Sketch2Pose: Estimating a 3D Character Pose from a Bitmap Sketch

KIRILL BRODT and MIKHAIL BESSMELTSEV, Université de Montréal, Canada

Fig. 1. Given a single natural bitmap sketch of a character (a), our learning-based approach allows to automatically, with no additional input, recover the 3D pose consistent with the viewer expectation (b). This pose can be then automatically copied a custom rigged and skinned 3D character (c) using standard retargeting tools (d). Input image © Olga Posukh.

Artists frequently capture character poses via raster sketches, then use these drawings as a reference while posing a 3D character in a specialized 3D software — a time-consuming process, requiring specialized 3D training and mental effort. We tackle this challenge by proposing the first system for automatically inferring a 3D character pose from a single bitmap sketch, producing poses consistent with viewer expectations. Algorithmically interpreting bitmap sketches is challenging, as they contain significantly distorted proportions and foreshortening. We address this by predicting three key elements of a drawing, necessary to disambiguate the drawn poses: 2D bone tangents, self-contacts, and bone foreshortening. These elements are then leveraged in an optimization inferring the 3D character pose consistent with the artist’s intent. Our optimization balances cues derived from artistic literature and perception research to compensate for distorted character proportions. We demonstrate a gallery of results on sketches of numerous styles. We validate our method via numerical evaluations, user studies, and comparisons to manually posed characters and previous work.

Code and data for our paper are available at http://www-labs.iro.umontreal.ca/ bmpix/sketch2pose/.

CCS Concepts: • Computing methodologies → Mesh geometry models.

Additional Key Words and Phrases: character posing, rigged and skinned characters, sketch-based posing, character sketches

ACM Reference Format:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery. 0730-0301/2022/7-ART85 $15.00 https://doi.org/10.1145/3528223.3530106

1 INTRODUCTION

“Art is not what you see, but what you make others see.”
— Edgar Degas

Artists routinely capture human poses via diverse drawings, from quick gestures to detailed character sketches. Sketching character poses is one of the core elements in artist training. In modern digital media production, artists often draw sketches of characters at the early storyboarding stage. Often drawn within tens of seconds each, those sketches serve as efficient and direct means to capture ideas and convey the poses to other team members. These natural, freely drawn sketches are then often used as a reference while manually posing a 3D character in an animation software.

The manual posing step, however, requires special training, and is tedious, time-consuming, taking up several minutes per pose even for a rough draft. For professionals, it is a frustrating task that may distract from the creative process and slow down the production pace; for classically trained artists with little knowledge of 3D animation software, it requires specialized training and thus may become an obstacle to implementing their ideas.

A system enabling artists to directly use a sketch as the only input to automatically pose a rigged and skinned 3D character thus would significantly simplify and democratize posing 3D characters, benefiting both novice and professional users. Creating such a system, unfortunately, faces significant challenges: First, contrary to the assumptions of previous work that targeted clean vector drawings of a single style [Bessmeltsev et al. 2016], sketches are often drawn on traditional media, such as pen and paper, or in a raster drawing software, are stylistically various, and are full of construction lines and extra strokes. Converting those images into clean vectorized drawings remains an open problem [Stanko et al. 2020]. More importantly, sketches are imprecise, incomplete, often contain occlusions, and may substantially distort character’s proportions, whether due to errors [Schmidt et al. 2009] or artistic license: As the Degas’s quote can be interpreted, drawings are not created to be a perfect
We further observe that depicted self-contacts (e.g., forearms and thighs in Fig. 1) are critical for understanding of a drawn pose and can be strong indicators of intended 3D bone direction. As outlined above, 2D joint positions themselves are unreliable, 2D projections of the 3D character onto the view plane (Fig. 5): drawn with a goal to capture and convey the essence of the pose, they are only approximate depictions of the character.

More specifically, drawings often significantly distort body proportions or depict characters with unrealistic body shapes (Fig. 4). While proportions are often distorted even for sketches of static poses, such distortion is a core and intentional component of gestures of dynamic poses [Kwon and Lee 2012]. Furthermore, artists use nonlinear [Singh 2002], grossly exaggerated or otherwise inaccurate perspective [Schmidt et al. 2009; Sudarsanam et al. 2008; Zhong et al. 2020], and incorrectly depict foreshortening (Fig. 6) [Wnuczko et al. 2016]. Naively ignoring these issues and applying standard optimizations relying on 2D joint positions only, may lead to wrong or imprecise results (Fig. 4).

We center our analysis on three elements of a character drawing that we believe are essential to resolving these issues: bone tangents, self-contacts, and foreshortening (Fig. 2). We observe that while, as outlined above, 2D joint positions themselves are unreliable, 2D bone tangents can be strong indicators of intended 3D bone direction. We further observe that depicted self-contacts (e.g., forearms and thighs in Fig. 1) are critical for understanding of a drawn pose and can be strong cues to disambiguate the unknown body part depth. Finally, we explicitly model and undo the distortions of the bone foreshortening using statistical analysis.

Overview. We introduce a novel system for inferring a 3D character from a single bitmap based a combination of optimization, deep learning, statistical analysis, and observations from perceptual and artistic literature. Our optimization is guided by three main subsystems predicting 2D bone tangents, self-contacts, and bone foreshortening. Equipped with these three predictions, we use a state-of-the-art optimization framework with a novel loss designed specifically to compensate for the inaccuracies in the natural drawings. Our optimization balances the pose realism against the image cues, while allowing body shape to change. This optimization enables our system to infer complex 3D poses with significantly distorted body lengths and proportions (Sec. 7). We infer the pose of a parameterized human model SMPL-X [Pavlakos et al. 2019], which can be automatically transferred to a custom character via standard animation software or modern retargeting systems [Aberman et al. 2020].

Contribution. Our contribution is two-fold:

- **We present the first large-scale dataset of 2D pose annotations on character sketches.** Our dataset includes 14,462 skeletons, each consisting of 18 manually annotated 2D labels with locations of joints. Around 1000 images also contain 2D locations of self-contacts. In addition, we have collected and annotated a smaller dataset containing 310 high-resolution raster character sketches, collected from a variety of artists highlighting different styles and characters, that we share with the most permissive usage license (CC-BY).

- **More importantly, we present the first framework that algorithmically reconstructs a 3D character pose directly from a single natural sketch.** Our framework supports many sketch styles, including gesture drawings.

We validate our system in a number of ways (Sec. 7). First, we present a gallery of 3D character poses computed automatically and without any additional input from natural sketches. Second, we perform user studies demonstrating the efficacy of our method. Finally, we quantitatively and qualitatively compare our algorithmic results to manually posed characters and previous work.
A common technique to pose a 3D character is via time-consuming
when an exact solution satisfying the 2D keypoints does not exist, may
(a) (b) (c)
original view alternate view original view alternate view original view alternate view
works. Input image (f) © Olga Posukh.

To our knowledge, no algorithm is capable of inferring a 3D char-
2 RELATED WORK
acter pose from a natural raster character sketch. The two closest
areas of relevant previous work are character posing interfaces and
human pose inference from an RGB photograph. We focus on the
most relevant works.

