Efficient and Generalized Sketch to 3D Human Pose Prediction

Anonymous ICCV submission

Paper ID 12236

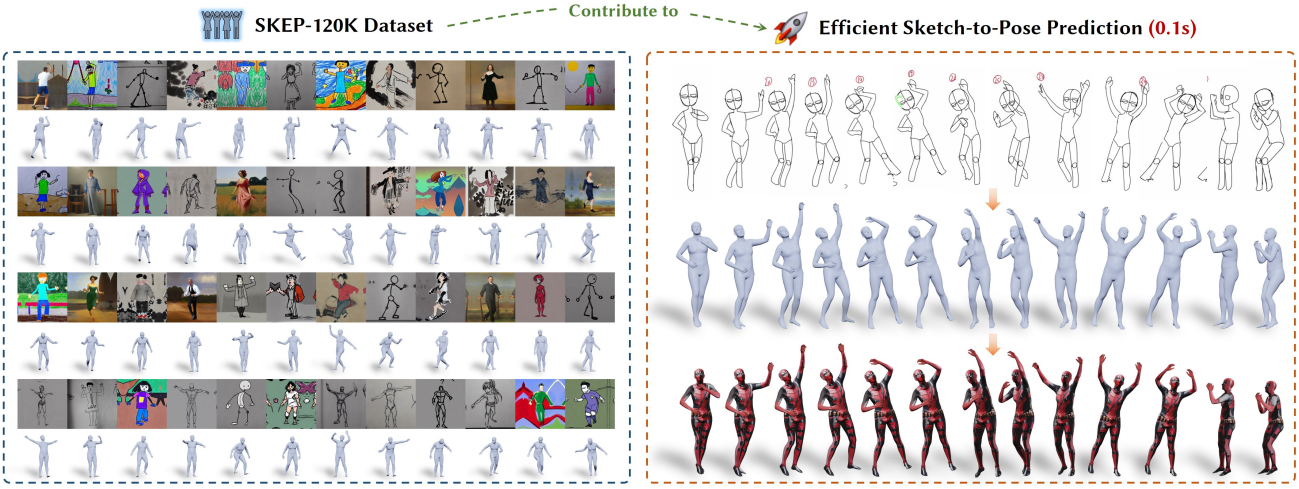

Figure 1. We present a novel approach for 3D human pose estimation from sketches. Benefiting from the large-scale SKEP-120K dataset, we are able to learn a data-driven sketch-to-pose model that exhibits improved generalization ability and efficient inference. The pose result can be automatically transferred to a custom-rigged and skinned 3D character.

Abstract

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. We first fine-tune an open-source image diffusion model on multiple sketch styles conditioned on 2D poses. This trained generator creates sketch images from 2D poses projected from 3D human poses. To mimic disproportionate cartoon-like structures, we perturb the pose structure by proportionally stretching specific body parts in the conditioning 2D poses. This process enables the creation of 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. We implement several heuristic losses to ensure geometric consistency between output 3D poses and detected 2D poses while maintaining faithful self-contacts. Through our diverse synthetic dataset and dedicated model architecture, we achieve state-of-the-art accuracy while substantially improving both inference speed and generalization capabilities. The code and data will be released upon publication.*

1. Introduction

Human pose estimation holds significant importance and finds widespread application across numerous scenarios. Among the various sources used for pose estimation, sketches emerge as a highly practical 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 [6], 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 sluggish 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. 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 iteratively 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 are still 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 [10, 14, 28, 32], silhouettes [52], and clean vector drawings [3]. These approaches, though efficient in constrained scenarios, are hindered by their reliance on unambiguous, clean inputs. For instance, Gesture3D [3] 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 [6] use deep learning 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 [12] and Motion Doodles [47] 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*. [49] and Schmitz *et al*. [42], 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 [1, 2, 39], 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 [48] being one of the first CNN-based approaches. This was followed by methods like Tekin *et al*. [46], which integrated CNNs with structured prediction to improve pose accuracy, and Martinez *et al*. [33], which proposed a fully connected network for 2D-to-3D lifting. Zhou *et al*. [57] added geometric constraints, while Pavlakos *et al*. [36] 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 [5] and SMPL-X [37] advanced pose and shape estimation with 3D human prior. Many methods [4, 13, 24, 27, 54] focus on improving the accuracy of human mesh recovery. Weakly supervised approaches like HMR [21] regressed SMPL parameters using 2D keypoints and adversarial losses. Kolotouros *et al*.'s SPIN [22] refined this approach with optimization, while EFT [19] fine-tuned SMPL predictions. And methods like 3DCrowedNet [9], JOTR [26], DPMesh [59] are designed to recover the occluded body mesh.

