INFLUPOSE: LEARNING STATE-CONDITIONED GEOMETRIC INFLUENCE FOR CATEGORY-LEVEL OBJECT POSE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Correspondence-based category-level pose estimation recovers an unseen object's pose and size from matches between its visible surface and canonical coordinates. Even reliable matches can differ in their geometric role. A broad surface may provide redundant evidence, whereas a small distinctive region can resolve an ambiguity. We introduce InfluPose, which learns this regional influence from interventions on paired complete training geometry. We delete and refill regions, measure the response of a frozen estimator, and combine the resulting changes with a dense geometric proxy. Four target channels describe orthogonal rotation, axial rotation, translation, and size. A point-cloud-only predictor learns these targets from partial observations. It combines influence with correspondence reliability in weighted Kabsch initialization and joint pose-and-size refinement. The pose objective trains the predictor to use this estimator-specific response field for recovery. Complete geometry is required only for training. On unseen object instances from REAL275 and HouseCat6D, InfluPose obtains the highest strict-pose mAP among the compared point-cloud methods. Matched REAL275 ablations show gains from regional supervision over learning the influence coefficients through the endpoint objective alone.

1 INTRODUCTION

Category-level object pose estimation must recover rotation, translation, and metric size for objects whose exact shapes were not seen during training. Correspondence-based methods address this problem by matching the visible 3D surface to canonical coordinates (Wang et al., 2019). After instance extraction, this alignment can use partial point-cloud geometry without RGB appearance features or the exact test-instance CAD model (Liu et al., 2026). Shape variation, occlusion, and sensor noise change the geometric evidence available for recovery. We evaluate this point-cloud setting on REAL275 and HouseCat6D.

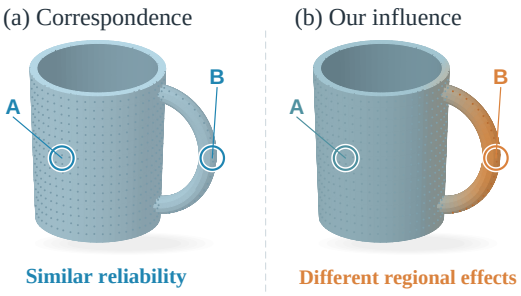

Figure 1: Similar correspondence reliability can coexist with different regional effects. Schematic regions A and B motivate learning influence alongside reliable matches.

Existing methods improve both the correspondences and their reliability. Canonical shape learning, prior deformation, and structure-guided adaptation accommodate instance variation (Chen et al., 2020; Tian et al., 2020; Chen & Dou, 2021). Learned keypoints, correspondence features, and shape-independent representations improve the evidence passed to alignment (Lin et al., 2024; Ren et al., 2025b;a). Confidence and uncertainty estimates then reduce the contribution of unreliable matches (Di et al., 2022; Ren et al., 2025b; You et al., 2024). These developments make geometric recovery more effective, but correspondence reliability alone does not describe the role of a region within the available surface support.

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

Consider a partly occluded mug with many reliable matches on its body and only a few on its handle. Neighboring body points may replace a missing body patch, while the visible handle may provide a distinctive orientation cue. The effect of either region therefore depends on the surrounding geometry and on the state component being estimated. Geometric fitting already uses spatial arrangement, but an endpoint pose loss does not explicitly supervise this regional dependence. We use the change in a recovered state after a regional intervention as an additional learning signal (Figure 1). This distinguishes the reliability of a match from the effect of its surrounding support on rotation, translation, or size.

We introduce **InfluPose**, a pose estimator that learns **geometric influence** through regional response supervision. A fixed estimator provides the reference for these responses. During training, we perturb regions of paired complete geometry and measure the resulting changes in its predictions. A partial-cloud network learns to predict this influence field from the visible observation. For a specified estimator and intervention, this field describes each region's effect on the predicted state. It complements correspondence reliability in learned geometric recovery.

Complete training geometry lets every intervention start from intact object support, before a particular view's occlusion has removed points. We delete a selected region, refill from retained points to preserve input cardinality, and compare the frozen estimator's output with its prediction on the intact cloud. The changes define four channels for orthogonal rotation, axial rotation, translation, and log-size. A finite set of probes supplies measured regional responses, while a dense geometric proxy covers locations between probes. We blend the two sources and transfer their targets to visible anchors. The partial-cloud predictor learns these targets from its incomplete observation.

The network predicts canonical correspondences, reliability, and influence over the same anchors, together with an initial metric size. Predicted influence enters weighted Kabsch initialization (Kabsch, 1976; Umeyama, 1991) and joint refinement of rotation, translation, and size. Direct supervision trains the response field, and the endpoint pose objective trains its use in recovery. This coupling makes regional supervision part of the estimation path. Inference requires one partial-cloud network pass, weighted Kabsch, and robust refinement. The complete-geometry estimator and regional interventions are needed only during training.

Our contributions are as follows.

- We introduce **InfluPose**, which learns state-conditioned geometric influence from complete training geometry and predicts it from partial point clouds for category-level pose recovery.
- We construct four-channel targets by combining finite deletion-and-refill responses with a dense geometric proxy, then couple predicted influence and correspondence reliability in weighted initialization and joint pose-and-size refinement.
- On unseen instances from REAL275 and HouseCat6D, InfluPose obtains the highest strict-pose mAP among the compared point-cloud methods. Matched REAL275 controls show gains from regional supervision across training seeds.

2 Related Work

Correspondence-based category-level pose recovery. NOCS defines a shared canonical coordinate space for recovering object pose and size from visible points (Wang et al., 2019). Canonical shape learning, prior deformation, and structural adaptation accommodate instance variation (Chen et al., 2020; Tian et al., 2020; Chen & Dou, 2021). Query6DoF and AG-Pose learn sparse queries or keypoints, while SpotPose and SpherePose improve correspondence features and representations (Wang et al., 2023; Lin et al., 2024; Ren et al., 2025b;a). Hierarchical point encoders and geometry-aware attention provide context for recovery (Qi et al., 2017; Zhao et al., 2021; Qin et al., 2022). Pose consistency and structural reasoning provide complementary constraints (Chen et al., 2021; Lin et al., 2021; Liu et al., 2022; Zheng et al., 2023). Alignment, deformation, registration, and completion address missing geometry (Lin et al., 2022; Zhang et al., 2022b; Zhou et al., 2023; Ren et al., 2026). Other approaches generate and compare pose hypotheses through multi-hypothesis consistency (Liu et al., 2025), diffusion, or flow (Zhang et al., 2023; Tang et al., 2026; Lee & Kim, 2025; Ouyang et al., 2025). We focus on the supervision supplied to point-cloud correspondence recovery. Regional interventions on complete geometry provide response targets that a partial-cloud predictor learns alongside correspondences.

Reliability, perturbation, and response-based supervision. Rotation confidence, correspondence outlier scores, and vote-level uncertainty help geometric estimators downweight unreliable evidence (Di et al., 2022; Ren et al., 2025b; You et al., 2024). InfluPose retains inlier probability and residual precision for this purpose. Its additional influence targets describe regional prediction responses. Occlusion and input-perturbation methods examine prediction changes after removing or modifying input regions (Zeiler & Fergus, 2014; Fong & Vedaldi, 2017). Influence functions analyze infinitesimal reweighting of training examples (Koh & Liang, 2017), while statistical leverage characterizes observations through fitted least-squares geometry (Hoaglin & Welsch, 1978; Drineas et al., 2012).

We use finite deletion-and-refill responses of a frozen complete-cloud estimator to supervise a partial-cloud predictor. A dense proxy extends supervision between probes, and endpoint training determines how the predicted field is used with reliability in recovery. This transfers estimator-specific regional responses into the pose-estimation path. Matched controls test the benefit of this supervision, while response and ground-truth-damage diagnostics test whether its predicted rankings identify consequential regions within the fitted model.

3 METHOD

3.1 OVERVIEW

Given a partial camera-frame point cloud $\mathcal{P} = \{\mathbf{x}_n\}_{n=1}^{N}$, we estimate rotation $\mathbf{R} \in \mathrm{SO}(3)$, translation $\mathbf{t} \in \mathbb{R}^3$, and metric size $\mathbf{s} \in \mathbb{R}^3_+$. During training, paired complete geometry provides regional intervention targets. The predictor learns these targets from the observed cloud and its derived geometric features. Symmetry metadata is used for symmetry-aware training supervision but is not an input to the learned encoder. Test-time rotation-representative selection uses detector class labels, without ground-truth mug-handle visibility. Correspondences define alignment, reliability weights the residuals, and response supervision trains the influence field used in initialization and joint refinement.

3.2 COMPLETE-GEOMETRY INTERVENTION SUPERVISION

Regional intervention targets. Deleting from a partial cloud measures a response after that sample's own occlusion pattern has already removed points. We instead probe paired complete training geometry so that each regional intervention starts from the intact object support. For canonical model $\mathcal{M} = \{\mathbf{m}_j\}_{j=1}^{N_c}$, scale a^*, and pose $(\mathbf{R}^*, \mathbf{t}^*)$, this gives $\mathcal{C} = \{\mathbf{R}^*(a^*\mathbf{m}_j) + \mathbf{t}^*\}_{j=1}^{N_c}$ and $\mathbf{s}^* = a^*(\max_j \mathbf{m}_j - \min_j \mathbf{m}_j)$. This is paired complete training geometry.

For each paired example, let $\mathbf{c}_P = N^{-1}\sum_n \mathbf{x}_n$ be the alignment center derived from the partial observation. We freeze a complete-cloud estimator G, select $M = 8$ regions by farthest-point sampling, remove the nearest ρN_c points with $\rho = 0.125$, and deterministically duplicate retained points to preserve 1,024 inputs. For each region m and refill offset r, we compare $\mathbf{d}_0 = G(\mathcal{C}; \mathbf{c}_P)$ with $\mathbf{d}_m^{(r)} = G(\mathcal{C}^{(-m,r)}; \mathbf{c}_P)$, where $\mathbf{d} = (\mathbf{R}, \mathbf{t}, \mathbf{s})$. The estimator, alignment center, input size, and offset are fixed within each comparison. Training cycles over three deterministic offsets. The complete-cloud center and normalization extent are recomputed after each intervention, while $\mathbf{c}_P$ stays fixed. The response therefore includes changes in geometric content, point multiplicity, and preprocessing. The refill is a fixed-cardinality intervention rather than a simulation of natural occlusion. Its target measures G's response to this specified change in support and multiplicity.

Suppressing r, define $\boldsymbol{\omega}_m = \log_{\mathrm{SO}(3)}(\mathbf{R}_m\mathbf{R}_0^\top)$ and $\mathbf{u}_0 = \mathbf{R}_0\mathbf{e}_y$. The regional response is

$$\tilde{\boldsymbol{\eta}}_m = \left[\|\boldsymbol{\omega}_m - (\boldsymbol{\omega}_m^\top\mathbf{u}_0)\mathbf{u}_0\|_2, |\boldsymbol{\omega}_m^\top\mathbf{u}_0|, \|\mathbf{t}_m - \mathbf{t}_0\|_2/\tau_t, \|\log(\mathbf{s}_m \oslash \mathbf{s}_0)\|_2/\tau_s\right]. \tag{1}$$

The channels describe the orthogonal-rotation, axial-rotation, translation, and log-size responses of G to regional deletion. With $\bar{\eta}^d = M^{-1}\sum_l \tilde{\eta}_l^d$, we set

$$\eta_m^{\mathrm{del},d} = \begin{cases} \mathrm{clip}(\tilde{\eta}_m^d/\bar{\eta}^d, 0, 10), & \bar{\eta}^d > 10^{-6}, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

and zero the axial target for continuously local-y symmetric instances. The scales τ_t and τ_s place the raw translation and log-size responses in reference units. They cancel in the per-channel ratio $\tilde{\eta}_m^d/\bar{\eta}^d$

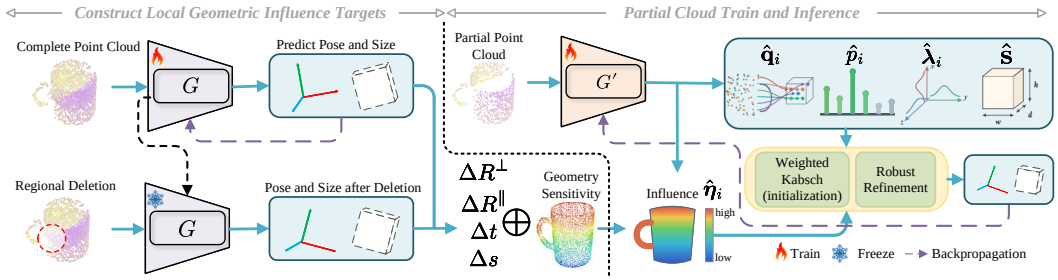

Figure 2: Target construction and partial-cloud training after pretraining G. Both complete-cloud branches use the same point cloud encoder G. Regional targets supervise the predicted influence, which enters joint recovery and receives endpoint gradients.

in Equation 2, so they do not change the normalized deletion targets while the channel stays active. Their effect on normalization is limited to the 10^{-6} activity threshold. They are retained when reporting raw response diagnostics. The normalized response represents relative regional influence within each state channel, rather than an absolute or cross-channel physical sensitivity.

