Sketch2PoseNet: Efficient and Generalized Sketch to 3D Human Pose Prediction

LI WANG, Nanjing University, China
YIYU ZHUANG, Nanjing University, China
YANWEN WANG, Nanjing University, China
XUN CAO, Nanjing University, China
CHUAN GUO, Snap Inc., USA
XINXIN ZUO, Concordia University, Canada
HAO ZHU*, Nanjing University, China

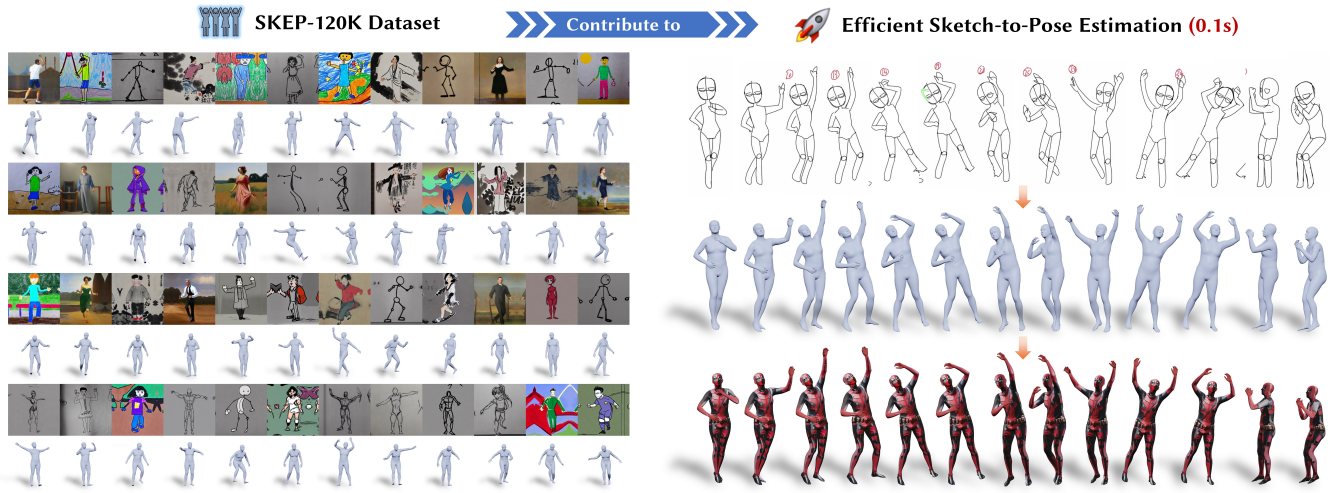

Fig. 1. We present a novel approach for 3D human pose estimation from sketches. Benefiting from the large-scale SKEP-120K dataset (left), we propose to learn a data-driven sketch-to-pose model that exhibits improved generalization ability and efficient inference (right).

3D human pose estimation from sketches has broad applications in computer animation and film production. Unlike traditional human pose estimation, this task presents unique challenges due to the abstract and disproportionate nature of sketches. Previous sketch-to-pose methods, constrained by the lack of large-scale sketch-3D pose annotations, primarily relied on optimization with heuristic rules—an approach that is both time-consuming and limited in generalizability. To address these challenges, we propose a novel approach leveraging a "*learn from synthesis*" strategy. Firstly, a diffusion model is learned to synthesize sketch images from 2D poses projected from 3D human poses, mimicking disproportionate human structures in sketches. This process enables the creation of a synthetic dataset, SKEP-120K , consisting of *120k* accurate sketch-3D pose annotation pairs across various sketch styles. Building on this synthetic dataset, we introduce an end-to-end data-driven framework for estimating human poses and shapes from diverse sketch styles. Our framework combines existing 2D pose detectors and generative diffusion priors for sketch feature extraction with a feed-forward neural network for efficient 2D pose estimation. Multiple heuristic loss functions have been incorporated to guarantee geometric coherence between the derived 3D poses and the detected 2D poses while preserving accurate self-contacts. Qualitative, quantitative, and subjective evaluations collectively affirm that our proposed model substantially surpasses previous ones in both estimation accuracy and speed for sketch-to-pose tasks.

CCS Concepts: • **Computing methodologies** → **Motion capture**; **Mesh geometry models**; **Shape inference**.

Additional Key Words and Phrases: Motion Capture, Sketch-based Modeling, Character Posing

ACM Reference Format:
Li Wang, Yiyu Zhuang, Yanwen Wang, Xun Cao, Chuan Guo, Xinxin Zuo, and Hao Zhu. 2025. Sketch2PoseNet: Efficient and Generalized Sketch to 3D Human Pose Prediction. In *SIGGRAPH Asia 2025 Conference Papers (SA*

*Corresponding Author.

Authors' Contact Information: Li Wang, Nanjing University, Nanjing, China, liwang1029@smail.nju.edu.cn; Yiyu Zhuang, Nanjing University, Nanjing, China, yiyu.zhuang@smail.nju.edu.cn; Yanwen Wang, Nanjing University, Nanjing, China, wangyanwen@smail.nju.edu.cn; Xun Cao, Nanjing University, Nanjing, China, caoxun@nju.edu.cn; Chuan Guo, Snap Inc., New York, USA, guochuan5513@gmail.com; Xinxin Zuo, Concordia University, Montreal, Canada, xinxin.zuo@concordia.ca; Hao Zhu, Nanjing University, Nanjing, China, zh@nju.edu.cn.

SA Conference Papers '25, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2137-3/25/12
https://doi.org/10.1145/3757377.3763855

Conference Papers '25), December 15–18, 2025, Hong Kong, Hong Kong. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3757377.3763855

1 Introduction

Human pose estimation holds significant importance and finds widespread application across numerous scenarios, including 3D human reconstruction [Tan et al. 2020; Zhu et al. 2016], 3D human generation [Wang et al. 2025; Zeng et al. 2023; Zhuang et al. 2025], view synthesis [Zhu et al. 2018], and animation [Liao et al. 2020; Zhu et al. 2024a]. Among the various sources used for pose estimation, sketches emerge as an efficient and versatile entity. Sketches are data that can be more easily designed by artists and are widely used in animation and film production. More broadly, the term 'sketch' encompasses a diverse range of graphical styles, including charcoal sketches, cartoons, stick figures, kids' drawings, oil paintings, ink paintings, and so forth.

Estimating human poses from sketches presents a significant challenge. Generalized photo-based pose estimation methods fall short in this task due to their exclusive training on realistic data. By contrast, sketches often disregard human proportionality and geometric perspective, opting for a more abstract representation of poses, thereby exacerbating the complexity of the sketch-to-pose conversion. To tackle this, Brodt *et al.*introduced Sketch2Pose [Brodt and Bessmeltsev 2022], which initializes by predicting 2D joint positions from sketches and subsequently aligns a 3D parametric human model to their bones via an optimization framework. Nonetheless, this method is slow and mostly tailored towards hand-drawn sketch lines. Pursuing a swift and highly generalized solution for the sketch-to-pose task remains an open problem.

To tackle this problem, we embraced a "learn from synthesis" strategy, which has been successfully applied in avatar modeling [Guo et al. 2023; Zhuang et al. 2024] and street view synthesis [Zhu et al. 2024c]. Starting from a modest quantity of sketches and corresponding 2D human pose datasets, a large-scale sketch-3D pose dataset is synthesized by a fine-tuned image generative model conditioned on human poses. Such data synthesis is tailored for the sketch-to-pose task. Specifically, we incorporated pose perturbations to create data representing disproportionate human figures and misaligned perspectives in sketches. Furthermore, we amassed a substantial collection of sketches encompassing diverse styles, conducted detailed categorical analyses, and thereby enriched the stylistic variety of the sketches we generated. Ultimately, we produced 120, 000 such high-quality sketch-pose data pairs.

Based on such a dataset, we introduce an end-to-end framework for estimating human mesh from various styled sketches. The generative diffusion prior is leveraged to extract human pose features in sketches and inject conditions that fit the drawing features to guide the denoising network. Unlike the iterative optimization strategy utilized by Sketch2Pose, we implement a neural network featuring a feed-forward architecture for almost 500 times faster pose estimation. A feature-extracting strategy tailored for sketches is introduced to boost the accuracy of 3D pose regression. Owing to our extensive dataset encompassing a wide range of styles and a meticulously designed loss function, our method achieves comparable pose estimation accuracy to Sketch2Pose, while significantly surpassing it in terms of speed and generalization capabilities.

