A Pose-Guided Forward-Kinematic Framework for Human Structure Prediction from Sparse 2D Joints

Sihan Liu[†], Natsumi Okatani [††], and Ryuji Shioya[†††]
s4b102400010@toyo.jp okatani@toyo.jp shioya@toyo.jp
[†]Graduate School of Information Sciences and Arts, Toyo University, Saitama, 350-8585 Japan
[††]Faculty of Information Sciences and Arts, Toyo University, Saitama, 350-8585 Japan
[†††]Faculty of Information Sciences and Arts, Toyo University, Saitama, 350-8585 Japan

Abstract
We evaluate a pose-guided system that predicts a kinematic human representation from sparse two-dimensional joints and renders a polyhedral wireframe. The main evidence comes from matched 50,000-update experiments on two synthetic capture sets, C1 and C2. On C1, FULL improves localization over FK-S3D with accurate input joints, but the Teacher-input differences are not resolved by source-bootstrap intervals. C2 compares DIRECT, FK-S3D, and FULL on 756 images, of which 747 admit Teacher inference. Under the same three-dimensional supervision, DIRECT yields lower localization and elbow/knee angle errors than FK-S3D in both tested seeds and input conditions, but larger bone-length errors. FULL uses the same architecture and FK as FK-S3D with additional supervision. Its comparison with DIRECT therefore changes both parameterization and losses, and their Teacher-input intervals include zero. Supervision deletions support a limited localization benefit from reprojection; rotation supervision affects limb-intersection indicators more consistently than mean joint error. Optional 80-step IK can improve projection fitting while worsening synthetic three-dimensional error under inaccurate observations. A historical evaluation of 16 anime images remains exploratory and lacks three-dimensional ground truth. These results identify supervision-dependent benefits and geometric trade-offs rather than a general accuracy advantage from FK.

Keywords:
human structure prediction; synthetic supervision; forward kinematics; sparse pose observations; anime illustration

1. Introduction

A useful body-structure reference should show connected limbs and trunk while retaining the illustrated character's pose. In anime images, clothing, exaggerated proportions, occlusion, and cropping can obscure the corresponding joints. A clean wireframe does not necessarily match the input pose; even a close two-dimensional projection can hide large depth errors. This distinction matters when a structure image is intended to support drawing or pose analysis rather than simply look plausible [1].

Our earlier experiments explored unpaired anime-to-structure translation with DiscoGAN [2]. Some outputs resembled simplified body drawings, but the experiments did not establish consistent pose correspondence. We therefore investigate an explicit route: predict a kinematic intermediate representation from synthetic joint correspondences, then render a deterministic polyhedral wireframe. The system does not recover a unique anatomical surface or the unavailable historical Maya rig.

The experiments address three linked questions. First, how does a constrained FK representation compare with direct XYZ regression when both receive the same three-dimensional supervision? Second, which additional losses affect localization or structural diagnostics when the architecture and FK remain fixed? Third, what changes when ideal joint projections are replaced by estimates, with or without post-hoc IK fitting? We evaluate these questions separately rather than inferring three-dimensional accuracy from rendered appearance.

This study implements an explicit pose-to-structure pipeline and evaluates its trade-offs on C1 and C2. C1 compares FK-S3D and FULL within a shared FK architecture; follow-up ablations examine the contributions of individual supervision terms. C2 adds DIRECT under the same three-dimensional objective used by FK-S3D. These comparisons distinguish changes in parameterization from changes in supervision: FULL adds losses, not a new FK architecture, and its comparison with DIRECT cannot isolate FK. The contribution is the implemented system and its controlled evaluation, rather than a new FK, IK, or rotation-representation algorithm. Input perturbations and the historical anime evaluation provide supporting evidence about limitations, not validation of reliable real-image three-dimensional recovery.

2. Related Work

2.1 Observation quality and pose lifting

Chen and Zwicker [1] study pose estimation for illustrated characters, including pose-guided retrieval. Martinez et al. [3] show how a regressor can lift detected two-dimensional joints to three dimensions. Our study likewise separates joint observation from structure

Manuscript received November 5, 2026
Manuscript revised November 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.11.1

prediction, but uses a different skeleton, synthetic training data, and a polyhedral rendering target. DIRECT is a controlled baseline built for these experiments, not a reproduction of [3] or a cross-paper performance comparison.

We use the existing Copillust Curriculum S2 checkpoint [4] as a frozen pose teacher. It is an RTMPose-m model [5] from an MMPose-based project, with HumanArt pretraining and Bizarre Pose fine-tuning; no teacher model is trained here. The fixed artifact is curriculum_s2/best_coco_AP_epoch_10.pth; its historical upstream code/configuration commit was not recorded. Section 3.1 specifies the local inference protocol. Thresholding controls which joints are observed and which are missing. Teacher responses are not treated as calibrated probabilities or verified anatomical landmarks.

2.2 Kinematic representations and fitting

Six-dimensional rotation regression [6] provides a continuous representation before orthogonalization. Related approaches use geometry in different ways: SPIN [7] integrates model fitting into training, HybrIK [8] combines analytical swing estimation with neural twist prediction, and ProtoRes [9] learns pose construction from sparse constraints. These precedents make clear that rotation prediction, kinematic connectivity, and learned initialization are established techniques. The contribution evaluated here is their particular system implementation and controlled supervision analysis.

Our predictor combines a residual MLP, training-derived rest templates, shared FK, and optional IK. Unlike SPIN, outputs of the evaluated IK stage are not fed back as training targets. We do not use the HybrIK or ProtoRes body models or their benchmark protocols.

VLPose [14] adapts 2D pose estimation across natural and artistic domains through vision–language tuning. Ludwig et al. [15] lift 2D pose sequences to 3D joint positions and rotations. Our system keeps its illustration teacher frozen and uses a single-pose MLP with template-based FK; the tasks and protocols differ.

2.3 Ambiguity and evaluation scope

Wehrbein et al. [10] model multiple plausible three-dimensional poses from a single image. That ambiguity remains relevant to our single-output predictor: visible joints cannot generally determine hidden body parts or depth. We therefore report synthetic three-dimensional joint error, wireframe overlap, intersection indicators, input fitting, and reviewed real-image landmark agreement as distinct measures. Improvement on one metric cannot establish improvement on the others.

3. Method

Figure 1 traces the input-to-structure representation for the fixed qualitative examples; the controlled input-perturbation results are reported in Figure 2.

3.1 Observations and normalization

In the original 26-joint framework, the input contains twelve two-dimensional body joints—bilateral shoulders, elbows, wrists, hips, knees, and ankles—and their validity indicators, for 36 scalar values in total. The predictor receives neither the target image nor a silhouette, query depth, ground-truth camera, or query-specific rest skeleton. All coordinates use a 256 × 256 image; anatomical left and right refer to the illustrated character.

For estimated inputs, the frozen Copillust S2 teacher processes the whole image once, with neither foreground recropping nor flip testing. The whole-image box is expanded by 1.25, matched to the 192:256 input aspect ratio, and bilinearly sampled with black padding. RGB means are (123.675, 116.28, 103.53), with standard deviations (58.395, 57.12, 57.375). SimCC axis outputs have 384/512 bins; argmax coordinates divided by two are mapped back to the 256 × 256 source image. A joint is valid only if both coordinates and the response are finite, the point lies within the image, and the smaller axis-maximum response reaches 0.3. The predictor requires at least four valid joints. Evaluation labels never replace missing observations.