Transfer to partial observations. For a probed training example, each visible anchor i inherits its nearest region $m(i)$. A training-only dense geometric proxy, computed from ground-truth canonical correspondences and metric size, supplies supervision between probes, giving the target

$$\boldsymbol{\eta}_i^* = \begin{cases} (1-\beta)\boldsymbol{\eta}_i^{\text{proxy}} + \beta\boldsymbol{\eta}_{m(i)}^{\text{del}}, & \text{probed}, \\ \boldsymbol{\eta}_i^{\text{proxy}}, & \text{otherwise}, \end{cases} \qquad \beta = 0.70. \tag{3}$$

The proxy encodes moment arms for rotation, uniform displacement for translation, and canonical distance for additive size perturbations. It covers every visible anchor, including locations between probes, with a uniform translation component. Deletion responses distinguish regions by their finite effect on G. Blending combines this estimator-specific information with dense geometric coverage on probed training examples. Both sources are normalized within each state channel, so their blend compares relative regional influence. Appendix A.1 specifies the proxy and probing schedule. Appendix A.2 describes normalization, refill, symmetry, and nearest-region transfer.

3.3 Partial-Cloud Prediction and Geometric Recovery

Influence prediction and learning. We select $K = 192$ anchors from the $N = 1{,}024$ observed points by farthest-point sampling in normalized coordinates. From these visible anchors, the partial-cloud network predicts

$$(\hat{\mathbf{q}}_i, \hat{p}_i, \hat{\boldsymbol{\lambda}}_i, \hat{\boldsymbol{\eta}}_i), \qquad \hat{\mathbf{s}}^{(0)} = H_s(\mathbf{f}_{\text{global}}), \tag{4}$$

where the outputs are canonical correspondence, inlier probability, axis-wise residual precision, four-channel influence, and initial metric size. Relation attention uses geometry-aware point-attention principles (Zhao et al., 2021; Qin et al., 2022). Its pairwise biases encode distance, normal, ray, and relative-depth differences in the camera observation. Here $\mathbf{f}_{\text{global}}$ is the pooled global token of the partial-cloud encoder. State conditioning distinguishes components of the recovered state. The field is predicted once from partial geometry and held fixed during refinement. Positive weight activations and a bounded metric-size head provide the inputs to recovery. Appendix A.1.2 gives their exact parameterizations, activation ranges, and solver clamps.

Let $\mathcal{A}$ collect valid anchor–active-channel triples and $\mathcal{E}$ the active example–channel pairs. With $\ell_\epsilon(x) = \log\max\{x, 10^{-5}\}$, let $\pi_{b,d}(\mathbf{v})$ denote the valid-anchor softmax of $\ell_\epsilon(\mathbf{v}_{b,:,d})/T$. The influence loss is

$$\begin{aligned} \mathcal{L}_{\text{inf}} = \Big\langle \mathrm{SL1}_{0.20}\big(\ell_\epsilon(\hat{\eta}_{bi}^d), \ell_\epsilon(\eta_{bi}^{*,d})\big) \Big\rangle_{\mathcal{A}} + \big\langle \mathrm{KL}(\pi_{b,d}(\boldsymbol{\eta}^*) \| \pi_{b,d}(\hat{\boldsymbol{\eta}})) \big\rangle_{\mathcal{E}} \\ + 0.25\,\mathrm{SL1}_{0.10}(\langle \hat{\eta}_{bi}^d \rangle_{\mathcal{A}}, \langle \eta_{bi}^{*,d} \rangle_{\mathcal{A}}). \end{aligned} \tag{5}$$

The regression term matches log-magnitudes, the KL term matches the spatial distribution across valid anchors, and the final term aligns the mean predicted and target influence. These terms supervise values and their allocation over the observed support. The endpoint objective provides a separate signal through the recovered state, allowing the field to serve geometric recovery as well as target prediction. Symmetry-inactive pairs are excluded. For continuous axial symmetry, the axial output remains an endpoint-trained axial influence channel (Appendix A.2.4). With the specified weights, $\mathcal{L} = \mathcal{L}_{\text{inf}} + \mathcal{L}_{\text{recovery}} + \mathcal{L}_{\text{aux}}$ groups regional supervision, recovered pose/size/box losses, and other prediction, cross-view consistency, and reference non-regression terms (Appendix A.1).

Influence-guided initialization. Pose and size are recovered together because metric correspondence alignment depends on the current size estimate. We first initialize alignment at the predicted size, then jointly adjust rotation, translation, and log-size using the residuals over the shared anchor set. Set $\mathbf{y}_i^{(0)} = \hat{\mathbf{s}}^{(0)} \odot \hat{\mathbf{q}}_i$ and

$$\bar{\eta}_i^{\text{pose}} = \tfrac{1}{3}(\hat{\eta}_i^{R_\perp} + \hat{\eta}_i^{R_\parallel} + \hat{\eta}_i^{t}), \quad \bar{\lambda}_i = \tfrac{1}{3}\sum_a \hat{\lambda}_{ia}, \quad \tilde{w}_i^{(0)} = \frac{\max(\hat{p}_i \bar{\lambda}_i \bar{\eta}_i^{\text{pose}}, \epsilon)}{\sum_j \max(\hat{p}_j \bar{\lambda}_j \bar{\eta}_j^{\text{pose}}, \epsilon)}. \tag{6}$$

Here $\epsilon = 10^{-5}$. Precision and influence are not upper-clipped in initialization. Weighted Kabsch then initializes the pose (Kabsch, 1976; Umeyama, 1991) by solving

$$(\mathbf{R}^{(0)}, \mathbf{t}^{(0)}) = \arg\min_{\mathbf{R}\in \mathrm{SO}(3), \mathbf{t}} \sum_i \tilde{w}_i^{(0)} \|\mathbf{R}\mathbf{y}_i^{(0)} + \mathbf{t} - \mathbf{x}_i\|_2^2. \tag{7}$$

State-conditioned robust refinement. The response-supervised field provides state-specific coefficients for joint refinement. These coefficients rescale the Jacobian's state blocks while preserving coupling through a shared normal equation. Direct response losses and endpoint gradients train the coefficients together. The rotation block also uses the current estimated object axis. Section 4.3 tests the four-channel design against shared influence and against coefficients learned without direct field supervision. For robust refinement, the fields below are clipped to $[10^{-4}, 1]$ for inlier probability, $[0.05, 50]$ for precision, and $[0.02, 8]$ for influence. At solver state k, let $\mathbf{s}^{(k)} = \text{clip}(\exp \boldsymbol{\ell}^{(k)}, 0.01, 0.80)$, $\mathbf{z}_i^{(k)} = \mathbf{s}^{(k)} \odot \hat{\mathbf{q}}_i$, and $\mathbf{r}_i^{(k)} = \mathbf{R}^{(k)}\mathbf{z}_i^{(k)} + \mathbf{t}^{(k)} - \mathbf{x}_i$. With $\boldsymbol{\Lambda}_i = \text{diag}(\hat{\boldsymbol{\lambda}}_i)$, the robust row weight is

$$a_i^{(k)} = \hat{p}_i \left(1 + \max\{\mathbf{r}_i^{(k)\top} \boldsymbol{\Lambda}_i \mathbf{r}_i^{(k)}, 10^{-12}\}\right)^{-1}. \tag{8}$$

For $\mathbf{u}^{(k)} = \mathbf{R}^{(k)}\mathbf{e}_y$, define $\mathbf{g}_{i,R}^{(k)} = \hat{\eta}_i^{R_\perp}(\mathbf{1}_3 - \mathbf{u}^{(k)} \odot \mathbf{u}^{(k)}) + \hat{\eta}_i^{R_\parallel}(\mathbf{u}^{(k)} \odot \mathbf{u}^{(k)})$ and $\mathbf{g}_i^{(k)} = [\mathbf{g}_{i,R}^{(k)}, \hat{\eta}_i^t \mathbf{1}_3, \hat{\eta}_i^s \mathbf{1}_3]$. The influence coefficients rescale the analytic joint Jacobian,

$$\mathbf{J}_i^{(k)} = \left[-[\mathbf{R}^{(k)}\mathbf{z}_i^{(k)}]_\times, \mathbf{I}_3, \mathbf{R}^{(k)} \,\text{diag}(\mathbf{z}_i^{(k)})\right], \qquad \widetilde{\mathbf{J}}_i^{(k)} = \mathbf{J}_i^{(k)} \,\text{diag}\left(\sqrt{\mathbf{g}_i^{(k)}}\right). \tag{9}$$

Column scaling assigns each anchor a distinct role in rotation, translation, and size recovery. It complements scalar residual weighting, which adjusts the reliability of a correspondence across all state components together. The residuals and $\boldsymbol{\Lambda}_i$ remain unchanged, while the scaled Jacobian defines the learned local recovery surrogate. Regional supervision and pose losses jointly train these state-specific coefficients. The rotational block uses the camera-basis diagonal approximation of the axial/orthogonal operator (Appendix A.2).

Joint robust refinement solves

$$\left(\sum_i a_i^{(k)} \widetilde{\mathbf{J}}_i^{(k)\top} \boldsymbol{\Lambda}_i \widetilde{\mathbf{J}}_i^{(k)} + \mu^{(k)} \mathbf{I}_9\right) \boldsymbol{\delta}^{(k)} = -\sum_i a_i^{(k)} \widetilde{\mathbf{J}}_i^{(k)\top} \boldsymbol{\Lambda}_i \mathbf{r}_i^{(k)}. \tag{10}$$

The $\mu^{(k)}\mathbf{I}_9$ term stabilizes the linear solve in the chosen local coordinates. Rotation, translation, and log-size increments use radians, meters, and dimensionless units, respectively. Bounded increments apply the solution while controlling the size of each refinement step. Appendix A.2 gives the Kabsch solution, exact surrogate, refinement rules and bounds, symmetry handling, and fixed-branch differentiability conditions.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and protocols. We evaluate InfluPose on REAL275 (Wang et al., 2019) and HouseCat6D (Jung et al., 2024). REAL275 contains six categories, 4.3K training images, and 2.75K test images from real scenes. The evaluation contains 15,576 detector inputs. HouseCat6D is a real-world benchmark with ten categories, approximately 20K training images from 34 scenes and 3K test images from five scenes.

Controlled ablations share the CAMERA+Real training mixture, 15,576-input REAL275 evaluation, and frozen estimator for deletion targets. Each changes the named training or recovery package. Ground-truth mug-handle visibility is used for symmetry-aware supervision during training. Test-time predictions do not use this annotation, symmetry-aware evaluation uses the official benchmark annotations (Wang et al., 2019).

Table 1 reports values to one decimal place. Prose differences use unrounded measurements. Method-comparison tables mark the best values in bold and the second-best with underlines. Ablation tables mark only the best values in bold. In Table 1, highlighting applies to accuracy metrics. Dashes denote unavailable comparable measurements.

Metrics and implementation. We report mean average precision (mAP, %) for pose and 3D bounding boxes. AP integrates a category's interpolated precision–recall curve in descending confidence order, and mAP averages category APs equally. IoU_x thresholds the benchmark boxes' intersection-to-union volume ratio at $x/100$. The n°/m cm criterion requires both symmetry-aware geodesic rotation error and Euclidean translation error to meet the stated thresholds. The REAL275 pose protocol first retains matches at 3D IoU ≥ 0.1. On REAL275, we use instance masks predicted by Mask R-CNN. On HouseCat6D, we follow the depth-instance-mask protocol used for AG-Pose-D (Ren et al., 2026). Inputs comprise 1,024 points, 16-neighbor normals, and 192 anchors. The complete estimator is frozen after 36 epochs. The partial network trains for 50 epochs with differentiable geometric recovery. Appendices A.1 and A.3.3 detail architecture, optimization, symmetry, and timing. All quantitative baseline rows use point-cloud or depth-only inference. Regional supervision uses $M = 8$, $\rho = 0.125$, and $\beta = 0.70$ (Appendix A.1.2).

4.2 BENCHMARK RESULTS

REAL275. In Table 1, Full obtains 57.73% at 5°/2 cm, the highest among the displayed methods. Relative to ComPose-D, the two 5° metrics improve by 2.13/1.39 points, while the two 10° metrics are lower by 2.53/3.44 points. Appendix Table 7 compares Full with the GT-complete-geometry oracle reference.