The contributions of our work can be summarized as follows:

- Using a learning-by-synthesizing strategy, we propose a novel approach to address the sketch-to-pose problem. This strategy involves synthesizing a large-scale, customized sketch-3D pose dataset, which substantially boosts the generalization capabilities of the sketch-to-pose estimator across diverse sketch styles.
- By developing a feed-forward structured network, we have significantly improved the speed of sketch-to-pose estimation, marking the 500 times faster than the prior SOTA sketch-to-pose estimator.
- Our meticulously designed network architecture and loss function have greatly enhanced the robustness of the prediction model, allowing it to accurately predict poses even in the presence of human proportion distortions and perspective inaccuracies that commonly exist in sketches. As a result, our method achieves state-of-the-art (SOTA) pose prediction accuracy.

2 Related Works

Sketching is widely regarded as an easy and accessible way to pose characters, catering to both professionals and non-artists. While notable progress has been made in related fields, such as sketch-based interfaces and image-based pose estimations, the unique challenges of handling abstract, disproportionate, and stylistically diverse sketches remain underexplored. This section reviews the most relevant works, categorized into sketch-based character posing and human pose estimation from a single photograph.

2.1 Sketch-Based Character Posing

Sketch-based character posing provides an intuitive means for users to manipulate 3D human poses, yet it introduces several significant challenges. Depth ambiguity, anatomical distortions, missing details, and diverse sketching styles make pose inference particularly difficult. Early works focused on stick figures [Davis et al. 2003; Hecker and Perlin 1992; Lin et al. 2010; Mao et al. 2005], silhouettes [Won and Lee 2016], and clean vector drawings [Bessmeltsev et al. 2016]. These approaches, though efficient in constrained scenarios, are hindered by their reliance on unambiguous, clean inputs. For instance, Gesture3D [Bessmeltsev et al. 2016] reconstructs poses from vector drawings but assumes minimal noise, precise connectivity, and no extra strokes—requirements that are rarely met by natural, user-drawn sketches. This reliance on specific input types significantly limits the usability of such systems, as users cannot freely use diverse sketching styles to specify desired poses.

Recent approaches like Sketch2Pose [Brodt and Bessmeltsev 2022] use a neural network to predict bitmap representations and optimize 3D model parameters for pose inference. However, due to the scarcity of sketch-to-3D model pairs for training, the method requires additional optimization to produce acceptable results. This not only introduces a significant computational burden, but also raises concerns about the reliability of the generated poses, which may lack naturalness or anatomical correctness.

An important application in this domain is the development of interactive systems for engaging with sketches. To achieve efficient

inference, systems like MonsterMash [Dvorožňák et al. 2020] and Motion Doodles [Thorne et al. 2004] offer fast, intuitive sketch-based interactions but are limited by strict input formats or detailed annotations. Systems designed for articulated human poses, like those by Unlu *et al.*[Unlu et al. 2022] and Schmitz *et al.*[Schmitz et al. 2023], impose further input constraints, requiring sketches to consist of 3D primitives. Previous methods have imposed a trade-off between efficiency and input diversity due to the lack of paired sketch-3D pose datasets. In contrast, our approach addresses this gap by proposing a large-scale dataset and building a pose estimation network that directly predicts poses from sketches. This enables efficient near-real-time performance while maintaining generalizability, offering a simple, direct, and scalable solution for sketch-to-pose estimation.

2.2 Human Pose from a Single Photograph

Estimating 3D human poses from a monocular image has been extensively studied in computer vision due to its significant applications in computer graphics, animation, and human-computer interaction. Early methods relied on handcrafted features [Andriluka et al. 2010; Balan et al. 2007; Ramanan 2011], using probabilistic models and tree-based structures. However, they struggled with occlusions, ambiguous poses, and appearance variations.

The introduction of deep learning shifted the field, with DeepPose [Toshev and Szegedy 2014] being one of the first CNN-based approaches. This was followed by methods like Tekin *et al.* [Tekin et al. 2016], which integrated CNNs with structured prediction to improve pose accuracy, and Martinez *et al.* [Martinez et al. 2017], which proposed a fully connected network for 2D-to-3D lifting. Zhou *et al.* [Zhou et al. 2017] added geometric constraints, while Pavlakos *et al.* [Pavlakos et al. 2017] used volumetric heatmaps for joint localization.

Despite progress, the need for large labeled datasets limited generalization, especially for non-photorealistic inputs. The introduction of parametric models like SMPL [Bogo et al. 2016] and SMPL-X [Pavlakos et al. 2019] advanced pose and shape estimation with 3D human priors. Many methods [Goel et al. 2023; Li et al. 2021, 2022; Zhang et al. 2021] focus on improving the accuracy of human mesh recovery. Weakly supervised approaches like HMR [Kanazawa et al. 2018] regressed SMPL parameters using 2D keypoints and adversarial losses, and HMD [Zhu et al. 2019, 2021] further refined detailed shape based on the predicted SMPL mesh. Kolotouros *et al.*'s SPIN [Kolotouros et al. 2019] refined this approach with optimization, while EFT [Joo et al. 2021] fine-tuned SMPL predictions. Methods like 3DCrowdNet [Choi et al. 2022], JOTR [Li et al. 2023b], and DPMesh [Zhu et al. 2024b] are designed to recover the occluded body mesh. Self-supervised methods, such as Wang *et al.* [Wang et al. 2019] and Novotny *et al.* [Novotny et al. 2019], reduced dependence on labeled data using geometric consistency.

Our method bridges human pose estimation and character drawing by leveraging visual priors in pre-trained diffusion models, as seen in VPD [Zhao et al. 2023]. Using our dataset, our fine-tuned network extracts structural and spatial features from the denoising U-Net for accurate human mesh recovery. By processing the input image in a single inference pass [Zhu et al. 2024b], our approach adapts diffusion priors to handle abstract sketches, ensuring reliable pose estimation.

3 SKEP-120K Dataset

It is widely recognized that the quality and abundance of the training data heavily influence the success of learning-based techniques for human mesh recovery. Surprisingly, we find that there is currently a notable absence of a large-scale, high-quality dataset containing sketches and 3D human poses. Though several available datasets [Brodt and Bessmeltsev 2022; Ju et al. 2023; Madhu et al. 2022; Smith et al. 2023] offer a substantial number of sketches, they provide only 2D pose labels and typically a single sketch style, making them insufficient for training a highly accurate, generalizable sketch-to-pose model.

Therefore, we propose a Sketch and 3D Pose dataset with 120k data pairs in various sketch styles, named as SKEP-120K dataset. As shown in Fig. 3, the dataset encompasses six styles according to artificial human scenes: *cartoons, oil paintings, ink paintings, charcoal sketches, stick figures, and kids' drawings*. Each style contains approximately 20,000 images. Our dataset provides human bounding boxes, 16 human joints (both 2D and 3D, with corresponding visible/invisible/included attributes), SMPL pose parameters, and text information. Due to the different definitions of the skeleton for a human pose in different datasets and considering the characteristics of gesture expression in the sketch, we define a new 3D skeleton to represent the human pose. Specifically, the body parts are based on MSCOCO database [Lin et al. 2014], and two additional joints regarding left and right toes are added to reflect fine-scale leg poses.

The creation process of our dataset is shown in Fig. 2. Firstly, we utilize VPoser [Pavlakos et al. 2019] to generate random SMPL models exhibiting diverse and plausible poses, where a variational autoencoder is designed to capture latent representations of human poses. Given the pervasive foreshortening in sketches, which makes the depicted human structure to deviate from the standard 3D-to-2D projection observed in real-captured images, we introduce skeletal-proportion perturbations during the 3D-to-2D mapping by adding random biases to the projected limb lengths as a skeleton-level data augmentation. This strategy yields skeletal configurations that better capture the characteristic proportional exaggerations and imbalances of hand-drawn sketches, thereby enhancing the pose diversity and accuracy of the sketch dataset and producing 2D joint annotations that are better aligned with sketch scenarios. Next, we consolidate data from the Sketch2Pose [Brodt and Bessmeltsev 2022], Human-Art [Ju et al. 2023], and Amateur Drawing datasets [Smith et al. 2023], unify their 2D keypoint annotations to our defined set of 16 joints, and partition them into six groups by drawing style. We then leverage BLIP2 [Li et al. 2023a] to generate appearance descriptions on sketch images and motion descriptions on SMPL rendering images. These descriptions, together with the target style labels, serve as conditioning prompts, yielding a sketch training set. On this basis, we train a text-conditioned image generation model following ControlNet [Zhang et al. 2023] to synthesize sketch data. During training, sketches from all six styles and their associated annotation parameters are jointly used to train the same model.