Let m_j indicate validity and $u_j = \left(x_j - 127.5\right)/128$. For $n = \sum_j m_j$, the observed centroid and scale are

$$c = \frac{1}{n}\sum_j m_j\, u_j,$$
$$s = \max\left(0.125,\ \sqrt{\frac{1}{n}\sum_j m_j \parallel u_j - c \parallel^2}\right). \tag{1}$$

The network receives the flattened $m_j\left(u_j - c\right)/s$ and the mask. Missing-coordinate contributions are explicitly zeroed. This normalizes observed spatial extent; it does not recover actual height or metric camera calibration.

3.2 Structure predictor and common FK

The original 26-joint FK predictor contains a 36-to-256 linear layer with SiLU activation, three width-256 residual blocks, a final LayerNorm, and a 177-dimensional output head. Each residual block is LayerNorm–Linear–SiLU–Linear with multiplier 0.5 on its residual branch. No image encoder or dropout is used. FK-S3D and FULL share this exact 451,761-parameter architecture.

The head predicts 26 local six-dimensional rotations, two root-translation values, one log scale, and 18 rest-template logits. Two predicted vectors are orthogonalized by Gram–Schmidt and a cross product [6], with normalization epsilon 10^{-8}; nodes without rotational degrees of freedom are fixed to identity. Softmax weights w_a define a whole-skeleton mixture $r_j = \sum_{a=1}^{18} w_a r_j^{(a)}$. All templates are extracted from training sources. A single weight vector mixes the complete templates rather than choosing an independent template for each joint.

For parent index $p(j)$, local rotation L_j, global rotation R_j, and position X_j, the shared FK is

$$R_j = R_{p(j)} L_j, \qquad X_j = X_{p(j)} + R_{p(j)} r_j, \qquad (2)$$

with $X_0 = 0$ and $R_0 = L_0$. After restoring observed center and scale, $t = s\hat{t} + c$ and $\ell = \mathrm{clip}\big(\mathrm{clip}(\hat{\ell}, -2, 2) + \log s, -2, 2\big)$. The projected representation is

$$Q_{j,xy} = e^{\ell} X_{j,xy} + t, \qquad Q_{j,z} = e^{\ell} X_{j,z}. \qquad (3)$$

The common FK representation maintains parent–child connectivity and the predicted rest-template bone lengths, which need not match the ground truth. It imposes neither anatomical joint-angle limits nor collision-free geometry. FK-S3D and FULL use identical FK; FULL changes supervision rather than model structure. FK is used to couple joint positions and rigid-part placement through a shared skeleton, rather than to guarantee lower localization error. This separates structural connectivity from localization accuracy; both must be evaluated.

The twelve possible body-input locations form a subset of the 26-node output hierarchy; an invalid observation does not remove its output node. The twenty rendering-control nodes are selected from that same hierarchy. Zero-length intermediate offsets can make distinct nodes coincide in projection. Supplementary Figs. S1–S3 and Tables S1–S2 provide the full joint projections, numbered parent graph and index mapping.

3.3 Matched direct-coordinate baseline

DIRECT uses the same 36-dimensional input, canonicalization, width-256 residual hidden network, batch ordering, and augmentation sequence as the corresponding FK-S3D seed. Its 78-dimensional head predicts canonical XYZ coordinates for all 26 joints. At inference, observed scale multiplies all coordinates, while observed center shifts XY only; ground-truth root, scale, rotation, and camera parameters are unavailable. DIRECT has 426,318 parameters, compared with 451,761 for FK. All 22 shared hidden tensors match the same-seed FK-S3D Step-0 initialization. The output head differs in size and instead uses training-derived mean XYZ biases and random weights scaled by 0.01. The two models therefore do not have identical full initialization or capacity.

DIRECT is optimized with the same full26 camera-coordinate SmoothL1 objective used by FK-S3D, making it a matched-budget comparison of parameterizations with different output heads. Comparison with FULL additionally changes supervision. DIRECT outputs joint coordinates but not local part rotations or a polyhedral mesh; its W80, mesh Dice, and mesh-intersection entries are N/A, not zero. Connecting DIRECT joints with the same parent graph yields shared endpoints, but does not verify bone lengths or depth. Both model families use predicted and target camera-scale coordinates for bone-length evaluation.

3.4 Supervision and controlled deletions

Synthetic targets provide all 26 joints Q^{gt}, local rotations L^{gt}, and rest offsets r^{gt}. Let $\rho(e)$ denote elementwise SmoothL1 with beta 1: half the squared error when its absolute value is below one, and the absolute error minus one half otherwise. The spatial loss is the mean of $\rho(Q - Q^{\mathrm{gt}})$ over the batch, all joints, and three coordinates, including synthetically hidden joints. The combined objective is

$$\mathcal{L} = \mathcal{L}_{3D} + a\mathcal{L}_{rep} + b\mathcal{L}_{rot} + c_b\mathcal{L}_{bone} + d\mathcal{L}_{sep}. \qquad (4)$$

For augmented input coordinates $\tilde{x}$, with body-to-full-skeleton index $b(j)$, the reprojection term is

$$\mathcal{L}_{rep} = \frac{\sum_{i,j} m_{ij}\ \mathrm{mean}_{xy}\rho\left(Q_{i,b(j),xy} - \frac{\tilde{x}_{ij} - 127.5}{128}\right)}{\max\left(\sum_{i,j} m_{ij}, 1\right)}. \qquad (5)$$

Its target is the augmented input, not the clean synthetic projection. Rotation supervision is the mean squared difference between all entries of the 26 local rotation matrices, not a geodesic-angle loss. For nonzero rest bones $\mathcal{B}$,

$$\mathcal{L}_{bone} = \mathrm{mean}_{i,j\in\mathcal{B}}\left(\frac{\| r_{ij} \| - \| r_{ij}^{\mathrm{gt}} \|}{\max\left(\| r_{ij}^{\mathrm{gt}} \|, 0.01\right)}\right)^2. \qquad (6)$$

The soft-separation term samples five equally spaced points on each of eight main limb segments, using body coordinates divided by e^{ℓ}. For segment pairs with no shared endpoint, d_{ab} is the minimum of the 25 sampled distances, computed with 10^{-8} inside the square root. The term averages $[\max(0, r_a + r_b - 0.035 - d_{ab})]^2$. Segment radii are (0.057, 0.046, 0.057, 0.046, 0.082, 0.058, 0.082, 0.058), ordered as left upper arm, left forearm, right upper arm, right forearm, left thigh, left shin, right thigh, and right shin. This is a sampled regularizer, not exact whole-body collision detection.

Table 1. FK-model supervision coefficients; common architecture and FK.

Condition	*3D*	*Rep.*	*Rot.*	*Bone*	*Sep.*
FK-S3D	1	0	0	0	0
FULL	1	1	0.05	0.1	0.01
−Reprojection	1	0	0.05	0.1	0.01
−Rotation	1	1	0	0.1	0.01
−Geometry	1	1	0.05	0	0

Geometry deletes both final terms; other weights are not retuned. Table 1 lists the FK supervision settings. DIRECT uses only the same three-dimensional loss as FK-S3D. Deletions keep all remaining coefficients fixed; the combined geometry deletion removes both bone-length and separation terms, so their individual contributions cannot be identified from that condition.

3.5 Optional original 80-step IK

The optional refinement, denoted W80, starts from the frozen network output and retains its rest skeleton and FK. Adam with learning rate 0.025 optimizes residual Euler angles, root translation, and log scale for 80 iterations without updating network parameters. Local rotations become $L_j^0 R_x(\alpha_{jx}) R_y(\alpha_{jy}) R_z(\alpha_{jz})$, where $\alpha_j = \tanh(\delta_j)\beta_j$ on rotatable nodes, $\beta_0 = 1.4$ radians and other bounds are 0.9. Root and log-scale changes are $0.5\tanh(\delta_t)$ and $0.5\tanh(\delta_k)$; log scale remains clipped to $[-2,2]$.