Recent advancements include MotionBERT [58], a transformer-based approach for long-range dependencies, and D3DP [16, 43], which used diffusion models for robust pose prediction. Self-supervised methods, such as Wang *et al*. [51] and Novotny *et al*. [35], 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 [56]. 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 [59], 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 [6, 20, 30, 44] offer a substantial number of sketches, they are limited to providing only 2D pose annotations and feature only a single sketch style. These previous sketch datasets are inadequate because of the lack of 3D poses and style diversity for training a highly accurate and 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 30,000 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 [29], 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 [37] to generate random SMPL models exhibiting diverse and plausible poses, where a variational autoencoder is designed to capture latent representations of human poses. Considering the foreshortening technique prevalent in sketch causes the difference between the character structure in sketch and that in the real-captured image, we add random biases to the generated bone length, then project them onto the 2D image planes. This process yields corresponding 3D and 2D joint annotations. Next, we aggregate data from the Sketch2Pose [6], HumanArt [20] and Amateur Drawing datasets [44]. The 2D keypoint annotations from these datasets are uniformed into our de-

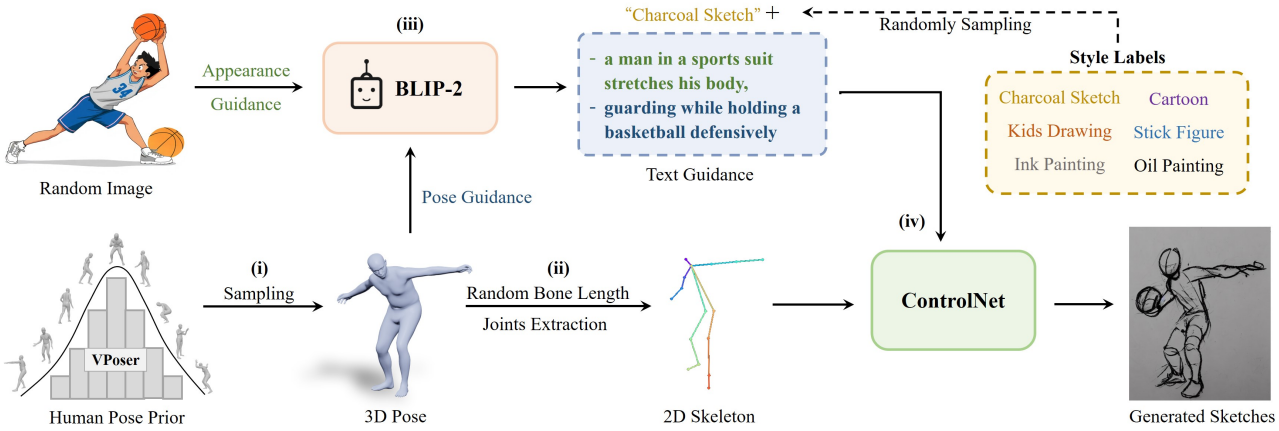

Figure 2. SKEP-120K Dataset Creation. Our dataset creation process involves the following stages: (i) generating diverse SMPL poses using VPoser; (ii) adding random biases to bone lengths and projecting them onto 2D planes; (iii) leveraging BLIP2 for text generation; (iv) training a text-conditioned image generation model to generate sketches.