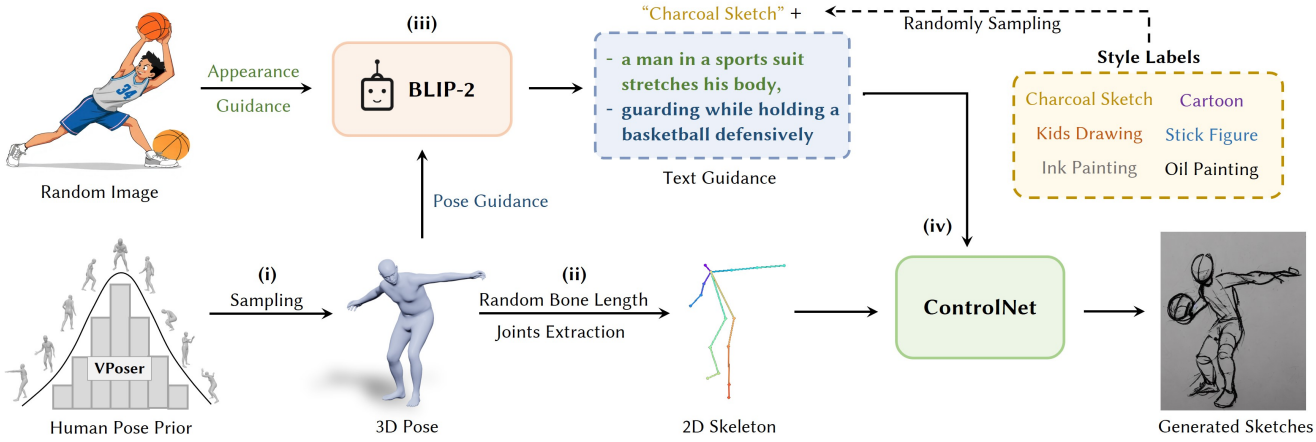

Fig. 2. SKEP-120K Dataset Creating Pipeline. Three stages are involved: (I) generating diverse 3D poses (as SMPL); (II) adding random biases to bone lengths and projecting to 2D poses; (III) generating diverse text guidance; (IV) training a text-conditioned image generator for sketch synthesis.

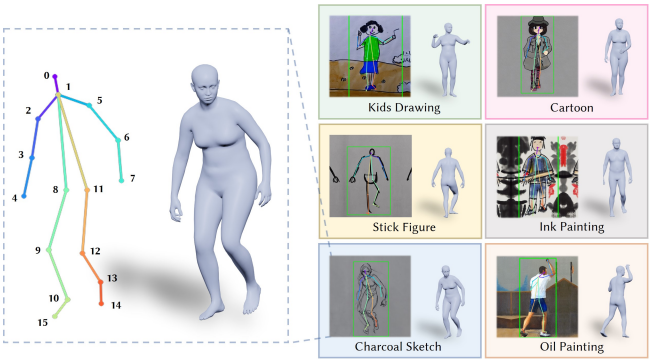

Fig. 3. Data Description. SKEP-120K dataset comprises six sketch styles: cartoons, oil paintings, ink paintings, charcoal sketches, stick figures, and kids' drawings. The provided 2D/3D joints are shown on the left.

After that, we manually curate the model-generated dataset to improve its quality further. Specifically, we invite experienced 3D modelers to filter out approximately 10% of the sketches in each style, including those with severely cluttered backgrounds that compromise subject visibility, those showing pose inconsistencies caused by overly challenging conditioning poses, and a subset with extremely distorted poses introduced by the added random biases. As a result, we generate character images across various styles with high accuracy and adherence to the 2D skeleton distribution. Given the diverse line-based nature of sketches, we use the off-the-shelf outline detector [Canny 1986] to extract line distributions within these images. A threshold is applied to identify the smallest region encompassing most lines, defining the character's bounding box. We compare the human bounding box with the bounding box generated from the detector of Human-Art, leaving a more accurate result, and manually filtering out the undetectable cases. We also detect the occlusion of each joint based on the occlusion relationship of the SMPL mesh, which is recorded as the label of each joint.

4 Method

Given the SKEP-120K dataset, our objective is to train a prediction model for recovering 3D human poses from sketches in varying styles. As shown in Fig. 4, our overall network consists of three modules: (I) a 2D guidance extractor (Sec. 4.1); a sketch feature extractor (Sec. 4.2); and an SMPL regressor (Sec. 4.3). From a probabilistic model perspective, the above process can be formulated as:

$$p_\phi(\boldsymbol{y}\,|\,\boldsymbol{x}) = p_{\phi_3}(\boldsymbol{y}\,|\,\mathcal{F})\,p_{\phi_2}(\mathcal{F}\,|\,\epsilon(\boldsymbol{x}),\mathcal{G})\,p_{\phi_1}(\mathcal{G}\,|\,\epsilon(\boldsymbol{x})), \tag{1}$$

where x denotes the input sketch; y is the 3D pose represented by SMPL parameters; $\mathcal{F}$ signifies informative feature maps extracted from sketches; $\mathcal{G}$ indicates the spatial guidance extracted from 2D poses; ϵ is a pre-trained image encoder network; p_{ϕ_1}, p_{ϕ_2} and p_{ϕ_3} correspond to the 2D guidance extractor, sketch feature extractor, and SMPL regressor, respectively. We will then explain each module and the objective functions in the following sections.

4.1 2D Guidance Extractor

Drawing inspiration from VPD [Zhao et al. 2023], our core idea involves extracting high-level pre-trained knowledge from a diffusion model. A fundamental prerequisite for achieving this is the extraction of 2D guidance.

The first step in extracting 2D guidance is to estimate 2D joints from input sketches. Leveraging our proposed SKEP-120K dataset, we have fine-tuned two state-of-the-art network models for human detection and 2D joint extraction from sketches. Specifically, the input sketches are first resized and padded to the resolution of 256×192 pixels to preserve the aspect ratio. Then, the human detector YOLOX [Redmon 2016] is fine-tuned by Human-Art dataset for human's bounding box detection in sketches, and the ViTPose [Xu et al. 2022] model is fine-tuned for 2D joints prediction from these bounded sketches. ViTPose utilizes a straightforward, non-hierarchical vision transformer as the encoder to capture human features in drawings, combined with a lightweight decoder that predicts body joints in a top-down approach. Finally, we obtain 2D joints $J^{2D} \in \mathbb{R}^{K\times 2}$ along with their corresponding confidence

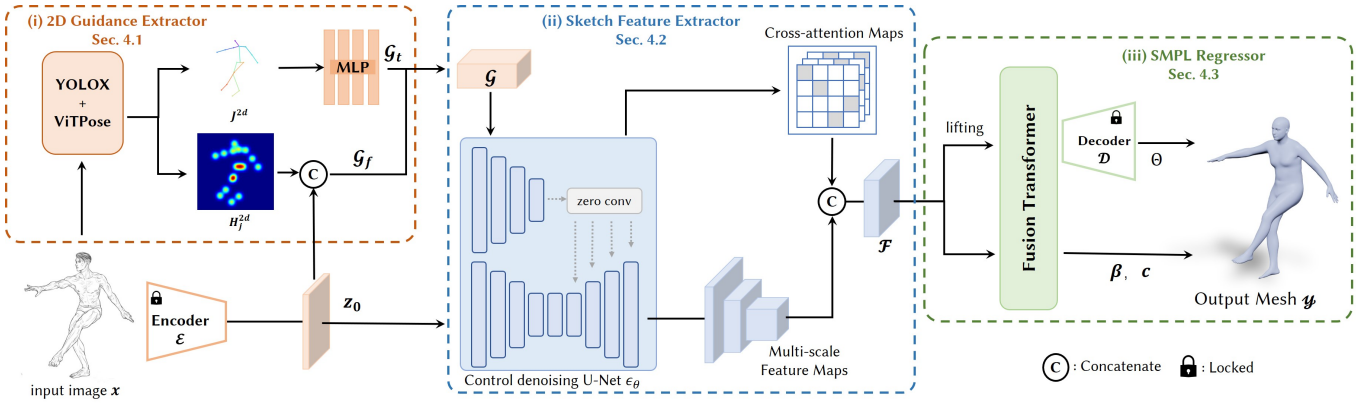

Fig. 4. Overall Pipeline. Given a sketch image as input, the network predicts 3D human poses represented by SMPL parameters. The overall network consists of three modules: a 2D guidance extractor as detailed in Sec. 4.1; a sketch feature extractor as detailed in Sec. 4.2; and an SMPL regressor as detailed in Sec. 4.3.

and transform them into heatmaps $H_j^{2D} \in \mathbb{R}^{K\times H'\times W'}$ using 2D Gaussian kernels [Cheng et al. 2020].

