alpha \sim \mathcal{U}(0.05, 0.25), \tag{3}$$

where the weakly conditioned branch is selected for 10% of the training samples. Unlike a completely unconditional branch, it retains a reduced sketch latent and therefore remains spatially anchored to the input body layout.

The latent Transformer receives the channel-wise concatenation of the structural condition and the current target state:

$$\hat{v}^\kappa_t = f_\phi([c^\kappa_{\text{str}}, z_t], t, c^\kappa_{\text{sem}}). \tag{4}$$

It is optimized using the flow-matching objective

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{x,y,t,\epsilon,\kappa}\left[\|\hat{v}^\kappa_t - v^\star\|_2^2\right]. \tag{5}$$

The latent Transformer contains 12 blocks with a hidden width of 512, eight attention heads, and 2×2 latent patches. Each block consists of self-attention, cross-attention, and a feed-forward network. The spatial input contains eight channels, consisting of four sketch-condition channels and four target-state channels. The network predicts a four-channel velocity field. The VAE and DINOv2 encoders remain frozen, while the latent Transformer is optimized for sketch-to-projection translation.

Input-anchored guidance and sampling. At inference, GeoInk starts from a Gaussian latent and integrates the learned velocity field from $t = 1$ to $t = 0$. We adapt the combination rule of classifier-free guidance (Ho & Salimans, 2022) to fully and weakly conditioned velocity predictions. The fully conditioned prediction v_c uses the complete sketch latent and semantic tokens, while the weakly conditioned prediction v_w uses a reduced sketch latent and zero semantic tokens. The guided velocity is

$$v_{\text{guide}} = v_w + s(v_c - v_w), \tag{6}$$

where s is the guidance scale. Both branches retain input-dependent spatial information.

For a linear Euler schedule $1 = t_0 > t_1 > \cdots > t_N = 0$, the latent state is updated as

$$z^{(i+1)} = z^{(i)} + (t_{i+1} - t_i)v_{\text{guide}}. \tag{7}$$

The final latent is decoded by the frozen VAE decoder to obtain the generated projection $\hat{y}$. The number of sampling steps, guidance strength, and adopted main configuration are specified in the experimental setup and evaluated in the sampling ablation.

3.3 Iterative Projection-to-SMPL Regression

The second stage estimates SMPL pose, shape, and weak-perspective camera from either a clean or generated SMPL projection. Since the projection already suppresses most nuisance appearance and exposes body-related spatial evidence, we use a conventional ResNet-34-based iterative regression backend rather than a specialized mesh architecture.

An ImageNet-pretrained ResNet-34 (He et al., 2016; Deng et al., 2009) extracts a 512-dimensional global feature $F(I)$, where I denotes a clean or generated projection. The regression state is defined as $q^k = [\boldsymbol{\theta}^k_{6\text{D}}, \boldsymbol{\beta}^k, \boldsymbol{\pi}^k]$. The pose is represented by 24 continuous 6D rotations, shape contains ten SMPL coefficients, and the weak-perspective camera contains scale and two-dimensional translation. The initial state is set to the mean pose, shape, and camera parameters.

Following iterative HMR regression (Kanazawa et al., 2018; Kolotouros et al., 2019a), the state is refined through three residual updates:

$$q^{k+1} = q^k + g_\psi\big([F(I), q^k]\big), \qquad k = 0, 1, 2. \tag{8}$$

The update function contains two fully connected layers followed by separate pose, shape, and camera heads. All layers and output heads are shared across the three iterations, so the process repeatedly applies one regressor rather than three independently parameterized stages.

The predicted 6D rotations are converted into rotation matrices through Gram–Schmidt orthogonalization (Zhou et al., 2019). A frozen neutral SMPL layer then produces 6,890 vertices and 24 joints. The predicted 2D joints are obtained with the same weak-perspective projection model used to construct the training targets.