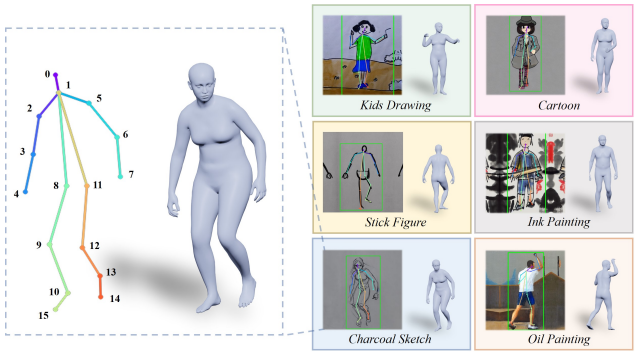

Figure 3. Data Description. SKEP-120K dataset comprises six sketch styles: cartoons, oil paintings, ink paintings, charcoal sketches, stick figures, and kids drawings. The provided 2D/3D joints are shown on the left.

fined 16 joints and are classified by drawing styles. Then, BLIP2 [25] is leveraged to generate appearance text information on sketch images and motion text information on SMPL rendering images, which we combine as the prompt condition for training a text-conditioned image generation model following ControlNet [55]. After training and manual selection, 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 [7] 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 HumanArt, 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 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:

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

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 as well as the objective functions in the following sections.

4.1. 2D Guidance Extractor

Drawing inspiration from VPD [56], 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 firstly resized and padded to the resolution of 256×192 pixels to preserve the aspect ratio. Then, YOLOX [40] is fine-tuned by Human-Art dataset for human's bounding box detection in sketches, and VITPose [53] model is

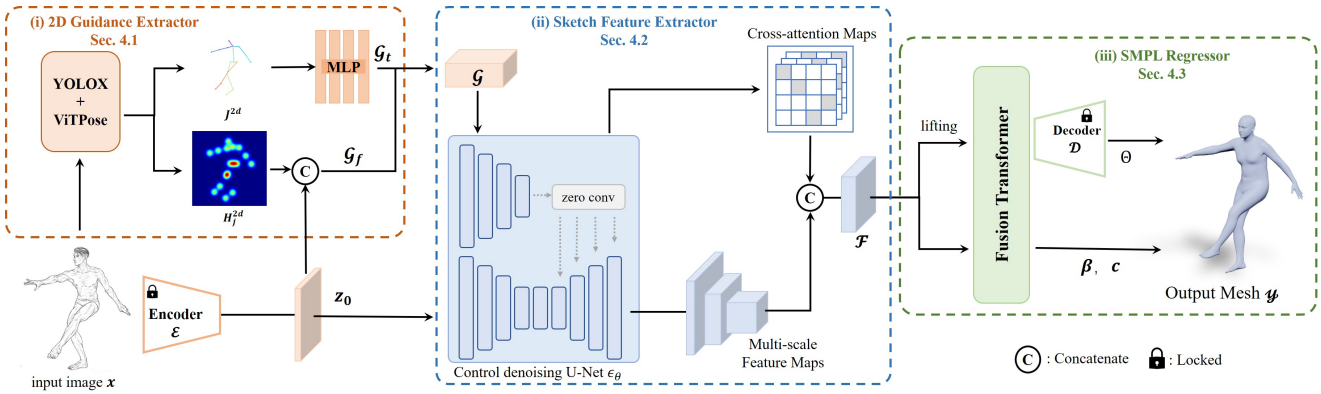

Figure 4. Overall Pipeline. Given a sketch image as input, the network predicts 3D human poses represented by SMPL parameters. The overall network consists of three modules: a 2D guidance extractor as detailed in Sec. 4.1; a sketch feature extractor as detailed in Sec. 4.2; and an SMPL regressor as detailed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

fine-tuned by SKEP-120K dataset 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 and transform them into heatmaps $H_j^{2D} \in \mathbb{R}^{K\times H^{'}\times W^{'}}$ using 2D Gaussian kernels [8].