2.1 Character Posing Interfaces
A common technique to pose a 3D character is via time-consuming
direct manipulation of joint angles (Forward Kinematics, FK) or
joint positions (Inverse Kinematics, IK) [Zhao and Badler 1994]. The
problem of IK is inherently underdetermined. In their pioneer work,
Grochow et al. [2004] address this issue via a Gaussian process–
based model trained on a motion capture dataset, requiring exact
and feasible positions of terminal joints.

Previous research explored a various range of alternative inputs
to a posing system, including stick figures [Davis et al. 2003; Hecker
and Perlin 1992; Lin et al. 2010; Mao et al. 2005], lines of actions [Guay et al.
2013, 2015] (a), or silhouettes [Won and Lee 2016] (c); or unambiguous, but
clean vector curve drawings [Bessmeltsev et al. 2016] (d), which are hard
if possible to obtain automatically from the raster drawings artists create.
Our framework allows to pose 3D characters directly via natural bitmap
character sketches of different styles, including rough gesture drawings (e)
and detailed character sketches (f), containing inaccuracies, ambiguities,
extra strokes, and rudimentary shading. We furthermore show that our
method works for rasterized clean vector drawings (d) explored in previous
works. Input image (f) © Olga Posukh.

Fig. 4. Distinguishing body part foreshortening, i.e. intentional out-of-plane
rotation, from atypical or inaccurate body proportions when no out-of-plane
rotation is intended, is ambiguous. An underestimation of body part length,
when an exact solution satisfying the 2D keypoints does not exist, may
lead to an inexpressive least-squares solution (a, see e.g. the character’s left
elbow); an overestimation leads to an unintended foreshortening (b, see e.g.
the character’s right knee). Our method addresses this ambiguity (c).

Fig. 3. The previous posing approaches were constrained to working with
either ambiguous inputs, such as stick figures [Davis et al. 2003; Hecker
2013, 2015] (b), or silhouettes [Won and Lee 2016] (c); or unambiguous, but
clean vector curve drawings [Bessmeltsev et al. 2016] (d), which are hard
if possible to obtain automatically from the raster drawings artists create.
Our framework allows to pose 3D characters directly via natural bitmap
character sketches of different styles, including rough gesture drawings (e)
and detailed character sketches (f), containing inaccuracies, ambiguities,
extra strokes, and rudimentary shading. We furthermore show that our
method works for rasterized clean vector drawings (d) explored in previous
works. Input image (f) © Olga Posukh.

Fig. 5. Unlike photographs, character sketches cannot be interpreted as
projections of a 3D character onto the screen. Left: input sketch, middle: a
character posed manually by an expert, right: our automatic result.

Fig. 6. One of the main sources of distortions in character sketches is
unreliably depicted body part lengths. In contrast to a perfect projection (c,
manually posed by an expert given image (b) as a reference), artist routinely
draw bones longer (d, red), often exceeding full their 3D length, or shorter
(d, blue) than their correct projection. In (d) we color the bones based on
the ratio of their depicted length and their correct 2D projection length in
(c), from blue to red.

[Glauer et al. 2016], and clean vector drawings [Bessmeltsev et al.
2016].

Stick figures (Fig. 3a) and silhouettes (Fig. 3b) [Won and Lee 2016]
are inherently ambiguous even for human observers [Bessmeltsev et al.
2016]. This ambiguity is often resolved via manual annotation
[Davis et al. 2003; Hecker and Perlin 1992; Mao et al. 2005], physical
constraints [Lin et al. 2010], or by putting user in the loop [Davis
et al. 2003]. Some works restrict output [Jain et al. 2012] or measure
proximity to human pose datasets [Choi et al. 2012; Wei and Chai
2011]. These methods are sensitive to inaccurate positioning of 2D stick figures [Davis et al. 2003]. Character sketches are imprecise
yet drawn to unambiguously convey a pose to a human observer
and thus contain the necessary cues for a reconstruction, which
we analyze and incorporate in our optimization, allowing us to
overcome the drawing distortions and inaccuracies.

In contrast to alternative posing interfaces, such as tangible de-
vice [Glauer et al. 2016], multi-view incremental approaches [Guay
et al. 2013], or sketch abstractions [Hahn et al. 2015], we infer a
3D pose directly from a single natural character sketch. Our sys-
tem thus allows artists to convert existing drawings, for instance,
the storyboards typically created during the ideation and planning
stage, into 3D poses without spending extra effort on posing.
Choi et al. [2016] introduce a motion editing system via sketch strokes, requiring an existing motion of a character as an input. Their system is complementary to ours, as we focus on reconstructing a static pose from a single natural sketch.

Gesture3D [Bessmeltsev et al. 2016] reconstructs a character pose from a clean vector drawing (Fig. 3d), assuming no extra strokes and precise connectivity (e.g. T-junctions). Sketches found in the wild are often rich with extra strokes, shading elements, and imprecise connections (Fig. 3e, f) and often cannot be automatically vectorized to that precision [Stanko et al. 2020]. Our system directly accepts such sketches as input. Furthermore, they minimize foreshortening, assuming “flat” 3D poses where each body part is nearly parallel to the screen; we explicitly predict foreshortening, lifting that assumption. We compare to Gesture3D in Section 7.

2.2 Human Pose Estimation from a Single Photograph

The problem of inferring a 3D human pose from a single RGB photograph, or monocular pose estimation, is an extensively studied topic in computer vision. Traditional approaches [Chen et al. 2011; Gall et al. 2010; Ionescu et al. 2014; Ramanan 2011; Sapp et al. 2010; Yang and Ramanan 2013] relied on custom image-based features and often used traditional machine learning techniques. Those approaches have been largely superseded by the deep learning-based approaches. Here we only outline the most relevant works for a survey, see, e.g. Chen et al. [2020].

3D Pose Estimation. Many learning-based approaches predict 3D human poses relying on large 3D datasets with the corresponding images [Pavlakos et al. 2017; Rogez et al. 2017; Tekin et al. 2016; Tomè et al. 2017; Toshev and Szegedy 2014; Zhou et al. 2016] or combining those with 2D in-the-wild pose datasets [Mehta et al. 2017a; Tekin et al. 2017; Yang et al. 2018; Zhou et al. 2017]. This line of work has been extended by enforcing skeletal consistency [Mehta et al. 2017b; Shi et al. 2020; Sun et al. 2017], joint constraints [Akhter and Black 2015; Mehta et al. 2020], or bone lengths [Dabral et al. 2017; Ramakrishna et al. 2012; Wang et al. 2019a]. All these methods require a significant amount of 3D labeled data. For our task, we would need thousands sketches and their corresponding 3D poses; no such dataset exists.

Some works sidestep this dependency on full 2D image – 3D skeleton annotations by using unpaired 2D-3D data [Tung et al. 2017] or by relying on well-established methods in 2D pose estimation [Cao et al. 2019; Carreira et al. 2016; Chen et al. 2018; Newell et al. 2016; Papandreou et al. 2017] and focusing on the 3D lifting in a supervised [Martinez et al. 2017] or self-supervised manner [Novotny et al. 2019]. Trained in a fully supervised manner, Supervised learning is infeasible in our context; unsupervised learning would require a model of how 3D joints get projected onto 2D labels. As discussed in Sec. 1, character sketches cannot be interpreted as perfect projections of 3D characters, but rather are artistic depictions of those. It is thus unclear how to define such projection models that support incorrectly drawn perspective and distortions of body proportions, typical for character sketches. Our system targets to infer the pose consistent with the artist intent, despite the distortions and ambiguities (Fig. 4c).

