Assessing Prediction Reliability for Probabilistic Pose Estimation

Omar Del-Tejo-Catala, Javier Perez Soler,
Nicolás García Sastre, Pau Garrigues Carbó
Instituto Tecnológico de Informática (ITI)
Valencia, 46022 Spain
e-mail: [odeltejo,javierperez,
ngarcia,pgarrigues]@iti.es

Jose-Luis Guardiola, Alberto J. Perez,
Juan-Carlos Perez-Cortes
Universitat Politècnica de València (UPV)
Valencia, 46022 Spain
e-mail: [joguagar,aperez,jcperez]@iti.es

***Abstract*—Ensuring 6D pose estimation models rely on semantically relevant visual cues is essential for robust estimations. We investigate explanation-based validation of pose predictions by extending Guided Grad-CAM and Guided Backpropagation to highlight regions driving rotation predictions. This enables analyzing whether the model attends to 3D keypoints rather than spurious background noise. We also explore synthetic-real distribution comparisons to filter predictions. We demonstrate that explanation quality can discard predictions relying on irrelevant evidence. Experiments show separation between low and high errors, achieving an Area Under the Receiver Operating Characteristic Curve (AUROC) of 0.846 for a spherical object and 0.968 for a cylinder. Crucially, these filtering strategies operate without ground-truth labels, enabling unsupervised validation at inference time.**

Keywords-guided grad-cam; explainable ai; guided backpropagation; probabilistic pose estimation; pose estimation.

I. Introduction

Estimating the six degrees of freedom (6D) pose of objects from images is a fundamental problem in computer vision, with applications in robotics —for instance, object pick-and-place problems—, augmented reality, and industrial quality inspection. Despite recent advances in deep learning-based methods, achieving reliable pose predictions remains challenging due to occlusions, clutter, and inherent object symmetries. Probabilistic models have recently been proposed [1] to address some of these challenges by representing uncertainty as distributions over the rotation space SO(3), the special orthogonal group of 3D rotations. However, regardless of the predictive framework employed, a central open question persists: are the predictions based on the correct visual evidence?

While pose estimators can produce high-confidence predictions, such predictions may still be unreliable if they are derived from spurious correlations in the background or irrelevant image regions. This issue is particularly problematic in approaches that train models using synthetic data to solve real-world use cases. This kind of training can bias the model toward synthetic-only cues and cause underperformance in real-world applications; this mismatch is known as the synthetic-real domain gap. Although domain adaptation techniques have been proposed to address this problem, it is also essential to ensure that unexpected anomalies in the object's appearance do not affect the predictions or, if they do, that these predictions can be filtered.

In the probabilistic pose estimation context, confidence metrics, such as likelihoods or entropy of predicted distributions, capture the model's internal uncertainty but provide no direct insight into whether the visual reasoning process is sound. To address this gap, eXplainable AI (XAI) techniques can serve as a powerful diagnostic tool.

In this work, we adapt a gradient-based explanation method to the probabilistic pose estimation setting. By generating saliency maps for the rotation predictions, we can visualize which regions of the input image most strongly influence the network's outputs. This representation enables a fine-grained inspection of whether the model focuses on the target object or instead relies on irrelevant cues, such as background textures or stains. Crucially, these explanations allow us to go beyond uncertainty quantification and introduce an additional filtering stage: pose predictions with unsatisfactory explanation patterns can be systematically identified and discarded.

Thus, the goals of this work are the following: (1) Investigate whether explainability techniques can reliably detect pose prediction errors without requiring ground-truth labels; (2) Propose and evaluate comparison strategies in three spaces (2D image, 3D model, rotation distribution) to filter unreliable predictions; and (3) Analyze the sensitivity of these methods to texture variability.

The remainder of the paper is organized as follows: Section II reviews the state of the art in explainable AI for pose estimation. Section III describes the material and methods employed. Section IV presents and discusses the experimental results. Section V concludes the paper and outlines future work.

II. State of the Art

XAI techniques produce saliency maps highlighting image regions driving a model's decision [2]. Gradient-based methods like Grad-CAM [3] yield coarse localization maps; combined with guided backpropagation [4], Guided Grad-CAM retains fine structural detail. Other methods include Integrated Gradients [5], DeepLIFT [6], and LRP [7]. Attention-based methods [8] leverage internal weights, while Score-CAM [9] uses forward-pass scores. SmoothGrad [10] reduces noise via averaged maps, whereas perturbation approaches like LIME [11] or SHAP [12] are computationally expensive.

