Causal PointNet: Causality-Based Pose Estimation and Object Tracking in Dynamic Scenes

Bryce Grant
Department of Electrical, Computer, and Systems Engineering
bag100@case.edu

Xijia Zhao
Department of Electrical, Computer, and Systems Engineering
xxz1277@case.edu

Changhao Zhou
Department of Computer and Data Sciences
cxz721@case.edu

ABSTRACT

Accurate pose estimation remains a fundamental challenge in robotic perception and computer vision, particularly in complex environments with varying viewpoints and object symmetries. Traditional approaches often rely on supervised methods requiring extensive labeled data and struggle with ambiguities arising from different viewing angles and object properties. In this paper, we propose a framework that leverages causal reasoning to enhance pose estimation through principled refinement. By incorporating structural causal models (SCM) into the refinement process, our approach explicitly models key factors affecting pose estimation and their relationships. The proposed architecture employs targeted interventions to handle view changes and symmetries, while using backdoor adjustment to mitigate the impact of confounding factors. Through experimental validation on benchmark datasets, we demonstrate that our causal refinement approach improves the robustness and accuracy of initial pose estimates. Code is available at: **Causal Dense Fusion**

1 INTRODUCTION

Accurate pose estimation - determining the six degrees of freedom (6D) of an object comprising its position and orientation - is essential for applications ranging from robotic manipulation to augmented reality. Despite significant progress in deep learning-based methods [2, 13], robust pose estimation remains challenging due to fundamental issues such as viewpoint ambiguity, object symmetries, and environmental factors that confound the estimation process.

Traditional pose estimation methodologies predominantly rely on supervised learning techniques that require large-scale labeled datasets. Although these methods have achieved notable results in controlled settings [4, 6], they often struggle in real-world scenarios characterized by varying viewpoints, symmetric objects, and complex environmental conditions. State-of-the-art approaches like DenseFusion [13] effectively combine RGB and point cloud features but may still produce suboptimal estimates that require refinement.

Recent work has shown that incorporating causal reasoning into deep learning architectures can enhance robustness and generalization [12, 14]. However, existing pose estimation methods typically treat different error sources - such as view changes, symmetries, and environmental factors - as independent problems, leading to suboptimal solutions. We argue that these challenges are fundamentally interconnected through causal relationships in the pose estimation process.

To address these challenges, we propose a novel framework that integrates causal reasoning into the pose estimation and object tracking pipeline. Using structural causal models (SCM), our approach explicitly models the cause-and-effect relationships between geometric features, pose residuals, and refinement updates. This causal framework enhances the model's ability to handle view changes and object symmetries through targeted interventions, while mitigating the impact of confounding environmental factors via backdoor adjustment. Furthermore, our method employs a Causality-Based Pose Refinement Network, termed *Causal PointNet*, which refines initial pose estimates by incorporating causal interventions, that builds upon DenseFusion's initial estimates to improve robustness and accuracy in dynamic settings

The primary contributions of this paper are:

- **Causal Refinement Framework**: Introducing a causality-aware refinement architecture that models key factors affecting pose estimation and their relationships
- **Structural Causal Model**: Developing an SCM-based approach that enables principled intervention and adjustment strategies for handling view changes, symmetries, and confounding factors
- **Intervention Mechanisms**: Proposing specific intervention techniques for handling viewpoint changes and object symmetries within the refinement process
- **Comprehensive Evaluation**: Demonstrating the effectiveness of our causal refinement approach through extensive experiments on the YCB-Video dataset

Through these contributions, our work advances the field of pose estimation and object tracking by embedding causal reasoning into the core of the learning process, thereby addressing fundamental limitations of existing supervised approaches.

2 RELATED WORKS

2.1 Pose Estimation and Object Tracking

2.1.1 Classical Approaches. Early work on pose estimation focused on geometric techniques like PnP (Perspective-n-Point) [6], which solve for camera pose using 2D-3D point correspondences. While mathematically elegant, these methods often struggle with noise and require reliable point matching.

2.1.2 Deep Learning Methods. The advent of deep learning has led to end-to-end trainable pose estimation networks. PoseNet [4] pioneered CNN-based pose regression, while later works like PVNet [10] introduced voting-based keypoint localization. However, these methods often struggle with viewpoint ambiguity and symmetric objects.