For branch $b \in \{c, g\}$, denoting clean and generated projection inputs, respectively, the regression loss is

$$\begin{aligned}\mathcal{L}_b = \mathcal{L}_{\text{pose}} + 0.01\rho_b^\beta \mathcal{L}_{\text{shape}} + 0.1\mathcal{L}_{\text{cam}} \\ + 5\mathcal{L}_{\text{2D}} + 5\mathcal{L}_{\text{3D}} + 0.1\rho_b^V \mathcal{L}_V.\end{aligned} \tag{9}$$

The pose, shape, and camera terms use mean-squared error. The 2D- and 3D-joint terms use Smooth-L_1 distances, while the vertex term uses an L_1 distance. Three-dimensional joints and vertices are pelvis-centered before their losses are computed. For clean projections, we set $\rho_c^\beta = \rho_c^V = 1$. Because generated projections may contain larger local contour and surface inconsistencies, we empirically use $\rho_g^\beta = \rho_g^V = 0.25$ for the generated branch while retaining the full pose, camera, 2D-joint, and 3D-joint weights.

3.4 Paired Mixed-Domain Adaptation

We pretrain the projection-to-SMPL regressor on clean projections, then adapt it with paired clean and generated projections sharing the same SMPL annotations. The generated inputs expose the regressor to the observations it receives at inference.

During adaptation, both domains are processed by the same ResNet-34 encoder and iterative regressor. The mixed-domain objective at epoch e, indexed from one, is

$$\mathcal{L}_{\text{mix}}(e) = \mathcal{L}_c + \lambda_e \mathcal{L}_g, \qquad \lambda_e = \min\left(1, \frac{e}{10}\right). \tag{10}$$

The generated-domain weight λ_e increases from 0.1 to 1.0 over the first ten epochs. Generated projections are cached before adaptation, so regression gradients do not propagate through GeoInk. At inference, only generated projections are passed to the regressor.

4 Experiments

4.1 Experimental Setup

Dataset and Protocol. We conduct all experiments on Sketch3D dataset (Wang et al., 2023), which provides freehand human sketches together with SMPL pose, shape, and weak-perspective camera annotations. After excluding samples that could not be processed by the complete sketch-to-mesh pipeline, the resulting dataset contains 12,399 sketch–body pairs. For each annotation, we render a clean grayscale SMPL projection using the neutral SMPL model, a scaled-orthographic camera, a white background, and a uniform gray body material. Both sketches and SMPL projections are resized to 256×256.

We use a seeded 90/10 split containing 11,159 training samples and 1,240 validation samples. The random seed is fixed to 42 for all experiments. After training GeoInk, we generate and cache one SMPL projection for every training and validation sketch. Unless otherwise specified, the main configuration uses 10 Euler sampling steps and a guidance scale of $s = 2$. All projection-to-SMPL models evaluated under the main setting receive the same generated projection cache.

Evaluation Metrics. We report mean per-joint position error (MPJPE), Procrustes-aligned MPJPE (PA-MPJPE), and per-vertex error (PVE), all measured in millimeters. MPJPE and PVE are calculated after subtracting the SMPL pelvis, whereas PA-MPJPE additionally applies an independent similarity alignment to each prediction. We evaluate all 24 joints returned by the SMPL layer and all 6,890 mesh vertices. Because many published HMR results use different 14- or 17-joint regressors, those values are not directly comparable with our 24-SMPL-joint protocol. We additionally evaluate generated projections against their paired clean SMPL renderings using LPIPS (Zhang et al., 2018) and foreground-mask intersection-over-union (IoU).

The 2D-joint supervision is recomputed from the annotated SMPL pose, shape, and weak-perspective camera using the same scaled-orthographic projection adopted by the renderer. We do not use legacy 2D keypoint fields defined in the original image coordinates because both sketches and projections are resized during preprocessing.