Attribution maps help verify if predictions rely on relevant cues. Quantitative metrics such as Over-MAP [13] measure attention–segmentation overlap, and PoseIG [14] uses specialized indices to quantify attribution focus. Filtering unreliable predictions based on these metrics is related to our approach. Similarly, UA-Pose [15] suppresses unreliable pose fits using geometric occlusion. However, saliency maps can sometimes misrepresent true decision drivers [16]–[18]. Extending XAI to 3D pose estimation remains largely unexplored [13]–[15]. Here, we adapt Guided Grad-CAM to a probabilistic 3D-pose

 ISBN: 978-1-68558-395-8

model, hypothesizing that predicted orientations correlate with extracted spatial saliency patterns.

III. Material and Methods

This section explores the datasets employed to measure the quality of our proposed prediction quality scores and the techniques employed. A schematic representation of the pipeline followed is shown in Figure 1.

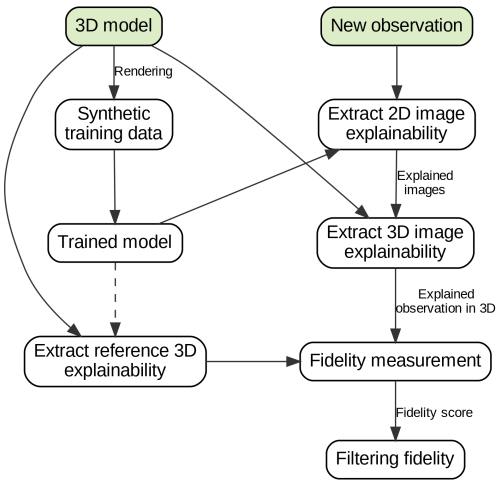


Figure 1. Pipeline of the prediction quality measurement.

A. Datasets

We selected two objects from prior work [1][19] to stress-test two distinct failure regimes that commonly arise in industrial 6D pose estimation: *(i)* a *geometric ambiguity* case and *(ii)* a *domain-gap* case. These regimes were chosen deliberately because they represent the two most prevalent sources of unreliable predictions: ambiguity in object geometry and mismatches between training and deployment domains. Furthermore, the selected objects—a cylinder and a sphere—are geometric primitives commonly encountered in industrial inspection and robotic manipulation scenarios.

Geometric Ambiguity (Cylinder): The cylinder features a square carving on one base and a triangle carving on the other base. This geometric configuration induces multi-modal rotation distributions per view, as different orientations may yield similar visual appearances. For example, 90° rotations around the cylinder's axis can leave the square carving visually unchanged. This regime tests whether explainability techniques can correctly identify when a prediction relies on insufficient geometric information.

Domain Gap (Sphere): The sphere has a "T"-shaped carving, but the real-world texture noise on its surface is absent from the synthetic training data. This regime tests the system's ability to detect when predictions are based on spurious features (texture noise) rather than the intended carving pattern. The noisy texture simulates real-world conditions where training data may not fully capture object appearance variations; the model may confuse backside texture patterns with the "T" carving, producing confident but incorrect 180° rotation predictions.

These objects were selected because they: (1) cover two fundamental failure modes in pose estimation: geometric ambiguity and domain mismatch; (2) allow clear evaluation of explainability techniques via identifiable visual carvings; and (3) have been validated in previous literature [1][19].

While these two objects cover important failure regimes, we acknowledge that they do not exhaustively represent all object types (e.g., texture-only objects without geometric keypoints, or highly articulated objects). Generalization to such categories remains future work.

Models are trained on synthetic [20][21] and CycleGAN domain-adapted renders [19]; real captures (Figure 2) are used for evaluation only.

B. Pose Estimation Model

The probabilistic pose estimation model from [1] predicts rotation probability distributions over the discretized SO(3) space. Translation is out of scope; it can be approximated from multicamera geometry.