After the 2D joints are obtained, pose features are extracted from the 2D joints J^{2D} and heatmaps H_j^{2D}, which provides spatial guidance for the denoising U-Net [Rombach et al. 2022] backbone $\boldsymbol{\epsilon_\theta}$. This process is referred as $p_{\phi_1}(\mathcal{G}\,|\,\boldsymbol{x})$ in Eq. 1. For the input image $\boldsymbol{x}$, we convert the cropped image $\boldsymbol{x} \in \mathbb{R}^{H\times W\times C}$ from pixel space to the latent space with frozen encoder $\mathcal{E}$ in the trained VQGAN from the Controlnet framework to obtain the latent representation $z_0 \in \mathbb{R}^{H'\times W'\times \mathcal{G}_z}$. Then, we concatenate the heatmap H_j^{2D} with the input image z_0 to obtain $\mathcal{G}_f \in \mathbb{R}^{(K+\mathcal{G}_z)\times H'\times W'}$. In most previous diffusion models [Dhariwal and Nichol 2021; Rombach et al. 2022; Zhang et al. 2023], the prompt guidance $\mathcal{G}_t$ usually relies on text embeddings derived from a frozen CLIP [Radford et al. 2021] model. In contrast, we replace the text with 2D joint positions J^{2D} as verified in [Zhu et al. 2024b]. To match the text token dimension D_j, a two-layer MLP is used to enhance the dimensionality of the 2D joint positions to 768 in the pre-trained diffusion model. This generates a spatial guidance $\mathcal{G}_t \in \mathbb{R}^{K\times D_j}$. The process can be expressed as follows:

$$\mathcal{G}_f = \text{Concat}(z_0, \mathbf{H}_j^{2D}), \tag{2}$$

$$\mathcal{G}_t = \text{MLP}(J_{2D}), \tag{3}$$

After that, $\mathcal{G}_f$ and $\mathcal{G}_t$ are injected into $\boldsymbol{\epsilon_\theta}$ through different channels, thus we obtain the 2D guidance $\mathcal{G}$.

4.2 Sketch Feature Extractor

Once the 2D guidance is obtained, our next objective is to extract informative features from the sketches for 3D pose estimation. A multi-scale features extractor is introduced based on the pre-trained denoising U-Net. Our key idea is to fully extract the pre-trained high-level knowledge from a pretrained diffusion model, named informative features $\mathcal{F}$, then utilize its learned knowledge to predict 3D human poses from sketches. We employ the denoising U-Net $\boldsymbol{\epsilon_\theta}$ as the image backbone, performing a single inference to extract features from image $\boldsymbol{x}$. To provide effective guidance, we utilize the conditional injection of human pose instead of the text condition, which makes the connection between these conditions and the input image such that the learned semantic information can be efficiently extracted.

Specifically, $p_{\phi_2}(\mathcal{F}\,|\,\epsilon(\boldsymbol{x}), \mathcal{G})$ is designed to extract hierarchical feature maps $\mathcal{F}$ from the input image $\boldsymbol{x}$ along with the 2D guidance $\mathcal{G}$. We observe that the pre-trained text-to-image diffusion model serves as an excellent initialization for p_{ϕ_2}, which has already established a connection between the vision and language domains.

It is also known that ControlNet leverages trainable copies of the encoding layers within the denoising U-Net, serving as a robust backbone to learn various conditional controls, significantly enhancing the fine-grained spatial controllability of the Latent Diffusion Model (LDM) [Rombach et al. 2022]. In our implementation, we utilize the ControlNet architecture to handle pose-conditioned information from the 2D guidance $\mathcal{G}$ and integrate it into the image features within the denoising U-Net $\boldsymbol{\epsilon_\theta}$. The output $\mathcal{F}$ in the decoding layers of $\boldsymbol{\epsilon_\theta}$ is expressed as:

$$\mathcal{F} = \mathrm{F}_n(z_0;\theta) + \mathrm{Z}(\mathrm{F}_n(\mathcal{G};\theta_c);\theta_z), \tag{4}$$

where $\mathrm{F}_n(\cdot;\theta)$ is a trained neural network, $\mathrm{Z}(\cdot;\cdot)$ denotes zero convolution layers with both weights and bias initialized to zeros, θ_c represents the parameters within ControlNet, θ_z is the parameters of zero convolution layers. We feed the latent feature map and the pose-conditioned inputs to the pre-trained $\boldsymbol{\epsilon_\theta}$ network and extract the multi-scale feature maps $\mathcal{F}_i$ from the last layer of each output block in different resolutions. Our experimental observations indicate that the extracted informative features represent more information about the structures in abstract sketches, enhancing the accuracy of the subsequent SMPL regression for sketch input.

In addition, we empirically find that the cross-attention maps $A_i \in \mathbb{R}^{|\mathcal{G}|\times H_i\times W_i}$ from the decoding layers of the U-Net $\boldsymbol{\epsilon_\theta}$ can provide occlusion-aware cues that indicate invisible parts and help to focus on the 2D skeleton condition information within the sketch. Therefore, we concatenate the feature maps with the cross-attention maps to generate the hierarchical feature maps $\mathcal{F} \leftarrow \{[\mathcal{F}_i, A_i]\}$, which incorporate explicit and implicit diffusion priors and thereby further enhance the performance of SMPL mesh regression.

4.3 SMPL Mesh Regressor

In the last stage, an SMPL mesh regressor is proposed to predict 3D poses from the previously extracted features $\mathcal{F}$. Specifically, $p_{\phi_3}(\boldsymbol{y}|\mathcal{F})$ refers to the prediction head that generates parameters of the body model from the hierarchical feature maps $\mathcal{F}$. We first lift the pose-guided 2D feature $\mathcal{F}$ to the 3D feature $\mathcal{F}_{3D}$, then extend the 2D features by incorporating 3D joint feature sampling. To integrate and align 2D and 3D features, we utilize a fusion transformer [Li et al. 2023b] to regress SMPL parameters. Moreover, we employ a pre-trained VQVAE [Van Den Oord et al. 2017], which is trained on a large-scale motion dataset AMASS [Mahmood et al. 2019] with extensive SMPL pose parameters to provide adequate human pose priors and preserve the correspondence of the VQGAN framework, which can obtain discrete representations of human poses. During the regression, the decoder of the VQVAE is utilized to get the pose parameters Θ, while the shape parameters β and camera parameters c are directly predicted using linear layers.

4.4 Objective Function

Unlike human pose estimation from real-captured photos, recovering human pose from artificial sketches in the literature is even more difficult due to the distorted proportions, perspective, and foreshortening. Specifically, sketches often depict characters with unrealistic body shapes or exaggerated body proportions. Therefore, standard optimization methods that depend solely on 2D joint positions can result in inaccurate or unnatural outcomes. Through observing the artwork of human character drawing, three key elements are identified as crucial for addressing these issues: joint angle, foreshortening, and self-contacts. Building on our proposed human drawing dataset with 3D pose annotations, we introduce new methods to address this.

Joint Angle. Character bones often appear longer than their actual length because of imprecision in drawings or the use of artistic interpretation [Hogarth 1970; Johnston and Thomas 1981; Stanchfield 2007] in human drawings. Due to inconsistent representations of bone length, directly utilizing absolute joint positions becomes impractical. And art literature has consistently highlighted the importance of accurately describing joint angles. Therefore, we expect the 3D bone projections to align with the bones depicted in 2D, ensuring that the reconstructed 3D joint angles have corresponding projections to the depicted 2D joint angles.

Our dataset provides precise annotations for the 2D joints x^{2D}, 3D joints x^{3D}, and exact SMPL pose parameters Θ of the characters in each drawing. For each bone i connecting joints j_1 and j_2, we represent its 3D vector as $\boldsymbol{b}_i^{3D} = x_{j_2}^{3D} - x_{j_1}^{3D}$, and its orthographic projection onto the screen as $\boldsymbol{b}_i^{2D}$. The 2D joints predicted by our algorithm are $\bar{x}^{2D}$, so the predicted vector corresponding to bone i is $\bar{\boldsymbol{b}}_{\boldsymbol{i}}^{\boldsymbol{2D}} = \bar{x}_{j_2}^{2D} - \bar{x}_{j_1}^{2D}$, and $\boldsymbol{n}$ represents the normal to the predicted bone $\bar{\boldsymbol{b}}_{\boldsymbol{i}}^{\boldsymbol{2D}}$. Guided by our principle of joint angle, the loss of parallelism between the projected 2D bones can be expressed as:

$$\mathcal{L}_{\text{parallel}} = \sum_i \left(\frac{\boldsymbol{b}_i^{2D}}{\|\boldsymbol{b}_i^{2D}\|} \cdot \boldsymbol{n} \right)^2, \tag{5}$$

This skeleton parallelism loss enables a more reasonable and natural alignment of human joints in sketches than joint position loss.