For each active sample, with normalized observations u_j, the objective is

$$\mathcal{E} = \frac{\sum_j m_j \parallel Q_{b(j),xy} - u_j \parallel^2}{\max(\sum_j m_j, 1)} + 0.002\,\mathrm{mean}(\alpha^2) + 0.003(\ell - \ell_0)^2 + 0.001\,\mathrm{mean}((t - t_0)^2). \quad (7)$$

Samples with an initial observed-point RMS error no greater than 0.25 pixels are not refined. The original W80 has no collision penalty, depth guard, multi-start procedure, or learned gate. We report network-only (NN) and post-IK results separately.

3.6 Deterministic polyhedral rendering

Twenty selected joints and their global rotations control procedural rigid parts representing the trunk, pelvis, head, neck, limbs, hands, feet, and joint shells. Most cross sections have four vertices; the head has an eight-vertex chamfered section. Head proportions and hand extensions are fixed by design, not inferred anatomical details. The rendering is neither a watertight skin surface nor a reconstruction of the earlier Maya model.

The deterministic renderer projects to 256 × 256 orthographic images, applies triangle z-buffering, and draws visible mesh edges with twofold supersampling, nominal one-pixel strokes, and a hidden-line tolerance of 0.004 synthetic units. Ink opacity is fixed at 0.5731573129161637 using historical T-training rendering statistics; root depth 4 is only a camera convention. Crisp lines reflect rendering choices and do not validate predicted depth.

In Fig. 1, the skeleton projection and polyhedral body depict the same stored FULL prediction. Body construction also uses its global rotations and apparent scale; predicted rest offsets are already reflected in the stored 3D joints, so the plotted 2D graph is not the renderer's complete input. Synthetic GT is an evaluation reference, not query geometry supplied to inference.

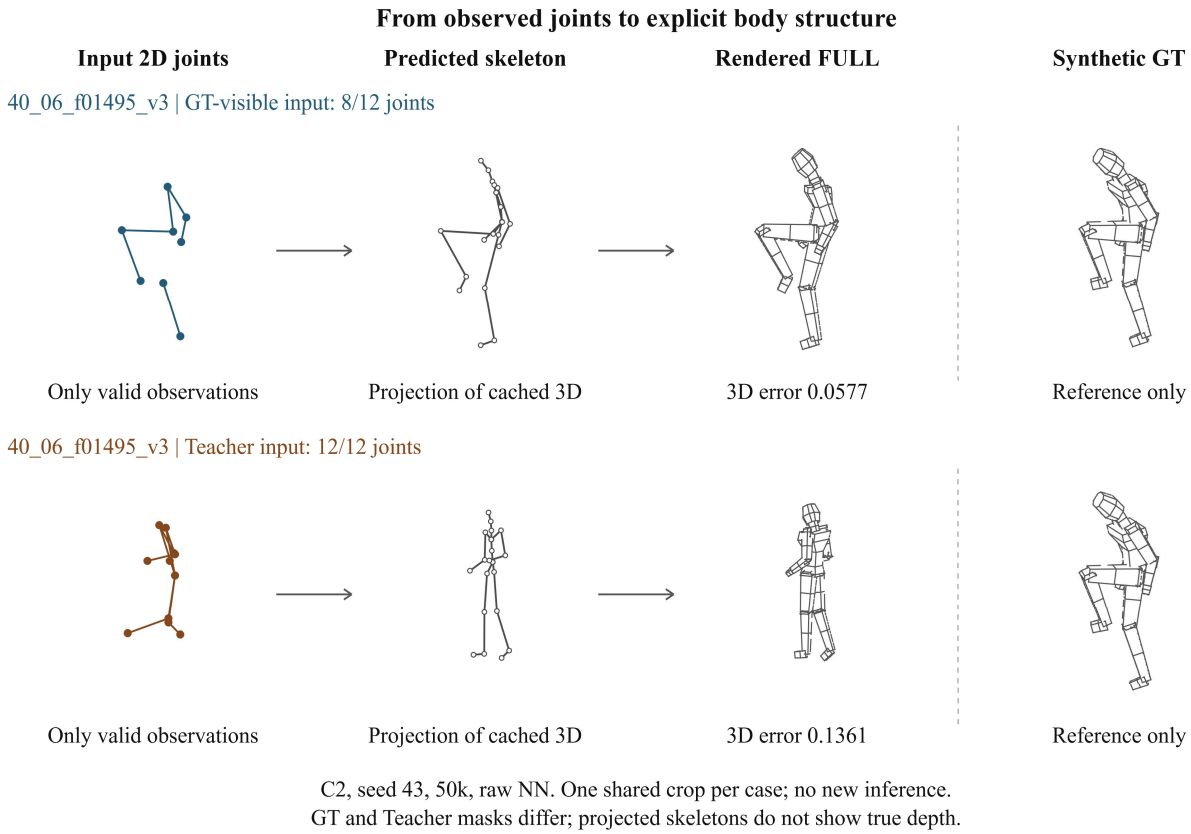


Figure. 1. Observation-to-structure trace for the central C2 case. Rows use GT-visible input with 8 valid body joints and Teacher input with 12. Columns show masked observations, the cached FULL 26-joint projection, its polyhedral body, and an evaluation-only synthetic GT reference. The renderer also uses global rotations and apparent scale; predicted rest offsets are reflected in the stored joints, so the 2D graph alone is insufficient. Predictions are seed-43, 50,000-update NN outputs without IK. The same case crop preserves relative placement.

4. Data and Experimental Protocol

4.1 Synthetic correspondence and source separation

Motion data originate from the CMU Motion Capture Database [11], using Bruce Hahne's BVH conversion at the fixed mirror revision [12]. Our code reads the numerical skeleton hierarchy and animation channels, normalizes the rig, and generates procedural body geometry together with aligned 2D/3D targets. The labels are exact within this synthetic system, but they are not independently measured anatomical surfaces or pose-teacher pseudo-labels.

The inventory comprises 4,078 synthetic images from 71 trials and 26 subject identifiers. Training uses 2,220 images from 39 trials, and development uses 1,080 from 19 trials. The remaining 778 images from 13 trials belong to viewed historical diagnostics, not new tests. Subject identifiers do not establish distinct individuals. The known same-session association between training trial 22_14 and historical trial 23_14 remains classified as train-related diagnostic evidence.

C1 and C2 have separate identities and evaluation histories (Table 2). C1 contains 768 images from twelve captures, with 762 queries shared by GT and Teacher evaluation. It was first opened after the original nine 50,000-update endpoints and evaluation rules were frozen; later seed-44 ablations are follow-up analyses on already viewed sources. C2 was registered separately, with twelve captures and 756 images; 747 admit Teacher inference. Its manifest preceded the model evaluation and excluded known source/content associations. C2 is not C1 with twelve images removed. New capture IDs and content checks cannot guarantee independent people, actions, or unknown recording sessions. Both sets share the same procedural rendering and filtering pipeline, which limits evaluation of extreme body contact.

Table 2. Distinct data roles and complete evaluation denominators.

Role	*Images*	*Groups*	*GT / Teacher*
Training	2,220	39 trials	Training only
Development	1,080	19 trials	1,080 / 1,053
C1: initial capture	768	12 captures	768 / 762
C2: later capture	756	12 captures	756 / 747
Perturbation subset	114	19 dev. trials	Coverage by level
Historical anime	16	16 images	No 3D GT; 13 outputs

C1 initial comparisons use 762 common queries; C2 primary GT/Teacher tables use all 756/747 inferable queries, with common-query results separately retained. For the historical 16-image evaluation in Table 6, PCK retains all 110 visible landmarks. The additional 84-image review and two six-image model-hidden follow-ups are evaluated separately in Sections 5.7 and 6.