C. 2D Explanation

Guided Grad-CAM is applied to interpret the Convolutional Neural Network (CNN)'s decision: Grad-CAM gradients of the rotation score with respect to the last convolutional layer localise relevant regions, and guided backpropagation refines them to edge-level detail. Since the rotation classes are not mutually exclusive, the technique is applied to each predicted rotation mode independently. Results for both objects are shown in Figures 3 and 4.

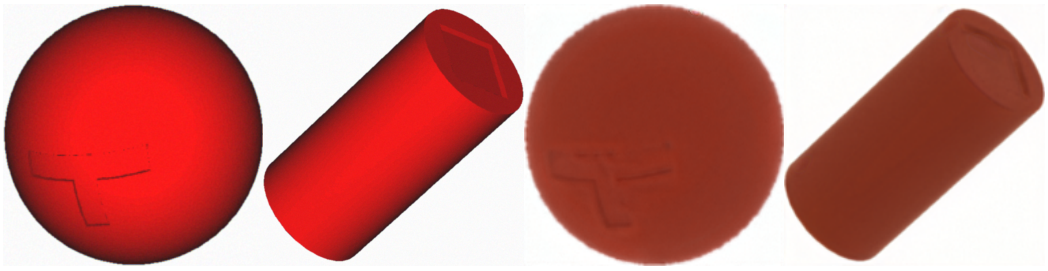
We chose Guided Grad-CAM and Guided Backpropagation over alternative methods for several practical reasons. Gradient-free methods, such as Score-CAM [9], require multiple forward passes per activation map channel, making them prohibitively expensive in our multi-view setup (4 cameras $\times$ multiple rotation modes per image). LRP [7] requires architecture-specific decomposition rules that are not readily available for the graph neural network components of our probabilistic model. Attention-based methods [8] assume transformer-like architectures with explicit attention weights, which our CNN-based backbone does not provide. SmoothGrad [10] could complement our gradient-based approach and is considered for future investigation, as it may reduce the noise observed in our saliency maps (see Section V).

D. 3D Explanation

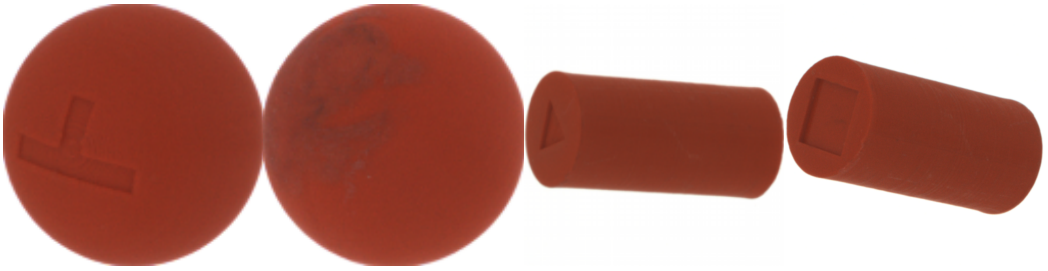
To extract the 3D explanation, the explanation process computes explainability at the pixel level using the 2D explanation method described above. Then, it uses the network's pose prediction to project the 2D explanations to the object's reference 3D model, which was used to train the pose prediction model. Due to the object's geometry and the camera system setup, all 3D points are visible to at least one camera. Many of them are captured by more than two cameras, so we can use several cameras to assign relevance values.

Gradient-based explainability methods often produce noisy activations, marking pixels as relevant even when they do not truly influence the prediction. To minimize the impact of a single camera's noise on the explained model, we assign the minimum value across all cameras. This approach implies that all cameras seeing a 3D point should agree that it is relevant.

Once a 3D model contains a relevance value per 3D point, the process compares the value obtained per 3D point against a reference explained 3D model.



(a) Synthetic capture of "T"-carved sphere (b) Synthetic capture of the carved cylinder (c) Realistic-transformed capture of "T"-carved sphere (d) Realistic-transformed capture of the carved cylinder



(e) Real capture of "T"-carved sphere (f) Real capture of sphere's noisy backside texture (g) Real capture of cylinder's triangle carving (h) Real capture of cylinder's square carving

Figure 2. Dataset samples showing synthetic, domain-adapted, and real captures for both objects.