Implementation Details. GeoInk contains a 54.96-million-parameter latent Transformer and is trained for 100,000 optimizer steps on one NVIDIA RTX 4090 GPU. The Stable-Diffusion-2.1 VAE and DINOv2-Base encoder remain frozen. We use a batch size of 32, bfloat16 arithmetic, and AdamW (Loshchilov & Hutter, 2019) with a learning rate of 2×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.01, gradient-norm clipping at 1.0, and EMA decay 0.9999. The latent Transformer has a hidden width of 512, 12 blocks, eight attention heads, and 2×2 latent patches.

All projection-to-SMPL regressors use bfloat16 arithmetic, AdamW with weight decay 10^{-4}, gradient-norm clipping at 1.0, and cosine learning-rate decay. The clean-projection regressor is initialized from an ImageNet-pretrained ResNet-34 and trained before mixed-domain adaptation. Mixed-domain adaptation is initialized from the clean-domain checkpoint and uses paired clean and generated projections. Checkpoints are selected according to generated-input validation MPJPE.

4.2 Comparison with Existing HMR Methods

Quantitative Comparison. We compare SketchHMR with representative human-mesh-recovery methods, including HMR (Kanazawa et al., 2018), MeshPose (Lê et al., 2024), Sketch2Pose (Brodt & Bessmeltsev, 2022), SPIN (Kolotouros et al., 2019a), CMR (Kolotouros et al., 2019b), MAED



Figure 3: Qualitative comparison on Sketch3D. Each row shows the input sketch, baseline predictions, the GeoInk intermediate projection, the SketchHMR result, and the ground-truth mesh. Red ellipses indicate representative reconstruction errors. MeshPose* uses its official pretrained checkpoint without fine-tuning, and its output is mapped to the full SMPL vertex space for visualization.

(Wan et al., 2021), HybrIK (Li et al., 2021), and TokenHMR (Dwivedi et al., 2024). Except for MeshPose, the comparison methods are trained on Sketch3D using the common data split, input preprocessing, supervision targets, checkpoint-selection criterion, and evaluation metrics, while retaining their corresponding project architectures. MeshPose is evaluated using the officially released pretrained checkpoint and inference pipeline without fine-tuning on Sketch3D, and is therefore included as an off-the-shelf zero-shot reference. Its native low-resolution mesh is converted to the full SMPL vertex space using the official mesh mapping before evaluation.

Table 1 reports the quantitative comparison. Under the adopted ten-step, guidance-scale-2 configuration, SketchHMR obtains 96.60 mm MPJPE. Among the comparison methods trained on Sketch3D, MAED achieves the lowest reconstruction errors. SketchHMR nevertheless obtains the best performance across all three metrics under the matched training and evaluation protocol.

Qualitative Comparison. Figure 3 compares the reconstruction results across a range of body orientations, articulation patterns, and self-occlusion conditions. Direct sketch-to-SMPL methods can recover plausible overall human shapes but may fail to preserve geometric evidence expressed by the sketch, including arm orientation, limb separation, body proportion, depth ordering, and local articulation. Red ellipses highlight representative regions in which the predicted mesh deviates from the input sketch or ground-truth body.

Across the displayed examples, SketchHMR more consistently preserves global articulation and the relative arrangement of body parts. The difference is particularly visible for side views, seated poses,

Table 1: Quantitative comparison on Sketch3D. MeshPose* uses the official pretrained checkpoint without fine-tuning; all other methods are trained under the common protocol. Errors are reported in millimeters.

Method	MPJPE ↓	PA-MPJPE ↓	PVE ↓
HMR (CVPR'18)	130.06	81.19	155.29
MeshPose* (CVPR'24)	244.04	145.88	285.18
Sketch2Pose (TOG'22)	116.35	75.28	139.44
SPIN (ICCV'19)	119.23	77.72	143.25
CMR (CVPR'19)	136.65	85.02	163.39
MAED (ICCV'21)	108.88	71.41	131.43
HybrIK (CVPR'21)	147.25	100.31	178.74
TokenHMR (CVPR'24)	230.29	129.75	279.76
Ours	**96.60**	**66.10**	**115.90**

Table 2: Ablation of the projection bridge and mixed-domain adaptation on Sketch3D. Direct HMR is trained and evaluated on sketches. Both bridge variants are evaluated on the same GeoInk-generated projections: clean-only uses clean-projection pretraining, while mixed-domain additionally uses paired clean/generated adaptation. Errors are in millimeters; lower is better.