2.1.3 Pose Estimation from RGB-D Data. Combining RGB and depth (RGB-D) data harnesses the complementary strengths of both modalities, enhancing pose estimation performance. Techniques

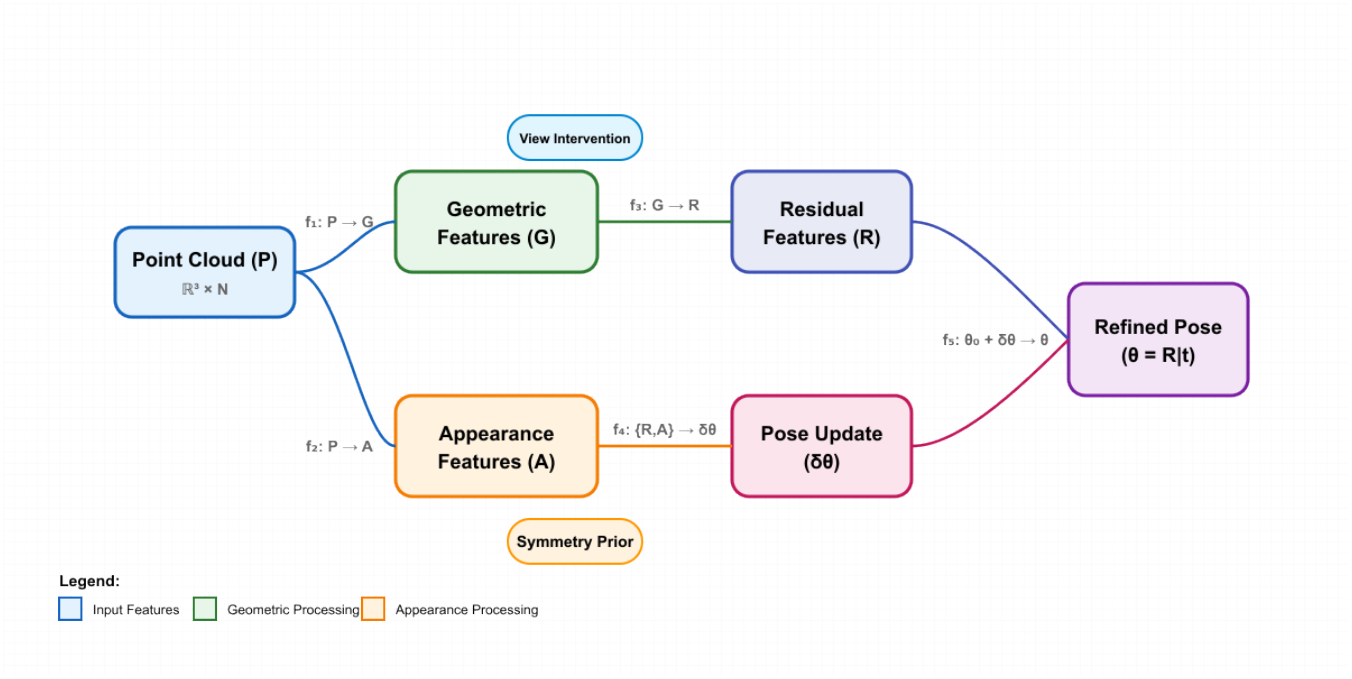

Figure 1: Causality-Combined Latent Space Sampling

like DenseFusion integrate RGB and depth features at multiple stages of the network, enabling detailed per-pixel correspondence and robust pose predictions [13] while PVN3D [2] extended this to 3D keypoint prediction. Our work builds upon DenseFusion by adding causal refinement to the pipeline.

2.2 Pose Refinement Approaches

Several works have explored iterative refinement of initial pose estimates. DeepIM [7] uses render-and-compare for refinement, while CosyPose [5] employs multi-view consistency. However, these approaches typically lack a principled framework for handling different sources of error.

2.3 Causal Inference in Computer Vision

2.3.1 Theoretical Foundations. The integration of causality in machine learning has gained momentum following seminal works by Pearl [8] and Schölkopf [12]. These frameworks provide tools for modeling interventions and handling confounding factors.