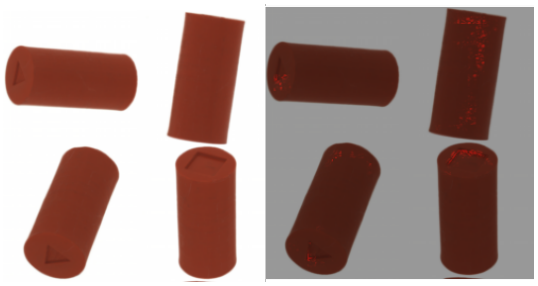


Figure 3. Visualization of Guided GradCAM activations over the cylinder.

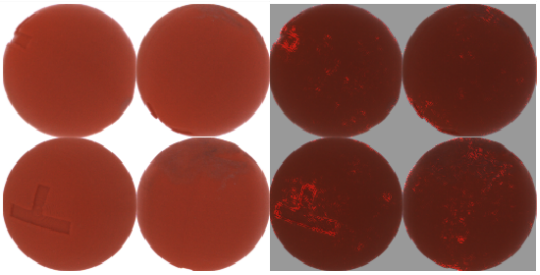


Figure 4. Visualization of Guided Backprop activations over the sphere.

1) Extracting the Reference Explainability: The reference model marks as relevant every 3D point belonging to a different orthogonal triangle of the training mesh, avoiding the noise introduced by averaging multiple explained synthetic batches.

2) Comparing Against Reference Explained 3D Model: At inference, an explained 3D model is extracted using the predicted (not ground-truth) pose; pixel–3D-point misalignment encodes the pose error signal. Three metrics are evaluated: Pearson correlation, dot product, and IoO (see Section III-E).

E. Theoretical Framework of the Explanation Filtering

The theoretical framework for filtering incorrect pose predictions using explainability compares an observed explanation with an ideal reference. We define the following notation:

- I: Ideal 3D explanation (reference model).
- O: Observed 3D explanation (from inference).
- ΔI: Noise in the ideal explanation generation process.
- ΔO: Noise in the observed explanation generation process.
- ΔR: Divergence in explanation due to rotation prediction error.

The similarity score S between reference and observation is computed as:

$$S = f(I + \Delta I, O + \Delta O + \Delta R) \tag{1}$$

where f is an evaluation metric (e.g., Pearson correlation, dot product, or IoO). The Intersection over Observation (IoO) metric quantifies how much of the observed explanation aligns with the reference. Unlike Intersection over Union (IoU), IoO does not penalize missing activations in the observation, focusing instead on whether observed activations fall within expected regions. This feature is advantageous because gradient-based explanations may under-activate some relevant regions while still being correct. The IoO metric is defined as:

$$IoO(I, O) = \frac{|I \cap O|}{|O|} \tag{2}$$

where $|I \cap O|$ represents the set of 3D points activated in both the reference and observation, and $|O|$ is the total activation in the observation.

IoO is well-suited to penalize False Positives (FPs) as background pixels that are activated but do not activate in the reference reduce the metric score. However, IoO does not penalize missing activations. While this is advantageous for incomplete pixel activations —having all reference points matched is unnecessary in some cases and the model is usually capable of adequately predicting using a subset of the relevant pixels— it can mask cases where the model fails to attend to critical keypoints altogether. For symmetric objects, the model may attend to only one symmetric feature (e.g., only the square

carving of the cylinder), yielding a high IoO while missing the discriminative triangle carving that would disambiguate rotations.

To evaluate this trade-off, we evaluate all predictions using three complementary metrics across three comparison spaces. This strategy helps identify cases where a single metric might be misleading.

Our goal is to threshold S to detect large values of ΔR. The key assumption is that ΔR correlates with the rotation prediction error: a larger error produces larger ΔR, because the misaligned pose causes the 2D explanation to project onto incorrect 3D surface regions. However, this correlation can weaken: (i) For the cylinder, for instance, rotations of 90° or 180° around its axis may leave one carving unchanged, producing similar explainability patterns despite non-zero error if the other carving is not attended to. (ii) For the sphere, backside texture noise may produce noisy activations that might be confused with the "T" carving, causing low ΔR at approximately 180° error.

The error progression experiments (Tables I and II) synthetically verify this correlation by perturbing pose predictions.