3D Shape Estimation. We are inspired by an alternative line of work that predicts pose and shape of a human simultaneously [Bogo et al. 2016; Dwivedi et al. 2021; Kanazawa et al. 2018; Madadi et al. 2018; Pavlakos et al. 2019, 2018; Xiang et al. 2019; Xu et al. 2019]. They represent the shape and pose via a parametric model of human body shapes, such as Skinned Multi-Person Linear Model (SMPL) [Loper et al. 2015] or Adam [Joo et al. 2018]. Madadi et al. [2018] predict SMPL parameters via first inferring 3D heatmaps, which are then processed via a denoising autoencoder and fed into an MLP. Kanazawa et al. [2018] regress SMPL parameters using weak supervision: they match the 2D keypoints locations and capitalize on the known structure of SMPL space that allows for efficient compact natural pose discriminators. Kolotouros et al. [2019] directly regress SMPL parameters on labels obtained with optimization-based approach of Bogo et al. [2016]. Joo et al. [2021] further improve results via using a fine-tuning framework based on a regression model. Müller et al. [2021] focus on predicting 3D shape and poses with self-contacts, using either SPIN [Kolotouros et al. 2019] or EFT [Joo et al. 2021] as the underlying framework. A recent continuation of this line of work by Dwivedi et al. [2021] focuses on clothed people. Specifically, they introduce a differentiable semantic rendering loss that distinguishes between clothed and minimally-clothed regions. Our self-contact detection uses the loss from Müller et al. [2021]. In contrast to their work that relies on 3D pose estimation to predict self-contacts, we predict the depicted self-contacts directly from the input image and map those onto the 3D model (Sec. 4.3). Its contemporaneous work [Fieraru et al. 2021] similarly relies on a dataset containing mesh regions in contact. Our dataset only contains image-space locations of self-contacts — more ambiguous yet easier to collect. We use the framework of Joo et al. [2021] with Kolotouros et al. [2019]. Our work, however, differs in two important aspects. First, [Joo et al. 2021] target reconstructing the 3D pose such that predicted 2D labels are projections of the 3D joints — a natural requirement for photographs. Our goal, however, is different: we aim to predict the artist-intended 3D pose whose projection may deviate significantly from the drawn sketch — this is reflected in our novel loss formulation (Sec. 5) that includes explicitly predicting bone foreshortening (Sec. 4.4). We compare with those reprojection-based approaches in Figs. 4, 14, Sec. 7, and Supplementary Materials.

3.1 Key Principles

Foreshortening. Artistic depiction of foreshortening is often far from accurate (Fig. 6). While drawing, artists do not use precise
While unreliable depiction of bone lengths invalidates the direct use of absolute joint positions, artistic literature repeatedly stresses the importance of correct depiction of joint angles. Angles are considered one of the key elements of the drawing, creating pose expressiveness and dynamism [Walt Stanchfield 2020]. We speculate therefore that in interpreting a drawn pose, human observers resolve the inaccuracies in absolute positions by relying on correctly drawn angles: both joint angles and bone tangents, i.e. angles they form with the coordinate axes. We therefore aim to preserve bone tangents, i.e. their angles with the coordinate axes. In other words, we expect the 3D bone projection to be parallel to the depicted bones in 2D, subject to regularity cues. Clearly, this also guides the reconstructed 3D joint angles to have similar projections to the depicted 2D joint angles.

**Perceived self-contacts.** Self-contacts, or contacts between different body parts, are key elements of many poses [Hogarth 1996]. Depending on a drawing, self-contacts may be explicitly drawn (Fig. 11, left) or somewhat ambiguously suggested (e.g. Fig. 7). We speculate that human observers use perceived self-contacts as one of the cues to resolve depth ambiguity, associating similar depths to touching body parts. Clearly, for a 3D character pose to be similar to the drawing in the original view, the depicted self-contacts must be preserved, regardless of the difference in character’s proportions. We therefore predict perceived contacts between different body parts using a neural network and enforce those during optimization. For each predicted contact in the drawing, we consider the participating body parts, and both preserve their 2D relative positions at the point of contact, as well as enforce true 3D contacts between them.

**Pose Naturalness and Regularity.** Finally, we observe, consistently with the previous work [Bessmeltsve et al. 2016; Xu et al. 2014], that human observers rely on Gestalt simplicity cues [Koffka 1955].

---

**Tangent Significance.** For orthographic or perspective cameras, the length of a bone’s 2D projection would never exceed its full length. For sketches, however, this is not so. Character bones often extend beyond their normal length on the drawings due to drawing inaccuracies or artistic license [Hogarth 1996; Thomas and Johnston 1981; Walt Stanchfield 2020]. In those cases, even when character’s body proportions are known, an exact 3D configuration for the given projection does not exist, and the least-squares solution fails to match the expressiveness of the drawing (Fig. 4a).
in interpreting drawings. We speculate that given the approximate nature of character sketches, viewers use regularity cues such as symmetry and parallelism, as well as expect the pose to be close to natural. We leverage regularity as one of the cues in our optimization and use an existing framework biasing the result towards natural poses [Joo et al. 2021].

3.2 Algorithm Overview
Given a single bitmap sketch of a character in a target pose (Fig. 7, left), our algorithm automatically infers a parametric human model SMPL [Loper et al. 2015] in the depicted pose (Fig. 7, right). The pose can then be transferred automatically onto a custom 3D character via standard animation software, such as Autodesk Maya or Blender, or via more advanced modern retargeting methods [Aberman et al. 2020].

We first predict three key elements of a character drawing: 2D bone tangents, body part contacts, and bone foreshortening. We use convolutional networks to predict 2D locations of main joints and image-space body part contacts; we then map the latter onto the 3D mesh (Sec. 4). We then use a nonlinear optimization using standard position-based $L_2$ reprojection loss to get a rough estimate of the pose, which we use to estimate foreshortening factor for each bone (Sec. 4.4) and contact vertices (Sec. 4.3). Finally, we use the three key elements in the nonlinear optimization with a novel loss that balances the perceptual cues, pose naturalness, and similarity to the input drawing, producing the final result (Sec. 5).

4 INFERRING KEY ELEMENTS OF A DRAWING
In the first stage of our algorithm, we infer the three elements of a drawing that we believe are key to its interpretation: 2D bone tangents, body part contacts, and bone foreshortening.

4.1 2D Joint Positions
We predict the 2D positions of the most important skeletal joints and rely on our final optimization (Sec. 5) to reconstruct the full 3D pose. In total, we predict 2D positions of $K = 18$ main joints (4 joints for each leg, 3 for each arm, 3 joints for the torso, 1 joint for the head).