Method	MPJPE ↓	PA-MPJPE ↓	PVE ↓
Direct HMR (sketch)	130.06	81.19	155.29
Bridge (clean-only)	111.76	73.92	132.75
Bridge (mixed-domain)	**96.60**	**66.10**	**115.90**

raised arms, crossed limbs, and poses involving strong self-occlusion, for which several direct methods merge adjacent limbs or recover an incorrect body orientation. The GeoInk column exposes the intermediate interpretation received by the second-stage regressor. When the generated projection preserves the silhouette and limb relationships expressed by the sketch, the iterative regressor generally produces a mesh close to the ground truth. Errors already present in the generated projection may instead propagate into the final reconstruction, showing that the explicit bridge improves interpretability without eliminating cascade errors.

4.3 Ablation Studies

Projection Bridge and Mixed-Domain Adaptation. Table 2 compares direct sketch regression with two SketchHMR variants. *Direct HMR* is trained and evaluated on sketches. *Bridge (clean-only)* uses GeoInk-generated projections at inference, but its regressor is trained only on clean rendered projections. *Bridge (mixed-domain)* starts from the same clean-pretrained regressor and further adapts it using paired clean and generated projections. Both bridge variants are evaluated on the same generated projections.

Direct HMR obtains 130.06 mm MPJPE, whereas Bridge (clean-only) obtains 111.76 mm. This result shows that the two-stage pipeline improves reconstruction over direct sketch regression under our evaluation protocol. Because the two systems differ in both input representation and training procedure, this comparison measures the benefit of the overall bridge pipeline.

Paired mixed-domain adaptation further reduces MPJPE from 111.76 to 96.60 mm, PA-MPJPE from 73.92 to 66.10 mm, and PVE from 132.75 to 115.90 mm. Since both bridge variants use the same generated projections at evaluation, these results support the benefit of adapting the regressor to GeoInk outputs.

Sampling Steps and Guidance Scale. We vary the number of Euler steps with the guidance scale fixed to $s = 2$, and then compare guidance scales of 1, 2, and 3 using 10 steps. The generator checkpoint, input sketches, initial noise, and downstream regressor are fixed across all settings. As shown in Table 3, increasing the number of steps from 10 to 25 or 50 increases latency without improving projection similarity or reconstruction accuracy. With 10 steps, $s = 2$ substantially

Table 3: GeoInk sampling ablation on Sketch3D. The upper block varies Euler steps with guidance fixed at 2, and the lower block varies guidance with 10 steps. Time is measured per image, and MPJPE is reported in millimeters.

Steps	Guidance	Time (ms) ↓	LPIPS ↓	IoU ↑	MPJPE ↓
10	2	**46.14**	**0.1287**	**0.6974**	**96.60**
25	2	84.15	0.1301	0.6957	97.09
50	2	139.97	0.1309	0.6952	97.59
10	1	48.63	0.1337	0.6928	106.94
10	2	**46.14**	**0.1287**	**0.6974**	96.60
10	3	46.85	0.1301	0.6970	**96.07**

Table 4: Backbone comparison for the mixed-domain projection regressor. Reconstruction errors are reported in millimeters, and lower is better.

Backbone	Params. (M)	MPJPE ↓	PA-MPJPE ↓	PVE ↓
ResNet-18	13.07	102.62	69.28	122.39
ResNet-34	23.18	**96.60**	**66.10**	**115.90**
ResNet-50	26.98	100.66	68.15	120.46

outperforms $s = 1$ and provides nearly the same MPJPE as $s = 3$ while achieving better LPIPS and mask IoU. We therefore adopt 10 steps and $s = 2$ as the default configuration.