The earlier O and T image collections contain 2,000 images each, but are not paired training data for the current structure predictor. Historical T images provide visual-style context, including 1,177 training images used only to determine fixed rendering statistics. The T collection's single-animation origin and mirrored duplicates do not imply independent actions. Real O images are reserved for exploratory evaluation.

4.2 Matched training and model identity

Original 26-joint training uses PyTorch [13] with Adam (learning rate 5×10^{-4}; $\beta_1 = 0.9$, $\beta_2 = 0.999$; $\varepsilon = 10^{-8}$; no weight decay), batch size 64, and StepLR decay by 0.5 every 20,000 updates. Each original 26-joint primary endpoint contains 50,000 updates. Output-head biases use training-only canonical rotation, root, and scale statistics; head weights are initially multiplied by 0.01.

Seeds 43 and 44 provide matched repetitions. Within each seed, FK-S3D, FULL, and deletion variants share the full Step-0 state and the sampling/augmentation sequence. DIRECT shares only the hidden-layer initialization (Section 3.3). Seed 42 is an audited historical reference, not a third new run. Reprojection and rotation deletions use seeds 43 and 44; joint geometry deletion uses seed 43 alone. All original 26-joint endpoints were trained for exactly 50,000 updates, including the first twenty validated steps of DIRECT. Primary comparisons use these fixed endpoints, not checkpoints chosen by development performance. The cross-view continuation is reported separately.

For input augmentation, a 0.6 probability triggers random joint removal while retaining at least four observed joints; Gaussian noise is sampled from standard deviations (0, 0, 2, 5) pixels. Missing inputs are set to zero, and synthetic 3D targets stay fixed. DIRECT retains the same full26 SmoothL1 objective and optimizer schedule. The FK launchers refer to the Stage J environment, whose archived audit records Python 3.12.6 and PyTorch 2.6.0+cu124. The exact CUDA runtime, driver, and cuDNN versions were not recorded for every FK training run; these are not inferred from the later cost environment.

4.3 Metrics, aggregation, and reviewed anime labels

The primary original 26-joint synthetic error is the mean Euclidean error over 12 root-relative body joints,

$$E_{3D} = \frac{1}{12} \sum_{j \in body12} \left\| \left(Q_j - Q_0\right) - \left(Q_j^{\mathrm{gt}} - Q_0^{\mathrm{gt}}\right) \right\|_2. \qquad (8)$$

Joint errors use normalized orthographic camera coordinates, not millimeters. The body12 measure does not summarize errors at every predicted full26 joint. Camera-scale relative bone-length error compares predicted and target parent–child distances. Wireframe Dice thresholds grayscale values below 216.75. The nonadjacent-limb intersection indicator is a restricted diagnostic, not a measure of anatomical correctness or exhaustive contact. Finite outputs that fail geometric checks remain included in error summaries.

GT inputs use accurate visible joint projections; Teacher inputs use predicted coordinates and their own validity masks. Historical development and C1 comparisons use 1,053 and 762 common queries, respectively. C2 reports all 756 GT-inferable or 747 Teacher-inferable queries for the primary condition-specific means; common-query analyses are kept separately. We first average within each capture source, then across sources with equal weight. Paired 95% intervals come from 5,000 source-level bootstrap resamples, not independent frames or between-seed variation. We distinguish the original frozen C1 evaluation from later analyses on those viewed sources.

Post-hoc C2 geometry diagnostics use the same frozen 50,000-step NN predictions and GT arrays, with identical queries across DIRECT, FK-S3D, and FULL within each input condition (756 GT or 747 Teacher queries from 12 sources). We measure absolute interior-angle error at both elbows and knees, averaged over four angles in degrees. For eight upper/lower arm and thigh/shank segments, relative length error is the absolute predicted-minus-GT length divided by GT length; the left and right summaries each use four segments. These

dimensionless limb measures differ from the all-bone measure in Table 3. They are averaged within query, then within source, then equally across sources. They do not measure anatomical angle-limit violations.

For descriptive source-paired effect sizes, let d_s be the mean error difference A minus B in source s. Cohen's d_z is mean(d_s) divided by the sample standard deviation of the 12 differences (11 degrees of freedom); negative values favor A. Seeds and input conditions are kept separate. These post-hoc descriptors supplement the 5,000-resample source-level intervals; they are neither frame-level effect sizes nor estimates of between-seed uncertainty.

The historical real-image evaluation in Table 6 includes 16 historical O-validation images with 110 definitely visible, coordinate-valid landmarks. The other 82 joint states comprise 11 occluded, 65 outside the frame, and six legacy not_visible records; the latter remain unresolved. The same annotator re-annotated or reviewed the points, sometimes with access to earlier labels. This was not an independent second review or a wholly blind annotation. The remaining 80 historical annotation records, including O-train32, are provisional.

Reported pixel error is computed first across eligible matched landmarks within an image and then averaged over images with model outputs. PCK uses all 110 reviewed visible joints, counting missing outputs as failures. Thresholds of 0.05 and 0.10 are applied separately using the image diagonal or the reviewed visible-joint bounding-box diagonal. Image-paired intervals use 5,000 bootstrap resamples. Reviewed labels evaluate fixed predictions only; they do not change observation, NN outputs, IK, or model selection. Because real-anime 3D ground truth is unavailable, no real-image 3D error is reported.

The controlled robustness subset comprises six fixed, evenly spaced images from each of 19 development sources (114 images), selected independently of model predictions. The tests separately add Gaussian coordinate noise of 0, 2, 5, or 10 pixels or remove 0, 1, 2, or 4 observed joints. Models receive identical perturbations. Noise does not change masks or clip out-of-frame values; deletion adds no noise. A sample with fewer than four remaining joints has no prediction but stays in the coverage denominator. Error summaries use eligible finite outputs, while paired differences require comparable queries. These perturbations do not reproduce the full range of teacher failures.

5. Results

5.1 Supervision and direct-coordinate comparison

On C1's original 762 common queries, accurate-input FULL-minus-FK-S3D error differences are −0.004638/−0.004172/−0.006425 for seeds 42/43/44; all source intervals favor FULL. Dice increases from 31.42–31.53% to 33.16–33.59%. With Teacher inputs, all three error intervals include zero. These original observations remain distinct from C2. Table 3 adds the matched DIRECT comparison on C2.

Table 3. C2 matched 50,000-update NN comparison: localization versus camera-scale bone accuracy.

Input / seed	*FK-S3D 3D ↓*	*DIRECT 3D ↓*	*FULL 3D ↓*	*DIRECT − FK-S3D [95% source CI]*	*Bone relative error ↓ FK-S3D / DIRECT / FULL*	*Queries*
GT / 43	0.11233	0.10879	0.10631	-0.00354 [-0.00553, -0.00162]	0.0610 / 0.0908 / 0.0593	756/756
GT / 44	0.11022	0.10820	0.10579	-0.00203 [-0.00357, -0.00043]	0.0638 / 0.0935 / 0.0603	756/756
Teacher / 43	0.23917	0.23484	0.23466	-0.00434 [-0.00666, -0.00174]	0.0974 / 0.1478 / 0.1011	747/756
Teacher / 44	0.23551	0.23316	0.23399	-0.00235 [-0.00363, -0.00096]	0.1007 / 0.1544 / 0.0976	747/756

C2: 12 sources, distinct from C1's 768/762 protocol. GT uses 756 queries; Teacher uses 747. Errors are source-macro in normalized camera units. DIRECT and FK-S3D share 3D supervision; FULL additionally changes supervision. Mesh metrics and W80 are N/A for DIRECT. Source intervals are not seed intervals.