Foreshortening. Empirically, artists typically do not rely on exact mathematical measurements for orthographic or perspective projections when creating drawings [Hogarth 1970; Stanchfield 2007]. Thus, directly reconstructing 3D poses from predicted 2D poses frequently results in highly inaccurate estimations of the angles formed between the bones and the screen. For bone i, the angle between the character in the drawing and the screen can be represented as the angle between the 3D vector of the skeleton and the 2D vector of its projection. The foreshortening loss for the skeleton can thus be formulated as:

$$\mathcal{L}_{\text{f}} = \sum_i \left(\frac{\|\boldsymbol{b}_i^{3D}\|}{\|\boldsymbol{b}_i^{2D}\|} - \frac{\|\bar{\boldsymbol{b}}_{\boldsymbol{i}}^{\boldsymbol{3D}}\|}{\|\bar{\boldsymbol{b}}_{\boldsymbol{i}}^{\boldsymbol{2D}}\|} \right)^2, \tag{6}$$

where $\mathcal{L}_{\text{f}}$ is the skeleton-screen angle's cosine.

Self-contacts. Self-contacts are prevalent in common human poses, which are revealed by prior works [Hogarth 1970]. We hypothesize that human observers often rely on perceived self-contacts to solve the problem of depth ambiguity and link touching body parts to similar depths. Previous works focus on optimizing regions based on manually annotated self-contact areas. They enforce physical contact between pairs of vertices by mapping each contact region onto the vertices of the roughly aligned SMPL mesh. In contrast, our dataset includes accurate SMPL pose parameters for the human body in sketches, which provides correct relative depth and joint positions of the character skeleton. In our method, we replace the previous self-contact loss with the SMPL pose parameter loss, which is calculated by L_1 loss between the predicted SMPL pose parameter and the ground-truth SMPL pose parameter, thus supervising the human mesh to recover the correct positions.

Instead of the previous position-based reprojection loss, our overall training objective in our method is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{parallel}} + \lambda_2 \mathcal{L}_{\text{f}} + \lambda_3 \mathcal{L}_{\text{pose}} + + \lambda_4 \mathcal{L}_{\text{shape}}, \tag{7}$$

where $\mathcal{L}_{\text{pose}}$ is the SMPL pose parameter loss, and $\mathcal{L}_{\text{shape}}$ is the SMPL shape parameter loss, λ_1, λ_2, λ_3, λ_4 are set to 3, 3, 2, 1.

5 Experiments

5.1 Implementation Details

Datasets. We utilize two datasets for performance evaluation:

- The artist-designed dataset is provided by Sketch2Pose [Brodt and Bessmeltsev 2022], which contains six sketches with corresponding 3D poses manually modeled by two artists that best align with the artist's intentions. The merit of this validation set lies in its accurate representation of the ideal 3D pose intended by the artists, whereas its limitation is the scant data volume, comprising merely six very challenging poses.
- The SKEP-120K validation set is created using the method outlined in Section 3. We invite experienced 3D modelers to manually sieve through and eliminate inaccurate data to guarantee high quality. This validation set contains 600 validation tuples, with 100 tuples for each of the six styles. Owing to its comprehensive coverage of

Table 1. Quantitative comparison on the artist-designed dataset and the SKEP-120K validation set.(Unit: mm)

Method	Expert1			Expert2			SKEP-120K		
	MPVE↓	MPJPE↓	PA-MPJPE↓	MPVE↓	MPJPE↓	PA-MPJPE↓	MPVE↓	MPJPE↓	PA-MPJPE↓
PyMAF [Zhang et al. 2021]	312.7	299.4	187.5	301.5	291.2	187.0	143.1	117.4	101.3
EFT [Joo et al. 2021]	144.6	144.4	98.9	168.8	158.9	103.4	158.7	133.1	111.6
HybrIK [Li et al. 2021]	348.7	345.5	199.6	365.2	352.0	208.5	211.0	177.7	144.0
CLIFF [Li et al. 2022]	186.4	181.5	142.2	217.3	201.3	144.5	137.3	113.3	98.0
HMR2.0 [Goel et al. 2023]	118.3	105.0	85.1	181.4	151.4	107.4	128.0	104.6	88.1
MotionBERT [Zhu et al. 2023]	170.6	165.2	120.1	189.1	172.2	127.6	124.4	99.4	83.6
DPMesh [Zhu et al. 2024b]	130.6	121.5	95.3	169.9	152.1	103.2	123.4	98.0	81.1
DPMesh(Retrained) [Zhu et al. 2024b]	127.7	121.4	94.1	166.4	147.1	94.3	122.6	97.3	80.6
Sketch2Pose [Brodt and Bessmeltsev 2022]	103.8	101.4	78.1	**145.5**	135.9	86.8	152.1	125.9	100.3
Ours	**103.1**	**95.7**	**77.4**	146.5	**131.5**	**84.3**	**106.7**	**87.7**	**72.6**

Table 2. Quantitative comparison of ablation study.

Method vs Expert 1	MPVE↓	MPJPE↓	PA-MPJPE↓
$w/o\ L_{\text{parallel}}$	169.8	165.7	102.5
$w/o\ L_{\text{f}}$	117.7	110.2	84.4
$w/o\ L_{\text{pose}}$	121.4	117.6	86.7
$w/o\ L_{\text{parallel,f,pose}}$	113.7	107.3	85.6
$w/o\ A_i$	104.9	97.1	79.6
$w/o\ J^{2D}$	117.4	113.9	89.6
w/o Data Curation	104.8	99.0	82.7
Ours (Full)	**103.1**	**95.7**	**77.4**

Table 3. Runtime of our method.

Method	I (Sec. 4.1)	II (Sec. 4.2)	III (Sec. 4.3)	Total Time
PyMAF*	0.09s	-	0.01s	0.10s
EFT*	0.34s	-	3.23s	3.57s
HyBrIK*	0.03s	0.06s	0.08s	0.17s
CLIFF*	2.60s	1.21s	0.04s	3.85s
HMR2.0*	0.52s	-	0.03s	0.55s
MotionBERT*	0.04s	-	0.15s	0.19s
DPMesh*	0.05s	0.06s	0.03s	0.14s
Sketch2Pose	4.75s	32.98s	30.15s	67.57s
Ours	0.04s	0.05s	0.03s	0.12s

* means the model is not for sketches but for regular photos.

various sketch styles and extensive data volume, it is well-suited for assessing the generalization capability.

Metrics. We adopt three standard metrics for 3D pose estimation: Mean Per Joint Position Error (MPJPE) and Procrustes-Aligned Mean Per Joint Position Error (PA-MPJPE) to evaluate the accuracy of the predicted 3D joint positions, and Mean Per Vertex Error (MPVE) to measure the accuracy of 3D mesh reconstruction in sketches. Quantitative metrics are only part of the evaluation for the prediction of 3D poses in sketches. We place greater emphasis on whether the visualized results align more closely with the artist's original intent.

Training Details. We train three models for data creation and sketch-to-pose prediction: ControlNet to synthesize sketch data, ViTPose for 2D keypoint detection, and our core sketch-to-pose network. For ControlNet, condition maps and sketches are padded to 512×512. Then, BLIP2 [Li et al. 2023a] is leveraged to generate prompt labels for the sketches. ViTPose is fine-tuned separately on our collected and synthesized sketch data and remains frozen during core-model training, serving solely as a 2D joints extractor. The core model training process consists of two stages. All training and experiments run on 4 Nvidia A6000 GPUs. The supplementary material provides more training details.

5.2 Quantitative Comparison

Quantitative comparisons on the artist-designed dataset and SKEP-120K validation set are shown in Tab. 1. The artist-designed dataset contains six real-world sketches with expert-annotated 3D poses, all of which are challenging non-daily poses. On this dataset, our model achieves the best overall performance of all metrics. Although gains on charcoal sketches are modest relative to Sketch2Pose and we even perform slightly worse on one expert-based metric, we reach this accuracy in roughly 1/500 of Sketch2Pose's runtime.

The SKEP-120K validation set contains sketches generated outside of our training set. It includes six sketch styles, each containing 100 images, all following a real-world sketch distribution. Across all styles, our method significantly outperforms prior approaches and achieves the best overall results.

