Novelty Assessment Report

Paper: Learning Robust Intervention Representations with Delta Embeddings

PDF URL: https://openreview.net/pdf?id=5d7prMWHNF

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-07

Abstract

Causal representation learning has attracted significant research interest during the past few years, as a means for improving model generalization and robustness. Causal representations of interventional image pairs (also called ``actionable counterfactuals'' in the literature), have the property that only variables corresponding to scene elements affected by the intervention / action are changed between the start state and the end state. While most work in this area has focused on identifying and representing the variables of the scene under a causal model, fewer efforts have focused on representations of the interventions themselves. In this work, we show that an effective strategy for improving out of distribution (OOD) robustness is to focus on the representation of actionable counterfactuals in the latent space. Specifically, we propose that an intervention can be represented by a Causal Delta Embedding that is invariant to the visual scene and sparse in terms of the causal variables it affects. Leveraging this insight, we propose a method for learning causal representations from image pairs, without any additional supervision. Experiments in the Causal Triplet challenge demonstrate that Causal Delta Embeddings are highly effective in OOD settings, significantly exceeding baseline performance in both synthetic and real-world benchmarks.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Learning Robust Representations of Interventions from Image Pairs**

A total of **21 papers** were analyzed and organized into a taxonomy with **15 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Causal Representation Learning from Interventional Image Pairs**
- **Causal Inference for Bias Mitigation and Robustness**
- **Causal Discovery from Visual Data**
- **Counterfactual Image Generation and Manipulation**
- **Algorithmic Bias Measurement via Controlled Experiments**
- **Causal Inference for Cross-Domain Adaptation**
- **Implicit Relation Modeling in Multimodal Matching**
- **Deformable Image Registration**
- **Template Matching for Medical Interventions**
- **Causal Role Studies in Psychology and Neuroscience**

Complete Taxonomy Tree

- Learning Robust Representations of Interventions from Image Pairs Survey Taxonomy
- Causal Representation Learning from Interventional Image Pairs
  - Intervention-Centric Causal Embeddings ★ (2 papers)
  - [0] Learning Robust Intervention Representations with Delta Embeddings (Anon et al., 2026) View paper
  - [21] Causal Triplet: An Open Challenge for Intervention-centric Causal Representation Learning (Liu Yue-jiang, 2023) View paper
  - Weakly Supervised Causal Variable Identification (1 papers)
  - [4] Weakly supervised causal representation learning (Brehmer, 2022) View paper
- Causal Inference for Bias Mitigation and Robustness
  - Confounder Removal via Causal Intervention (3 papers)
  - [1] Clothes-invariant feature learning by causal intervention for clothes-changing person re-identification (Li Xulin, 2023) View paper
  - [5] Intra-and Inter-Image Causal Intervention for Robust Semantic Segmentation in Remote-Sensing Images (Lei Yu, 2024) View paper
  - [14] Learning High-Order Features for Fine-Grained Visual Categorization with Causal Inference (Yuhang Zhang, 2025) View paper
  - Debiasing via Backdoor Adjustment (1 papers)
  - [10] CABIN: Debiasing Vision-Language Models Using Backdoor Adjustments (Pang Bo, 2025) View paper
  - Synthetic Data Augmentation for Robustness (1 papers)
  - [8] Not Just Pretty Pictures: Text-to-Image Generators Enable Interpretable Interventions for Robust Representations (Yuan, 2022) View paper
- Causal Discovery from Visual Data
  - Pairwise Causal Direction Classification (1 papers)
  - [3] Discovering causal signals in images (David Lopezâ□□Paz, 2017) View paper
  - Causal Graph Construction from Pairwise Features (1 papers)
  - [13] From Causal Pairs to Causal Graphs (Rezaur Rashid, 2022) View paper