Inference speed. The same configuration reaches **144.5 FPS** on an RTX 4090 with batch size one, taking 6.920 ms per object for feature construction, network prediction, Kabsch initialization, and robust refinement. The complete-cloud estimator and regional probes are used only during training. FPS is $1000/L$ for mean latency L in milliseconds. Timing excludes detection, data loading, and transfer (Appendix A.3.3). Table 1 reports FPS where comparable measurements are available.

HouseCat6D. Full leads the displayed methods at IoU_{25}/IoU_{50} (90.4%/65.8%) and the two 5° thresholds (12.5%/15.1%), exceeding ComPose-D by 8.8/0.7 box-mAP and 0.7/2.4 strict-pose points (Table 2). Full ranks second at the two 10° thresholds (28.6%/34.1%), 6.2/4.8 points below ComPose-D.

The benchmark comparisons evaluate the complete method. To isolate the contribution of influence supervision, we next use matched REAL275 controls. Appendix A.3 reports per-category results.

4.3 EVALUATING REGIONAL SUPERVISION FOR RECOVERY

Influence supervision and recovery. We compare direct regional supervision with training the influence coefficients through the endpoint objective. The *No direct supervision* control in Table 3 retains the four-channel head, parameter count, initialization, and refinement path, but disables the

Table 1: Point-cloud pose estimation on REAL275 (mAP, %, and FPS). AG-Pose-D† denotes the reproduction by Ren et al. (2026).

Method	IoU		Pose				FPS ↑
	50 ↑	75 ↑	5°/2 cm ↑	5°/5 cm ↑	10°/2 cm ↑	10°/5 cm ↑	
FS-Net (Chen et al., 2021)	81.1	63.5	19.9	33.9	–	69.1	71.0
GPV-Pose (Di et al., 2022)	83.0	64.4	32.0	42.9	55.0	73.3	20.6
SAR-Net (Lin et al., 2022)	79.3	62.4	31.6	42.3	50.3	68.3	39.9
SSP-Pose (Zhang et al., 2022b)	82.3	66.3	34.7	44.6	–	77.8	–
RBP-Pose (Zhang et al., 2022a)	–	67.8	38.2	48.1	63.1	79.2	24.7
HS-Pose (Zheng et al., 2023)	82.1	74.7	46.5	55.2	68.6	82.7	90.1
DR-Pose (Zhou et al., 2023)	78.9	68.2	41.7	46.0	67.7	76.3	25.0
Query6DoF (Wang et al., 2023)	82.5	76.1	49.0	58.9	68.7	83.0	87.2
GenPose (Zhang et al., 2023)	–	–	52.1	60.9	72.4	84.0	17.1
HRC-Pose (Li et al., 2025)	**83.4**	**77.8**	49.8	58.6	72.5	**85.4**	–
AG-Pose-D† (Ren et al., 2026)	83.2	75.6	48.8	58.8	68.5	80.8	–
ComPose-D (Ren et al., 2026)	82.1	77.0	55.6	61.3	**77.8**	85.0	–
InfluPose (Full)	82.1	76.6	**57.7**	**62.7**	75.3	81.6	144.5

Table 2: Point-cloud pose estimation on HouseCat6D (mAP, %). AG-Pose-D† denotes the reproduction by Ren et al. (2026).

Method	IoU		Pose			
	25 ↑	50 ↑	5°/2 cm ↑	5°/5 cm ↑	10°/2 cm ↑	10°/5 cm ↑
FS-Net (Chen et al., 2021)	74.9	48.0	3.3	4.2	17.1	21.6
GPV-Pose (Di et al., 2022)	74.9	50.7	3.5	4.6	17.8	22.7
AG-Pose-D† (Ren et al., 2026)	81.4	59.9	9.7	10.6	25.9	29.7
ComPose-D (Ren et al., 2026)	81.6	65.1	11.8	12.7	**34.8**	**38.9**
InfluPose (Full)	**90.4**	**65.8**	**12.5**	**15.1**	28.6	34.1

Table 3: Influence-supervision and recovery ablations on REAL275 (mAP, %).

(a) Influence supervision

Setting	IoU_{75} ↑	5°/2 cm ↑
No learned influence	74.82	52.70
No direct supervision	75.53	54.90
Proxy only	74.49	47.56
Deletion only	75.44	51.69
Shuffled deletion	75.60	52.13
Full	**76.58**	**57.73**

(b) Recovery path

Setting	IoU_{75} ↑	5°/2 cm ↑
Trained Kabsch endpoint	68.55	50.35
Uniform influence	73.84	50.12
Shared influence	74.60	55.65
Full directional operator	75.76	53.56
Stop-gradient	63.08	21.32
Full	**76.58**	**57.73**

direct influence losses (Appendix A.1.2). Its influence coefficients still learn through the endpoint objective. Across three training seeds (Table 4), this control reaches $54.79 \pm 0.34\%$ at 5°/2 cm, compared with $57.50 \pm 0.24\%$ for Full (mean ± sample standard deviation). Full gains 2.82, 2.83, and 2.45 points, averaging 2.70 points over endpoint training alone. All controls retain correspondence, inlier, precision, and initial-size prediction.

Proxy only and Deletion only reach 47.56% and 51.69% strict AP, both below the 54.90% No direct supervision control. Combining the two target sources yields higher strict AP than either source alone. Proxy only and Shuffled deletion reach 74.49% and 75.60% at IoU_{75}, respectively, below Full's 76.58%. Combined targets give the highest IoU_{75} and strict AP among these supervision controls.

Panel (b) tests how influence enters recovery. Trained Kabsch and uniform influence coefficients reach about 50%. Shared influence reaches 55.65%, leaving a 2.07-point gap to the four-channel model. The full directional operator reaches 53.56%, below the diagonal configuration (Appendix A.3.1). Removing gradients through the geometric solver lowers strict AP to 21.32%.

Results across training seeds. Table 4 also reports no learned influence ($52.46 \pm 0.60\%$) and shared influence ($55.44 \pm 0.41\%$) across seeds 260907, 2027, and 1337. Full exceeds every control in all three runs. The no-influence, no-direct-supervision, and shared-influence controls assess the combined design, the regional labels, and the component-specific channels, respectively.

Table 4: Strict-pose mAP across three training seeds on REAL275 (%).

Setting Training seed	Run 1 260907	Run 2 2027	Run 3 1337	Mean ± std
No learned influence	52.70	51.78	52.91	52.46 ± 0.60
No direct supervision	54.90	54.41	55.07	54.79 ± 0.34
Shared influence	55.65	55.69	54.97	55.44 ± 0.41
Deletion only	51.69	51.71	51.35	51.58 ± 0.20
Full	**57.73**	**57.24**	**57.52**	$\mathbf{57.50 \pm 0.24}$

Contribution of individual influence channels. Table 5 removes each channel from both supervision and recovery weighting. Every removal lowers strict-pose AP by 3.96–7.06 points. Removing size influence increases median size error from 7.62% to 7.99% and lowers strict AP to 50.67%.

Table 5: Influence-channel ablations on REAL275. State-error columns report medians defined in Equation 32.

Setting	Rot. (deg) ↓	Trans. (cm) ↓	Size (%) ↓	5°/2 cm ↑
Full	**3.27**	**0.96**	**7.62**	**57.73**
Without $R_\perp$	3.42	1.01	8.59	53.77
Without $R_\parallel$	3.39	1.03	8.83	52.37
Without translation influence	3.35	1.01	9.32	52.17
Without size influence	3.50	1.00	7.99	50.67

Learning the targets from partial geometry. For the fixed Full model, we compute target–prediction Spearman correlation over the 192 anchors within each observation. We average the correlations over active observations per channel, then average the four channels equally. The resulting agreement is 0.481 on 360 class-balanced instances, with 180 active examples for axial rotation. All channelwise intervals are positive (Appendix Table 14). This measures how well partial geometry predicts the training targets.

Regional prediction response. We next test whether the predicted rankings identify regions that affect the fitted model's output. Figure 3 uses 360 correct-class REAL275 instances matched at bbox IoU ≥ 0.5 (60 per category). Each cloud has ten nearly equal-size regions along its first PCA axis, ranked by mean anchor scores. At $q \in \mathcal{Q} = \{0.1, 0.2, 0.3\}$, we delete the top one, two, or three regions, or equally many random regions, and refill to 1,024 points by duplication using eight paired offsets. Relative to the undeleted prediction, responses are $\Delta_R = d_{\mathrm{q}}(\mathbf{R}', \mathbf{R}_0)/5^\circ$, $\Delta_t = \|\mathbf{t}' - \mathbf{t}_0\|_2/(2\,\mathrm{cm})$, and $\Delta_s = \|\log \mathbf{s}' - \log \mathbf{s}_0\|_2/0.10$, where d_{q} respects benchmark symmetry. For ranking a and output b,

$$\mathrm{AUC}_{a\to b} = \frac{1}{|\mathcal{Q}|} \sum_{q\in\mathcal{Q}} \left\langle \Delta_b^{\mathrm{top}(a,q)} - \Delta_b^{\mathrm{random}(q)} \right\rangle_{o,r}. \tag{11}$$

The brackets average observations o and paired refills r equally. Response AUC is this discrete budget mean, not ROC-AUC, and can lie outside $[0, 1]$. Positive values mean greater response than

random deletion. Δ response AUC subtracts a comparator's AUC from learned influence's. The six paired contrasts compare three states against inlier probability and precision. All use 50,000 shared bootstrap draws of 17 physical objects and a conservative Bonferroni divisor of 12 for 95% simultaneous coverage.

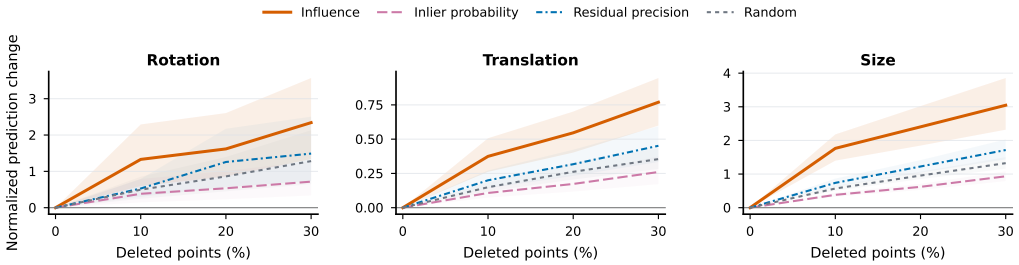

Figure 3: Prediction changes under regional deletion on 360 correctly classified REAL275 instances, bbox-matched at IoU ≥ 0.5 (60 per category). Each change is measured from the undeleted prediction. Curves average observations equally over eight shared refill offsets, matching Table 6. Bands are pointwise 95% object-cluster intervals. Paired contrasts use the joint correction.

Table 6 reports influence–response Spearman ρ_S. We correlate average-tie ranks of region scores and refill-averaged single-region responses within each observation, exclude constant fields, and average the correlations across observations. Rank intervals are pointwise 95%, and † marks an interval containing zero. The paired response-AUC contrasts show that influence-selected regions produce greater responses than inlier-selected regions for all three states, and than precision-selected regions for translation and size. The rotation–precision interval includes zero.

Table 6: Regional response association and influence-minus-reliability contrasts on REAL275.

Field	Influence–response ρ_S	Δ response AUC (Inf.–Inlier)	Δ response AUC (Inf.–Prec.)
Rotation	0.378 [0.325, 0.433]	1.218 [0.522, 2.373]	0.671 [-0.597, 2.007]†
Translation	0.311 [0.246, 0.377]	0.383 [0.236, 0.537]	0.241 [0.110, 0.390]
Size	0.423 [0.371, 0.467]	1.759 [1.093, 2.528]	1.179 [0.586, 1.882]

Estimation damage. On the same 360 observations, deleting high-influence regions at a 20% budget increases ground-truth error more than matched random deletion in all three states (Appendix Table 15). The pointwise 95% object-cluster intervals are positive. Rotation and translation correspond to 3.04° and 0.31 cm of additional mean error. Thus the learned rankings identify regions whose removal degrades the fixed estimator's accuracy.

On REAL275, direct regional supervision improves recovery across training seeds, and the partial-cloud network predicts the regional target rankings. In the fixed estimator, deleting high-ranked regions produces larger prediction changes and greater ground-truth error than matched random deletion.

5 Conclusion

We presented InfluPose, a category-level pose estimator that learns regional geometric influence from complete training geometry and predicts it from partial point clouds. Deletion-and-refill responses and a dense proxy supervise four state channels that complement correspondence reliability in weighted initialization and joint refinement. On unseen instances from REAL275 and HouseCat6D, InfluPose obtains the highest strict-pose mAP among the compared point-cloud methods. Matched REAL275 training controls show gains from direct regional supervision, and deletion tests link the predicted rankings to changes in the fixed estimator's output and accuracy.