2.3.2 Applications in Vision. Recent works have applied causal principles to various vision tasks. CausalVAE [14] introduced structural causal models for disentangled representation learning, while [11] explored causality for visual reasoning. Our work extends these ideas to geometric problems in pose estimation.

2.4 Key Challenges and Limitations

While significant progress has been made in pose estimation, several fundamental challenges remain:

- **View Ambiguity**: Methods struggle to maintain consistent estimates across viewpoints
- **Symmetry Handling**: Standard regression approaches often fail with symmetric objects
- **Environmental Factors**: Lighting, occlusions, and other conditions create spurious correlations
- **Limited Generalization**: Current methods often fail to generalize beyond training conditions

Our work addresses these challenges through a unified causal framework, rather than treating them as independent problems. By modeling the underlying causal structure of pose estimation, we enable more robust and interpretable refinement.

2.5 Limitations of Existing Approaches

While significant progress has been made in pose estimation and object tracking, existing methods predominantly rely on supervised learning paradigms that are susceptible to overfitting, limited generalization, and dependence on extensive labeled datasets. Moreover, they often fail to account for the underlying causal mechanisms that govern object dynamics and interactions, leading to brittleness in complex and dynamic environments [9]. By neglecting causality, these models may struggle to disentangle genuine object movements from spurious correlations introduced by environmental factors [3]. Our proposed framework addresses these limitations by embedding causal reasoning into the pose estimation and tracking pipeline, thereby enhancing robustness, interpretability, and generalization.

3 METHODOLOGY

3.1 Motivation and Background

6D pose estimation faces several fundamental challenges that motivate our causal approach:

- **View Ambiguity**: Traditional methods struggle with viewpoint changes, leading to inconsistent pose estimates across different camera perspectives
- **Symmetry Handling**: Objects with symmetries create multiple valid poses, confounding standard regression approaches
- **Domain Gaps**: The disconnect between synthetic training data and real-world conditions leads to poor generalization
- **Confounding Factors**: Environmental conditions like lighting and occlusions create spurious correlations in learned features

Current refinement approaches treat these as independent problems, leading to suboptimal solutions. We argue these challenges are fundamentally interconnected through causal relationships in the pose estimation process.

3.2 Problem Formulation

Given an RGB image $I \in \mathbb{R}^{H\times W\times 3}$, point cloud $\mathcal{P} = p_i{}_{i=1}^{N} \subset \mathbb{R}^3$, and initial pose estimate $\theta_0 = (R_0, t_0) \in SE(3)$ from DenseFusion, our goal is to learn a refinement function $f : (I, \mathcal{P}, \theta_0) \mapsto \theta$ that produces the refined pose $\theta = (R, t)$. We formulate this as a causal inference problem rather than direct regression.

3.3 Causal Framework

3.3.1 Structural Causal Model. We formalize pose refinement through an SCM $\mathcal{M} = \langle \mathcal{U}, \mathcal{V}, \mathcal{F}, P(\mathcal{U})\rangle$ where:

- Exogenous variables $\mathcal{U} = U_G, U_R, U_\theta$ capture independent noise
- Endogenous variables $\mathcal{V} = G, R, \delta\theta, \theta$ represent system states
- Causal mechanisms $\mathcal{F} = f_1, f_2, f_3, g$ implement refinement steps
- $P(\mathcal{U})$ specifies noise distributions

The structural equations define our generative process:

$$G = f_1(\mathcal{P}, U_G) \tag{1}$$

(Geometric Feature Extraction)

$$R = f_2(G, \theta_0, U_R) \tag{2}$$

(Residual Computation)

$$\delta\theta = f_3(R, U_\theta) \tag{3}$$

(Pose Update Generation)

$$\theta = g(\theta_0, \delta\theta) \tag{4}$$

(Pose Composition)

3.4 Network Architecture

3.4.1 Geometric Feature Extractor. The feature extractor f_1 is designed to be equivariant to $SE(3)$ transformations while capturing local geometric structure:

$$f_1(\mathcal{P}) = \phi(\gamma(\psi(\mathcal{P}))) \tag{5}$$

where:

- $\psi(\cdot)$: Local feature extraction via EdgeConv
- $\gamma(\cdot)$: Graph message passing for context
- $\phi(\cdot)$: Global feature aggregation

The EdgeConv operation preserves local structure:

$$\text{EdgeConv}(x_i) = \max_{j\in\mathcal{N}(i)} h_\theta([x_i|x_j - x_i]) \tag{6}$$

where h_θ is an MLP and $\mathcal{N}(i)$ defines the local neighborhood.