These results show that our method balances high accuracy with strong efficiency. Its generalization surpasses both the generic pose estimation algorithm and Sketch2Pose algorithm. To ensure a fair evaluation, all methods use the same 2D inputs produced by our trained ViTPose model. For generic pose estimation, we select the current leading method DPMesh and retrain it with our data. The performance gains verify the effectiveness of our dataset. Moreover, with identical training data, our method still outperforms alternative approaches. Furthermore, its fast inference enables efficient application to video, beyond static images.

5.3 Qualitative Comparison

We have visualized comparisons across diverse sketch styles in Fig. 6. Notably, (a) - (b) demonstrate that our model can predict poses consistent with real human body proportions even for cartoons with exaggerated proportions. (c) - (f) highlight that our method yields more accurate predictions for cartoons, children's drawings, stick figures, ink paintings, and oil paintings. Sketch2Pose struggles

beyond charcoal sketches, and the generic image-to-pose baseline DPMesh degrades markedly when dealing with various styles. By contrast, only our method sustains high performance across multiple sketch styles. We attribute this to our three-stage pose prediction network design for sketch feature extraction and synthesized dataset with perturbations.

Fig. 7 presents the predictions for challenging poses. Specifically, (g) illustrates a scenario where one of the character's arms is fully occluded, yet our method infers a plausible pose. In (h), overlapping character lines with a handheld object simplify structure, but the pose is still correctly predicted. These results indicate that our approach better meets artists' creative needs and predicts 3D human poses more accurately. We also retarget predicted poses to custom characters as shown in Fig. 1. Our method facilitates frame-by-frame prediction due to its efficient inference speed, thereby significantly improving the applicability and effectiveness of sketch-to-pose.

Fig. 8 presents the frame-by-frame results of our method applied to continuous line animations, demonstrating the generalization capability for pose estimation from videos. In addition, The supplementary material provides a user study for subjective evaluation.

5.4 Ablation Study

We perform an ablation study on the artist-designed dataset to evaluate the efficacy of our proposed loss terms. Specifically, we ablate the loss terms $\mathcal{L}_{\text{parallel}}$, $\mathcal{L}_{\text{f}}$, and $\mathcal{L}_{\text{pose}}$ from Eq. (4.4), and replace all proposed loss terms with the joint-distance-based loss used in DPMesh, retraining the model with all other settings the same. The evaluation results reported in Tab. 2 reveal a performance degradation across all ablation settings, indicating the indispensable contribution of each ablated loss term to the overall model performance.

Next, we perform further ablations on the network design. Specifically, we remove the 2D joint prediction J^{2D} from the 2D guidance extractor and exclude the cross-attention maps A_i from the sketch feature extractor. Tab. 2 shows a decline in model performance, demonstrating the effectiveness of our network design. Moreover, the quality of generated sketches is important for model performance. We fine-tune ControlNet on real sketches to synthesize high-quality training data, and subsequently apply manual curation to further enhance data quality. To validate the effect of sketch quality, we retrain our model on the pre-curation set and find that training on the curated set yields superior performance, showing the importance of generated-sketch quality for overall effectiveness.

5.5 Runtime Analysis

The step-by-step runtimes of our method and Sketch2Pose are reported in Table 3. Sketch2Pose involves a three-stage process that differs from our design, including: (I) a 2D pose and initial 3D mesh estimator; (II) an optimizer to correct errors in the estimated 3D pose; and (III) a final refinement to better align the result with the stylistic and structural constraints of the sketch. Since Sketch2Pose leverages iterative optimization approaches to solve all three stages, it incurs substantial computational cost. By contrast, our approach achieves an over 500× speedup while delivering superior performance compared to Sketch2Pose. This improvement is primarily attributed to our efficient feed-forward neural network design. In addition to the comparison with Sketch2Pose, we also compare our method with other approaches used for human pose estimation from regular photos. The results demonstrate that our method achieves a significant advantage in runtime, matching or even surpassing the methods of pose estimation from regular photos.

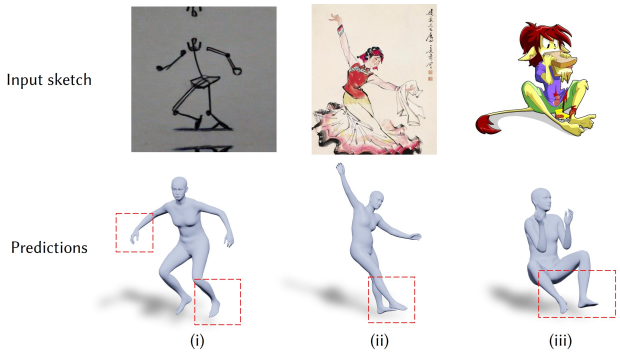

Fig. 5. Failure cases. Our model may predict inaccurate terminal joints.

6 Conclusion

We present a novel approach to estimating human poses from sketches. By adopting a learning-by-synthesizing strategy, we have synthesized a large-scale, customized sketch-pose dataset tailored for this task, significantly enhancing our model's generalization capabilities across various sketch styles. Furthermore, the proposed feed-forward structured network has markedly improved the speed of sketch-to-pose estimation.

Limitations. Terminal joints (e.g., those in the hands and feet) in sketch-based human pose estimation remain particularly challenging to predict, owing to multiple factors. Firstly, sketches are inherently abstract and often lack fine structural detail. For instance, a single limb stroke may ambiguously depict either the contour of a forearm or an extended hand as shown in Fig. 5 (i), providing only limited discriminative cues. Secondly, terminal joints in sketches are often occluded or have ambiguous depth cues. While our method leverages a pretrained diffusion model to infer plausible positions for occluded terminal joints, its predictions remain unnatural as shown in Fig. 5 (ii). Thirdly, the hierarchical structure of the human skeleton means that minor errors at proximal joints can propagate along the kinematic chain, leading to disproportionately large errors at distal joints as shown in Fig. 5 (iii). Collectively, these factors result in lower estimation accuracy for terminal joints.

Our current work focuses on sketch-driven 3D human pose estimation. Subject-specific body-shape reconstruction is out of the scope of our research. We have conducted further analysis of this limitation in the supplementary materials, and addressing this issue has become a priority for future research.

Acknowledgments

This study was funded by NKRDC 2022YFF0902200 and Jiangsu Broadcasting Corporation.

References

Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. 2010. Monocular 3D Pose Estimation and Tracking by Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 623–630.

Alexandru O Balan, Leonid Sigal, Michael J Black, James E Davis, and Horst W Haussecker. 2007. Detailed Human Shape and Pose from Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–8.

Mikhail Bessmeltsev, Nicholas Vining, and Alla Sheffer. 2016. Gesture3D: Posing 3D Characters via Gesture Drawings. *ACM Transactions on Graphics (TOG)* 35, 6, Article 165 (Nov. 2016).

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. 2016. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 561–578.

Kirill Brodt and Mikhail Bessmeltsev. 2022. Sketch2Pose: Estimating a 3D Character Pose from a Bitmap Sketch. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–15.

John Canny. 1986. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 6 (1986), 679–698.

Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. 2020. HigherHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5386–5395.

Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. 2022. Learning to Estimate Robust 3D Human Mesh from In-the-Wild Crowded Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1475–1484.

James Davis, Maneesh Agrawala, Erika Chuang, Zoran Popović, and David Salesin. 2003. A Sketching Interface for Articulated Figure Animation. *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)* (2003), 320–328.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion Models Beat Gans on Image Synthesis. *Advances in Neural Information Processing Systems (NeurIPS)* 34 (2021), 8780–8794.

Marek Dvorožňák, Daniel Sỳkora, Cassidy Curtis, Brian Curless, Olga Sorkine-Hornung, and David Salesin. 2020. Monster Mash: a Single-View Approach to Casual 3D Modeling and Animation. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–12.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. 2023. Humans in 4D: Reconstructing and Tracking Humans with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 14783–14794.

Longwei Guo, Hao Zhu, Yuanxun Lu, Menghua Wu, and Xun Cao. 2023. RAFaRe: Learning Robust and Accurate Non-parametric 3D face Reconstruction from Pseudo 2D&3D Pairs. In *Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence (AAAI)*, Vol. 37. 719–727.

Ronie Hecker and Kenneth Perlin. 1992. Controlling 3D Objects by Sketching 2D Views. *Proceedings of the International Society for Optics and Photonics (SPIE)* 1828 (1992), 46–48.

Burne Hogarth. 1970. Dynamic Figure Drawing. *(No Title)* (1970).