AI USE STATEMENT

Generative AI was used for assistance with manuscript writing, language refinement, consistency checks, and discussion of experimental design considerations. All research ideas, experimental protocols, implementations, experiments, mathematical derivations, analyses, and conclusions were developed and verified by the authors. The authors reviewed all AI-assisted suggestions and take full responsibility for the final manuscript and reported results.

REPRODUCIBILITY STATEMENT

The anonymous project page for the paper and code is `https://iclr2027-anonymous.github.io/InfluPose`. Section 4 specifies dataset coverage, inputs, and metrics. Appendix A.1 gives architecture, training, symmetry, and seed provenance. Appendix A.2 specifies recovery properties and branch assumptions. Appendix A.3.1 reports additional comparisons, and Appendix A.3.3 specifies runtime boundaries.

REFERENCES

Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6D object pose and size estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11970–11979, 2020. doi: 10.1109/CVPR42600.2020.01199.

Kai Chen and Qi Dou. SGPA: Structure-guided prior adaptation for category-level 6D object pose estimation. In *IEEE/CVF International Conference on Computer Vision*, pp. 2773–2782, 2021.

Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Aleš Leonardis. FS-Net: Fast shape-based network for category-level 6D object pose estimation with decoupled rotation mechanism. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1581–1590, 2021.

Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. GPV-Pose: Category-level object pose estimation via geometry-guided point-wise voting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6771–6781, 2022.

Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, 13 (111):3475–3506, 2012. URL `https://jmlr.org/papers/v13/drineas12a.html`.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *IEEE International Conference on Computer Vision*, pp. 3429–3437, 2017. doi: 10.1109/ICCV.2017.371.

David C. Hoaglin and Roy E. Welsch. The hat matrix in regression and ANOVA. *The American Statistician*, 32(1):17–22, February 1978. ISSN 1537-2731. doi: 10.1080/00031305.1978.10479237. URL `http://dx.doi.org/10.1080/00031305.1978.10479237`.

HyunJun Jung, Shun-Cheng Wu, Patrick Ruhkamp, Guangyao Zhai, Hannah Schieber, Giulia Rizzoli, Pengyuan Wang, Hongcheng Zhao, Lorenzo Garattoni, Sven Meier, Daniel Roth, Nassir Navab, and Benjamin Busam. HouseCat6D - a large-scale multi-modal category level 6D object perception dataset with household objects in realistic scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22498–22508, 2024. doi: 10.1109/CVPR52733.2024.02123.

W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976. doi: 10.1107/S0567739476001873.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894. PMLR, 2017. URL `https://proceedings.mlr.press/v70/koh17a.html`.

Seunghyun Lee and Tae-Kyun Kim. Joint learning of pose regression and denoising diffusion with score scaling sampling for category-level 6D pose estimation. In *IEEE/CVF International Conference on Computer Vision*, 2025. doi: 10.1109/ICCV51701.2025.00545. URL https://arxiv.org/abs/2510.04125.

Zhujun Li, Shuo Zhang, and Ioannis Stamos. Learning point cloud representations with pose continuity for depth-based category-level 6D object pose estimation. In *IEEE/CVF International Conference on Computer Vision Workshops*, 2025. doi: 10.1109/ICCVW69036.2025.00254. URL https://arxiv.org/abs/2508.14358.

Haitao Lin, Zichang Liu, Chilam Cheang, Yanwei Fu, Guodong Guo, and Xiangyang Xue. SAR-Net: Shape alignment and recovery network for category-level 6D object pose and size estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717, 2022.

Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. DualPoseNet: Category-level 6D object pose and size estimation using dual pose network with refined learning of pose consistency. In *IEEE/CVF International Conference on Computer Vision*, pp. 3540–3549, 2021.

Xiao Lin, Wenfei Yang, Yuan Gao, and Tianzhu Zhang. Instance-adaptive and geometric-aware keypoint learning for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21040–21049, 2024. doi: 10.1109/CVPR52733.2024.01988.

Jian Liu, Wei Sun, Chongpei Liu, Hui Yang, Xing Zhang, and Ajmal Mian. MH6D: Multi-hypothesis consistency learning for category-level 6-D object pose estimation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):4820–4833, 2025. doi: 10.1109/TNNLS.2024.3360712.

Jian Liu, Wei Sun, Hui Yang, Zhiwen Zeng, Chongpei Liu, Jin Zheng, Xingyu Liu, Hossein Rahmani, Nicu Sebe, and Ajmal Mian. Deep learning-based object pose estimation: A comprehensive survey. *International Journal of Computer Vision*, 134(2):81, 2026. doi: 10.1007/s11263-025-02646-6.

Jierui Liu, Zhiqiang Cao, Ying-Jhe Tang, Xilong Liu, and Min Tan. Category-level 6D object pose estimation with structure encoder and reasoning attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(10):6728–6740, 2022. doi: 10.1109/TCSVT.2022.3169144.

Wenzhe Ouyang, Qi Ye, Jinghua Wang, Zenglin Xu, and Jiming Chen. RFMPose: Generative category-level object pose estimation via riemannian flow matching. In *Advances in Neural Information Processing Systems*, volume 38, pp. 75284–75303, 2025. doi: 10.52202/085713-2267.

Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5099–5108, 2017.

Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11143–11152, 2022.

Huan Ren, Wenfei Yang, Xiang Liu, Shifeng Zhang, and Tianzhu Zhang. Learning shape-independent transformation via spherical representations for category-level object pose estimation. In *International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2503.13926.

Huan Ren, Wenfei Yang, Shifeng Zhang, and Tianzhu Zhang. Rethinking correspondence-based category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1170–1179, 2025b. doi: 10.1109/CVPR52734.2025.00117.

Huan Ren, Yihan Chen, Chuxin Wang, Nailong Liu, Wenfei Yang, and Tianzhu Zhang. ComPose: A unified completion-pose framework for robust category-level object pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14315–14324, 2026.

Yingbo Tang, Zhiqiang Cao, Peiyu Guan, Xurong Gong, Mengyao Wang, and Junzhi Yu. Diff-COPE: Diffusion-based category-level 6D object pose estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(3):3556–3572, 2026. doi: 10.1109/TCSVT.2025.3613733.

Meng Tian, Marcelo H. Ang Jr, and Gim Hee Lee. Shape prior deformation for categorical 6D object pose and size estimation. In *European Conference on Computer Vision*, pp. 530–546, 2020. doi: 10.1007/978-3-030-58589-1_32.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991. doi: 10.1109/34.88573.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6D object pose and size estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2637–2646, 2019.

Ruiqi Wang, Xinggang Wang, Te Li, Rong Yang, Minhong Wan, and Wenyu Liu. Query6DoF: Learning sparse queries as implicit shape prior for category-level 6DoF pose estimation. In *IEEE/CVF International Conference on Computer Vision*, pp. 14009–14018, 2023.

Yang You, Wenhao He, Jin Liu, Hongkai Xiong, Weiming Wang, and Cewu Lu. CPPF++: Uncertainty-aware sim2real object pose estimation by vote aggregation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9239–9254, 2024. doi: 10.1109/TPAMI.2024.3419038.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833, 2014. doi: 10.1007/978-3-319-10590-1_53.

Jiyao Zhang, Mingdong Wu, and Hao Dong. GenPose: Generative category-level object pose estimation via diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 54627–54644, 2023.

Ruida Zhang, Yan Di, Zhiqiang Lou, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. RBP-Pose: Residual bounding box projection for category-level pose estimation. In *European Conference on Computer Vision*, pp. 655–672, 2022a. doi: 10.1007/978-3-031-19769-7_38.

Ruida Zhang, Yan Di, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. SSP-Pose: Symmetry-aware shape prior deformation for direct category-level object pose estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 7452–7459, 2022b.

Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip H. S. Torr, and Vladlen Koltun. Point transformer. In *IEEE/CVF International Conference on Computer Vision*, pp. 16239–16248, 2021. doi: 10.1109/ICCV48922.2021.01595.

Linfang Zheng, Chen Wang, Yinghan Sun, Esha Dasgupta, Hua Chen, Aleš Leonardis, Wei Zhang, and Hyung Jin Chang. HS-Pose: Hybrid scope feature extraction for category-level object pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17163–17173, 2023.

Lei Zhou, Zhiyang Liu, Runze Gan, Haozhe Wang, and Marcelo H. Ang Jr. DR-Pose: A two-stage deformation-and-registration pipeline for category-level 6D object pose estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1192–1199, 2023. doi: 10.1109/IROS55552.2023.10341552.

A APPENDIX

The appendix provides implementation details, properties of geometric recovery, and additional experiments on recovery accuracy, target transfer, and regional prediction response.

A.1 IMPLEMENTATION DETAILS

A.1.1 FROZEN COMPLETE-GEOMETRY RESPONSE ESTIMATOR

The first training stage constructs the complete-geometry response estimator G used in Section 3.2. Given a complete camera-frame cloud, its normals, and the paired partial-cloud centroid as a translation anchor, it predicts rotation, translation, and metric size through one shared latent representation. After training, its parameters are fixed. Changes in its predictions after regional deletion provide the influence targets.

Inputs and encoders. The inputs are 1,024 complete points, one normal per point, and the centroid of the paired partial observation. The encoders center and scale the complete cloud with a shared normalization rule while preserving camera-frame directions. The geometric branch follows the hierarchical point-set organization of PointNet++ (Qi et al., 2017), using farthest-point sampling and radius neighborhoods at three scales. Its local blocks add learned geometric relation biases, related to position-aware point attention in Point Transformer (Zhao et al., 2021). The second branch retains 256 sampled tokens and applies a standard Transformer encoder (Vaswani et al., 2017) to signed point–normal and directional descriptors. The two pooled features are combined with the complete-to-partial centroid offset before decoding. Intact and perturbed clouds use the same preprocessing rule.

Shared latent and pose-and-size decoder. The fused features are mapped to a shared posterior mean $\boldsymbol{\mu} \in (-1, 1)^{128}$. A common decoder and three linear state heads jointly predict a six-dimensional rotation representation, a translation residual, and metric size through

$$\mathbf{R} = \mathrm{Rot6D}(\mathbf{o}_R), \qquad \mathbf{t} = \mathbf{c}_P + 0.30 \tanh(\mathbf{o}_t), \qquad \mathbf{s} = 0.01 + 0.79\sigma(\mathbf{o}_s), \tag{12}$$

with translation and size in meters. Thus each translation component lies within 0.30 m of its anchor and each size component lies in $(0.01, 0.80)$ m. All state components share the same latent and decoder trunk. Category labels and symmetry metadata are used by training losses and evaluation, rather than by the encoders or prediction heads.

Measuring deletion responses. The estimator is trained with pose, size, box, reconstruction, view-consistency, and latent-regularization objectives. To measure deletion responses, the decoder uses the posterior mean $\mathbf{z} = \boldsymbol{\mu}$ with fixed parameters and no random sampling or dropout. Thus the measured change comes from the input perturbation. The partial-cloud anchor $\mathbf{c}_P$ stays fixed, while each intact or perturbed complete cloud is centered and scaled using the same preprocessing rule.

Estimator optimization. The complete-geometry estimator is trained with AdamW and a cosine learning-rate schedule for 36 epochs. Weight decay and gradient clipping are used throughout. Its parameters remain fixed when generating influence targets.

We check response consistency across deterministic refills and association with increases in ground-truth estimation error. The pooled refill and damage correlations are 0.920 and 0.638 on REAL275, and 0.815 and 0.330 on HouseCat6D, respectively. The HouseCat6D evaluation uses a smaller cohort, so its correlations are interpreted descriptively.

Dataset-specific complete geometry. NOCS and HouseCat6D use the annotated complete-geometry construction in Section 3.2, with the same geometric augmentation for paired partial and complete clouds. On HouseCat6D, bottle, can, and glass have continuous local-y symmetry, while tube is asymmetric. Complete models and their annotations are used only during training.

A.1.2 PARTIAL-CLOUD PREDICTION AND TRAINING

Influence targets. In Equation 1, $\tau_t = 0.02\,\mathrm{m}$ and $\tau_s = 0.10$. For metric canonical coordinates $\mathbf{g}_i = \mathbf{s}^* \odot \mathbf{q}_i^*$, $r_i = \sqrt{g_{ix}^2 + g_{iz}^2}$, and $h_i = |g_{iy}|$, the dense proxy is

$$\tilde{\boldsymbol{\eta}}_i^{\mathrm{proxy}} = \left[\max\left\{\sqrt{h_i^2 + r_i^2/4}, 10^{-4}\right\},\ \max\{r_i, 10^{-4}\},\ 1,\ \max\{\|\mathbf{q}_i^*\|_2, 10^{-4}\}\right]. \tag{13}$$

The proxy uses ground-truth canonical correspondences and metric size only during training. It is normalized per channel over visible anchors and clipped to $[0, 10]$ before blending. Every two batches, the frozen estimator probes half the examples in the selected batch. The targets follow Equation 3.