To this end, we train a top performing deep convolutional 2D pose estimation network [Sun et al. 2019; Wang et al. 2019b; Xiao et al. 2018] on the dataset of sketches with their 2D skeletal annotations we collected (Sec. 6). We resize the input drawing preserving the aspect ratio and pad them to the resolution of 384x288 pixels. The network outputs a 96x72 pixel heatmap for each of the $K$ joints, showing the per-pixel confidence score of the chosen joint location. The joint position is then taken as the maximum point on the heatmap. For the details on architecture please refer to the original paper [Sun et al. 2019].

The output of this stage of our algorithm is the 2D positions $\hat{x}^{2D}_j$ of $K$ skeletal joints in the image coordinate frame (Fig. 7a). 2D bone tangents are then defined as differences of those 2D positions.

4.2 Initial Alignment of the 3D Model
We leverage these 2D positions in an initial optimization that produces a SMPL model roughly aligned with the drawing (Fig. 7c) via a state-of-the-art framework Exemplar Fine-Tuning (EFT) [Joo et al. 2021]. Starting with a pre-trained network performing regression from the input image into the space of the SMPL parameters [Kolotouros et al. 2019], EFT optimizes the weights of the network to minimize the $L_2$-reprojection distance between 2D joint positions $\hat{x}^{2D}_j$ and 2D orthographic projections $\hat{x}^{2D}_j$ of the 3D SMPL joints:

$$E_2D = \sum_j w_j ||\hat{x}^{2D}_j - \hat{x}^{2D}_j||^2.$$  (1)

We train the regression network on the poses produced by Pavlakos et al. [2019], which uses a pose naturalness prior. We thus inherit naturalness of poses as an implicit prior. For all details, please refer to the original paper [Joo et al. 2021]. We run their method for 150 iterations, with default parameters. The output of this stage is a set of 85 parameters encoding a human in a pose roughly similar to the drawing, including $24 \times 3$ parameters for body pose, 10 for body shape, 2 for camera 2D translation and 1 for uniform scale. The input RGB images are 224 x 224px.

4.3 Detecting Self-Contacts
We then detect depicted self-contacts in the image space and map those onto the vertices of the roughly aligned mesh (Fig. 8). The vertices will be then used in the final optimization that enforces contacts between some of them (Sec. 5).

Our sketch dataset contains 2D positions of perceived self-contacts. We use it to train a 2D contact prediction network, outputting 2D contact heatmaps. The network has the same architecture as in Sec. 4.1.

The self-contact heatmap predicts areas of potential contacts in the image space. We first need to filter out noise and separate different contact regions, which we map to separate groups of SMPL mesh vertices that each should have at least one pair of vertices touching. To this end, we first threshold the self-contact heatmap with a conservative threshold of 0.5, and compute the connected components over the thresholded heatmap, forming contact regions.

Our next step is mapping each contact region onto the vertices of the roughly aligned SMPL mesh. Note that a straightforward approach of simply projecting each connected component onto the
We then use the linear parameterization of each bone to transfer the depicted body part lengths often lead to unexpected angles with the screen. We correct these distortions (b).

In this step, our goal is to compensate for the distortions in the depicted foreshortening, introduced by artist inaccuracies, exaggerated perspective, and proportions mismatch (Fig. 10a). The output of this step, bone foreshortening under orthographic projection, will inform the angles between the 3D bones and the screen (Fig. 10b).

A straightforward solution would be to use a ground-truth dataset with correspondences between depicted 2D poses and intended 3D poses; such dataset, unfortunately, does not exist. Instead, our insight is that while the correspondences are unknown, the reconstructed and intended angles the bones make with the screen should be equally distributed.

As a proxy for the unknown distribution of the intended angles, we take the distribution $\mathcal{A}_i$ of angles each bone $i$ in a motion capture dataset [Mahmood et al. 2019] makes with an appropriate view plane (Fig. 9a). Ideally, the choice of the view planes should capture the drawing angles the artists choose for a given pose; as an approximation, we use the dataset’s default camera plane. Note that while this computation can easily be extended to multiple view planes, we did not find it necessary, since the dataset already provides enough pose variety even from the default viewpoint. We then compute the distribution $\mathcal{B}_i$ of angles each bone $i$ makes with the screen after performing rough optimization (Sec. 4.2) for all the images in our drawing dataset (Fig. 9b).

Our goal is now to find a function that, for each bone $i$, transforms the reconstructed angles $\beta_i \sim \mathcal{B}_i$ such that their distribution matches $\mathcal{A}_i$ as closely as possible. As noted in Wnuczko et al. [2016], the accuracy in observer’s perception of 3D directions seems to vary with the foreshortening angle; we conjecture the same is true for artists depicting foreshortened 3D bones. We furthermore observe that (1) artists seem to exaggerate perspective for foreshortened lines; (2) even when intending no foreshortening, artists often draw slightly shorter bones due to inaccuracies or weak but inaccurate lines; (2) even when intending no foreshortening, artists often draw slightly shorter bones due to inaccuracies or weak but inaccurate perspective (e.g. Fig. 6). Guided by these observations, we first represent the distributions of angles $\mathcal{A}_i$ and $\mathcal{B}_i$ via histograms with 10 equal bins, from $0^\circ$ to $90^\circ$ (Fig. 9a, b). We then model the transformation as a cubic polynomial:

$$f(\beta) = a\beta^3 + b\beta^2 + c\beta + d,$$

(2)

with initially unknown parameters $a, b, c, d$, and $\beta \in [0, \pi/2]$. To make the distributions of $f(\mathcal{B}_i)$ and $\mathcal{A}_i$ similar, we find the unknown parameter values by minimizing a sum of Earth Mover’s Distances [Rubner et al. 2000] between the two angle distributions for each bone, discretized as histograms:

$$\min_{a,b,c,d}\sum_i \inf_{\alpha,\beta} \sum_{\xi} \mathbb{E}(\alpha,\beta) - \| (\alpha - f(\beta) )\|_\Pi(\mathcal{A}_i, \mathcal{B}_i),$$

(3)

where $\Pi(\mathcal{A}_i, \mathcal{B}_i)$ is the set of all joint distributions whose marginals are $\mathcal{A}_i$ and $\mathcal{B}_i$. As we observed above, foreshortening is typically exaggerated, so we add constraints $0 \leq f(\beta) \leq \beta$. We would like to preserve the bones parallel to the screen to be still parallel after the transformation, i.e. $\beta = 0$, so we set $d = 0$. Minimizing the energy in Eq. 3, we get $a = 0.312$, $b = -0.448$, $c = 0.503$.

After this transformation, the angle distributions are better aligned (Fig. 9). This optimization is done only once for the dataset. We have additionally tested other classes of functions (cubic splines and piecewise linear functions). They result in similar, somewhat more complex functions that have little effect on the results, so we chose the cubic function as the simplest option.

At test time, for a given image, after the roughly aligned 3D skeleton is computed (Sec. 4.2), we calculate the angles $\beta_i$ between each bone and the screen. We note that since SMPL is limited to human proportions, atypical proportions of the depicted character likely

---

**Fig. 9.** In the foreshortening transformation stage, we design a function (c) that transforms the bone-screen angles after the initial optimization closer to the ground-truth angles. To this end, we find the transformation bringing the distribution of such angles after the initial optimization $\mathcal{B}$ (b) closer to the distribution of ground-truth angles $\mathcal{A}$ (a), yielding the transformed angle distribution $f(\mathcal{B})$ (d). The final angles are then used to compute the target foreshortening of each body part used in the final optimization. Here we display the histograms for the left thigh bone.