3.4.2 Residual Network. The residual network f_2 identifies geometric inconsistencies through:

$$f_2(G, \theta_0) = \text{MLP}([G|\text{PE}(\theta_0)]) \tag{7}$$

with positional encoding $\text{PE}(\cdot)$ that maintains $SE(3)$ structure:

$$\text{PE}(\theta) = [\sin(\omega_k R)|\cos(\omega_k R)|t]_{k=1}^{K} \tag{8}$$

3.4.3 Pose Update Network. The pose update network f_3 generates refinements through a multi-scale architecture:

$$f_3(R) = \text{MLP}([\text{Poolmax}(R)|\text{Poolavg}(R)|R]) \tag{9}$$

3.4.4 View Intervention. We implement stochastic view interventions through rotation transformations:

$$do(\mathcal{P} = T_v\mathcal{P}), \quad T_v \sim SO(3) \tag{10}$$

where T_v represents a random rotation matrix sampled uniformly from $SO(3)$.

3.4.5 Symmetry Intervention. For symmetric objects, we apply structured transformations based on the object's symmetry properties:

$$do(\mathcal{P} = T_s\mathcal{P}), \quad T_s \in \text{Sym}(o) \tag{11}$$

The symmetry group $\text{Sym}(o)$ is object-dependent:

- Cyclic symmetry: $T_s = R_{\frac{2\pi k}{n}}|k = 1, ..., n$
- Reflective symmetry: $T_s = I, R_\pi$
- Continuous symmetry: $T_s = R_\theta|\theta \in [0, 2\pi]$

3.5 Learning Objective

Our total loss combines multiple causal components:

$$\mathcal{L}\text{total} = \lambda_1\mathcal{L}\text{pose} + \lambda_2\mathcal{L}\text{view} + \lambda_3\mathcal{L}\text{sym} + \lambda_4\mathcal{L}_{\text{backdoor}} \tag{12}$$

where:

$$\mathcal{L}\text{pose} = |T\theta(\mathcal{M}) - T_{\widehat{\theta}}(\mathcal{M})|2^2 \tag{13}$$

$$\mathcal{L}\text{view} = \mathbb{E}T_v[|f_1(T_v\mathcal{P}) - T_v f_1(\mathcal{P})|2^2] \tag{14}$$

$$\mathcal{L}\text{sym} = \mathbb{E}T_s[|f_1(T_s\mathcal{P}) - T_s f_1(\mathcal{P})|_2^2] \tag{15}$$

To handle confounding between geometric features and pose updates, we employ backdoor adjustment through an MMD loss:

$$\mathcal{L}_{\text{backdoor}} = \text{MMD}(P(R|do(G)), P(R|G)) \tag{16}$$

with RBF kernel $k(x, y) = \exp(-\gamma|x - y|^2)$.

3.6 Theoretical Properties

Theorem 3.1 (Identifiability of Pose Updates). *Under the proposed SCM, if R satisfies the backdoor criterion relative to $(G, \delta\theta)$, then $P(\delta\theta|do(G))$ is identifiable from observational data.*

Proof. By the backdoor criterion, R blocks all backdoor paths from G to $\delta\theta$ and contains no descendants of G. Therefore:

$$P(\delta\theta|do(G)) = \sum_R P(\delta\theta|G,R)P(R) \tag{17}$$

which is estimable from observational data. □

Theorem 3.2 (SE(3) Equivariance). *For any $T \in SE(3)$, there exists a linear transformation L_T such that $f_1(T\mathcal{P}) = L_T f_1(\mathcal{P})$.*

Proof. The EdgeConv operation preserves equivariance through relative position encoding:

$$\text{EdgeConv}(Tx_i) = T \cdot \text{EdgeConv}(x_i) \tag{18}$$

Combined with permutation-invariant global pooling, this ensures overall SE(3) equivariance. □