Regressor Backbone. We evaluate the influence of the projection-regressor backbone while keeping the iterative regression architecture, loss functions, generated projection inputs, and mixed-domain training protocol unchanged.

As shown in Table 4, ResNet-34 reduces MPJPE from 102.62 to 96.60 mm compared with ResNet-18, corresponding to a relative improvement of 5.87%. ResNet-50 has more parameters than ResNet-34 but performs worse across all three reconstruction metrics. These results indicate that increasing backbone capacity does not necessarily improve reconstruction performance under the current training setting. We therefore adopt ResNet-34 as the default projection-regressor backbone.

5 Conclusion

We presented SketchHMR, a two-stage framework that recovers a parametric human body from a freehand sketch through a generated, camera-aligned SMPL projection. Paired mixed-domain adaptation improves the regressor's robustness to generated observations, enabling SketchHMR to outperform direct regression baselines on Sketch3D. The explicit projection bridge also makes intermediate interpretation errors observable, although such errors may propagate to the final reconstruction. Future work will investigate more efficient sampling and multi-hypothesis projections for ambiguous sketches.

AI use statement

We used an LLM-based coding assistant (Codex) for implementation and debugging, data preparation and alignment, baseline adaptation, automation of training and ablation runs, and preparation of evaluation tables and visualizations. Codex was not used to develop the research methodology or design the experiments. The authors reviewed all AI-assisted code, validated the implementation through training and evaluation, and checked the experimental configurations and reported results against experiment records. The authors take full responsibility for the methodology, implementation, experimental results, interpretations, and final content of this work.

REFERENCES

Kirill Brodt and Mikhail Bessmeltsev. Sketch2Pose: Estimating a 3d character pose from a bitmap sketch. *ACM Transactions on Graphics*, 41(4):85:1–85:15, 2022. doi: 10.1145/3528223.3530106.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J. Black. TokenHMR: Advancing human mesh recovery with a tokenized pose representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1323–1333, 2024.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14783–14794, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7122–7131, 2018.

Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11127–11137, 2021.

Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2252–2261, 2019a.

Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4501–4510, 2019b.

Eric-Tuan Lê, Antonis Kakolyris, Petros Koutras, Himmy Tam, Efstratios Skordos, George Papandreou, Rıza Alp Güler, and Iasonas Kokkinos. MeshPose: Unifying DensePose and 3d body mesh reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybrIK: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3383–3393, 2021.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics*, 34(6):248:1–248:16, 2015. doi: 10.1145/2816795.2818013.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model-based human pose and shape estimation. In *International Conference on 3D Vision (3DV)*, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Synthetic training for accurate 3D human pose and shape estimation in the wild. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2020.

Gizem Ünlü, Mohamed Sayed, and Gabriel J. Brostow. Interactive sketching of mannequin poses. In *International Conference on 3D Vision*, pp. 700–710, 2022. doi: 10.1109/3DV57658.2022.00080.

Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13033–13042, 2021.

Fei Wang, Kongzhang Tang, Hefeng Wu, Baoquan Zhao, Hao Cai, and Teng Zhou. SketchBodyNet: A sketch-driven multi-faceted decoder network for 3d human reconstruction. In *Pacific Graphics Short Papers and Posters*, pp. 11–19. The Eurographics Association, 2023. doi: 10.2312/pg.20231266.

Li Wang, Yiyu Zhuang, Yanwen Wang, Xun Cao, Chuan Guo, Xinxin Zuo, and Hao Zhu. Sketch2PoseNet: Efficient and generalized sketch to 3D human pose prediction. In *SIGGRAPH Asia 2025 Conference Papers*. Association for Computing Machinery, 2025. doi: 10.1145/3757377.3763855.

Yuanlu Xu, Song-Chun Zhu, and Tony Tung. DenseRaC: Joint 3D pose and shape estimation by dense render-and-compare. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753, 2019.