The following properties hold: (1) $\Delta I < \Delta O$, as ideal explanations can average over multiple synthetic batches; (2) $\Delta I = 0$ is achievable via manual annotation; (3) exact alignment implies $\Delta R = 0$; (4) if $O \subseteq I \Rightarrow f(I, O) = 1$; and (5) $O \cap I = \emptyset \Rightarrow f(I, O) = 0$.

In practice, we can perform this comparison either in the image space or by projecting the image attributions to the training reference model. We measured the results for both approaches. In 3D space, we measured the Pearson correlation between the extracted 3D reference explanation and the observed explanation projected to the 3D model. In 2D image space, we render the 3D reference explanation using the predicted pose, and compute the Pearson correlation between the raw observed image attributions and the rendered reference explanation.

F. Filtering Predictions Using a Distribution Comparison with the Reference

Another filtering strategy compares probability distributions: (1) infer the observation's rotation/translation, (2) render the reference model in the predicted pose, (3) infer the rendered images' distribution, and (4) measure difference. We use two metrics: thresholded blob matching and Pearson correlation.

The approach of comparing distributions using thresholding comprises the following steps. First, the distributions are thresholded utilizing a fraction of the distribution's maximum confidence (between a fifth and a fifteenth), seeking to maximize the number of remaining blobs. Then, we reduce the mask to the likeliest rotations for each thresholded distribution's blob. We then dilate the points around the most likely rotations before comparing the overlap between each mask in the reference and observation distributions. The metric is the ratio of the intersection of blobs in the reference and observation distributions to the total number of blobs. A sample of the thresholding process is shown in Figure 5.

Regarding the second metric, the Pearson correlation, it does not require thresholding; thus, it is applied directly to compare the reference and observation distributions. Therefore, two rotation distributions, such as the ones shown in Figure 5(b), are directly compared using Pearson correlation. The distributions are compared in their original SO3 space, not in the image space used to represent them in Figure 5.

Figure 5 visualizes rotation distributions as unwrapped 3D unitary spheres: red is X-axis, green is Y-axis, intensity is likelihood. Views of the square yield 4 Y-axis solutions; views of the triangle yield 3. Views without features assign equal likelihood at 90 degrees.

IV. RESULTS

The goal of our evaluation is to determine whether the proposed metrics can separate predictions with low rotation error from those with high rotation error, without access to ground-truth labels. We quantify this separation using the Area Under the Receiver Operating Characteristic Curve (AUROC). A prediction is labelled *positive* (valid) when its rotation error is below 10° and *negative* (invalid) otherwise. An AUROC of 1.0 indicates perfect separation—i.e., the metric can distinguish all correct predictions from incorrect ones—while 0.5 corresponds to random guessing.

We evaluate two complementary experimental setups. First, real-world test predictions use actual sensor captures with ground-truth annotations to assess performance under realistic conditions. Second, error progression experiments synthetically perturb ground-truth poses with increasing rotation offsets (from 0° to 180°) and measure the metric response. These controlled perturbations ensure coverage of the full rotation-error spectrum and allow verifying the monotonic relationship between rotation error and metric degradation, independently of the model's natural error distribution.

In the following figures, we refer to the comparison scores described in Sections III-E and III-F as the "Similarity Score." Each scatter plot shows individual predictions, with the x-axis representing rotation error and the y-axis representing the similarity score. An effective filtering metric should show a clear downward trend: high similarity scores for low rotation errors and low scores for high errors.

1) Cylinder: This section presents the cylinder object results. Table I summarises the AUROC values for each comparison method.

TABLE I. CYLINDER AUROC COMPARISONS.

Space	Comparison	Real AUROC	Syn. Prog. AUROC
3D model	Pearson	0.571	-
3D model	Dot product	**0.968**	-
3D model	IoO	**0.968**	0.965
Image	Pearson	0.524	-
Image	IoO	**0.937**	0.970
Dist.	Pearson	0.825	**1.**
Dist.	Blob matching	**0.873**	-

Filtering Using Fidelity: We will begin evaluating the explainability results for the cylinder. Figure 6 shows the quality of fidelity metrics to separate between low and high rotation losses. Regarding AUROC, both 3D-space and image-space metrics achieve large AUROC values. In both cases, the Pearson comparison technique achieved the worst results, with the dot product and IoO being the best comparison techniques for the cylinder, achieving 0.968 AUROC for both the dot product and IoO in 3D space and 0.937 for IoO in image space.