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 [41] backbone $\boldsymbol{\epsilon_\theta}$. This process is referred as $p_{\phi_1}(\mathcal{G} \mid \boldsymbol{x})$ 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 [11, 41, 55], the prompt guidance $\mathcal{G}_t$ usually rely on text embeddings derived from a frozen CLIP [38] model. In contrast, we replaces the text with 2D joint positions J^{2D} as verified in [59]. 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, $p_{\phi_2}(\mathcal{F} \mid \epsilon(\boldsymbol{x}), \mathcal{G})$ 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) [41]. 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:

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

where $\text{F}_n(\cdot;\theta)$ is a trained neural network, $\text{Z}(\cdot;\cdot)$ denotes zero convolution layers with both weights and bias initial-

ized to zeros, θ_c represent 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.

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, $p_{\phi_3}(\boldsymbol{y} \mid \mathcal{F})$ refers to the prediction head that generates parameters of the body model from the hierarchical feature maps $\mathcal{F}$. We first lift pose-guided 2D feature $\mathcal{F}$ to 3D feature $\mathcal{F}_{3D}$, then extend 2D features by incorporating 3D joint feature sampling. To integrate and align 2D and 3D features, we utilize a fusion transformer [26] to regress SMPL parameters. Moreover, we employ a pre-trained VQ-VAE [50], which is trained on a large-scale motion dataset AMASS [31] 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: bone tangents, foreshortening and self-contacts. Building on our proposed human drawing dataset with 3D pose annotations, we introduce new methods to address this.

Bone Tangent Character bones often appear longer than their actual length because of imprecision in drawings or the use of artistic interpretation [15, 18, 45] 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 is $\bar{x}^{2D}$, so the predicted vector corresponding to bone i is $\bar{\boldsymbol{b}}_{\boldsymbol{i}}^{\boldsymbol{2D}} = \bar{x}_{j_2}^{2D} - \bar{x}_{j_1}^{2D}$, $\boldsymbol{n}$ represents the normal to the predicted bone $\bar{\boldsymbol{b}}_{\boldsymbol{i}}^{\boldsymbol{2D}}$. Guided by our principle of bone tangent, 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 Artists typically do not rely on exact mathematical measurements for orthographic or perspective projections When creating drawings [15, 45]. 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 cosine of the angle between the skeleton and the screen.

Perceived self-contacts Self-contacts, or contacts between different human body parts, are essential components of numerous poses [15]. 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 allows us to obtain the correct relative depth and joint positions of the character skeleton. In our method, we replace the previous self-contact loss with the SMPL

Table 1. Quantitative comparison on the artist-designed dataset (Unit: mm)

Method	Expert1			Expert2		
	MPVE↓	MPJPE↓	PA-MPJPE↓	MPVE↓	MPJPE↓	PA-MPJPE↓
PyMAF [54]	312.7	299.4	187.5	301.5	291.2	187.0
EFT [19]	144.6	144.4	98.9	168.8	158.9	103.4
HybrIK [24]	348.7	345.5	199.6	365.2	352.0	208.5
CLIFF [27]	186.4	181.5	142.2	217.3	201.3	144.5
HMR2.0 [13]	118.3	105.0	85.1	181.4	151.4	107.4
MotionBERT [58]	170.6	165.2	120.1	189.1	172.2	127.6
DPMesh [59]	127.7	121.4	94.1	166.4	147.1	94.3
Sketch2Pose [6]	103.8	101.4	78.1	**145.5**	135.9	86.8
Ours	**103.1**	**95.7**	**77.4**	146.5	**131.5**	**84.3**

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:

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

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 [6], 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. The difference is that we enlist artists 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 various sketch styles and extensive data volume, it is well-suited for assessing the generalization capability.

Metrics We adopt three common 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.