Geometric interpretation of the proxy. For the point derivative $\mathbf{n} \times \mathbf{g}$ under rotation about a unit axis $\mathbf{n}$,

$$\|\mathbf{e}_y \times \mathbf{g}\|_2 = r, \qquad \mathbb{E}_\phi \|(\cos\phi, 0, \sin\phi) \times \mathbf{g}\|_2^2 = h^2 + \tfrac{1}{2} r^2, \tag{14}$$

where ϕ is uniform on $[0, 2\pi)$. Before the lower bound, the squared orthogonal proxy $h^2 + r^2/4$ lies between one half and one times this mean square. Translation has derivative $\mathbf{I}_3$ at every point. For additive size,

$$\frac{\partial[\mathbf{R}(\mathbf{s} \odot \mathbf{q})]}{\partial \mathbf{s}} = \mathbf{R}\,\mathrm{diag}(\mathbf{q}), \qquad \|\mathbf{R}\,\mathrm{diag}(\mathbf{q})\|_F = \|\mathbf{q}\|_2. \tag{15}$$

The log-size Jacobian instead contains $\mathbf{s} \odot \mathbf{q}$. The size proxy therefore acts as a canonical geometric prior. Central differences on 128 points and 64 horizontal axes agree with these derivatives to 1.4×10^{-10} in double precision.

Partial observation network. The network receives 1,024 points and selects $K = 192$ anchors by farthest-point sampling in normalized coordinates. A 14-dimensional descriptor combines the normal, three local covariance ratios, camera ray, signed and absolute incidence, relative depth, density, and depth spread. The anchor exchange is related to geometry-aware point attention (Zhao et al., 2021) and geometric relation encoding (Qin et al., 2022). Four relation attention blocks encode partial camera observations through pairwise biases for point distance, normal agreement, ray agreement, and relative depth.

Prediction-head parameterization and ranges. Linear heads map each fused anchor token to one inlier, three precision, and four influence logits, activated as

$$\hat{p}_i = \sigma(o_i^p), \qquad \hat{\lambda}_{ia} = \mathrm{softplus}(o_{ia}^\lambda) + 0.05, \qquad \hat{\eta}_i^d = \mathrm{softplus}(o_i^{\eta,d}) + 0.02. \tag{16}$$

Here σ is sigmoid and $\mathrm{softplus}(x) = \log(1 + e^x)$. Finite logits give $\hat{p}_i \in (0, 1)$, $\hat{\lambda}_{ia} \in (0.05, \infty)$, and $\hat{\eta}_i^d \in (0.02, \infty)$, with finite endpoints attainable by floating-point saturation. Kabsch uses Equation 6. Refinement clamps are specified in Section 3.3. Across all 15,576 REAL275 detector inputs, the largest normalized anchor weight is 0.018. The effective anchor count $1/\sum_i (\tilde{w}_i^{(0)})^2$ has median 175.3 of 192 and minimum 102.8.

A linear head maps the global token to $\mathbf{o}_s \in \mathbb{R}^3$, giving

$$\hat{\mathbf{s}}^{(0)} = 0.01 + 0.79\sigma(\mathbf{o}_s), \qquad \boldsymbol{\ell}^{(0)} = \log\max\{\hat{\mathbf{s}}^{(0)}, 10^{-4}\} = \log\hat{\mathbf{s}}^{(0)}. \tag{17}$$

Operations are elementwise. Size lies in $(0.01, 0.80)$ meters, with endpoints attainable by saturation, so the logarithm floor is inactive. It initializes metric size for Kabsch and log-size for refinement.

Residual precision is trained with a diagonal Gaussian negative log likelihood,

$$\mathcal{L}_{\mathrm{prec}} = \left\langle \frac{1}{6} \sum_{a \in \{x,y,z\}} \left[\bar{\lambda}_{ia} (r_{ia}^*)^2 - \log \bar{\lambda}_{ia}\right] \right\rangle_{i \in \mathcal{V}_{\mathrm{in}}}, \quad \bar{\lambda}_{ia} = \mathrm{clip}(\hat{\lambda}_{ia}, 0.05, 50), \tag{18}$$

where $\mathbf{r}_i^* = \mathbf{R}^*(\mathbf{s}^* \odot \hat{\mathbf{q}}_i) + \mathbf{t}^* - \mathbf{x}_i$. The residuals in this loss and robust refinement are in meters, with no additional unit normalization. For a single-axis residual below $1/\sqrt{50} \simeq 0.141\,\mathrm{m}$, this term alone favors the upper precision clamp. Precision and Cauchy weights control residual contributions to recovery. They do not estimate calibrated uncertainty.

Optimization. The partial-cloud predictor is trained for 50 epochs with AdamW, cosine decay, mixed precision, global gradient clipping, and an exponential moving average (EMA) for evaluation. During Stage 2 training, we randomly replace some partial-cloud points with samples drawn uniformly from a surrounding box before computing normals and observation features. Losses cover correspondence, inlier, precision, influence, quotient-rotation, translation, log-size, box, view-consistency, and non-regression terms. Ablations and additional seeds share the schedule, loss coefficients, augmentation, and default REAL275 evaluation.

Additional supervision and directional controls. *No direct supervision* retains the four-channel head, parameter count, weighted Kabsch, refinement coefficients, target construction, and other losses, while disabling direct influence regression, spatial KL, and mass losses. The head still learns through pose and size losses. *Full directional operator* keeps the direct losses and replaces diagonal rotational weighting with the full axial/orthogonal operator (Appendix A.3.1). Both follow the shared training and evaluation protocol on all 15,576 detector inputs.

Training configuration and reporting. The default settings are $M = 8$, $\rho = 0.125$, and $\beta = 0.70$. These specify eight probe centers, deletion of 128 of the 1,024 complete-cloud points per probe, and response/proxy weights of 0.70/0.30. Table 13 reports the sensitivity study.

Training stability. On the same augmented input, the non-regression loss penalizes deterioration beyond a tolerance in rotation error, translation error, and box IoU relative to a frozen copy of the predictor.

Symmetry metadata. During training, ground-truth mug-handle visibility determines the symmetry status of mug instances for symmetry-aware supervision. At test time on REAL275, rotation-representative selection uses detector class labels without ground-truth handle visibility. Symmetry-aware evaluation uses the official mug-handle visibility annotations Wang et al. (2019). Symmetry selection follows geometric recovery and preserves the camera-frame local-y axis, translation, size, and predicted correspondence and influence fields. Appendix A.2 gives the quotient-class statement.

A.2 Properties of the Targets and Solver

We derive properties of the perturbation targets, Kabsch initialization, and robust refinement under explicit assumptions.

A.2.1 Fixed cardinality perturbation and target construction

Deletion and refill. Let $N_r = \max\{1, \mathrm{round}_{\mathrm{py}}(\rho N_c)\}$, with ties-to-even rounding and $1 \le N_r < N_c$. We retain the farthest $N_c - N_r$ points from each region center and append N_r point–normal pairs from a circularly shifted prefix of the retained set. Removed source indices never reappear, although coincident tuples at other indices may remain. This preserves N_c inputs. The reported $N_c = 1024$ and removal fractions give $N_r \in \{64, 128, 256\}$. The estimator and alignment center remain fixed as in Section 3.2.

The $\beta = 1$ and *Deletion only* controls. In Equation 3, $\beta = 1$ uses deletion responses for probed samples and retains the proxy for unprobed samples. *Deletion only* instead assigns zero targets to unprobed samples and excludes them from the direct influence loss. They still contribute through pose, box, and auxiliary losses. The two settings therefore differ on unprobed samples under the schedule in Appendix A.1.

Nearest-region transfer. For distinct centers $\mathbf{c}_1, \ldots, \mathbf{c}_M$, a unique nearest assignment $m(\mathbf{x})$ has margin

$$\Delta(\mathbf{x}) = \min_{j \neq m(\mathbf{x})} \big(\|\mathbf{x} - \mathbf{c}_j\|_2 - \|\mathbf{x} - \mathbf{c}_{m(\mathbf{x})}\|_2 \big) > 0. \tag{19}$$

The assignment is unchanged for $\|\mathbf{x}' - \mathbf{x}\|_2 < \Delta(\mathbf{x})/2$ by the triangle inequality. At Voronoi boundaries, however, transferred values can jump.

Normalization and pooling. For finite nonnegative responses $a_1, \dots, a_M$ with mean $\bar{a} = M^{-1} \sum_m a_m > 10^{-6}$, the pre-clipping values $b_m = a_m / \bar{a}$ satisfy

$$\frac{1}{M} \sum_m b_m = 1, \qquad 0 \le b_m = \frac{M a_m}{\sum_l a_l} \le M. \tag{20}$$

The activity threshold follows Equation 2. With $M = 8$, the upper clip at 10 is inactive. For $M > 10$, it can reduce the mean. Proxy clipping, nearest-region transfer, blending, and prediction also need not preserve unit means. The equal channel coefficients in Equation 6 combine standardized targets without physical calibration across channels.

A.2.2 Weighted Kabsch initialization

For finite points and positive normalized weights $\tilde{w}_i^{(0)}$, set $\mathbf{y}_i^{(0)} = \hat{\mathbf{s}}^{(0)} \odot \hat{\mathbf{q}}_i$, $\bar{\mathbf{y}} = \sum_i \tilde{w}_i^{(0)} \mathbf{y}_i^{(0)}$, $\bar{\mathbf{x}} = \sum_i \tilde{w}_i^{(0)} \mathbf{x}_i$, and

$$\mathbf{A} = \sum_i \tilde{w}_i^{(0)} (\mathbf{y}_i^{(0)} - \bar{\mathbf{y}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top, \qquad \sigma_1 \ge \sigma_2 \ge \sigma_3 \ge 0. \tag{21}$$

The standard reflection-corrected solution (Kabsch, 1976; Umeyama, 1991) is

$$\mathbf{R}^{(0)} = \mathbf{V} \operatorname{diag}(1, 1, \det(\mathbf{V}\mathbf{U}^\top)) \mathbf{U}^\top, \qquad \mathbf{t}^{(0)} = \bar{\mathbf{x}} - \mathbf{R}^{(0)} \bar{\mathbf{y}}. \tag{22}$$

It minimizes Equation 7 in exact arithmetic. Degenerate covariance spectra can make the solution non-unique, and the float32 SVD is sensitive near rank deficiency or repeated singular values.

A.2.3 Analytic Jacobian and influence modulation

Residual derivative and its domain. For finite states and correspondences, $\mathbf{R} \in \mathrm{SO}(3)$, and $\exp(\ell_j) \in (0.01, 0.80)$, differentiating the local perturbation $\mathbf{R}' = \mathrm{Exp}(\boldsymbol{\delta}_R)\mathbf{R}$, $\mathbf{t}' = \mathbf{t} + \boldsymbol{\delta}_t$, and $\boldsymbol{\ell}' = \boldsymbol{\ell} + \boldsymbol{\delta}_\ell$ gives Equation 9. At a size-clamp boundary, the solver retains this interior Jacobian as a local approximation. Influence column scaling then defines the surrogate used below.

Camera-basis rotation weighting. The rotational scaling uses $g_{i,R,j} = (1 - u_j^2)\hat{\eta}_i^{R_\perp} + u_j^2 \hat{\eta}_i^{R_\parallel}$, the diagonal of

$$\hat{\eta}_i^{R_\perp} \mathbf{I}_3 + (\hat{\eta}_i^{R_\parallel} - \hat{\eta}_i^{R_\perp}) \mathbf{u}\mathbf{u}^\top. \tag{23}$$

Each entry interpolates between axial and orthogonal channels in the camera basis. The coordinate-dependent diagonal approximation can mix their contributions, and their spatial support can overlap.