**Fig. 10.** Without the foreshortening transformation (a), distortions in the depicted body part lengths often lead to unexpected angles with the screen. We correct these distortions (b).

mesh may lead to suboptimal results, since the mesh often deviates significantly from the drawing (Fig. 8c). Instead, we use the predicted 2D skeleton as a proxy to find this mapping. We first compute convex hull of each contact region for robustness, then intersect each hull with the 2D skeleton, forming a set of 2D bone segments (Fig. 8b). We then use the linear parameterization of each bone to transfer the segments in contact onto the 3D SMPL skeleton (Fig. 8c), and finally mark all mesh vertices skinned to these 3D segments as contact vertices for this contact region (Fig. 8d).

### 4.4 Foreshortening Transformation

In this step, our goal is to compensate for the distortions in the depicted foreshortening, introduced by artist inaccuracies, exaggerated perspective, and proportions mismatch (Fig. 10a). The output of this step, bone foreshortening under orthographic projection, will inform the angles between the 3D bones and the screen (Fig. 10b).

A straightforward solution would be to use a ground-truth dataset with correspondences between depicted 2D poses and intended 3D poses; such dataset, unfortunately, does not exist. Instead, our insight is that while the correspondences are unknown, the reconstructed and intended angles the bones make with the screen should be equally distributed.

As a proxy for the unknown distribution of the intended angles, we take the distribution $\mathcal{A}_i$ of angles each bone $i$ in a motion capture dataset [Mahmood et al. 2019] makes with an appropriate view plane (Fig. 9a). Ideally, the choice of the view planes should capture the drawing angles the artists choose for a given pose; as an approximation, we use the dataset’s default camera plane. Note that while this computation can easily be extended to multiple view planes, we did not find it necessary, since the dataset already provides enough pose variety even from the default viewpoint. We then compute the distribution $\mathcal{B}_i$ of angles each bone $i$ makes with the screen after performing rough optimization (Sec. 4.2) for all the images in our drawing dataset (Fig. 9b).

Our goal is now to find a function that, for each bone $i$, transforms the reconstructed angles $\beta_i \sim \mathcal{B}_i$ such that their distribution matches $\mathcal{A}_i$ as closely as possible. As noted in Wnuczko et al. [2016], the accuracy in observer’s perception of 3D directions seems to vary with the foreshortening angle; we conjecture the same is true for artists depicting foreshortened 3D bones. We furthermore observe that (1) artists seem to exaggerate perspective for foreshortened lines; (2) even when intending no foreshortening, artists often draw slightly shorter bones due to inaccuracies or weak but inaccurate perspective (e.g. Fig. 6). Guided by these observations, we first represent the distributions of angles $\mathcal{A}_i$ and $\mathcal{B}_i$ via histograms with 10 equal bins, from $0^\circ$ to $90^\circ$ (Fig. 9a, b). We then model the transformation as a cubic polynomial:

$$f(\beta) = a\beta^3 + b\beta^2 + c\beta + d,$$

(2)

with initially unknown parameters $a, b, c, d$, and $\beta \in [0, \pi/2]$. To make the distributions of $f(\mathcal{B}_i)$ and $\mathcal{A}_i$ similar, we find the unknown parameter values by minimizing a sum of Earth Mover’s Distances [Rubner et al. 2000] between the two angle distributions for each bone, discretized as histograms:

$$\min_{a,b,c,d}\sum_i \inf_{\alpha,\beta} \sum_{\xi} \mathbb{E}(\alpha,\beta) - \| (\alpha - f(\beta) )\|_\Pi(\mathcal{A}_i, \mathcal{B}_i),$$

(3)

where $\Pi(\mathcal{A}_i, \mathcal{B}_i)$ is the set of all joint distributions whose marginals are $\mathcal{A}_i$ and $\mathcal{B}_i$. As we observed above, foreshortening is typically exaggerated, so we add constraints $0 \leq f(\beta) \leq \beta$. We would like to preserve the bones parallel to the screen to be still parallel after the transformation, i.e. $\beta = 0$, so we set $d = 0$. Minimizing the energy in Eq. 3, we get $a = 0.312$, $b = -0.448$, $c = 0.503$.

After this transformation, the angle distributions are better aligned (Fig. 9). This optimization is done only once for the dataset. We have additionally tested other classes of functions (cubic splines and piecewise linear functions). They result in similar, somewhat more complex functions that have little effect on the results, so we chose the cubic function as the simplest option.

At test time, for a given image, after the roughly aligned 3D skeleton is computed (Sec. 4.2), we calculate the angles $\beta_i$ between each bone and the screen. We note that since SMPL is limited to human proportions, atypical proportions of the depicted character likely
cause discrepancies in body shape estimation, and, as a result, in angles $\beta_j$. For many characters, we observe that these mismatches can be explained by an atypical scale of upper body relative to the lower body. To alleviate this issue and avoid incorrectly foreshortened bones, we subtract the minimal angle with the screen, computed separately over the upper body and lower body:

$$\beta'_j = \beta_j - \min_{j \in B} \beta_j,$$

where $B$ is the set of upper or lower body bones.

Finally, the predicted foreshortening is represented by $q_i = \cos f(\beta'_i)$, the target foreshortening for bone $i$ used in the final optimization.

5 3D POSE OPTIMIZATION

The second stage of our system is the optimization that starts with the roughly aligned pose (Sec. 4.2) and finds the artist-intended 3D pose of the given character. To this end, the optimization leverages the 2D bone tangents, body part contacts, and bone foreshortening computed in the previous stage. As a framework, we use EFT [Joo et al. 2021] which optimizes over the weights $w$ of the regression neural network, initialized by the rough alignment stage (Sec. 4.2). We follow their optimization process (Adam algorithm, default PyTorch parameters, learning rate of $10^{-3}$). We perform a fixed number of 60 iterations.

Within that framework, instead of the traditional position-based $L_2$ reprojection loss (Eq. 1), we propose the following novel loss for our task, based on our principles (Sec. 3.1):

$$\min_{w} E_{\text{parallel}} + \lambda_{E_{f}} + E_{\text{contacts}} + E_{\text{reg}}. \quad (4)$$

We denote $x_{j}^{3D} \in \mathbb{R}^3, j = 1, \ldots, K$ the 3D joint positions of the SMPL model. Note that these positions are, for a fixed input image, functions of the neural network weights $w$. For each bone $i$ connecting joints $j_1$ and $j_2$, we denote its 3D vector as $b_i^{3D} = x_{j_2}^{3D} - x_{j_1}^{3D}$, and its orthographic projection onto the screen as $b_i^{2D}$. Details about the individual terms in Eq. 4 are below, in the order they appear:

- **Parallelism.** Guided by our principle of tangent significance, we favor parallelism between the projected 3D bones and their 2D depictions:

$$E_{\text{parallel}} = \sum_{i} \left( \frac{\|b_i^{2D}\|}{\|b_i^{3D}\|} \cdot n \right)^2,$$

where $n$ is a normal to the depicted bone $z_{j_1}^{2D} - z_{j_2}^{2D}$.