Training Details We trained three models in the data creation and sketch-to-pose prediction: ControlNet for generating sketch data, ViTPose for 2D keypoint prediction from sketches, and the core sketch-to-pose prediction model. For ControlNet, the condition maps and sketch images are padded to a resolution of 512×512. Then, BLIP2 [25] is leveraged to generate prompt labels for the sketch images. The ControlNet is trained for 80,000 steps with a batch size of 16. Next, we resize the collected sketch images to a resolution of 256×192 and employ it to train the ViTPose, trained for 210 epochs with a batch size of 32. The core model training process consists of two stages: In the first stage, we pre-train the model on the Human3.6M [17], MuCo-3DHP [34], MSCOCO, and CrowdPose [23] datasets for 30 epochs with a batch size of 32, using absolute joint positions as supervision to help the model learn the human pose information. In the second stage, we fine-tune the model on the SKEP-120K dataset with our designed objective function as Eq. (7) for 20 epochs with a batch size of 32, enabling the model to learn various sketch styles and can recover natural and accurate 3D pose of sketch images. All training and experiments are conducted on 4 NVIDIA A6000 GPUs.

5.2. Qualitative Comparison

The quantitative comparisons on the artist-designed dataset and SKEP-120K validation set are shown in Tab. 1 and Tab. **??**, respectively.

The Artist-designed dataset only contains six samples in charcoal sketch style, all of which are challenging non-daily poses. On this dataset, our model achieves the best overall performance of all metrics. Although our accuracy improvement on charcoal sketches is modest compared to Sketch2Pose, and we even perform slightly worse on one metric related to an expert's annotation, it is noteworthy that

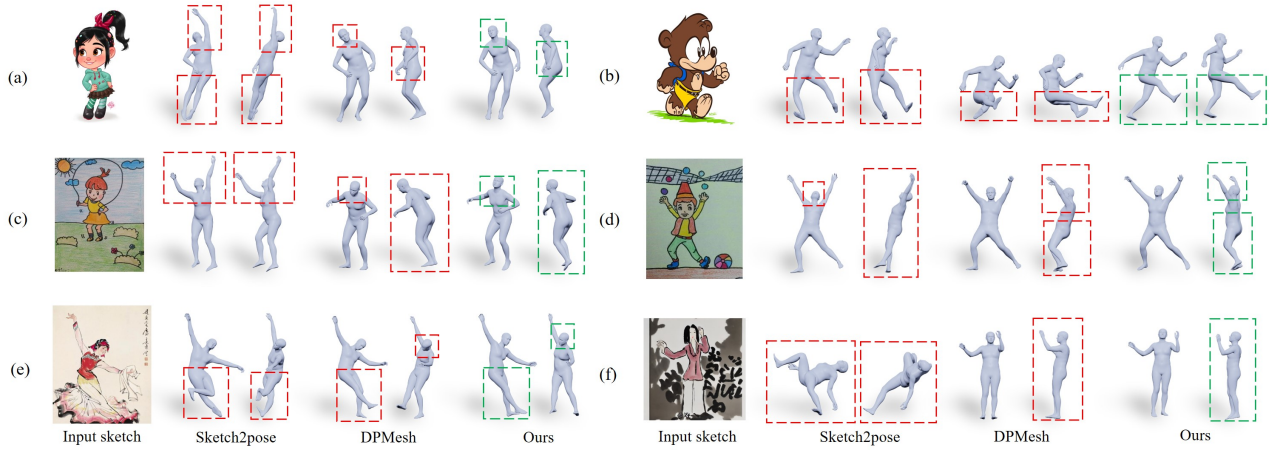

Figure 5. **Qualitative Comparison of Multiple Sketch Styles.** Our proposed model accurately predicts real human body proportions in cartoon images and outperforms other methods in various sketch styles. Its high performance across multiple sketch styles is attributed to our three-stage pose prediction network design and diverse dataset with perturbations. The red dashed box highlights the unreasonable 3D human pose estimation.