On C2, DIRECT has lower body12 error than FK-S3D in all four seed–input comparisons; all source-paired intervals for DIRECT minus FK-S3D lie below zero (Table 3). The gain comes with larger camera-scale relative bone-length error: 0.0908–0.0935 for DIRECT versus 0.0610–0.0638 for FK-S3D with GT inputs, and 0.1478–0.1544 versus 0.0974–0.1007 with Teacher inputs. FK parameterization alone therefore does not improve joint localization. FULL has lower GT error than DIRECT for both seeds, although the DIRECT-minus-FULL interval excludes zero only for seed 43. Under Teacher inputs, the differences are +0.000173 [−0.002006, +0.001853] and −0.000827 [−0.004280, +0.001851]; neither interval separates the models. FULL also uses additional supervision, so that comparison cannot isolate FK.

The frozen-cache geometry check further qualifies this trade-off. In DIRECT/FK-S3D/FULL order, elbow/knee angle MAEs are 14.97/15.92/14.52 degrees for seed 43 with GT inputs and 15.06/15.54/13.91 for seed 44. With Teacher inputs they are 19.53/21.08/20.01 and 19.35/20.38/19.14 degrees, respectively. FK-S3D has larger angle error than DIRECT in all four conditions; fixed connectivity therefore does not imply more accurate joint geometry.

For seed 43, left/right relative limb-length errors are 0.0890/0.0871 for DIRECT, 0.0515/0.0574 for FK-S3D, and 0.0479/0.0527 for FULL under GT inputs; under Teacher inputs they are 0.1576/0.1535, 0.0929/0.0984, and 0.0918/0.1004. For seed 44, the corresponding GT values are 0.0894/0.0932, 0.0547/0.0595, and 0.0488/0.0527; Teacher values are 0.1601/0.1560, 0.0976/0.1014, and 0.0883/0.0962. Thus lower limb-length error coexists with worse geometric-angle error for FK-S3D. No anatomical-validity claim follows from either metric.

For DIRECT minus FK-S3D, source-paired d_z values for body12 localization error are −0.984/−0.691 for GT seeds 43/44 and −0.953/−0.975 for Teacher seeds 43/44. For the eight-segment relative length error, the corresponding values are +2.520/+2.480 and +5.082/+6.290. The opposite signs summarize a consistent localization–length trade-off across these 12 sources, not universal superiority or practical significance. C2 is an already-viewed synthetic collection, and these descriptors do not add independent real-anime validation.

For the matched seed43/44 comparison, C2 results favor FULL over FK-S3D under accurate inputs in both tested seeds (43 and 44), with source intervals
below zero. Teacher-input point estimates favor FULL, but the seed-44 interval includes zero. Development, C1, C2, common-query, and full26 analyses have distinct scopes; we do not combine their best values. Historical seed42 FULL–FK-S3D results are retained separately and are not counted as a new matched replication.

5.2 Replicated supervision ablations

Removing reprojection supervision increases accurate-input NN error for both seeds on development and C1, with four source intervals above zero. Table 4 reports the C1 results. C2 shows increases of 0.006897 and 0.004451 for seeds 43 and 44, with intervals above zero; under Teacher inputs the increases are 0.006161 and 0.005062. These repetitions support a limited benefit of reprojection supervision, not robustness to arbitrary observation error. The supervised 2D target is the augmented input, not the clean synthetic projection.

Table 4. C1 follow-up NN deletion effects: same-seed FULL reference, shared FK and 50,000 updates.

Set	*Input*	*Seed*	*Δ3D: −Reprojection [paired 95% CI]*	*Δ3D: −Rotation [paired 95% CI]*	*ΔIntersection −Rotation (pp)*
C1	GT	43	+0.00726 [+0.00443, +0.00994]	-0.00198 [-0.00554, +0.00093]	+28.32
C1	GT	44	+0.00822 [+0.00620, +0.01029]	-0.00080 [-0.00311, +0.00122]	+30.84
C1	Teacher	43	+0.00427 [+0.00230, +0.00630]	-0.00103 [-0.00422, +0.00175]	+32.39
C1	Teacher	44	+0.00544 [+0.00302, +0.00787]	-0.00168 [-0.00479, +0.00125]	+33.07

Δ = deletion − FULL; 3D is normalized camera error; pp = percentage points. C1 follow-up uses 762/768 queries across 12 viewed sources, not a new blind test. This table reports C1 NN differences; Section 5.3 examines W80 separately. Geometry deletion remains seed-43-only.

Rotation-supervision deletion does not consistently raise or lower mean joint error. All eight NN source intervals from the original development/C1 analyses include zero. Under Teacher inputs, however, removing this term increases the limb-intersection indicator by 29.80/28.16 percentage points on development and 32.39/33.07 points on C1 (seeds 43/44). On C2, seed 43's GT-input joint error decreases by 0.001967 [−0.003899, −0.000194] when rotation supervision is removed; seed 44 is inconclusive. The stronger evidence for rotation supervision thus concerns this structural-risk indicator rather than a uniform improvement in localization.

The combined bone-length/separation deletion was evaluated only for seed 43. On C1 with GT inputs, it gives NN error 0.127841 and W80 Dice 49.10%, compared with FULL's 0.129035 and 48.67%. Deleting both terms does not uniformly worsen the results, and this experiment cannot separate their individual effects.

5.3 Observation errors and optional IK

W80 lowers error with accurate observations but can worsen 3D predictions when fitted to teacher estimates. For FULL on C1, GT-input error falls from 0.1272–0.1290 to 0.1141–0.1152 after W80, whereas Teacher-input error rises in all nine initial endpoints. Table 5 reports those

original results. On C2, the four FK-S3D/FULL combinations for seeds 43 and 44 improve by 0.012220–0.014714 under GT inputs yet worsen by 0.022165–0.024024 with Teacher inputs. All corresponding source intervals lie below and above zero, respectively.

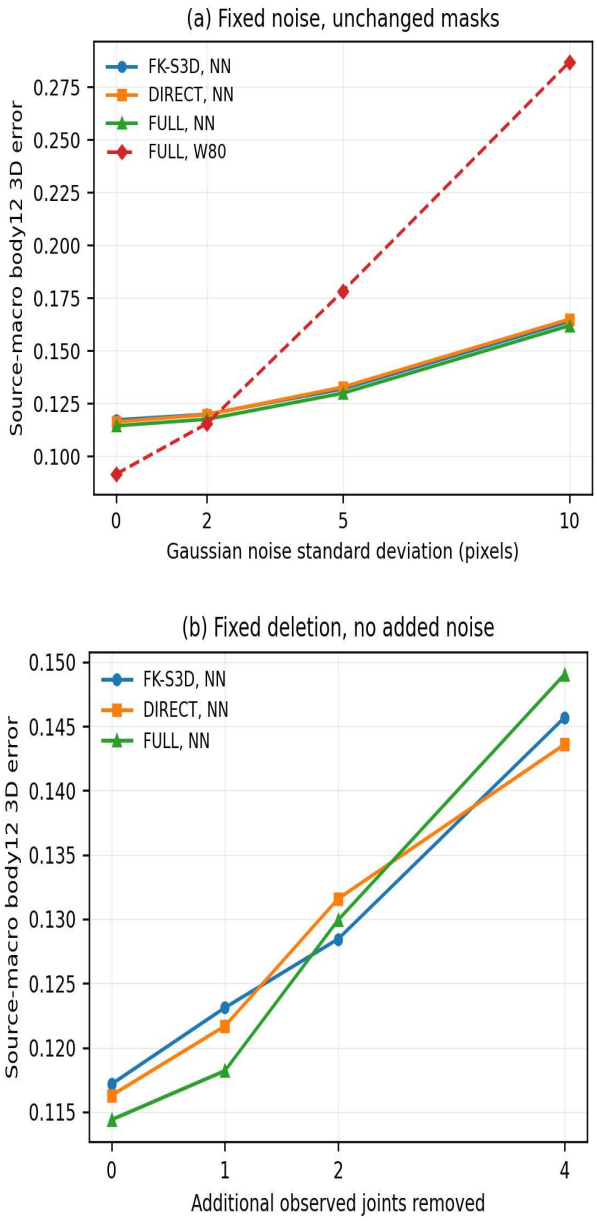


Figure. 2. Fixed perturbations on 114 historical-development queries from 19 sources. Seed 43 is a fixed display seed; the panels show the plotted methods only. (a) Noise leaves masks unchanged, with 114/114 inferable at every level; W80 is shown only for FULL. (b) Deletion adds no noise; errors use eligible outputs (114, 114, 114, 105), while coverage always retains denominator 114. Plot lines connect measured conditions, not additional experiments.