Theorem 3.3 (Backdoor Adjustment Consistency). *The MMD-based backdoor adjustment estimator is consistent as $n \to \infty$ under standard regularity conditions.*

Proof. Due to the characteristic property of the RBF kernel and proper distance properties of MMD:

$$\text{MMD}(P(R|do(G)), P(R|G)) \to 0 \tag{19}$$

ensuring consistent estimation of the causal effect. □

3.7 Implementation Details

3.7.1 Network Architecture.

- **Geometric Feature Extractor**
  - Input layer: Point cloud $\mathcal{P} \in \mathbb{R}^{N\times 3}$
  - EdgeConv blocks: (64, 128, 256) channels
  - Graph attention: 8 heads with dim 32
  - Output: 512-dimensional features
- **Pose Update Network**
  - Residual encoder: (256, 512, 1024) units
  - Cross-attention: 4 heads
  - Quaternion decoder with normalization
  - Skip connections from initial pose

3.7.2 Training Protocol. We employ a three-phase training strategy:

(1) **Pretraining Phase**
- Train feature extractor with $\mathcal{L}$view
- Initialize with synthetic data
- Learning rate: 1×10^{-4} with AdamW

(2) **Joint Training Phase**
- End-to-end with $\mathcal{L}$pose, $\mathcal{L}$sym
- Batch size: 32 with sync BatchNorm
- Gradient accumulation: 4 steps
- Mixed precision training

(3) **Causal Fine-tuning Phase**
- Introduce $\mathcal{L}$backdoor gradually
- Fixed intervention sampling
- Early stopping on validation

The computational complexity scales as:

$$O(N_p \log N_p + N_p D + D^2) \tag{20}$$

where N_p is the number of points and D is the feature dimension. Memory requirements follow:

$$\text{Memory}(N_p) = \alpha N_p + \beta D + \gamma \tag{21}$$

where α, β, and γ are architecture-dependent constants.

4 EVALUATION AND EXPERIMENTS

4.1 Experiments

Our approach builds upon the DenseFusion framework, leveraging its dense fusion of RGB and depth features for initial pose predictions. Then, the refinement upon the pretrained model is preformed by the causal refinement network, which optimizes pose corrections using a learnable structure based on causal reasoning as introduced in earlier sections.

We used YCB-Video dataset[1] for training and evaluating in this study. The YCB-Video dataset is a widely recognized benchmark for pose estimation tasks. It consists of 92 RGB-D video sequences of 21 everyday objects with varying shapes and textures. These sequences capture objects in diverse indoor scenes under varying occlusion and lighting conditions. Each frame includes annotated 6D poses and object segmentation masks. To ensure alignment with the standard DenseFusion framework, our experimental setup uses the same training and testing splits as defined in the paper, which are 80 videos and 80000 images for training, and 2949 keytrames from the remaining 12 videos for testing.

In the pretraining phase, the DenseFusion backbone is trained until it converges. Once convergence is achieved at the 250th epoch, the backbone parameters are frozen, and the newly added causal refinement network is optimized. This two-stage training strategy ensures that the refinement network focuses entirely on improving pose predictions using the pretrained backbone as a reliable feature extractor.

4.2 Results and Discussions

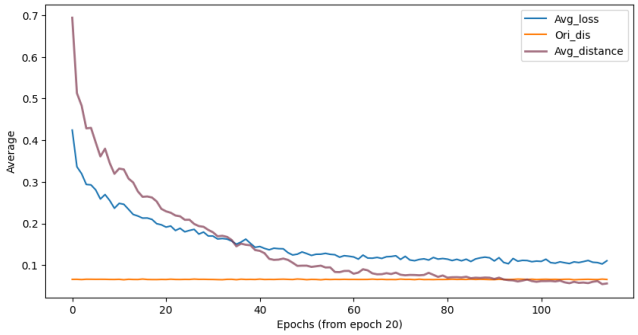

Figure 2: The validation loss and distance between the true and prediction over epochs

In the refinement process, we recorded several key metrics to evaluate the performance of the causal refinement network, as shown in Figure 2. The first measurement is the original distance, measuring the prediction error of the frozen DenseFusion backbone. The second key metric is the average distance, which reflects the