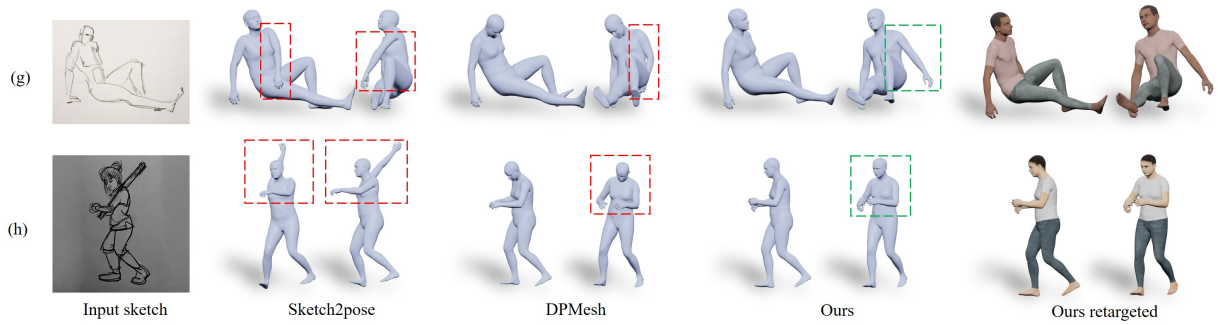

Figure 6. **Qualitative Comparison of Challenging Poses.** Our method successfully recovers plausible poses even when character parts are obscured or lines overlap, as illustrated in specific scenarios. These results demonstrate that our approach better meets artists' needs and predicts 3D human poses 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.

Table 2. Quantitative comparison of methods on the SKEP-120K validation set.

Method	MPVE↓	MPJPE↓	PA-MPJPE↓
PyMAF [54]	143.1	117.4	101.3
EFT [19]	158.7	133.1	111.6
HybrIK [24]	211.0	177.7	144.0
CLIFF [27]	137.3	113.3	98.0
HMR2.0 [13]	128.0	104.6	88.1
MotionBERT [58]	124.4	99.4	83.6
DPMesh [59]	122.6	97.3	80.6
Sketch2Pose [6]	165.6	146.4	126.6
Ours	**106.7**	**87.7**	**72.6**

Table 3. Quantitative comparison of ablation study.

Method vs Expert 1	MPVE↓	MPJPE↓	PA-MPJPE↓
Ours (w/o L_{parallel})	169.8	165.7	102.5
Ours (w/o L_{f})	117.7	110.2	84.4
Ours (w/o L_{pose})	121.4	117.6	86.7
Ours	**103.1**	**95.7**	**77.4**

Table 4. Runtime of our method.

Phase	I (Sec. 4.1)	II (Sec. 4.2)	III (Sec. 4.3)	Total Time
Time (ms)	44.3	49.1	31.8	125.2

our method achieves this level of accuracy in roughly $1/500$ of the time taken by Sketch2Pose.

The SKEP-120K validation set contains sketches in a

larger amount and in more diverse styles, including six different sketch styles. On this dataset, our model outperforms Sketch2Pose and DPMesh by a large margin in all six sketch styles.

The above quantitative experiments demonstrate that our method balances high accuracy with remarkable efficiency. The generalization performance of our model far exceeds not only the generic pose estimation algorithm (like DPMesh, MotionBERT), but also the Sketch2Pose algorithm for sketches. Furthermore, its rapid inference speed allows sketch-based 3D pose estimation to extend beyond static images and effectively process video data.

5.3. Qualitative Comparison

We visualized the comparison results on the SKEP-120K dataset, artist-designed dataset, and Internet images and videos. Fig. 5 presents the prediction results for sketches of different styles. Notably, (a) - (b) demonstrate that our proposed model can accurately predict poses that adhere to real human body proportions, even in cartoon images where the body proportions deviate significantly from reality. (c) - (f) highlight that our method yields more accurate predictions for cartoons, children's drawings, stick figures, ink paintings, and oil paintings. We can see that Sketch2Pose encounters difficulties with inputs beyond charcoal sketches, while DPMesh, a general-purpose image-to-human pose prediction network, experiences a significant drop in prediction accuracy when dealing with various sketch styles. By contrast, our method is the only model capable of maintaining high performance across multiple sketch styles. We attribute our model's effective feature extraction and enhanced generalization to our three-stage pose prediction network design for sketch feature extraction and our synthesized dataset with perturbations.