- Counterfactual Image Generation and Manipulation
    - Structural Causal Model-Based Counterfactual Generation (2 papers)
    - [7] Benchmarking counterfactual image generation (Nefeli Gkouti, 2024) View paper
    - [9] Counterfactual Generative Modeling with Variational Causal Inference (Wu, 2024) View paper
    - Affective Image Manipulation (1 papers)
    - [2] Emoedit: Evoking emotions through image manipulation (Jingyuan Yang, 2025) View paper
- Algorithmic Bias Measurement via Controlled Experiments (1 papers)
    - [6] Benchmarking algorithmic bias in face recognition: An experimental approach using synthetic faces and human evaluation (Hao Liang, 2023) View paper
- Causal Inference for Cross-Domain Adaptation (1 papers)
    - [18] Spatial-temporal Causal Inference for Partial Image-to-video Adaptation (Chen Jin, 2021) View paper
- Implicit Relation Modeling in Multimodal Matching (1 papers)
    - [19] How to Understand" Support"? An Implicit-enhanced Causal Inference Approach for Weakly-supervised Phrase Grounding (Luo Jiamin, 2024) View paper
- Deformable Image Registration (2 papers)
    - [11] Robust non-rigid registration through agent-based action learning (Julian Krebs, 2017) View paper
    - [12] Deformable image registration based on similarity-steered CNN regression (Xiaohuan Cao, 2017) View paper
- Template Matching for Medical Interventions (1 papers)
    - [20] Toward Robust Partial-Image Based Template Matching Techniques for MRI-Guided Interventions. (Eung Joo Lee, 2023) View paper
- Causal Role Studies in Psychology and Neuroscience (3 papers)
    - [15] Self-images play a causal role in social phobia (C. Hirsch, 2003) View paper
    - [16] Causal inference and the evolution of opposite neurons (S. Badde, 2021) View paper
    - [17] The nature of phonological processing and its causal role in the acquisition of reading skills. (Richard K. Wagner, 1987) View paper

Narrative

Core task: learning robust representations of interventions from image pairs. The field encompasses diverse approaches to understanding how images change under interventions, spanning causal representation learning, bias mitigation, discovery methods, and counterfactual generation. The taxonomy reveals several major branches: some focus on extracting causal structure directly from visual data (e.g., Causal Representation Learning from Interventional Image Pairs, Causal Discovery from Visual Data), while others emphasize generating or manipulating images to reflect hypothetical changes (Counterfactual Image Generation and Manipulation). Additional branches address practical concerns such as measuring algorithmic bias through controlled experiments, adapting models across domains using causal principles, and specialized applications in medical image registration or psychological studies. Works like Causal Signals Images[3] and Weakly Supervised Causal[4] illustrate early efforts to identify causal relationships in visual settings, whereas recent methods such as Causal Intervention Segmentation[5] and Counterfactual Generative Modeling[9] demonstrate growing sophistication in leveraging interventional data for downstream tasks.

Within the intervention-centric causal embeddings cluster, a key theme is how to encode the effect of an intervention—rather than just the before-and-after states—into a reusable representation. Delta Embeddings[0] sits squarely in this line of work, emphasizing robust encoding of intervention effects from paired images. This contrasts with nearby efforts like Causal Triplet[21], which also explores intervention representations but may differ in architectural choices or the granularity of causal assumptions. Across related branches, open questions persist around disentangling confounders from true causal effects, scaling to complex real-world scenarios with limited supervision, and bridging the gap between controlled experimental settings (as in Benchmarking Algorithmic Bias[6]) and naturalistic image distributions. The original paper contributes to this active area by proposing methods that prioritize intervention robustness, positioning it among works that treat interventions as first-class objects worthy of their own learned embeddings.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Causal Triplet: An Open Challenge for Intervention-centric Causal Representation Learning

Authors: Liu Yue-jiang, Alahi, Alexandre, Russell, Chris, et al. (13 authors total) | **Year/Venue**: 2023 | **URL**: View paper

Abstract

Recent years have seen a surge of interest in learning high-level causal representations from low-level image pairs under interventions. Yet, existing efforts are largely limited to simple synthetic settings that are far away from real-world problems. In this paper, we present Causal Triplet, a causal representation learning benchmark featuring not only visually more complex scenes, but also two crucial desiderata commonly overlooked in previous works: (i) an actionable counterfactual setting, w...

Relationship Analysis

Both papers belong to the Intervention-Centric Causal Embeddings category, focusing on representing interventions as embeddings from image pairs. They overlap in addressing the core task of learning robust intervention representations that are invariant to scene content and sparse in affected causal variables. The key difference is that the original paper (Delta Embeddings) proposes a specific method using delta embeddings with contrastive and sparsity losses to learn intervention representations, while the candidate paper (Causal Triplet) introduces a benchmark challenge for evaluating intervention-centric causal representation learning methods across synthetic and real-world datasets with compositional and systematic distribution shifts.

Contributions Analysis

Overall novelty summary. The paper proposes Causal Delta Embeddings (CDE) to represent interventions as scene-invariant, sparse transformations in latent space, learning from image pairs without additional supervision. It resides in the 'Intervention-Centric Causal Embeddings' leaf, which contains only two papers total (including this work and one sibling). This represents a relatively sparse research direction within the broader taxonomy of 21 papers across causal representation learning, suggesting the specific focus on intervention embeddings rather than causal variable identification remains underexplored.