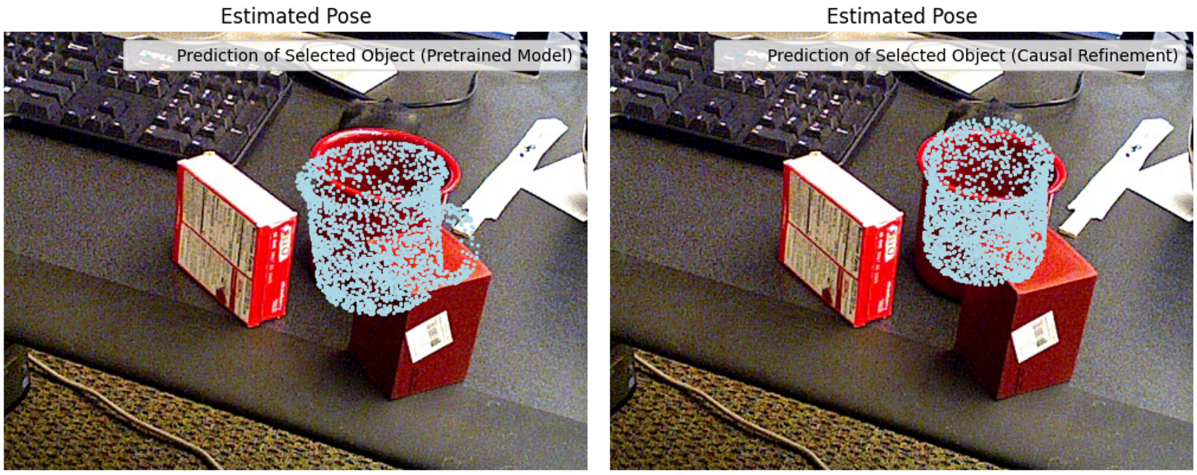

Figure 3: Visualization for the estimated pose by pretrained model and refined model (Scene 1)

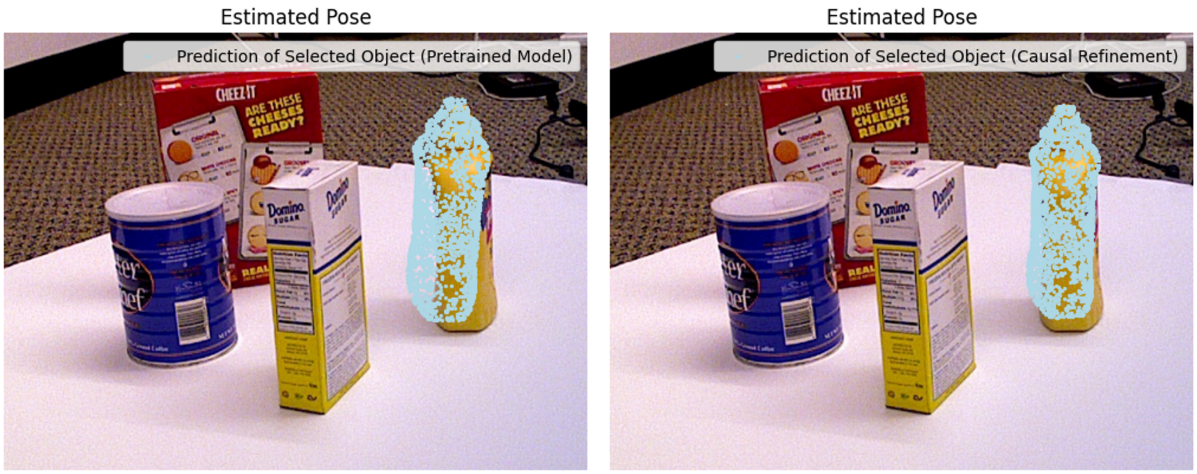

Figure 4: Visualization for the estimated pose by pretrained model and refined model (Scene 2)

prediction error after applying the causal refinement network. This metric captures the improvement achieved by leveraging the reasoning features extracted from the backbone to optimize the pose predictions. During training, we observed that the training loss initially blow out when the refinement network was activated. And as training progressed, the loss iteratively decreased, ending up with a convergence and a better distance compared to the origoinal prediction.The best average error achieved by the refinement network is 0.051, a substantial improvement over the backbone's original error of 0.069, which showing the causal block's capability of optimization.