Fig. 6 shows the prediction results for some challenging human poses. Specifically, (g) illustrates a scenario where one of the character's arms is completely obscured, yet our method successfully recovers the most plausible pose. In (h), the lines of the characters and a handheld object overlap, presenting a simplified structural line. Despite this, our method accurately predicts the correct pose. These visualized results indicate that our approach better meets artists' creative needs and predicts 3D human poses more accurately. We then present the results of retargeted pose to custom characters. As shown in Fig. 1, our method facilitates frame-by-frame prediction due to its efficient reasoning speed, thereby significantly improving the applicability and effectiveness of sketch-to-pose.

Fig. **??** presents the frame-by-frame prediction results of our method applied to continuous line animations. This demonstrates the method's generalization capability to extract human poses from sketch videos.

5.4. Ablation Study

We conduct an ablation study on the artist-designed dataset to verify the effectiveness of our proposed loss terms. Specifically, we ablate the loss terms L_{parallel}, L_{f}, and L_{pose} from Eq. (4.4) and retrain the model with all other settings the same. The performance of these ablated models is reported in Tab. 3. We can see that the ablation clearly degrades the performance, and the full objective function turns out to be the most effective design.

5.5. Runtime Analysis

We present the step-by-step runtime of our method in Table 4. Our algorithm outperforms Sketch2Pose by approximately 500 times in speed while achieving superior performance, which is attributed to our efficient feed-forward neural network design.

6. Conclusion

We present a novel approach to the challenging problem of human pose estimation 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. Our method does have some limitations. Firstly, the predicted terminal joints in the pose, like the wrist, ankle, and head, are relatively less accurate. We attribute this primarily to the fact that most sketches convey fewer details of these subtle poses, making the prediction even more difficult. Additionally, excessively complex backgrounds can disrupt the method's predictions, highlighting the need to enhance its resilience to environmental interference.

References

[1] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Monocular 3d pose estimation and tracking by detection. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 623–630. Ieee, 2010. 3

[2] Alexandru O Balan, Leonid Sigal, Michael J Black, James E Davis, and Horst W Haussecker. Detailed human shape and pose from images. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 3

[3] Mikhail Bessmeltsev, Nicholas Vining, and Alla Sheffer. Gesture3d: Posing 3d characters via gesture drawings. *ACM Trans. Graph.*, 35(6), 2016. 2

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a

single image. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 561–578. Springer, 2016. 3

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image, 2016. 3

[6] Kirill Brodt and Mikhail Bessmeltsev. Sketch2pose: estimating a 3d character pose from a bitmap sketch. *ACM Transactions on Graphics (TOG)*, 41(4):1–15, 2022. 2, 3, 7, 8

[7] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986. 4

[8] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020. 5

[9] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. 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*, pages 1475–1484, 2022. 3

[10] James Davis, Maneesh Agrawala, Erika Chuang, Zoran Popović, and David Salesin. A Sketching Interface for Articulated Figure Animation. *Proc. Symposium on Computer Animation*, pages 320–328, 2003. 2

[11] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 5

[12] Marek Dvorožňák, Daniel Sỳkora, Cassidy Curtis, Brian Curless, Olga Sorkine-Hornung, and David Salesin. Monster mash: a single-view approach to casual 3d modeling and animation. *ACM Transactions on Graphics (ToG)*, 39(6):1–12, 2020. 3

[13] 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*, pages 14783–14794, 2023. 3, 7, 8

[14] Ronie Hecker and Kenneth Perlin. Controlling 3d objects by sketching 2d views. *Proc. SPIE*, 1828:46–48, 1992. 2

[15] Burne Hogarth. Dynamic figure drawing. *(No Title)*, 1970. 6

[16] Karl Holmquist and Bastian Wandt. Diffpose: Multi-hypothesis human pose estimation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15977–15987, 2023. 3

[17] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 7

[18] Ollie Johnston and Frank Thomas. *The illusion of life: Disney animation*. Disney Editions New York, 1981. 6

[19] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. *2021 International Conference on 3D Vision (3DV)*, pages 42–52, 2021. 3, 7, 8