Ollie Johnston and Frank Thomas. 1981. *The Illusion of Life: Disney Animation.* Disney Editions New York.

Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. 2021. Exemplar Fine-Tuning for 3D Human Model Fitting Towards In-the-Wild 3D Human Pose Estimation. *Proceedings of the International Conference on 3D Vision (3DV)* (2021), 42–52.

Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. 2023. Human-Art: A Versatile Human-Centric Dataset Bridging Natural and Artificial Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 618–629.

Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2018. End-to-End Recovery of Human Shape and Pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 7122–7131.

Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. 2019. 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 (ICCV)*. 2252–2261.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models. In *Proceedings of the International Conference on Machine Learning (ICML)*. 19730–19742.

Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. 2021. 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 (CVPR)*. 3383–3393.

Jiahao Li, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. 2023b. JOTR: 3D Joint Contrastive Learning with Transformers for Occluded Human Mesh Recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 9110–9121.

Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. 2022. CLIFF: Carrying Location Information in Full Frames into Human Pose and Shape Estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 590–606.

Miao Liao, Sibo Zhang, Peng Wang, Hao Zhu, Xinxin Zuo, and Ruigang Yang. 2020. Speech2Video Synthesis with 3D Skeleton Regularization and Expressive Body Poses. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

Juncong Lin, Takeo Igarashi, Jun Mitani, and Greg Saul. 2010. A Sketching Interface for Sitting-Pose Design. In *Proceedings of the Sketch-Based Interfaces and Modeling Symposium (SBIM)*. 111–118.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 740–755.

Prathmesh Madhu, Angel Villar-Corrales, Ronak Kosti, Torsten Bendschus, Corinna Reinhardt, Peter Bell, Andreas Maier, and Vincent Christlein. 2022. Enhancing Human Pose Estimation in Ancient Vase Paintings via Perceptually-Grounded Style Transfer Learning. *ACM Journal on Computing and Cultural Heritage (JOCCH)* 16, 1 (2022), 1–17.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of Motion Capture as Surface Shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5442–5451.

C Mao, S F Qin, and D K Wright. 2005. A Sketch-Based Gesture Interface for Rough 3D Stick Figure Animation. *Proceedings of the Sketch-Based Interfaces and Modeling Symposium (SBIM)* (2005).

Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. 2017. A Simple yet Effective Baseline for 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2640–2649.

David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. 2019. C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure from Motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 7688–7697.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10975–10985.

Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. 2017. Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7025–7034.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*. 8748–8763.

D. Ramanan. 2011. Part-Based Models for Finding People and Estimating Their Pose. (2011), 199–223.

J Redmon. 2016. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 779–788.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

Caro Schmitz, Constantin Rösch, Domenic Zingsheim, and Reinhard Klein. 2023. Interactive Pose and Shape Editing with Simple Sketches from Different Viewing Angles. *Computers & Graphics (CG)* 114 (2023), 347–356.

Harrison Jesse Smith, Qingyuan Zheng, Yifei Li, Somya Jain, and Jessica K Hodgins. 2023. A Method for Animating Children's Drawings of the Human Figure. *ACM Transactions on Graphics (TOG)* 42, 3 (2023), 1–15.

Walt Stanchfield. 2007. Gesture Drawing for Animation. *Washington: Leo Brodie* 1 (2007).

Feitong Tan, Hao Zhu, Zhaopeng Cui, Siyu Zhu, Marc Pollefeys, and Ping Tan. 2020. Self-Supervised Human Depth Estimation from Monocular Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 650–659.

Bugra Tekin, Artem Rozantsev, Vincent Lepetit, and Pascal Fua. 2016. Direct Prediction of 3D Body Poses from Motion Compensated Sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 991–1000.

Matthew Thorne, David Burke, and Michiel Van De Panne. 2004. Motion Doodles: An Interface for Sketching Character Motion. *ACM Transactions on Graphics (TOG)* 23, 3 (2004), 424–431.

Alexander Toshev and Christian Szegedy. 2014. Deeppose: Human Pose Estimation via Deep Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1653–1660.

Gizem Unlu, Mohamed Sayed, and Gabriel Brostow. 2022. Interactive Sketching of Mannequin Poses. In *Proceedings of the International Conference on 3D Vision (3DV)*.

700–710.

Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural Discrete Representation Learning. *Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017).

Keze Wang, Liang Lin, Chenhan Jiang, Chen Qian, and Pengxu Wei. 2019. 3D Human Pose Machines with Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2019).

Yanwen Wang, Yiyu Zhuang, Jiawei Zhang, Li Wang, Yifei Zeng, Xun Cao, Xinxin Zuo, and Hao Zhu. 2025. TeRA: Rethinking Text-driven Realistic 3D Avatar Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Jungdam Won and Jehee Lee. 2016. Shadow Theatre: Discovering Human Motion from a Sequence of Silhouettes. *ACM Transactions on Graphics (TOG)* 35, 4, Article 147 (July 2016).

Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. *Advances in Neural Information Processing Systems (NeurIPS)* 35 (2022), 38571–38584.

Yifei Zeng, Yuanxun Lu, Xinya Ji, Yao Yao, Hao Zhu, and Xun Cao. 2023. AvatarBooth: High-Quality and Customizable 3D Human Avatar Generation. *arXiv preprint arXiv:2306.09864* (2023).

Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. 2021. PyMAF: 3D Human Pose and Shape Regression with Pyramidal Mesh Alignment Feedback Loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 11446–11456.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3836–3847.

Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. 2023. Unleashing Text-to-Image Diffusion Models for Visual Perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5729–5739.

Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. 2017. Towards 3D Human Pose Estimation in the Wild: A Weakly-Supervised Approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 398–407.

Hao Zhu, Yebin Liu, Jingtao Fan, Qionghai Dai, and Xun Cao. 2016. Video-Based Outdoor Human Reconstruction. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)* 27, 4 (2016), 760–770.

Hao Zhu, Hao Su, Peng Wang, Xun Cao, and Ruigang Yang. 2018. View Extrapolation of Human Body from a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4450–4459.

Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. 2019. Detailed Human Shape Estimation from a Single Image by Hierarchical Mesh Deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4491–4500.

Hao Zhu, Xinxin Zuo, Haotian Yang, Sen Wang, Xun Cao, and Ruigang Yang. 2021. Detailed Avatar Recovery from Single Image. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 44, 11 (2021), 7363–7379.

Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. 2024a. Champ: Controllable and Consistent Human Image Animation with 3D Parametric Guidance. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 145–162.

Shenhao Zhu, Li Wang, Xun Cao, Ruigang Yang, Xinxin Zuo, and Hao Zhu. 2024c. StreetSyn: A Full Radiance Field Solution for Street and Vehicle Free-View Synthesis. In *Proceedings of the International Conference on Neural Information Processing (ICONIP)*. 404–419.

Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. 2023. MotionBERT: A Unified Perspective on Learning Human Motion Representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 15085–15099.

Yixuan Zhu, Ao Li, Yansong Tang, Wenliang Zhao, Jie Zhou, and Jiwen Lu. 2024b. DPMesh: Exploiting Diffusion Prior for Occluded Human Mesh Recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1101–1110.

Yiyu Zhuang, Yuxiao He, Jiawei Zhang, Yanwen Wang, Jiahe Zhu, Yao Yao, Siyu Zhu, Xun Cao, and Hao Zhu. 2024. Towards Native Generative Model for 3D Head Avatar. *arXiv preprint arXiv:2410.01226* (2024).

Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. 2025. IDOL: Instant Photorealistic 3D Human Creation from a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 26308–26319.