It should be noted that only 3 evaluated real-world samples exceeded 10° error, limiting statistical reliability. However, high

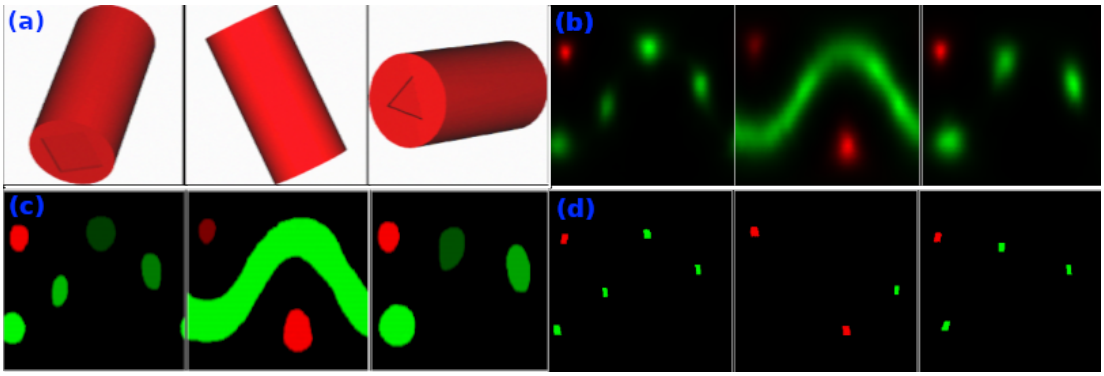


Figure 5. Process of filtering predictions using a distribution comparison. (a) Multi-camera object captures. (b) Rotation distributions. (c) Thresholded distribution blobs. (d) Maximum dilated blobs.

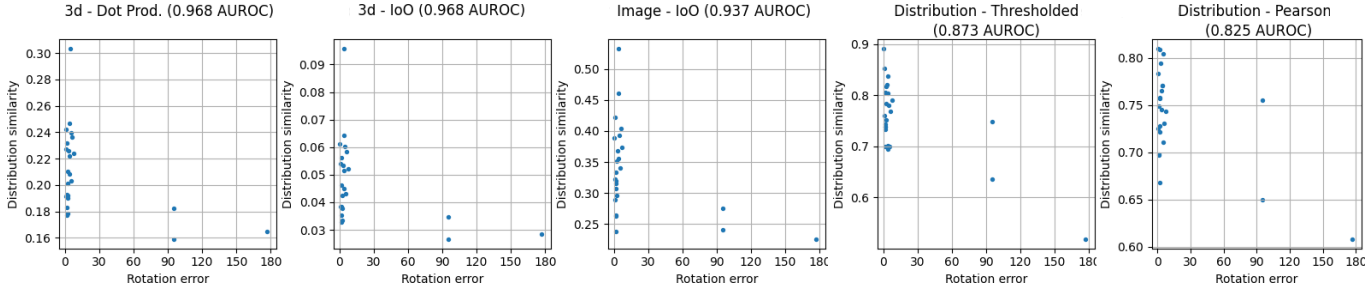


Figure 6. Some of the fidelity and distribution-space metric scores for the cylinder object's model.

values (0.968) indicate strong separation. Error progression experiments confirm monotonic relationships between error and metric score.

Filtering Using Reference Predictions: This section measures the quality of filtering predictions by comparing the observation's probability distribution estimate with the corresponding estimate from the reference model, as explained in Section III-F.

As shown in Figure 6, filtering using the reference predictions would outperform methods employing fidelity metrics, were it not for an observation at 90 degrees that achieved a higher metric score. Both comparison approaches, i.e., thresholding and Pearson, achieve similar results (0.873 and 0.825 AUROC, respectively).

2) Sphere: This section presents sphere object results for the challenging domain-gap scenario. Table II summarises the AUROC values.

TABLE II. SPHERE AUROC COMPARISONS.