Table 5. FULL, Teacher inputs: original C1 W80 trade-off.

Seed	*3D NN / W80*	*Input px NN / W80*	*95% CI: Δ3D*
42	0.25846 / 0.27978	6.693 / 1.090	[+0.01704, +0.02534]
43	0.26069 / 0.28228	6.734 / 1.085	[+0.01790, +0.02527]
44	0.25914 / 0.27956	6.746 / 1.086	[+0.01623, +0.02456]

W80 − NN; C1, 762/768 queries, 12 sources. Better input fitting coexists with increased synthetic 3D error. For C1 seed 43, W80 reduces the supplied-input residual from 6.734 to 1.085 pixels, while synthetic 3D error rises from 0.260694 to 0.282281 and depth error from 0.146916 to 0.175768. Controlled perturbations show the same risk (Fig. 2a): with noise of 0/2/5/10 pixels, FULL43 NN errors are 0.1144/0.1175/0.1298/0.1619, compared with 0.0915/0.1154/0.1782/0.2870 after W80. At 5 and 10 pixels, refinement amplifies 3D error. Removing four observed joints leaves 105/114 cases inferable (92.1%), while the nine insufficient cases remain in the coverage denominator (Fig. 2b). These are fixed development diagnostics rather than a comprehensive simulation of teacher error.

5.4 Exploratory application to reviewed anime images

Of the 16 historical anime images, 13 contain at least four valid teacher joints. Images 1785, 1871, and 885 have insufficient observations and produce no structural prediction. Model outputs cover 104 of the 110 reviewed visible joints; the teacher covers 103, on a different set of points. Missing outputs count as failures in the full PCK denominator. Table 6 reports all six fixed FK endpoints without selecting a preferred seed.

Table 6. Reviewed historical anime evaluation; original W80, not a later IK variant

Model / seed	*NN px ↓*	*W80 px ↓*	*PCK-I % NN / W80*	*PCK-B % NN / W80*
FK-S3D 42	24.126	14.659	45.45 / 69.09	8.18 / 39.09
FULL 42	21.986	14.093	45.45 / 70.91	9.09 / 40.00
FK-S3D 43	22.724	14.845	45.45 / 70.00	15.45 / 36.36
FULL 43	23.247	14.131	45.45 / 71.82	8.18 / 37.27
FK-S3D 44	23.256	14.886	49.09 / 70.91	11.82 / 36.36
FULL 44	21.599	14.916	48.18 / 70.00	6.36 / 40.00

Pixel errors are image-macro over 13 inferable images/104 points. Both PCK columns retain the complete 110-point denominator (six missing predictions fail); I = image diagonal, B = reviewed visible-joint bounding-box diagonal, threshold 0.05. Same-annotator review; no real 3D ground truth.

Relative to FK-S3D, FULL's NN pixel-error differences on the reviewed anime images are −2.140, +0.523, and −1.657 for seeds 42, 43, and 44. All image-paired intervals include zero. Only seed 43 has a post-IK paired interval excluding zero, which is insufficient for a cross-seed claim. IK lowers two-dimensional pixel error for all six endpoints; this measures projection fitting, not verified depth recovery.

A same-image, same-landmark comparison to the frozen teacher uses 13 images and 99 common points. Teacher error is 12.6666 pixels; post-IK model differences are +0.0419 to +0.6473 pixels, with intervals containing zero. The results do not establish better localization than the teacher or reliable recovery of teacher-missing joints.

The evaluation changed through label review, not model improvement. Seventeen of the previous 127 visible points are no longer included; the remaining 110 coordinates shift by a mean of 7.268 pixels. On common points with unchanged model outputs, the revised coordinates raise pixel error by approximately 0.81–1.80 pixels. DIRECT was not added to this frozen real-image quantitative comparison.

5.5 Limited cross-view exploration

A limited cross-view candidate continued from FULL50k for 20,000 additional updates in each of two seed-43 arms. At the matched 70,000-update endpoints, Teacher-input NN error on historical development data was 0.251050 for Treatment and 0.251415 for Control, a relative reduction of 0.145%. This did not meet the predefined 2% progression threshold. No seed-44 candidate replication or subsequent candidate-specific capture evaluation was performed. C2 evaluates the frozen main models, not this closed candidate.

5.6 Measured computational cost

With resident weights, one CPU thread, Windows 10, and PyTorch 2.6.0+cu124, the representative FULL43 path takes 0.741 s per image on CPU or 0.644 s with an RTX 2060 running the teacher and NN while W80 and rendering remain on CPU. Each figure is the median of three runs after one warm-up on a fixed image. The timed input is an already converted RGB float tensor in CPU memory; model loading, disk I/O, input conversion, and PNG encoding are excluded. GPU boundaries are synchronized and include transfers. These values describe neither cold start nor file-to-file latency. Isolated single-image CPU NN calls take 2.694 ms for FULL and 0.643 ms for DIRECT (ten repetitions after two warm-ups). Batch-16 figures are amortized throughput, not interactive single-image delay.

5.7 Exploratory evaluation on additional anime images

We fixed a 100-image O-validation review queue comprising the 16 historical images and 84 additional images. The additional images followed a predetermined hash order fixed before this round's prediction analysis, without ranking Teacher or model quality; the original O-test images were not used. The frozen historical cohort contains 16 reviewed images and 110 definitely visible, coordinate-valid landmarks. Of the 84 additional images, 84 have completed reviews containing 737 definitely visible, coordinate-valid landmarks; 0 remain without a completed review. The historical and added cohorts are reported separately; uncompleted drafts are not treated as invisible ground truth or included in an accuracy denominator.

Across the full queue, the frozen Teacher supplies 900 valid observations out of 1,200 possible body joints; 94 of 100 images meet the prespecified minimum of four valid joints for structure inference. Validity requires finite in-frame coordinates and a smaller axis-maximum response of at least 0.3. These are machine-observation and inference-coverage counts, not human visibility or landmark accuracy. Missing predictions count as failures

in full human-visible PCK denominators; unreviewed points supply no reference labels. There is no real O-domain 3D ground truth. The added-cohort results below keep both seeds and NN versus original W80 separate; W80 is inapplicable to XYZ-only DIRECT.

Historical same-annotator review does not constitute an independent second-annotator study. In this cached extension, human labels are used only to evaluate already saved predictions and do not change observation masks, network parameters, NN/W80 outputs or model selection. The added visible-point records contain 336 manual clicks, 0 manual drags, 401 explicit acceptances of machine candidates and 0 points with other or unknown provenance. Teacher display preceded 654 visible-point actions; for 83 it was not displayed in this tool, and 0 have unknown exposure. Candidate acceptance is an assisted human-confirmed reference, not independent ground truth, and a manual click or drag does not establish blinding. Raw provenance, known display exposure and uncompleted/provisional states remain separate. All six cached network endpoints spanning DIRECT, FK-S3D and FULL with seeds 43 and 44 are retained without choosing a preferred seed. This extension does not modify the original frozen Table 6 or its historical-cohort definition.