A.2.4 Damped geometric recovery

Bounded increments. For the raw increment $\boldsymbol{\delta}^{(k)} = [\boldsymbol{\delta}_R^{(k)}, \boldsymbol{\delta}_t^{(k)}, \boldsymbol{\delta}_\ell^{(k)}]$ from Equation 10, define

$$\mathcal{C}_\rho(\mathbf{v}) = \begin{cases} \mathbf{v}, & \|\mathbf{v}\|_2 \le \rho, \\ \rho \mathbf{v} / \|\mathbf{v}\|_2, & \|\mathbf{v}\|_2 > \rho, \end{cases} \qquad \mathcal{C}_\ell(\mathbf{v}) = \mathrm{clip}(\mathbf{v}, -\ell_{\max}, \ell_{\max}), \tag{24}$$

where the latter acts componentwise. The limits are $\theta_{\max} = 8^\circ$, $d_{\max} = 0.015\,\mathrm{m}$, and $\ell_{\max} = 0.12$. The axis–angle conversion Φ forms

$$\mathbf{q}(\mathbf{v}) = [\cos(a/2),\, \xi(a)\mathbf{v}], \quad a = \|\mathbf{v}\|_2, \qquad \xi(a) = \begin{cases} \sin(a/2)/a, & a \ge 10^{-6}, \\ 1/2 - a^2/48, & 0 \le a < 10^{-6}, \end{cases} \tag{25}$$

and applies the scale-invariant quaternion-to-matrix formula. Recovery applies the bounded increments as

$$\begin{aligned} \mathbf{R}^{(k+1)} &= \Phi\Big(\gamma^{(k)} \mathcal{C}_{\theta_{\max}}(\boldsymbol{\delta}_R^{(k)})\Big)\, \mathbf{R}^{(k)}, \\ \mathbf{t}^{(k+1)} &= \mathbf{t}^{(k)} + \gamma^{(k)} \mathcal{C}_{d_{\max}}(\boldsymbol{\delta}_t^{(k)}), \\ \boldsymbol{\ell}^{(k+1)} &= \boldsymbol{\ell}^{(k)} + \gamma^{(k)} \mathcal{C}_\ell(\boldsymbol{\delta}_\ell^{(k)}). \end{aligned} \tag{26}$$

The scaled Jacobian and step scale use the refinement clamps in Section 3.3. Kabsch retains the distinct scalar-weight rule in Equation 6.

For K solver anchors, the normal matrix and damping are

$$\mathbf{H}^{(k)} = \sum_{i=1}^{K} a_i^{(k)} \widetilde{\mathbf{J}}_i^{(k)\top} \mathbf{\Lambda}_i \widetilde{\mathbf{J}}_i^{(k)}, \qquad \mu^{(k)} = 10^{-3} + 10^{-4}\,\mathrm{mean}(\mathrm{diag}(\mathbf{H}^{(k)})). \tag{27}$$

The ridge is numerical in the mixed-unit coordinates of Section 3.3. The confidence- and conditioning-dependent step scale is

$$\begin{aligned} c^{(k)} &= \left(\frac{1}{K}\sum_i \hat{p}_i\right)\left(\frac{1}{4K}\sum_{i,d} \hat{\eta}_i^d\right), \\ \gamma^{(k)} &= \sigma\Big(8(c^{(k)} - 0.20)\Big)\,\sigma\Big(2\left[7 - \log_{10}\max(\kappa(\mathbf{H}^{(k)}), 1)\right]\Big), \end{aligned} \tag{28}$$

with eigenvalues floored at 10^{-9} in κ.

Scope of the quadratic subproblem. For finite residuals and effective Jacobians, nonnegative row weights, and positive diagonal precisions, the raw normal-equation solution minimizes

$$Q_k(\boldsymbol{\delta}) = \frac{1}{2}\sum_i a_i^{(k)} \left\| \mathbf{\Lambda}_i^{1/2}(\mathbf{r}_i^{(k)} + \widetilde{\mathbf{J}}_i^{(k)}\boldsymbol{\delta}) \right\|_2^2 + \frac{\mu^{(k)}}{2}\|\boldsymbol{\delta}\|_2^2. \tag{29}$$

In exact arithmetic, $\mathbf{H}^{(k)}$ is positive semidefinite and $\mu^{(k)} \geq 10^{-3}$, giving a unique minimizer. Finite-value safeguards, clipping, and adaptive scaling can change that solution. The scaled Jacobian is a surrogate, so refinement does not guarantee descent of the original nonlinear residual objective.

Conditional step bounds. Assume finite current states and repaired increments, $\mathbf{R}^{(k)} \in \mathrm{SO}(3)$, and a finite step scale $0 \leq \gamma^{(k)} \leq 1$. In exact arithmetic,

$$d_{\mathrm{SO}(3)}(\mathbf{R}^{(k+1)}, \mathbf{R}^{(k)}) \leq 8^\circ, \quad \|\mathbf{t}^{(k+1)} - \mathbf{t}^{(k)}\|_2 \leq 0.015\ \mathrm{m}, \quad \|\boldsymbol{\ell}^{(k+1)} - \boldsymbol{\ell}^{(k)}\|_\infty \leq 0.12. \tag{30}$$

Clipping bounds translation and log-size, and both branches of Φ produce a rotation angle no larger than the clipped input norm. The final finite size output is clamped to $[0.01, 0.80]^3$. These bounds constrain update magnitude, not estimation error or convergence. They require finite values and apply before symmetry representative selection, which can change the ordinary rotation distance afterward.

Continuous symmetry and the axial influence channel. For continuous axial symmetry, the axial target is zeroed and excluded from direct influence loss and target-agreement statistics. Canonical correspondence supervision uses horizontal radial distance, and the final rotation is mapped to a benchmark-equivalent representative. Complete-cloud deletion uses the frozen estimator's raw deterministic representatives, so the orthogonal-rotation response still depends on their choice.

Pose and box losses train the axial prediction. It enters the Kabsch channel average (Equation 6), camera-basis rotational weighting (Equation 9), and adaptive step scale (Equation 28), affecting the observable axis, translation, and size. A positive value does not imply observability of the symmetry-equivalent axial angle. A zero head output is raised to the refinement clamp's 0.02 floor, and zeroing one diagonal gate channel is not a quotient-tangent projection.

Continuous-symmetry representative. For $\mathbf{R} \in \mathrm{SO}(3)$ and $\mathbf{u} = \mathbf{R}\mathbf{e}_y$, the representative map chooses $\mathbf{e} = \mathbf{e}_z$ if $|u_x| > 0.90$ and $\mathbf{e} = \mathbf{e}_x$ otherwise, then forms

$$\mathbf{x}_0 = \mathbf{e} - (\mathbf{e}^\top \mathbf{u})\mathbf{u}, \quad \mathbf{x} = \frac{\mathbf{x}_0}{\|\mathbf{x}_0\|_2}, \quad \mathbf{z} = \frac{\mathbf{x} \times \mathbf{u}}{\|\mathbf{x} \times \mathbf{u}\|_2}, \tag{31}$$

recomputes $\mathbf{x} = (\mathbf{u} \times \mathbf{z})/\|\mathbf{u} \times \mathbf{z}\|_2$, and returns $\mathbf{R}_{\mathrm{can}} = [\mathbf{x}, \mathbf{u}, \mathbf{z}]$. The first normalization is safe in exact arithmetic because $\|\mathbf{x}_0\|_2^2 \geq 1 - 0.90^2$ on the $\mathbf{e}_x$ branch and $> 0.90^2$ on the $\mathbf{e}_z$ branch. The resulting basis is right-handed and orthonormal, and $\mathbf{Q}_y = \mathbf{R}^\top \mathbf{R}_{\mathrm{can}}$ fixes $\mathbf{e}_y$. Consequently $\mathbf{R}_{\mathrm{can}} = \mathbf{R}\mathbf{Q}_y$ preserves the rotation equivalence class in $\mathrm{SO}(3)/\mathrm{SO}(2)_y$. The representative can change nonsmoothly at the reference-axis switch. Its ordinary rotation distance from the pre-gauge representative need not obey the step bound.

A.2.5 DIFFERENTIATION THROUGH THE SOLVER

Pose and box losses backpropagate through Kabsch and robust refinement. For fixed shapes, sampled anchors, masks, and frozen quantities, local differentiability requires finite tensors, nonzero normalization denominators, a full-rank Kabsch covariance with distinct singular values, and a fixed reflection branch. The conditioning factor requires simple extremal eigenvalues. The damped solve must be finite with safeguards inactive. Clamps, norms, symmetry selection, and pose/box losses must remain on smooth branches. In particular, symmetric yaw maximizers must be unique or tied functions must agree locally.

Under these conditions, the composed computation is differentiable. Paired SVD sign ambiguity cancels in Kabsch. At branch boundaries, autograd may select a subgradient or one branch without classical differentiability. Saturation and stationarity can give zero gradients. This real-valued statement excludes floating-point rounding. The stop-gradient ablation in Table 3 measures removing solver gradients together with the resulting optimization trajectory.

A.3 ADDITIONAL EXPERIMENTAL RESULTS

The benchmark comparisons, supervision controls, target-transfer measurements, and solver analysis use the shared training and evaluation setup below. Trained REAL275 variants use the same CAMERA+Real training mixture, frozen response estimator, and 15,576-input REAL275 protocol (Appendix A.1). Table conventions follow Section 4. Inference timing is specified in Appendix A.3.3.

Baseline scope and protocols. Published baseline rows use the configurations and evaluation protocols of their cited sources. The quantitative tables compare point-cloud and depth-only methods on REAL275 and HouseCat6D. Matched REAL275 ablations vary influence supervision within the same recovery system.

Table 7 compares partial-cloud recovery with prediction from GT-aligned complete geometry. The first-stage result is an oracle reference, since its input is unavailable in the partial-cloud inference task.

Table 7: Partial-cloud recovery and the complete-geometry oracle on REAL275.

Configuration	IoU		Pose				FPS ↑
	50 ↑	75 ↑	5°/2 cm ↑	5°/5 cm ↑	10°/2 cm ↑	10°/5 cm ↑	
Full	82.13	76.58	57.73	62.69	75.27	81.56	144.5
Stage1 (GT complete geometry)	80.14	77.87	79.97	84.34	87.92	92.38	–

Strict and relaxed pose thresholds. Full has higher strict-pose AP and lower relaxed-pose AP than ComPose-D on both benchmarks (Tables 1 and 2). The controls below examine how influence supervision and refinement affect recovery accuracy.

The following tables break down the HouseCat6D comparison in Table 2 by category, separating box mAP from strict-pose mAP. The main HouseCat6D table includes the depth-only AG-Pose-D reproduction reported by Ren et al. (2026).

HouseCat6D. The same inference pipeline uses partial point clouds on the HouseCat6D test split. The *box* category reaches 93.9% at IoU_{25} and 2.3% at $5°/2\,\text{cm}$. The *cup* category reaches 99.4% and 1.2%, respectively. The four pose-mAP averages are 12.5%, 15.1%, 28.6%, and 34.1% across the ten categories. *Tube* has the lowest IoU_{50} at 42.3%. Box overlap and the joint $5°/2\,\text{cm}$ criterion measure different aspects of recovery. IoU_{25} captures coarse extent alignment, while strict pose AP requires accurate rotation and translation together. These category results locate the remaining strict-pose errors. We next examine the supervision and recovery design through the REAL275 controls.

Table 8: Per-category IoU_{25}/IoU_{50} AP on HouseCat6D (%).

(a) Bottle–Remote

Method	Mean ↑	Bottle ↑	Box ↑	Can ↑	Cup ↑	Remote ↑
FS-Net (Chen et al., 2021)	74.9/48.0	65.3/45.0	31.7/1.2	98.3/73.8	96.4/68.1	65.6/46.8
GPV-Pose (Di et al., 2022)	74.9/50.7	66.8/45.6	31.4/1.1	98.6/75.2	96.7/69.0	65.7/46.9
InfluPose (Full)	**90.4/65.8**	**85.6/75.9**	**93.9/61.5**	**98.7/83.2**	**99.4/79.4**	**90.0/61.7**

(b) Teapot–Shoe

Method	Teapot ↑	Cutlery ↑	Glass ↑	Tube ↑	Shoe ↑
FS-Net (Chen et al., 2021)	69.9/59.8	71.0/51.6	99.4/32.4	79.7/**46.0**	71.4/55.4
GPV-Pose (Di et al., 2022)	75.4/61.6	70.9/**52.0**	**99.6/62.7**	76.9/42.4	67.4/50.2
InfluPose (Full)	**96.7/68.9**	**82.1**/50.5	83.1/55.7	**82.5**/42.3	**92.2/79.1**

Table 9: Per-category box and pose AP of InfluPose on HouseCat6D (%). Overall averages the ten categories equally.

Category	N	IoU_{75} ↑	5°/2 cm ↑	5°/5 cm ↑	10°/2 cm ↑	10°/5 cm ↑
Bottle	2,929	48.3	25.4	25.5	46.8	47.4
Box	2,929	12.0	2.3	2.3	6.9	7.0
Can	2,929	34.1	22.6	22.8	51.2	52.9
Cup	2,929	11.7	1.2	1.3	6.2	7.8
Remote	2,929	9.9	14.3	15.2	30.9	36.4
Teapot	2,929	14.0	3.5	4.1	12.1	18.3
Cutlery	2,929	19.6	7.0	13.0	28.3	38.9
Glass	2,929	21.9	21.7	39.1	33.4	57.7
Tube	2,929	14.8	5.1	5.1	13.0	13.1
Shoe	2,929	52.7	21.6	22.3	57.4	61.0
Overall	29,290	23.9	12.5	15.1	28.6	34.1

A.3.1 Expanded Ablation Metrics

We report additional metrics for the supervision and recovery controls, then examine anchor context, influence channels, and regional probes. All controls follow the REAL275 protocol in Appendix A.1.