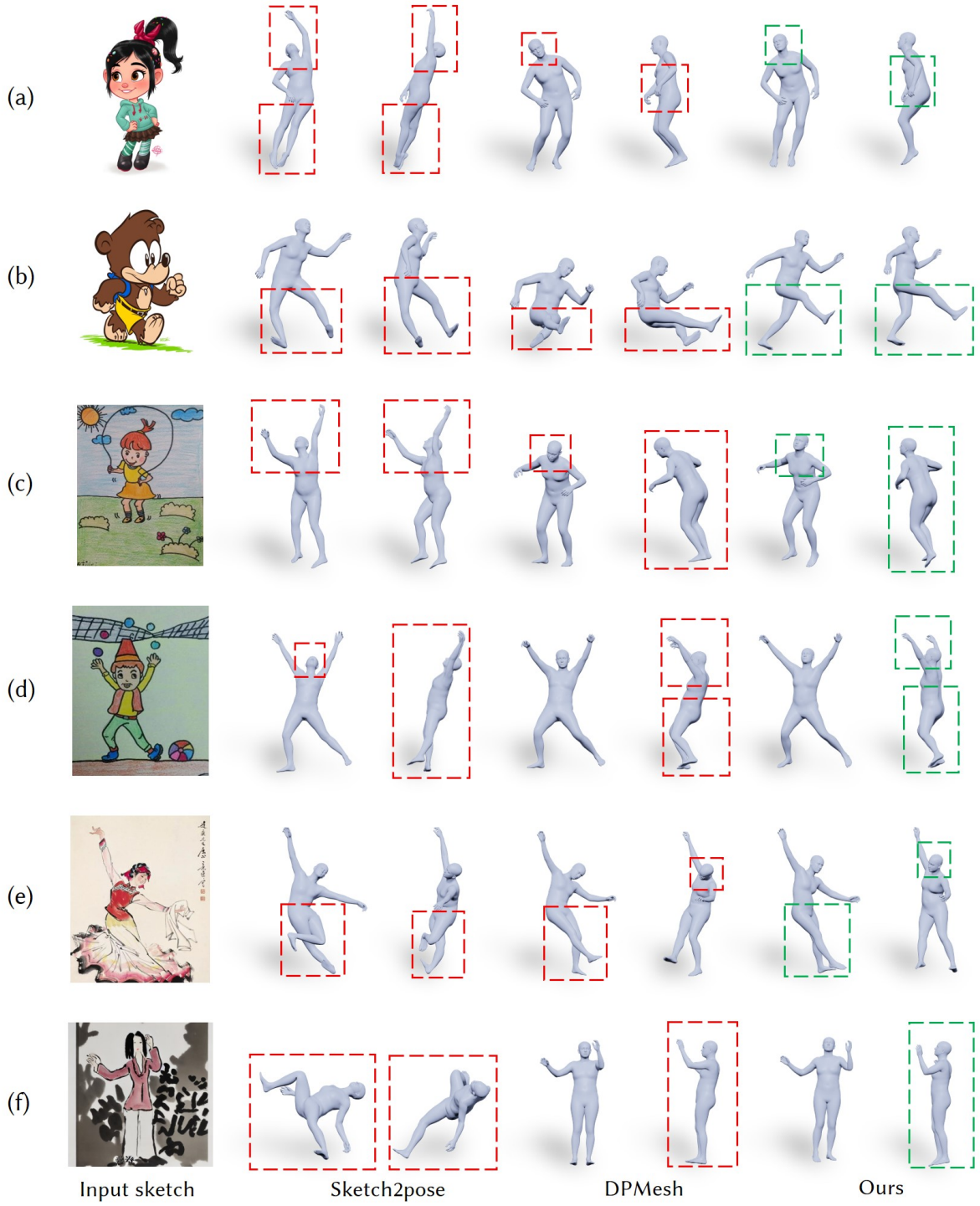

Fig. 6. **Qualitative Comparison of Multiple Sketch Styles.** Our proposed model accurately predicts real human body proportions in cartoon images and outperforms other methods in various sketch styles. Its high performance across multiple sketch styles is attributed to our three-stage pose prediction network design and diverse dataset with perturbations. The red dashed box highlights the unreasonable 3D human pose estimation.

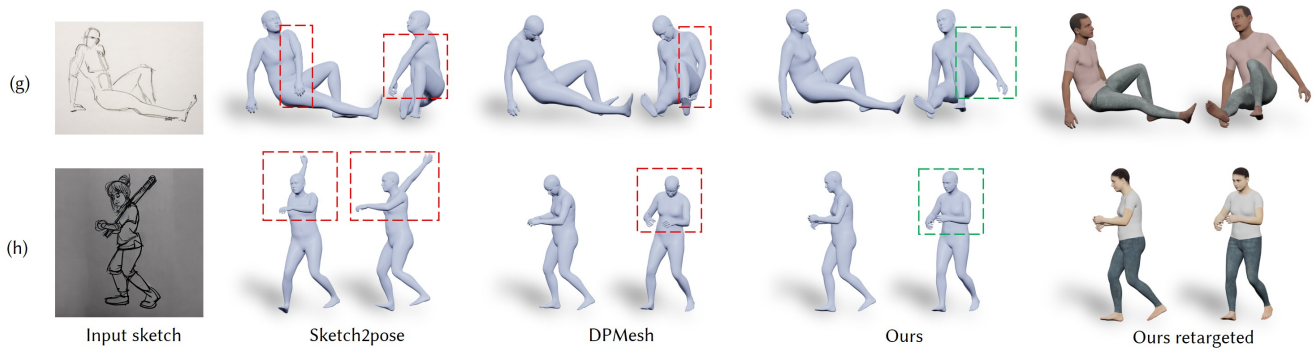

Fig. 7. **Qualitative Comparison of Challenging Poses.** Our method successfully recovers plausible poses even when character parts are obscured or lines overlap, as illustrated in specific scenarios. These results demonstrate that our approach better meets artists' needs and predicts 3D human poses with higher accuracy. The red dashed box highlights the unreasonable 3D human pose estimation. And our results can be seamlessly applied to a custom character using standard tools.

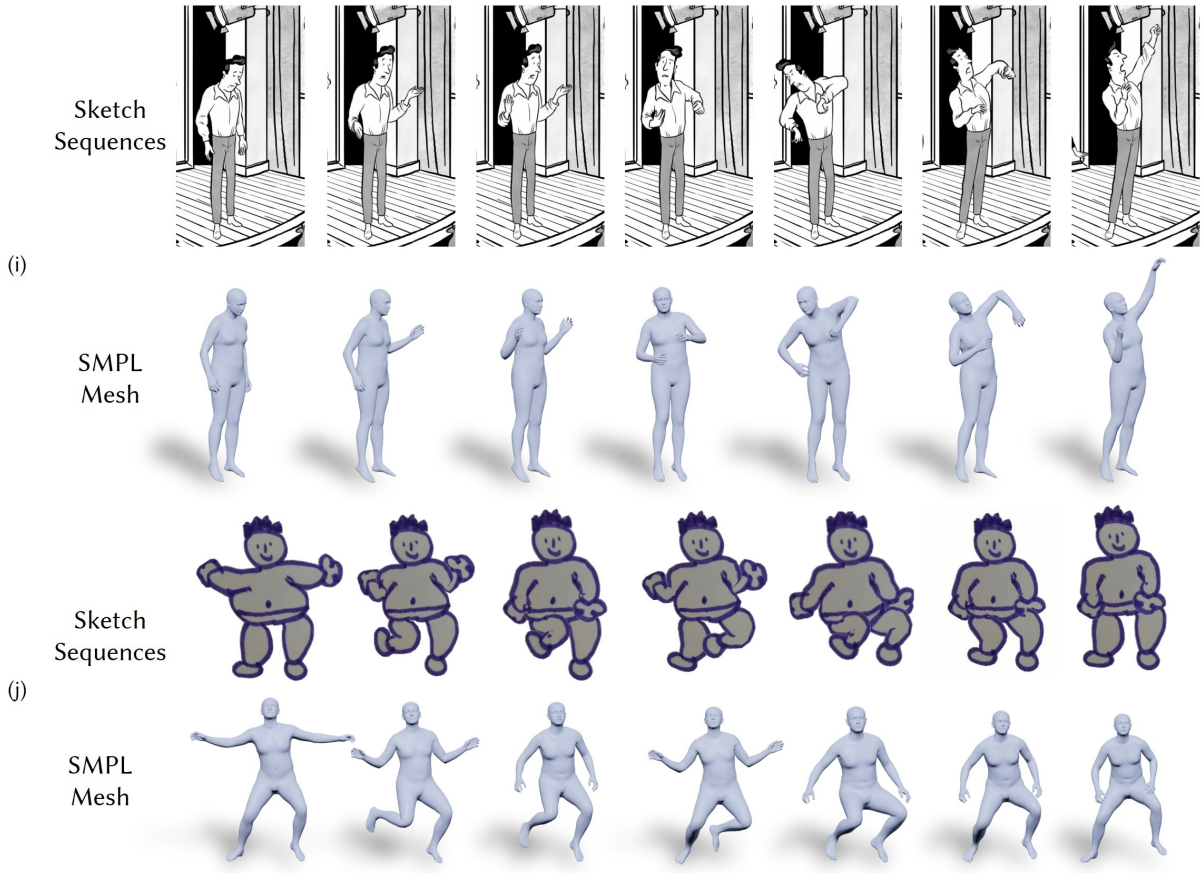

Fig. 8. **Results on Sketch Video Input.** The frame-by-frame prediction results show our method's generalization capability in extracting human poses from sketch videos.