For the added cohort, image-macro pixel errors were DIRECT NN 19.232/18.868, FK-S3D NN 24.391/23.750, FK-S3D W80 12.888/12.850, FULL NN 22.021/21.765, FULL W80 12.529/12.574 (seed 43/44). Each reported network/stage condition has finite predictions on 81/84 images and 727/737 visible landmarks. In the same method order, full-visible PCK at 0.05 of the image diagonal is 65.26/67.30%, 52.92/55.22%, 77.07/77.48%, 56.85/57.80%, 76.66/77.20% (seed 43/44); all 737 landmarks remain in each denominator, including the ten missing predictions as failures. These are two-dimensional assisted-reference measures, not depth accuracy or independent ground truth.

The added cohort remains a provisional descriptive sample, with all cached endpoints retained without selecting a winner. Image-bootstrap intervals describe the reviewed-image sample; they do not establish between-annotator reliability, population-wide generalization or real-anime 3D accuracy.

5.8 Exploratory template-count comparison on C2

We compared training-derived rest libraries with K = 1, 4, 8, and 18 using seeds 43 and 44. The single-template model uses the training-template medoid at index 13. The K = 4 and K = 8 libraries use deterministic farthest-point prefixes starting from the same medoid; validation and test poses do not select templates. Their four runs each received 50,000 updates under the original FULL protocol. The K = 1 and K = 18 conditions reuse the fixed-rest and historical FULL endpoints. K = 32 was not tested because the training pool contains only 18 unique rest templates.

Evaluation used the previously viewed C2 collection: 756 GT-input and 747 Teacher-input queries across twelve capture sources. Body12 error is measured in normalized camera units. ALL19 relative length error uses max(GT bone length, 0.01) as the denominator. Paired 95% intervals were computed from 5,000 capture-source bootstrap resamples, with seeds and input conditions kept separate.

For each K = 1, 4, or 8 versus K = 18 comparison, all four body12 error-difference intervals included zero; this does not establish equivalence. Length-error differences depended on the condition. K = 1 had higher error in two conditions and unresolved differences in two. K = 4 had lower error in three conditions and an unresolved difference in one; K = 8 had lower error in one and unresolved differences in three.

No preferred library was selected from these results. Changing K also changes template coverage, mixture-head capacity, and initial rest predictions, so the comparison does not isolate template count. The intervals describe variation across capture sources, not training seeds. These NN diagnostics establish neither W80 improvements nor anatomical correctness.

5.9 Post-hoc common-supervision diagnostic on C2

A post hoc 2-by-2 diagnostic compares DIRECT and FK under either 3D-only supervision or a shared objective: full26 SmoothL1 plus unit-weight masked SmoothL1 reprojection to the actually augmented input coordinates. Reprojection targets use the input noise and missing-point mask, not corrected ground-truth projections. The shared objective excludes FULL's rotation, bone and capsule terms. Each new common-objective model starts from its architecture's registered seed-43 Step-0 state and receives 50,000 updates with matched sampling and augmentation. DIRECT and FK retain different heads and parameter counts. These models are distinct from the original FULL and do not isolate a pure FK algorithm effect.

On the previously inspected C2 sources, NN body12 error for GT/Teacher input is 0.10879/0.23484 for DIRECT-3D, 0.11233/0.23917 for FK-3D, 0.10069/0.23268 for DIRECT-common and 0.10192/0.23518 for FK-common. Means are capture-source-equal over twelve sources, with 756 GT-input and 747/756 inferable Teacher-input queries; finite failed outputs remain included. Errors use normalized camera units, not millimeters. Paired 95% intervals from 5,000 source-bootstrap replicates for common minus 3D-only lie below zero for both architectures and inputs. Under the shared objective, FK minus DIRECT is +0.00123 [-

0.00090, 0.00353] for GT and +0.00249 [-0.00008, 0.00499] for Teacher; these intervals do not establish equivalence.

With the shared objective, FK reduces ALL19 relative bone-length error by 0.02787 [0.02115, 0.03498] for GT and 0.05366 [0.04983, 0.05760] for Teacher, but raises GT elbow/knee angle error by 1.72 [0.84, 2.68] degrees. The Teacher angle interval includes zero. The nineteen-source historical development set also gives higher FK-common localization error than DIRECT-common: +0.00770 [0.00345, 0.01328] for GT and +0.00463 [0.00185, 0.00782] for Teacher. Thus, the stronger DIRECT reference and adverse development results are retained. This single-seed, unadjusted diagnostic supports supervision and geometry trade-offs, not FK superiority or independent confirmation. FK-common also increases XY and GT-projection errors in both input conditions.

6. Discussion and Limitations

Under matched 3D supervision on C2, DIRECT predicts more accurate joint positions than FK-S3D but incurs larger bone-length errors. FULL's gains over FK-S3D with accurate inputs therefore reflect added supervision within the same FK architecture, rather than a new kinematic model. Comparing FULL against DIRECT changes both parameterization and losses; their Teacher-input intervals do not clearly separate the methods and do not prove equivalence. Rotation predictions also control rigid body parts beyond joint positions, so joint error and intersection indicators can move in different directions. The additional elbow/knee diagnostics reinforce this distinction: FK-S3D reduces limb-length error but increases angle error relative to DIRECT. The framework is therefore an inspectable structure parameterization with measured trade-offs, not evidence that hard skeletal connectivity alone produces anatomically correct poses.

FK preserves predicted rest-bone lengths during rotation, but does not ensure correct segment directions. In the recursion in Section 3.2, each joint position depends on rotations along its ancestral chain. An upstream rotation error can displace several descendant joints, with positional effects that depend on pose and lever arms. Errors at different joints may reinforce or partly cancel. This propagation concerns positions: rotating both segments at an elbow or knee together leaves their included angle unchanged. Angle errors instead depend on relative segment directions and the predicted rest geometry. Smaller bone-length errors can therefore coexist with larger angle MAE. DIRECT predicts coordinates without explicitly composing local rotations or enforcing rest-bone lengths. These differences offer a possible geometric explanation for the observed trade-off, but the experiments do not establish rotation-error propagation as its sole cause.

Twelve observed landmarks cannot uniquely determine depth or unobserved head and neck structure. Rest-template mixtures constrain proportions, and procedural rigid shells do not represent anatomical surfaces. DIRECT retains a shared-parent skeleton but lacks part rotations and meshes, leaving mesh-specific metrics and W80 inapplicable. The intersection indicator excludes adjacent segments and does not detect all contact. W80 has no collision constraint and can overfit incorrect two-dimensional joints. Clean wireframe edges address none of these limitations.

C1's first frozen evaluation, later follow-up ablations on those viewed sources, and separately registered C2 have different evidential roles and denominators. Both capture sets share one synthetic generator; unknown session aliases, possible preprocessing overlap, and limited extreme-pose coverage remain. DIRECT and the main loss deletions were tested with two seeds, while geometry deletion has one. Source-bootstrap intervals characterize variability across the sampled sources, not between-seed uncertainty. The perturbation study covers 114 historical queries, and timings describe a small fixed workload rather than general deployment performance. Section 5.8 reports the K = 1, 4, 8, and 18 comparisons. Changes in template coverage and model capacity prevent these results from isolating the effect of template count.

The historical real-anime evaluation in Table 6 uses 16 previously viewed images with incomplete teacher coverage, same-annotator label review, unknown overlap with teacher pretraining, and no three-dimensional ground truth. Independent landmark review and actual 3D or multi-view evidence would be needed to support stronger reconstruction claims. Neither improved rendering sharpness nor closer agreement with the teacher can substitute for such evidence.