- **Foreshortening.** We use the transformed bone foreshortening calculated in Sec. 4.4 to guide the target length of each bone’s projection:

$$E_{f} = \sum_{i} (\|b_i^{2D}\| - L_i q_i)^2,$$

where $L_i$ is the length of the bone $i$, as estimated by the rough alignment stage (Sec. 4.2). Note that here we use a fixed bone length $L_i$, as opposed to $\|b_i^{3D}\|$ that can vary during the optimization. In our experiments, we found that otherwise the optimization often exploits that dependency to minimize the energy, adjusting the bone lengths instead of the angles between screen and bones.

- **Contacts.** For each set of contact vertices computed in Sec. 4.3 (Fig. 8), we enforce physical contact between at least one pair of vertices. We define $E_{\text{cont3D}}$ as the sum of four energy terms from [Müller et al. 2021], aimed at minimizing Euclidean distances between contact vertices, aligning their normals, and avoiding self-collisions. Please see Appendix A for details. Furthermore, as outlined in Sec. 3.1, we aim to preserve the relative positions of the bones and joints in each contact region. To this end, for each contact region we find points on the 2D skeleton that are the closest to the contact, compute their 2D positions relative to each other, and aim to preserve those 2D positions between the same points on the 3D SMPL skeleton. Precisely, to determine those points, we select the local maxima of the heatmap over each 2D bone. We then connect each such point with all others within the same contact region, forming vectors $\hat{e}_i^{2D}$, which capture the relative position of such point with respect to another one. We aim to preserve these vectors exactly for the 3D pose, when projected onto the original view. For each point on the 2D skeleton, we find its corresponding point on the 3D skeleton by using the linear (arclength) parameterization of each bone and simply taking the 3D point with the same parameter value along the same bone. Denoting the vectors connecting the corresponding 3D skeleton points after projection as $\hat{e}_i^{2D}$, we set:

$$E_{\text{cont2D}} = \sum_{i} \|\hat{e}_i^{2D} - \hat{e}_i^{2D}\|^2.$$

The final term is $E_{\text{contacts}} = \lambda_{\text{cont3D}} E_{\text{cont3D}} + \lambda_{\text{cont2D}} E_{\text{cont2D}}$.

- **Regularity.** As suggested by perception studies and previous work [Bessmeltsev et al. 2016; Xu et al. 2014], we speculate that human observers leverage regularity cues when interpreting sketches. In particular, the viewers expect nearly parallel 2D bones to stay parallel in 3D; feet nearly parallel to the floor to be standing on the floor. In enforcing this, we utilize perception research–indicated angle threshold [Hess and Field 1999] of 17°, below which we consider 2D bones to be parallel to each other or the floor. $E_{\text{reg}}$ thus is a simple sum of squared differences between the corresponding normalized bone directions $b_i$ and their target directions.

The naturalness of the poses is enforced implicitly by the EFT framework itself, as discussed in 4.2. For all the results presented in the paper, we use $\lambda_{E_{f}} = 10, \lambda_{\text{cont3D}} = 0.06, \lambda_{\text{cont2D}} = 10^{-3}$.

6 DATASET

We propose two novel datasets: the first large-scale dataset of 2D pose annotations for character sketches (D1), and a smaller dataset of high-quality character sketches (D2). Dataset D1 contains more than 3K images containing one or more sketched characters in articulated poses with 2D position annotated for each key joint (up to 18 per skeleton), in total containing 14,462 skeletons. Each image was annotated by a single annotator only; human annotations of sketches, however, are largely consistent, as shown by previous work [Bessmeltsev et al. 2016]. Dataset D2 contains 310 high-quality character sketches, with a very permissive usage license (CC BY 2.5). We use D1 to train and validate our 2D keypoint prediction
network. We show a few examples from D1 in the Supplementary Materials; all the input images of the results in the paper, unless indicated otherwise, are from D2, and thus are not used in training.

Data Collection and Annotation. For D1, as the first step of data collection, we manually query search images via engines such as Google, Bing, and Baidu for character sketches and filter out irrelevant images. We additionally collect images with similar queries from Flickr and Pinterest and removed duplicates. We then hired an annotation service that marked up to 18 joint locations for each drawn character, with particular instructions to annotate occluded or not explicitly drawn joints if their position is clear from context, but skip the ones that are ambiguous. Naturally, this dataset contains sketches of numerous styles and complexities, greyscale and color, digitally drawn and scanned pen-and-paper drawings.

For D2, we collect high-resolution scans or photographs of character sketches from artists of different backgrounds. The sketches contain gestures, contour drawings, and detailed character sketches of humans in various, often highly articulated poses. The sketches are done in a variety of techniques on paper (pencil, pens, watercolor).

7 RESULTS AND VALIDATION

So far we have shown many examples of 3D characters, algorithmically posed via a single bitmap sketch (Fig. 1, 4, 5, 7). Our learning-based solution allows to pose 3D characters via natural, noisy, incomplete, and inaccurate character sketches, inaccessible to previous work. Our novel optimization allows us to successfully resolve ambiguities, inaccuracies, and distortions typical for character sketches, see e.g. Fig. 11, 12 for additional results. Our method robustly handles occlusions (e.g. the left arm in Fig. 12, second row) and altogether missing body parts (Fig. 12, top), typical for incomplete quick sketches or gestures. For all the examples, our method convincingly recreates the drawn poses in 3D.

Note that we only target estimating body pose, not its shape. Therefore, after our optimization we set the shape to the SMPL default for all our results.

We validate the key aspects of our method in a number of ways. The questionnaires used in the evaluations and detailed results are included in our supplementary.

Ablation Study. We perform an ablation study of our method (Fig. 13). We demonstrate results on a challenging example, each time skipping one component of our algorithm by disabling the corresponding loss term. For reference, we show a reprojection-based method [Müller et al. 2021] (Fig. 13a), highly sensitive to the depicted bone lengths, introducing strong unexpected foreshortening. Disabling the foreshortening transformation (Fig. 13b) also leads to a foreshortened pose, albeit slightly less (right shin) due to focusing on parallelism instead of 2D joint positions. Optimization without contacts results in an incorrect depth prediction of the left hand (Fig. 13c). Disabling the regularity term results in a left knee with a bent, invisible from the front (Fig. 13d). Please see the supplementary materials for the ablation study on the other inputs.

Foreshortening Transformation Function. We evaluate the robustness of Eq. 3 by fitting the function to smaller random subsets $A$ (10%, 1%, 0.1%, and 0.05% of full training dataset). We get the same of similar coefficients of $a$, $b$ and $c$, within the tolerance of 0.2, as for full training dataset. We further test robustness by fitting our foreshortening transformation function on a subset of standing poses only, obtaining $a' = 0.95, b' = 0.12, c' = 0.23$, as opposed to the original values in Sec. 4.4 ($a = 0.312, b = -0.448, c = 0.503$). Despite the difference in coefficients, the two polynomials $f(\beta)$ (Eq. 2) with those coefficients are virtually identical over the interval we are interested in, i.e. $\beta \in [0, \pi/2]$.