Space	Comparison	Real AUROC	Syn. Prog. AUROC
3D model	Dot product	0.788	0.788
3D model	IoO	0.776	-
Image	Pearson	**0.846**	0.856
Image	IoO	0.814	-
Dist.	Pearson	0.788	**0.990**
Dist.	Blob matching	**0.849**	-

Filtering Using Fidelity: Guided Backprop provides the best activations for the sphere. As shown in Figure 4, noisy patterns in the real samples are inappropriately considered relevant because they were unseen during training.

Results for the fidelity metric tracking — both in 3D and image spaces — can be seen in Figure 7. Rotation losses are more distributed than in the cylinder's case, as the object exhibits more texture noise and does not have evenly

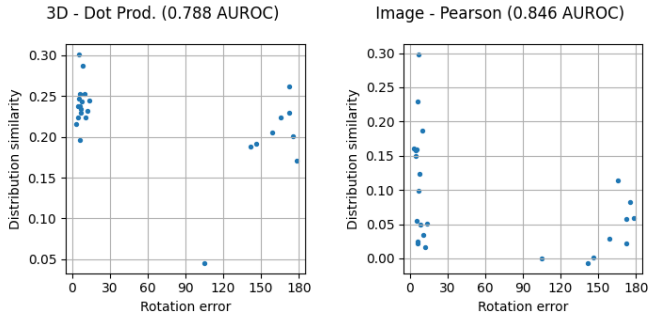


Figure 7. Some of the fidelity metric scores for the sphere object's model.

spaced possible solutions. Rotation losses near 180° represent predictions where the model confused the noisy backside texture with the "T" carving: the explanation activates the backside texture region, which spatially aligns with where the "T" carving would be if the object were rotated 180°, producing a misleadingly high similarity score.

The results over the sphere are less separable than in the cylinder's case, reflecting the additional challenge posed by domain-gap noise. The 3D-space fidelity metrics achieved similar AUROC values (0.788 for dot product and 0.776 for IoO), both lower than for the cylinder. This degradation occurs because projecting noisy 2D explanations to the 3D model amplifies spurious activations through the min-aggregation process (Section III.C). Image-space metrics outperformed 3D-space ones, with Pearson correlation achieving the highest fidelity-based AUROC (0.846). This reversal compared to the cylinder indicates that when texture noise dominates the explanation, direct image-space comparison avoids the projection-induced artifacts that degrade 3D metrics. The noisy texture provokes the spike in metric scores at approximately

160° rotation loss, as visualized in Figure 4.

Filtering Using Reference Predictions. Employing predictions instead of fidelity to filter predictions yields better results, as an AUROC of 0.849 can be achieved using blob matching (Figure 8), outperforming the fidelity-based metrics but only by a small margin. Although it does not achieve a perfect separation of low and high rotation loss predictions, the decreasing trend is more perceptible than in the fidelity case.

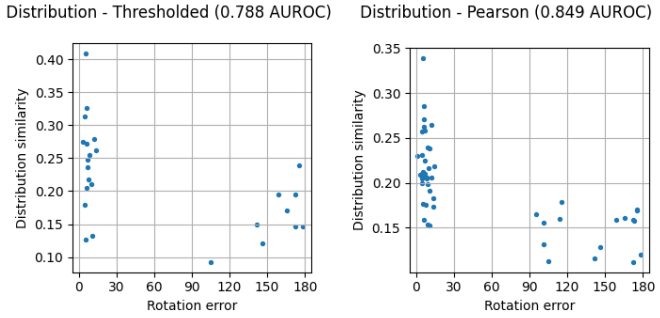


Figure 8. Distribution-space metrics for the sphere object's model.

A. Discussion and Method Selection

The experimental results reveal distinct performance patterns across the two object types and three evaluation spaces. For the cylinder (geometric ambiguity regime), 3D-space metrics (IoO and dot product) achieve the highest AUROC values (0.968). The strong performance stems from the deterministic projection of keypoints. Image-space metrics achieve similar results (0.937) but avoid intermediate 3D projection artifacts. Distribution-space metrics achieve perfect monotonic separation in synthetic experiments.

Conversely, for the sphere (domain-gap regime), 3D-space metrics perform significantly worse (0.776–0.788) because projection amplifies texture noise. Here, image-space Pearson correlation outperforms 3D metrics (0.846), indicating robustness to domain-gap artifacts. Furthermore, distribution comparison (blob matching) achieves 0.849 AUROC, outperforming all fidelity methods by bypassing the explanation pipeline entirely to directly compare prediction distributions.