The domain gap has two distinct components. The synthetic set couples projected joints to a procedural body-and-rendering pipeline, while anime observations may contain stylized body proportions, foreshortening, severe cropping, and occlusion by hair or clothing. In addition, the frozen teacher may miss or misplace those joints. C2 changes capture sources within the same synthetic rendering pipeline; it does not test a new artistic image domain. Teacher-input experiments expose observation errors on synthetic images, but they do not reproduce all anime errors or constitute domain adaptation. No synthetic-plus-anime fine-tuning of the reported structure predictor is claimed. The historical 16-image reviewed evaluation cannot establish real-anime depth accuracy. The automatic observation-mask and framing statistics compare machine detections with synthetic masks without using human visibility labels. These diagnostics are

separate from the provisional 2D references in the completed-review cohorts. Synthetic geometric/random-dropout masks and response-thresholded Teacher masks have different mechanisms; differences in valid-joint counts or bounding boxes therefore do not isolate a causal domain-gap effect.

A provenance-stratified sensitivity analysis examines 83 visible manual-click landmarks across ten added-cohort images. These selected point actions were recorded without prior Teacher display in the annotation tool. Nine images, containing 79 of those landmarks, support structure inference; PCK retains all 83 points, counting missing outputs as failures. W80 minus NN image-macro error in the 256-pixel interface ranges from -8.96 to -6.22 pixels across both seeds and FK configurations; all four paired image-bootstrap intervals are below zero. FULL-minus-FK-S3D pixel intervals include zero. Comparisons with the Teacher instead use 72 common finite points on nine images. Nonrandom provenance groups and tool-level nonexposure do not establish a causal display effect or fully independent annotation.

A separate model-hidden annotation follow-up contains six new images from four source families and 72 definitely visible, manually clicked landmarks. The same annotator completed these labels after all model outputs had been frozen; the tool displayed no Teacher, network or IK predictions. All eleven applicable Teacher/network/stage conditions provide finite outputs on all six images and all 72 points. In native-image coordinates, the Teacher has 13.77-pixel image-macro error and PCK-I/PCK-B of 69/72 and 61/72 at 0.05 of the respective diagonals. The four FK W80 conditions have 14.50-14.71-pixel error and 69/72 PCK-I, improving over their own NN projections; their paired image- and family-bootstrap error intervals are below zero. FULL-versus-FK-S3D differences depend on the small-sample aggregation and remain inconclusive. Native pixel errors are resolution-dependent and are not directly pooled with the 100-image cohort's 256-pixel errors. Six images and four families provide only exploratory 2D evidence, not a second-annotator study, population-level validation or real-anime 3D accuracy. Overlap with Teacher pretraining is unknown.

An additional model-hidden follow-up used six new Kasuga images with 72 visible landmarks. They share one artist family with the earlier follow-up and were annotated by the same person. Frozen Teacher predictions cover 70/72 points; all ten applicable network/stage conditions cover 72/72. Full-denominator PCK-I/PCK-B at 0.05 are 68/72 and 62/72 for Teacher, versus 68/72 and 61/72 for all four W80 conditions. W80-minus-NN native-pixel differences range from -33.76 to -29.83, with all four paired image-bootstrap intervals below zero. FULL-minus-FK-S3D pixel intervals on 72 common points and W80-minus-Teacher pixel intervals on 70 common points include zero; this does not establish equivalence. Source-cluster intervals are undefined with one family, and Teacher-pretraining overlap remains unknown. These exploratory 2D results remain separate from the 100-image cohort and earlier six-image study and do not establish real-anime 3D accuracy.

7. Conclusion

The C1/C2 experiments show a trade-off between joint accuracy and bone-length consistency, rather than a general advantage of FK parameterization. Under matched three-dimensional supervision on C2, DIRECT predicts joint positions and elbow/knee angles more accurately than FK-S3D, while FK-S3D has lower bone-length error. C1 and C2 support accurate-input localization gains for FULL over FK-S3D within their shared architecture. This is a comparison of supervision, not evidence for a new FK model. FULL and DIRECT are not clearly separated under Teacher inputs, and their comparison changes both parameterization and losses. The ablations provide limited support for reprojection supervision and a more consistent effect of rotation supervision on limb-intersection indicators than on mean joint error. W80 further shows that improved projection fitting can accompany worse three-dimensional predictions. The historical anime evaluation remains exploratory. These findings support an explicit structure representation with measured trade-offs, not unconditional superiority over coordinate regression or reliable three-dimensional recovery from anime images.

Acknowledgments

The motion data originate from the CMU Graphics Lab Motion Capture Database, created with support from NSF EIA-0196217. The cited NSF support concerns creation of the CMU database, not funding of this study. This study received no external funding.

References

[1] S. Chen and M. Zwicker, “Transfer Learning for Pose Estimation of Illustrated Characters,” Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 793–802, 2022.

[2] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, “Learning to Discover Cross-Domain Relations with Generative Adversarial Networks,” Proc. ICML, PMLR, vol. 70, pp. 1857–1865, 2017.

[3] J. Martinez, R. Hossain, J. Romero, and J. J. Little, “A Simple yet Effective Baseline for 3D Human Pose Estimation,” Proc. IEEE International Conference on Computer Vision (ICCV), pp. 2640–2649, 2017.

[4] Akatuki25, “Copillust—Illustration Pose Estimation,” software repository, Curriculum S2 checkpoint best_coco_AP_epoch_10.pth. Accessed October 28, 2026.
[5] T. Jiang et al., “RTMPose: Real-Time Multi-Person Pose Estimation based on MMPose,” arXiv:2303.07399, 2023.
[6] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, “On the Continuity of Rotation Representations in Neural Networks,” Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5745–5753, 2019.
[7] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, “Learning to Reconstruct 3D Human Pose and Shape via Model-Fitting in the Loop,” Proc. IEEE/CVF ICCV, pp. 2252–2261, 2019.
[8] J. Li, C. Xu, Z. Chen, S. Bian, L. Yang, and C. Lu, “HybrIK: A Hybrid Analytical-Neural Inverse Kinematics Solution for 3D Human Pose and Shape Estimation,” Proc. IEEE/CVF CVPR, pp. 3383–3393, 2021.
[9] B. N. Oreshkin, F. Bocquelet, F. G. Harvey, B. Raitt, and D. Laflamme, “ProtoRes: Proto-Residual Network for Pose Authoring via Learned Inverse Kinematics,” International Conference on Learning Representations (ICLR), 2022.
[10] T. Wehrbein, M. Rudolph, B. Rosenhahn, and B. Wandt, “Probabilistic Monocular 3D Human Pose Estimation With Normalizing Flows,” Proc. IEEE/CVF ICCV, pp. 11199–11208, 2021.
[11] CMU Graphics Lab, “Carnegie Mellon University Motion Capture Database.” Database and acknowledgment information, accessed October 28, 2026.
[12] B. Hahne, “CMU Motion Capture BVH Conversion,” 2010 MotionBuilder-friendly release, mirrored in una-dinosauria/cmu-mocap, revision 09a07f54f3bbb58797325f009282d0b2048a2871.
[13] A. Paszke et al., “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” Advances in Neural Information Processing Systems, vol. 32, 2019.
[14] J. Li, P. Chen, X. Ju, H. Xu, and J. Jia, “VLPose: Bridging the Domain Gap in Pose Estimation with Language-Vision Tuning,” arXiv:2402.14456v1, 2024. https://arxiv.org/abs/2402.14456v1
[15] K. Ludwig, Y. Oksymets, R. Schön, D. Kienzle, and R. Lienhart, “Efficient 2D to Full 3D Human Pose Uplifting including Joint Rotations,” arXiv:2504.09953v1, 2025. https://arxiv.org/abs/2504.09953v1