Qualitative Evaluation. We asked 2 artists and 7 non-professionals to comment on the results of our algorithm. We showed them each pair of input and our algorithmic result and asked to comment on the following statement, separately for each pair, "This 3D character pose captures the artist intended drawn pose", with 5 Likert-type reply options: "Strongly disagree" (-2), "Disagree" (-1), "Neither disagree nor agree" (0), "Agree" (1), "Strongly Agree" (2). On average, the respondents agreed with the statement (avg = 1.06, std = 0.38). The layout of the study and the results are presented in the Supplementary.

Comparison to Prior Art. In Figure 15, we compare our method to Gesture3D [Bessmeltsev et al. 2016]. Their method relies on having a clean vector drawing, including inferring joint depth order from clean vector T-junctions, and assuming all the terminal joints are clearly visible and outlined (a). Our method handles a much wider variety of inputs, including natural bitmap sketches found in the wild. Automatically vectorizing such sketches to get a similar quality vectorization is an open problem (Figure 15c). Even supplied with correct 2D labels, Gesture3D does not capture the notion of pose naturalness, often resulting in unnatural poses (bottom middle). Furthermore, Gesture3D is designed for drawings with minimal foreshortening, producing flat, static poses (Figure 15, top middle). Our method successfully captures complex poses with significant foreshortening (Figure 15, top right).

We compare with reprojection-based methods in Fig. 4, 14, and Supplementary Materials. We use the implementations provided by the authors. Whenever a method accepts 2D labels as input, we supply the 2D labels predicted by our 2D network for a fair comparison. We first train SPIN [Kolotouros et al. 2019] on the results of Pavlakos et al. [2019], then improve those with the system of Joo et al. [2021], and fine-tune the SPIN regression network on those poses for better quality. For the methods directly predicting a 3D pose from an image, we retrained them using our dataset, following their training protocol. We run our method directly on the input sketch.

These methods do not aim to capture the artist-intended pose, rather focusing on the task of finding a natural pose with projections of 3D joints close to the 2D joint positions. In the presence of distortions, typical for character sketches, such as incorrect depiction of perspective and bone lengths, such approach often leads to exaggerated foreshortening (Fig. 4, 14), irregularities or unnatural poses (Fig. 12). Our method successfully reconstructs poses close to natural in the presence of such distortions and inaccuracies for both standard proportions (Fig. 14, top) and non-humanoid or characters with unrealistic or non-human proportions (Fig. 14, bottom).
Qualitative Comparison. We validate the quality of our results by comparing them to the state of the art alternative [Müller et al. 2021] via a comparative perceptual study. Study participants were shown input sketches, together with our algorithmic posing result and an alternative posing result. The layout of the study is presented in the Appendix. The input sketch was shown on the top, marked as A, and the two posing results were placed at the bottom in random order and marked as “B” and “C”. Participants were then asked “Which of the poses below, B or C, more accurately captures the drawn pose A on top? If both are equally acceptable, choose ‘Both’. If neither, select ‘Neither’.” We included 12 questions. We collected answers for each query from 14 different participants, including 5 males and 9 females, age ranging from 21 to 32 years; 3 were artists. The study data is presented in the supplementary.

To avoid the influence of the body shape on the study results, we reset body parameters to the average body shape for both methods. Similarly, since neither hands nor turn of the head are guided by the input sketch and are only controlled by their respective priors, we reset these parameters to their default values.

Fig. 16 summarizes the results. Participants preferred our results over the one of 64% of the time, ranked our methods on par 10% of the time, and preferred the alternative only 8% of the time. This study convincingly demonstrates that the 3D poses we produce are more consistent with viewer expectation than the ones produced by previous approaches.

Hand poses and head turn. Our system does not capture hand poses and the turn of the head; inferring those features from incomplete drawings proved to be a challenge. As a follow-up to our study (Fig. 16), we have asked users who selected ‘Neither’ for their comments, and most of the comments addressed hand poses and the turn of the head. Instead of relying on heuristics, we allow for a simple user interaction: the user is able to choose one of the predefined hand poses (fist, flat palm, palmar flexion) for each hand and turn the head around its axis (Fig. 17). For this figure, a user adjusted the hand poses and the head turn within a few seconds. All the other results were processed in a fully automatic way; comparison with the previous work was done with automatically computed results.

Comparison with manually posed characters. We provided six of our input sketches and the SMPL model in a neutral pose to two 3D modeling experts and asked them to manually pose the characters into the poses drawn on the sketches. The artists took roughly from 5 to 15 minutes to pose the character for each drawing, while our algorithm inferred each pose in 1.5 minutes on average (Sec. 7). We have furthermore performed a qualitative comparison user study with the same layout as for the comparison with the previous work, each time presenting our results and manually posed results in a random order. We asked 6 participants. Participants preferred our results 27% of the time, ranked both our and manual results as equally good 18% of the time, and preferred the manually posed characters 44% of the time. The participants chose “Neither” 11% of the time, disagreeing with both manually posed and our results.

Finally, we have quantitatively compared our algorithmic results with the manually posed 3D characters, as shown in Table 1. With respect to the standard MPJPE and PA-MPJPE metrics, we show that our results have smaller or equal errors than the previous work, similar to the natural variation between different experts. Those standard metrics, however, are not perception-based and thus are not necessarily indicative of user preferences.
Table 1. Error metrics on the manually posed characters.

<table>
<thead>
<tr>
<th></th>
<th>MPJPE</th>
<th>PA-MPJPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pavlakos et al. 2019 vs Expert 1</td>
<td>266</td>
<td>232</td>
</tr>
<tr>
<td>Pavlakos et al. 2019 vs Expert 2</td>
<td>282</td>
<td>237</td>
</tr>
<tr>
<td>Kolotouros et al. 2019 vs Expert 1</td>
<td>189</td>
<td>152</td>
</tr>
<tr>
<td>Kolotouros et al. 2019 vs Expert 2</td>
<td>223</td>
<td>153</td>
</tr>
<tr>
<td>Joo et al. 2021 vs Expert 1</td>
<td>107</td>
<td>93</td>
</tr>
<tr>
<td>Joo et al. 2021 vs Expert 2</td>
<td>136</td>
<td>100</td>
</tr>
<tr>
<td>Müller et al. 2021 vs Expert 1</td>
<td>126</td>
<td>105</td>
</tr>
<tr>
<td>Müller et al. 2021 vs Expert 2</td>
<td>150</td>
<td>116</td>
</tr>
<tr>
<td>Ours vs Expert 1</td>
<td>103</td>
<td>78</td>
</tr>
<tr>
<td>Ours vs Expert 2</td>
<td>126</td>
<td>88</td>
</tr>
<tr>
<td>Expert 1 vs Expert 2</td>
<td>116</td>
<td>79</td>
</tr>
</tbody>
</table>

Fig. 12. Our algorithmic results can be automatically transferred onto a given custom character using standard tools (we used Blender’s Animation Retargeting feature). We show the results of Müller et al. [2021] for comparison. Input image (top) © Rafianimates, (third row) © Brad Regier, (bottom) © Zoska Leutina.