Direct influence supervision. Across three seeds, direct supervision improves strict-pose mAP by an average of 2.70 points (Table 4). Table 10 expands the seed-260907 comparison to all six benchmark metrics. The control keeps the learnable four-channel influence head and its use in recovery, but removes the direct influence losses. Without direct supervision, strict AP is 54.90%, versus 57.73% for Full. Full leads across all six metrics, with a 2.82-point gain at the strict threshold. The individual target sources are examined in Table 3.

Table 10: Supervision and directional-operator controls on REAL275 (mAP, %).

Configuration	IoU_{50}	IoU_{75}	5°/2 cm	5°/5 cm	10°/2 cm	10°/5 cm
Full	**82.13**	**76.58**	**57.73**	**62.69**	**75.27**	**81.56**
No direct influence supervision	81.25	75.53	54.90	60.49	73.46	81.07
Full directional operator	81.75	75.76	53.56	58.85	72.63	80.81

Table 11 reports median state errors on the same 14,182 correctly classified, bbox-matched instances from 17 physical objects for Full and the controls. Ground-truth dimensions include the annotated uniform scale. For each matched instance, the three state errors are

$$e_R = d_{\mathrm{q}}(\hat{\mathbf{R}}, \mathbf{R}^*) \text{ in degrees}, \quad e_t = 100\|\hat{\mathbf{t}} - \mathbf{t}^*\|_2 \text{ in cm}, \quad e_s = \frac{100}{3}\sum_{j=1}^{3}\left|\frac{\hat{s}_j}{s_j^*} - 1\right| \text{ in \%}, \quad (32)$$

with translations and sizes expressed in meters. We take the median of each error over the matched instances, using the lower of the two middle values when the count is even. These medians summarize individual state errors, whereas pose AP requires both rotation and translation to meet their thresholds. Full has the lowest median error for each state among these controls. Compared with no direct supervision, translation error decreases from 0.981 to 0.963 cm and size error from 7.759% to 7.619%, consistent with the higher joint pose AP in Table 10.

Table 11: Median state errors on 14,182 matched REAL275 observations.

Configuration	Rotation (°) ↓	Translation (cm) ↓	Size (%) ↓
Full	**3.268**	**0.963**	**7.619**
No direct influence supervision	3.330	0.981	7.759
Full directional operator	3.531	1.037	8.568

Full directional rotation operator. For $\mathbf{u} = \mathbf{R}\mathbf{e}_y$, the control uses

$$\mathbf{G}_{i,R}^{1/2} = \sqrt{\hat{\eta}_i^{R_\perp}}(\mathbf{I}_3 - \mathbf{u}\mathbf{u}^\top) + \sqrt{\hat{\eta}_i^{R_\parallel}}\mathbf{u}\mathbf{u}^\top \quad (33)$$

and right-multiplies the rotation Jacobian by this matrix during refinement. The estimated axis follows the current rotation. Translation and size influence, row weights, damping, and clipping bounds remain unchanged. Unlike diagonal weighting, this operator does not become isotropic when the squared axis coordinates are equal. The control reaches 53.56% strict AP, 4.16 points below Full. Both box APs and all four pose APs are lower than Full's. These results favor Full's diagonal weighting under this training recipe. Appendix A.2 describes the diagonal approximation.

Context-aware anchor prediction. Table 12 compares anchor prediction with and without cross-anchor context and explicit geometric bias. Rank ρ is the mean target–prediction Spearman correlation over the four state channels on the same balanced set of 60 instances for every row. We first correlate average-tie ranks across anchors, average over active observations within each channel, and then average the four channel means, as in Section 4.3. Constant target or prediction fields are excluded. The main transfer experiment evaluates Full on 360 observations. Each context row retrains a complete encoder with its associated parameter count and optimization trajectory.

Table 12: Context ablation on REAL275.

Setting	IoU_{75} ↑	5°/2 cm ↑	Rank ρ ↑
Independent anchors	26.69	3.32	0.375
Attention without geometric bias	74.52	47.74	0.405
Full	**76.58**	**57.73**	**0.507**

Compared with attention without geometric bias, Full increases target–prediction rank agreement from 0.405 to 0.507 and 5°/2 cm AP from 47.74% to 57.73%. Independent anchors reach 3.32% strict AP. Full has the highest rank agreement and strict-pose AP among these retrained encoders.

State-conditioned influence channels. Table 5 in the main text removes each influence channel from both direct supervision and recovery weighting, keeping the same anchors and the remaining channels. Its rotation, translation, and size columns use the per-instance errors and median convention in Equation 32.

Every channel-removal row is 3.96–7.06 points below Full at 5°/2 cm. The corresponding median state error increases for both rotation removals and for translation removal. Removing size influence

increases median relative size error from 7.62% to 7.99% and lowers strict pose AP from 57.73% to 50.67%. Full has the lowest rotation, translation, and size error medians and the highest strict AP among these controls. Each ablation changes both supervision and recovery weighting, assessing their combined contribution to joint recovery.

Regional-probe sensitivity. Full uses $M = 8$, $\rho = 0.125$, and $\beta = 0.70$. The geometry rows vary M at fixed $\rho = 0.125$, then vary ρ at fixed $M = 8$. The blend rows vary β only for probed examples. The $\beta = 1$ setting retains dense-proxy supervision for unprobed examples, as specified in Appendix A.2.1.

Table 13: Regional-probe and target-blend sensitivity on REAL275 (mAP, %).

Setting	$5°/2$ cm ↑	$10°/5$ cm ↑	IoU_{50} ↑
(a) Probe geometry			
$M = 4,\ \rho = 0.125$	53.22	78.55	81.52
$M = 8,\ \rho = 0.0625$	54.63	80.94	**82.29**
Full (default)	**57.73**	**81.56**	82.13
$M = 8,\ \rho = 0.25$	55.41	79.67	81.48
$M = 16,\ \rho = 0.125$	52.20	78.06	81.98
(b) Target blend			
$\beta = 0.30$	52.01	78.30	81.97
Full (default)	**57.73**	**81.56**	**82.13**
$\beta = 1.00$ (probed branch)	55.30	78.90	81.20

Full has the highest $5°/2$ cm and $10°/5$ cm AP in both grids. The smaller deletion fraction $\rho = 0.0625$ gives the highest IoU_{50} in the probe-geometry grid, while Full leads all three metrics in the blend grid. Changing M from 8 to 4 or 16 reduces strict-pose AP by 4.51 or 5.53 points, respectively. At fixed ρ, each intervention deletes the same number of points. Varying M changes region-center coverage and the assignment of regional responses to anchors. The region-count controls therefore vary coverage, target correlation, and supervision allocation jointly. The ρ and β rows separately vary the deletion-and-refill pattern and target blending.

A.3.2 Influence Diagnostics

Uncertainty in target transfer. For each observation and channel, we compute Spearman correlation as the Pearson correlation between target and predicted ranks over the 192 anchors, assigning ties their average rank. We exclude pairs with target or prediction standard deviation below 10^{-6}, including symmetry-inactive axial fields. We average the remaining correlations equally within each channel, then average the four channel means for the scalar agreement in the main text. Table 14 reports two bootstrap intervals from the same 360 target–prediction pairs. The physical-object bootstrap uses 50,000 draws of objects with replacement, keeping all their views. The observation bootstrap uses 10,000 draws of individual views. The axial channel uses 180 active observations from nine objects. The other channels use 360 observations from 17 objects. Both procedures recompute the observation mean and report its 2.5th and 97.5th percentiles. All pointwise 95% channel intervals are positive.

Table 14: Target–prediction rank agreement with bootstrap intervals.

Channel	Observations / objects	Mean ρ	Observation 95% CI	Object-cluster 95% CI
Orthogonal rotation	360 / 17	0.636	[0.617, 0.654]	[0.593, 0.683]
Axial rotation	180 / 9	0.393	[0.350, 0.432]	[0.294, 0.482]
Translation	360 / 17	0.411	[0.383, 0.438]	[0.320, 0.498]
Size	360 / 17	0.484	[0.456, 0.512]	[0.373, 0.578]

Regional response diagnostics. We test whether deleting regions ranked highly by predicted influence causes larger prediction changes in a fixed partial-cloud estimator. Figure 4 summarizes rank–response association and cross-output interventions. Figure 3 compares influence with reliability and random rankings. Both use the same observation weighting and pointwise 95% intervals from physical-object resampling.

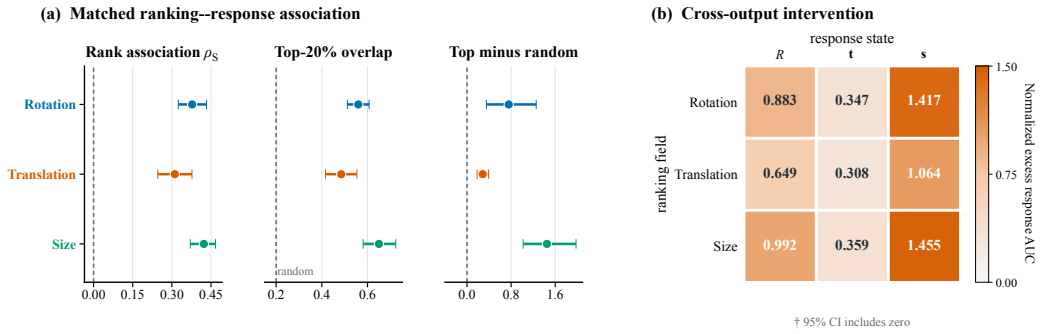

Figure 4: Finite influence diagnostics under refill-paired regional deletion on 360 REAL275 instances. Panel (a) reports single-region rank association, top-20% overlap, and top-minus-random response at 20% deletion. Panel (b) reports cross-output response AUC, with unadjusted 95% clustered-bootstrap intervals.

We use the Response AUC and normalized output changes defined in Equation 11, including its equal weighting over observations, deletion budgets, and paired refills. Positive values indicate larger responses for high-ranked deletion than for matched random deletion. A region score is the mean of its anchor scores. For rotation ranking, continuously local-y symmetric instances use $\hat{\eta}^{R_\perp}$, whereas asymmetric instances use the mean of $\hat{\eta}^{R_\perp}$ and $\hat{\eta}^{R_\parallel}$.

We partition each 1,024-point cloud into ten nearly equal-cardinality bins (six of size 102 and four of size 103) along the first PCA axis. This partition differs from the eight FPS-centered regions used to construct training targets. We evaluate deletions at $\mathcal{Q} = \{0.1, 0.2, 0.3\}$ on 360 correctly classified, IoU-≥ 0.5 bbox-matched detector instances from REAL275, sampled uniformly across six categories. Learned and matched-random rankings use the same eight fixed refill offsets. Association and cross-output intervals use 1,000 bootstrap resamples of the 17 physical objects. Ground truth supplies matching, symmetry-aware scoring, object identities, and error calculation, but does not enter the prediction heads.

For each observation, we average single-region responses over eight refills and compute Spearman correlation with the ten region scores, using average ranks for ties. Constant score or response vectors are excluded. Top-20% overlap uses the two highest-response regions, with random expectation $2/10 = 0.2$. The cross-output matrix evaluates every ranking and response pair under the same response-AUC definition. Figure 4 shows positive rank–response association for all three state channels, strongest for size, alongside the cross-output responses of this fixed model.

Ground-truth error degradation. We use the same model, observations, rankings, deleted sets, and eight refill offsets to measure changes in benchmark pose and size error. Damage is the normalized error after deletion minus the error before deletion. Negative values indicate improvement and are retained. Table 15 compares damage from high-influence and matched random deletion at 20%, averaging refills and then observations. The pointwise 95% intervals use 1,000 draws that resample physical objects with all their observations. All three differences have positive intervals. Rotation and translation correspond to 3.04° and 0.31 cm of additional mean error, respectively. High-influence deletion also increases mean error relative to the undeleted prediction in all three states. The rankings therefore identify regions whose removal degrades this estimator on average.

Paired comparison with reliability rankings. We score each region by its mean predicted inlier probability or its mean predicted residual precision. Precision is averaged over anchors and axes before the refinement clamp. Table 6 reports paired comparisons using the same three deletion budgets, eight refill offsets, and observation weighting. Intervals use 50,000 bootstrap draws of

Table 15: Ground-truth error degradation under 20% regional deletion.

State	Error normalizer	Top minus random	Pointwise 95% CI
Rotation	5°	0.607	[0.271, 1.006]
Translation	2 cm	0.155	[0.083, 0.242]
Size	0.10 log-size norm	1.068	[0.540, 1.756]

whole objects with a conservative Bonferroni divisor of 12 (two-sided tail probability $0.05/(2\times12)$), also applied to the six displayed contrasts. Influence has positive response-AUC differences over inlier ranking for all three states and over precision for translation and size. The rotation–precision interval includes zero.