[20] Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, and Lei Zhang. 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*, pages 618–629, 2023. 3

[21] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7122–7131. IEEE Computer Society, 2018. 3

[22] 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 International Conference on Computer Vision*, 2019. 3

[23] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019. 7

[24] 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*, pages 3383–3393, 2021. 3, 7, 8

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 4, 7

[26] Jiahao Li, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, and Yi Yang. Jotr: 3d joint contrastive learning with transformers for occluded human mesh recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9110–9121, 2023. 3, 6

[27] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pages 590–606. Springer, 2022. 3, 7, 8

[28] Juncong Lin, Takeo Igarashi, Jun Mitani, and Greg Saul. A sketching interface for sitting-pose design. In *Proc. Sketch-Based Interfaces and Modeling Symposium*, pages 111–118, 2010. 2

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[30] Prathmesh Madhu, Angel Villar-Corrales, Ronak Kosti, Torsten Bendschus, Corinna Reinhardt, Peter Bell, Andreas

Maier, and Vincent Christlein. Enhancing human pose estimation in ancient vase paintings via perceptually-grounded style transfer learning. *ACM Journal on Computing and Cultural Heritage*, 16(1):1–17, 2022. 3

[31] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5442–5451, 2019. 6

[32] C Mao, S F Qin, and D K Wright. A sketch-based gesture interface for rough 3D stick figure animation. *Proc. Sketch Based Interfaces and Modeling*, 2005. 2

[33] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 3

[34] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 7

[35] David Novotný, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3DPO: canonical 3d pose networks for non-rigid structure from motion. *CoRR*, abs/1909.02533, 2019. 3

[36] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7025–7034, 2017. 3

[37] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 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*, pages 10975–10985, 2019. 3

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5

[39] D. Ramanan. Part-based models for finding people and estimating their pose. *Springer*, pages 199–223, 2011. 3

[40] J Redmon. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 4

[41] 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*, pages 10684–10695, 2022. 5

[42] Caro Schmitz, Constantin Rösch, Domenic Zingsheim, and Reinhard Klein. Interactive pose and shape editing with simple sketches from different viewing angles. *Computers & Graphics*, 114:347–356, 2023. 3

[43] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14761–14771, 2023. 3

[44] Harrison Jesse Smith, Qingyuan Zheng, Yifei Li, Somya Jain, and Jessica K Hodgins. A method for animating children's drawings of the human figure. *ACM Transactions on Graphics*, 42(3):1–15, 2023. 3

[45] Walt Stanchfield. Gesture drawing for animation. *Washington: Leo Brodie*, 1, 2007. 6

[46] Bugra Tekin, Artem Rozantsev, Vincent Lepetit, and Pascal Fua. Direct prediction of 3d body poses from motion compensated sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[47] Matthew Thorne, David Burke, and Michiel Van De Panne. Motion doodles: an interface for sketching character motion. *ACM Transactions on Graphics (ToG)*, 23(3):424–431, 2004. 3

[48] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 3

[49] Gizem Unlu, Mohamed Sayed, and Gabriel Brostow. Interactive sketching of mannequin poses. In *2022 International Conference on 3D Vision (3DV)*, pages 700–710. IEEE, 2022. 3

[50] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 6

[51] Keze Wang, Liang Lin, Chenhan Jiang, Chen Qian, and Pengxu Wei. 3d human pose machines with self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 3

[52] Jungdam Won and Jehee Lee. Shadow theatre: Discovering human motion from a sequence of silhouettes. *ACM Trans. Graph.*, 35(4), 2016. 2

[53] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022. 4

[54] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11446–11456, 2021. 3, 7, 8

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 4, 5

[56] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5729–5739, 2023. 3, 4

[57] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild:

A weakly-supervised approach. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[58] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15085–15099, 2023. 3, 7, 8

[59] Yixuan Zhu, Ao Li, Yansong Tang, Wenliang Zhao, Jie Zhou, and Jiwen Lu. Dpmesh: Exploiting diffusion prior for occluded human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1101–1110, 2024. 3, 5, 7, 8