Method Selection Recommendation: We recommend a two-stage filtering approach. First, for objects with clear geometric keypoints, use 3D-space fidelity metrics (IoO or dot product) as the primary filter. Second, for objects with texture variability or domain gaps, prioritize distribution-space comparison (blob matching) or image-space metrics (Pearson correlation) to handle explanation noise.

V. Conclusion and Future Work

We presented a framework for filtering pose estimation predictions by leveraging XAI and distribution comparison without ground-truth labels. For the cylinder, 3D projected fidelity metrics demonstrated reliable error detection from saliency alignment. For the sphere, gradient-based attributions were dominated by domain-shifted texture noise, making distribution comparison more robust.

Together, these results suggest a two-stage reliability pipeline: first validate whether attributions are stable and object-centred, then apply the appropriate filter. Future work should explore: (i) extracting the 3D reference from validation data, enabling texture-keypoint objects; (ii) improving attribution precision through alternative methods like SmoothGrad [10], Score-CAM [9], LRP [7], and attention mechanisms [8]; and (iii) extending evaluation to other objects.

Acknowledgment. This work has been carried out within the framework of project GUARDIANES with grant number CER-20251017, funded by the *Centro para el Desarrollo Tecnológico Industrial* (CDTI).

References

[1] O. Del-Tejo-Catala *et al.*, "Probabilistic pose estimation from multiple hypotheses", *IEEE Access*, vol. 11, no. April, pp. 64 507–64 517, 2023.

[2] T. I. Amosa *et al.*, "Multi-camera multi-object tracking: A review of current trends and future advances", *Neurocomputing*, vol. 552, p. 126 558, 2023.

[3] R. R. Selvaraju *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization", in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[4] T. J. Springenberg *et al.*, "Striving for simplicity: The all convolutional net", *CoRR*, vol. abs/1412.6806, 2014.

[5] M. Sundararajan *et al.*, "Axiomatic Attribution for Deep Networks", in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 3319–3328.

[6] A. Shrikumar *et al.*, "Learning Important Features Through Propagating Activation Differences", in *Proceedings of the 34th International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 3145–3153.

[7] A. Binder *et al.*, "Layer-wise relevance propagation for neural networks with local renormalization layers", Apr. 2016.

[8] H. Zhang *et al.*, "Diverse Attention for Explanations and Robustness", in *International Conference on Learning Representations (ICLR)*, 2021.

[9] M. Chen *et al.*, "Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks", in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[10] D. Smilkov *et al.*, "SmoothGrad: removing noise by adding noise", *arXiv preprint arXiv:1706.03725*, 2017.

[11] M. Ribeiro *et al.*, ""why should i trust you?": Explaining the predictions of any classifier", in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[12] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions", *arXiv preprint arXiv:1705.07874*, 2017.

[13] C. Kantor *et al.*, "Over-map: Structural attention mechanism and automated semantic segmentation ensembled for uncertainty prediction", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, pp. 15 316–15 322, May 2021.

[14] Q. He *et al.*, "Analyzing and diagnosing pose estimation with attributions", in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 4821–4830.

[15] M.-F. Li *et al.*, "Ua-pose: Uncertainty-aware 6d object pose estimation and online object completion with partial references", in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 1180–1189.

[16] J. Adebayo *et al.*, "Sanity checks for saliency maps", *Advances in Neural Information Processing Systems*, vol. 31, no. NeurIPS, pp. 9505–9515, 2018.

[17] P. J. Kindermans *et al.*, "The (un)reliability of saliency methods", *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11700, pp. 267–280, 2019.

[18] A. Ghorbani *et al.*, "Interpretation of neural networks is fragile", *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 3681–3688, 2019.

[19] O. Del-Tejo-Catala *et al.*, "Synthetic-real domain adaptation for probabilistic pose estimation", *Computer Science Research Notes*, vol. 31, no. 1-2, pp. 127–136, 2023.

[20] Y. Xiang *et al.*, *PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes*, Nov. 2018.

[21] W. Kehl *et al.*, "SSD-6D: Making RGB-Based 3D Detection and 6D Pose Estimation Great Again", in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, Nov. 2017, pp. 1530–1538.