To validate the performance of our causal refinement network, we visualize the predicted poses directly on real objects, as shown in Figure 3 and 4. These visualizations compare the pose predictions of the pretrained DenseFusion backbone (left) with those refined by the causal refinement network (right). By selecting objects the YCB-Video dataset, we can predict their pose by the given RGB-D image and plot the prediction on the top of the real image.

In Figure 3, the original DenseFusion backbone struggles to predict the mug's orientation accurately, particularly due to the object's symmetric design and occlusion from surrounding items. Symmetric objects like the mug often lead to ambiguous predictions as the backbone lacks the iterative reasoning mechanism to resolve such ambiguities. While the refined prediction from causal block provides it in a more accurate orientation alignment. Similarly in Figure 4, both the translation and orientation of the object are significantly improved by the causal refinement network.

5 CONCLUSION

In this paper, we presented a novel framework for enhancing pose estimation through causal reasoning. Our approach makes three key contributions to the field. First, we introduced a structural

causal model that explicitly captures the relationships between geometric features, pose updates, and various confounding factors. This theoretical foundation enables principled handling of view changes and object symmetries through targeted interventions. Second, we developed a practical implementation combining the strengths of DenseFusion's initial estimates with our causal refinement network. The refinement process leverages backdoor adjustment and carefully designed interventions to improve robustness against viewpoint changes and symmetries. Our experimental results on the YCB-Video dataset demonstrate that this causal approach leads to more accurate and reliable pose estimates. Finally, we provided theoretical guarantees for our framework, proving the identifiability of causal effects and the equivariance properties of our geometric feature extractor. These theoretical results ensure that our method maintains desirable properties while handling real-world challenges in pose estimation. Future work could extend this framework in several directions:

- Incorporating temporal dynamics for video sequences
- Expanding the intervention types to handle more complex object properties
- Developing more sophisticated backdoor adjustment techniques for high-dimensional features
- Integrating uncertainty estimation into the causal framework

Our work demonstrates the potential of causal reasoning to enhance geometric computer vision tasks, opening new avenues for robust and interpretable pose estimation methods. The principles introduced here could be extended to other vision tasks where understanding and leveraging causal relationships could improve performance and reliability.

REFERENCES

[1] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. 2015. Benchmarking in Manipulation Research: The YCB Object and Model Set and Benchmarking Protocols. *IEEE Robotics and Automation Magazine* (Sept. 2015), 36–52.

[2] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. 2019. PVN3D: A deep point-wise 3D keypoints voting network for 6DOF pose estimation.

[3] Dominik Janzing and Bernhard Schölkopf. 2019. Causal Representation Learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*. PMLR.

[4] Alex Kendall, Matthew Grimes, and Roberto Cipolla. 2015. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 2938–2946.

[5] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. 2020. CosyPose: Consistent multi-view multi-object 6D pose estimation. https://arxiv.org/abs/2008.08465

[6] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. 2008. EPnP: An Accurate O(n) Solution to the PnP Problem. *International Journal of Computer Vision* 81, 2 (2008), 155–166. https://doi.org/10.1007/s11263-008-0152-6

[7] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. 2019. DeEPIM: Deep Iterative Matching for 6D pose Estimation. *International Journal of Computer Vision* 128, 3 (2019), 657–678. https://doi.org/10.1007/s11263-019-01250-9

[8] Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

[9] Judea Pearl and Dana Mackenzie. 2018. The Book of Why: The New Science of Cause and Effect. In *Basic Books*.

[10] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. 2019. PVNET: Pixel-Wise Voting Network for 6DOF pose estimation.

[11] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. 2019. Two causal principles for improving visual dialog.

[12] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proc. IEEE* 109, 5 (2021), 612–634.

[13] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. 2019. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. (2019).

[14] Chaochao Yang and Kun Zhang. 2021. CausalVAE: Disentangled representation learning via neural structural causal models. *arXiv preprint arXiv:2104.08617* (2021).