Table 1. Error metrics on the manually posed characters.

Fig. 13. An ablation study of our algorithm. (a) Result of a reprojection-based method of Müller et al. [2021]. (b) Our optimization without foreshortening transformation (b), without contacts (c), without the regularity energy (d), and our final result (e). For each pose, the original view is on the left, the alternate view is on the right.

Parameter Sensitivity. We show (Fig. 19) that our method produces plausible results for a range of parameters. Naturally, changing $\lambda_f$ provides a way to balance trusting the depicted foreshortening.
Fig. 14. Additional comparisons with reprojection-based approaches [Joo et al. 2021; Kolotouros et al. 2019; Müller et al. 2021], which fail to reconstruct plausible poses due to the typical inaccuracies and distortions of a character sketch. Our method correctly recovers the intended 3D poses (right).

Fig. 15. Gesture3D [Bessmeltsev et al. 2016] only accepts clean vector drawings (a), unable to find 2D joint locations otherwise. Given a noisy sketch (b), modern vectorization methods often produce noisy vectorizations (c), incompatible with Gesture3D. Even provided with 2D labels (d), Gesture3D may produce implausible or static and flat poses (middle). Our method directly successfully infers 3D poses from a variety of sketches, including rasterized clean vector drawings like (a) and noisy raster drawings (b), producing realistic, expressive, and dynamic poses (right).

Fig. 16. Summary of comparative preferences in our perceptual study. Participants strongly preferred our results over the state-of-the-art alternative [Müller et al. 2021].

Similarly, increasing $\lambda_{\text{cont2D}}, \lambda_{\text{cont3D}}$ prioritizes self-contacts in the final pose.

2D Keypoint Detection Validation. We evaluate the performance of 2D keypoint detection on our validation dataset, consisting of 882 drawings each containing a single character (roughly 6% of our dataset). We use a standard Percentage of Correct Keypoints (PCK@0.5) metric on this dataset, as well as mean Average Precision (mAP) metric on a range of Object Keypoint Similarity (OKS) thresholds. Overall, the 2D keypoint detection network reaches 0.891 of PCK@0.5 and 0.854 of mAP, which is substantial considering the

Fig. 17. We allow for a simple user interaction to edit the pose features our system does not infer: selecting from a small set of predefined hand poses and adjusting the turn of the head. This interaction typically takes a few seconds.

Fig. 18. Our algorithmic results (right) are often visually comparable with the characters manually posed by experts (middle). Our computations are roughly 3-10 times faster than manual posing.

Fig. 19. Our method is robust to significant changes in the parameters. Our result with the default parameter values is on the right.
We have presented and validated the first method to infer a 3D humanoid character pose from a single bitmap sketch and introduced the first large-scale dataset of 2D skeletal joint annotations for bitmap sketches. Our system combines a modern deep learning framework with an optimization, guided by observations on the nature of sketches. Our method can process drawings of many different styles with occlusions, distorted proportions, and extra strokes or elements of shading, allowing to directly use natural drawings without any preprocessing or cleanup. We confirm that the poses our framework produces agree with the observers' expectations, by a significant margin more than the previous work.

Our work raises many directions for future research. First of all, we hope that the introduction of the 2D joint labels dataset will inspire follow-up research in 2D character inbetweening, segmentation, or consolidation of character sketches, among other possibilities. An interesting extension of our work would be to generalize it to arbitrary non-humanoid skeletons, where pose datasets are unavailable, via physics-based animation systems. Finally, an important line of research would generalize our method to non-skeletal rigs, supporting facial animation and nonlinear deformations.

ACKNOWLEDGMENTS
We would like to thank Oleg Muranov and Anastasia Seretkina for insightful discussions and data. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant No.: RGPIN-2019-05097 (“Creating Virtual Shapes via Intuitive Input”), the Fonds de recherche du Québec - Nature et technologies (FRQNT) under Grant No.: 2020-NC-270087, and the NSERC-FRQNT NOVA Grant No. 314090. We also thank Maria Myronova and Elena Uretskaya for sharing their sketches.

REFERENCES


A DETAILS ON THE CONTACT TERM $E_{cont3D}$

In this section, we precisely follow Müller et al. [2021]; this is given only for completeness, all the missing details can be found in the original paper.

**Definition A.1.** Given a mesh $M$, we define two vertices $v_i$ and $v_j \in M$ to be in self-contact, if $||v_i - v_j|| < r_{self} \text{ and } geo(v_i, v_j) > r_{geo}$, where $geo(v_i, v_j)$ is the geodesic distance between $v_i$ and $v_j$. We use $r_{geo} = 30cm$ and $r_{self} = 2cm$.

We denote all vertex pairs in $M$ satisfying this definition as $M_{C}$. We further define an operator $\mathcal{H}()$ that returns a set of unique vertices in $M_{C}$, and an operator $\mathcal{f}_d()$ that takes $\bar{v}_i$ as input and returns the Euclidean distance to the nearest $v_j$ that is far enough in the geodesic sense.

We set

$$E_{cont3D} = L_C + L_P + L_A,$$

where

$$L_C = \frac{1}{|\mathcal{H}(M_{C})|} \sum_{v_i \in \mathcal{H}(M_{C})} \tanh \mathcal{f}_d(v_i)$$

encourages every vertex in the predicted set of contact vertices $\bar{M}_{C}$ to be in contact. Vertices in contact are pulled together via a contact term $L_C$. To prevent self-intersections, vertices inside the mesh are pushed to the surface via a pushing term $L_P$. Finally, $L_A$ aligns the surface normals of two vertices in contact.

To compute self-contact terms, we first find which vertices are inside, $M_I \subset M$ via generalization winding numbers [Jacobson et al. 2013]. SMPL-X is not a closed mesh; this complicates the self-intersection tests. We close it by adding a vertex at the back of the mouth. In addition, SMPL and SMPL-X often self-intersect by default, e.g. torso and upper arms. We identify such common self-intersections and filter them out from $M_I$. To capture fine-grained contact, we map the union of inside and contact vertices onto the HD SMPL-X surface, i.e. $M_{D_{CI}} = H_D(M_I \cap M_{C})$, which is further segmented into an inside $M_{D_{I}}$ and outside $M_{D_{CI}}$ subsets via intersection tests. The objectives are defined as

$$L_C = \alpha_1 \sum_{p_i \in M_{D_{I}}} \tanh^2 \frac{\mathcal{f}_d(p_i)}{\alpha_1},$$

$$L_P = \beta_1 \sum_{p_i \in M_{D_{I}}} \tanh^2 \frac{\mathcal{f}_d(p_i)}{\beta_2},$$

$$L_A = \sum_{(p_i, p_j) \in M_{D_{CI}}} 1 + (N(p_i), N(p_j)).$$

Here $M_{D_{CI}}$ is the subset of vertices in contact in $M_{D_{I}}$. We use the same parameter values as in the original paper, $\alpha_1 = \alpha_2 = 0.005$, $\beta_1 = 1$, and $\beta_2 = 0.04$. 

---

**References**


---

**ACM Trans. Graph., Vol. 41, No. 4, Article 85. Publication date: July 2022.**