A.3.3 Recovery Controls and Inference Efficiency

We examine how predicted precision and Cauchy weights affect recovery from fixed network predictions, then describe end-to-end inference timing. The trained supervision and rotation-operator controls are defined in Appendix A.3.1.

Precision and Cauchy interventions. Table 16 reports mAP (%) while keeping Full's predictions and raw-precision Kabsch initialization fixed. Cauchy weights downweight large standardized residuals, while predicted precision controls their direction-dependent scaling. Removing Cauchy weights lowers strict-pose AP by 0.44 points. Fixing refinement precision to $50\mathbf{I}_3$ lowers it by 14.25 points. IoU_{75} also decreases by 0.27 and 1.29 points, respectively. The fixed-precision control tests the joint effect of whitening and residual standardization. Precision retains an unsaturated entry in 273 of 360 audited observations. These results support retaining predicted precision, with a smaller gain from Cauchy weighting.

Table 16: Inference-time precision and Cauchy controls on REAL275 (mAP, %). Δ AP is the change in 5°/2 cm AP relative to Full, in percentage points.

Configuration	$\text{IoU}_{75}\uparrow$	5°/2 cm $\uparrow$	Δ AP
Full	**76.58**	**57.73**	–
Without Cauchy	76.31	57.28	-0.44
Precision fixed to $50\mathbf{I}_3$	75.29	43.48	-14.25

End-to-end inference timing. InfluPose takes 6.920 ms per object (144.5 FPS) on a single RTX 4090 with batch size one, measured over 100 timed runs after 30 warm-up runs. The timed region starts before geometric feature construction and ends after the final robust-refinement update, and therefore includes feature construction, partial-cloud network inference, weighted Kabsch initialization, and one robust-refinement update. Detection, data loading, host-to-device transfer, and metric computation are excluded. The complete-geometry estimator and regional probes are used only during training.

We re-profiled FS-Net, GPV-Pose, HS-Pose, SAR-Net, RBP-Pose, DR-Pose, Query6DoF, and GenPose on one RTX 4090 for Table 1, following each method's timing protocol. Symmetry handling follows Appendix A.1.

A.4 Additional Visualizations

Training dynamics. Figure 5 shows epoch-averaged rotation and translation errors for the complete-geometry estimator and partial-cloud network over 36 and 50 epochs, respectively. The curves show unsmoothed epoch means of prediction errors from the training logs.

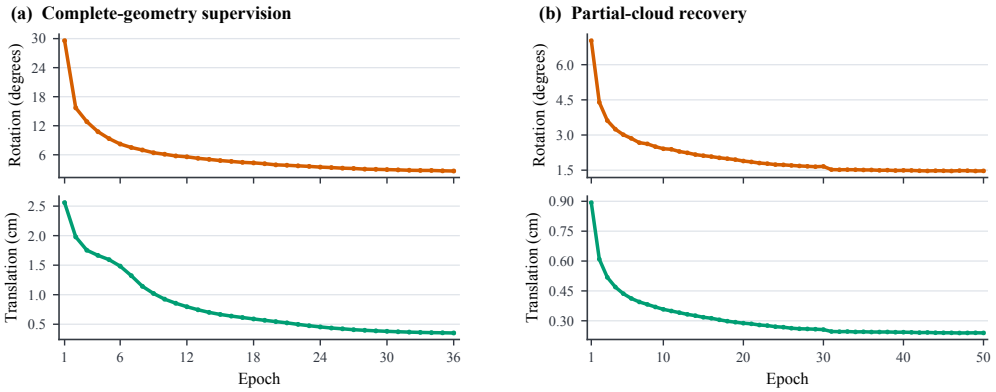

Figure 5: Training dynamics of the complete-geometry estimator over 36 epochs and the partial-cloud network over 50 epochs. Curves show unsmoothed epoch means for rotation error in degrees and translation error in centimeters.

Influence fields across object categories. Figure 6 visualizes complete-geometry influence targets and the corresponding partial-cloud predictions across the six CAMERA25 object categories. Each target–prediction pair is evaluated on the same visible partial point cloud. The four channels induce distinct spatial rankings over the observed surface, indicating that the learned influence is state-conditioned rather than a single generic importance map. For example, on the mug, the target assigns high axial-rotation and size influence to the handle, while orthogonal-rotation influence is concentrated more strongly around the rim and outer body. The partial-cloud prediction preserves these characteristic high-ranking regions despite having access only to the visible observation.

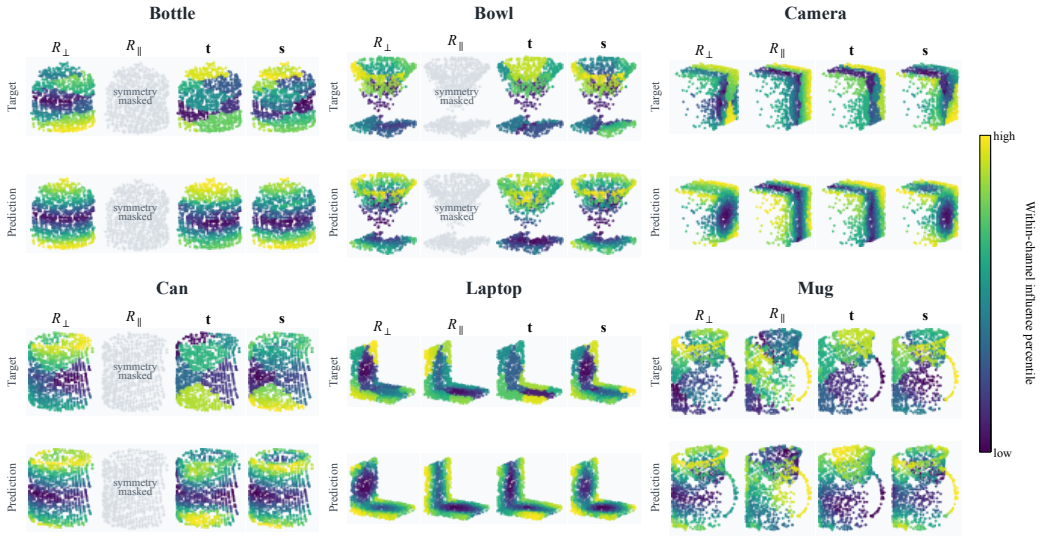

Figure 6: Complete-geometry targets (top) and partial-cloud predictions (bottom) for four state-conditioned influence channels. Warmer colors indicate higher within-channel relative influence; gray $R_{\parallel}$ panels denote continuously symmetric categories.

Qualitative comparison with existing methods. Figure 7 compares saved DPDN and AG-Pose predictions with InfluPose on the same REAL275 frames. These RGB-D methods provide qualitative context. The quantitative tables cover point-cloud/depth-only methods. The selected views illustrate pose and box alignment alongside the aggregate benchmark results.

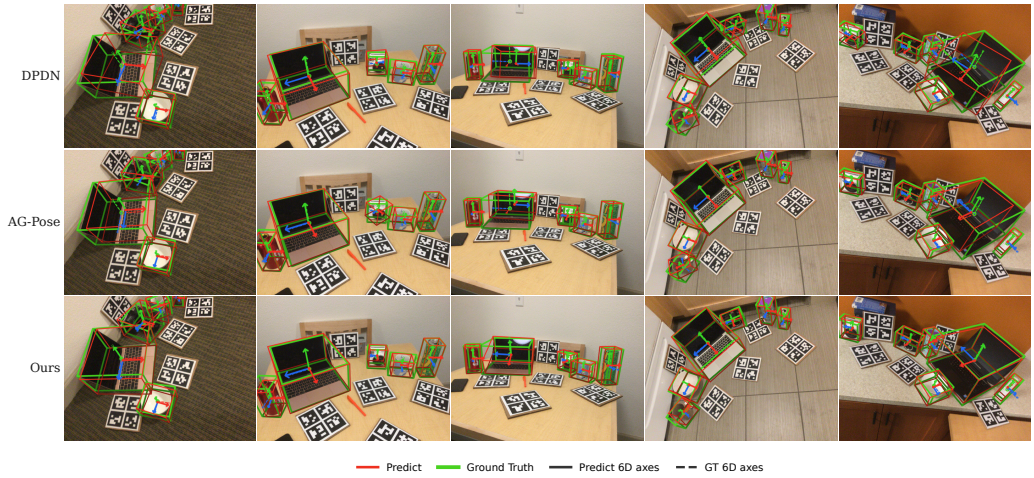

Figure 7: Qualitative pose comparisons on REAL275. Green boxes denote ground truth and red boxes denote predictions. Axes indicate orientation.

Qualitative results across scenes. Figure 8 shows InfluPose predictions in six REAL275 scenes from multiple viewpoints. The predicted 3D boxes align with the ground-truth extents in these examples despite changes in arrangement, clutter, and partial occlusion. The views complement the benchmark results with examples of recovery from different visible surfaces.

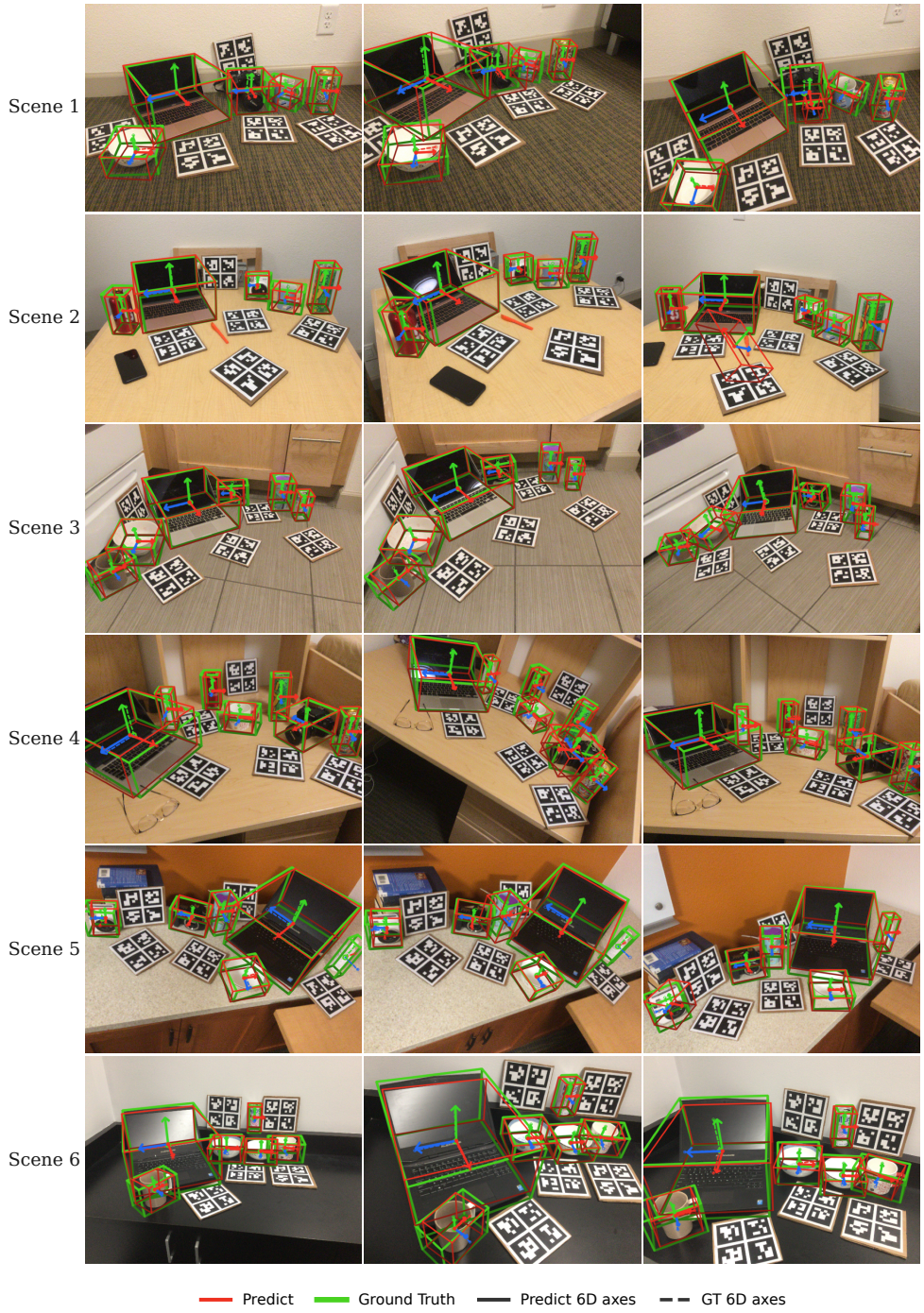

Figure 8: Qualitative results of InfluPose across six representative scenes from REAL275. Green boxes indicate ground truth and red boxes indicate our predictions. The corresponding 6D axes are shown for reference.