The taxonomy reveals that most neighboring work concentrates on identifying causal variables from paired data (Weakly Supervised Causal Variable Identification) or applying causal reasoning for bias mitigation and robustness. The sibling paper in the same leaf likely shares the intervention-centric perspective but may differ in architectural or methodological details. Nearby branches address counterfactual generation using structural causal models and confounder removal via intervention modeling, indicating the field has explored related but distinct angles—generating counterfactual images versus learning reusable intervention representations.

Among 23 candidates examined across three contributions, none were flagged as clearly refuting the proposed work. The CDE framework examined 10 candidates with zero refutable overlaps, the multi-objective loss examined 3 candidates with zero refutations, and the patch-wise extension examined 10 candidates with zero refutations. This limited search scope suggests that within the top-K semantic matches and citation expansion, no prior work directly anticipates the combination of scene-invariant intervention embeddings with unsupervised learning from image pairs, though the analysis does not claim exhaustive coverage of all relevant literature.

Based on the available signals, the work appears to occupy a relatively novel position within a sparse research direction, though the literature search examined only 23 candidates. The taxonomy structure and contribution-level statistics indicate limited direct prior work on intervention-centric embeddings, but broader themes around causal representation learning and counterfactual reasoning are well-established in neighboring branches. A more comprehensive search beyond top-K semantic matches would be needed to fully assess novelty across the entire field.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Causal Delta Embedding (CDE) framework

Description: The authors propose a framework that represents interventions as delta vectors in latent space, satisfying properties of independence, sparsity, and object invariance. This enables robust generalization to out-of-distribution samples by learning intervention representations that are invariant to visual scene context.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Weakly supervised causal representation learning

URL: View paper

Brief Assessment

Weakly Supervised Causal[4] focuses on learning causal variables and graph structure from intervention pairs using variational autoencoders, not on learning intervention representations as delta vectors in latent space. The candidate addresses causal representation learning of system states, while the original addresses learning representations of the interventions themselves.

2. Counterfactual explanations as interventions in latent space

URL: View paper

Brief Assessment

Counterfactual Latent Interventions[35] focuses on generating counterfactual explanations for model predictions in feature space by learning interventions in a latent space derived from structural causal models. The ORIGINAL paper's CDE framework learns delta embeddings representing interventions between pre/post-intervention image pairs for action recognition, which is a fundamentally different task and application domain.

3. Counterfactual image editing with disentangled causal latent space

URL: View paper

Brief Assessment

Counterfactual Causal Latent[34] focuses on counterfactual image editing using backdoor disentangled causal latent spaces for pixel-level image manipulation, not on learning generalizable intervention representations in latent space for action prediction from image pairs as in the original paper.

4. Interventional causal representation learning

URL: View paper

Brief Assessment

Interventional Causal Learning[37] focuses on identifying latent causal variables from interventional data using geometric signatures of support independence, not on learning delta vector representations of interventions in latent space for robust generalization to OOD samples.

5. Drivedreamer: Towards real-world-drive world models for autonomous driving

URL: View paper

Brief Assessment

Drivedreamer[36] focuses on world models for autonomous driving video generation using diffusion models, not on learning generalizable disentangled representations of interventions in latent space for causal representation learning.

6. Universal visual decomposer: Long-horizon manipulation made easy

URL: View paper

Brief Assessment

Universal Visual Decomposer[40] focuses on task decomposition for long-horizon manipulation using pre-trained visual representations to discover subgoals, not on learning generalizable disentangled representations of interventions in latent space as delta vectors with independence, sparsity, and object invariance properties.

7. Linear causal disentanglement via interventions

URL: View paper

Brief Assessment

Linear Causal Disentanglement[32] focuses on identifiability of linear causal models from interventional data using matrix decomposition techniques, not on learning intervention representations as delta vectors in latent space for visual generalization tasks.

8. Nonparametric identifiability of causal representations from unknown interventions

URL: View paper

Brief Assessment

Nonparametric Causal Identifiability[38] focuses on identifying latent causal variables and their causal graphs from interventional data in a nonparametric setting, not on learning intervention representations as delta vectors in latent space for robust generalization.

9. Identifiability guarantees for causal disentanglement from soft interventions

URL: View paper

Brief Assessment

Soft Interventions Identifiability[33] focuses on identifying latent causal variables and their causal structure from interventional data in genomics, not on learning intervention representations as delta vectors in latent space for visual action recognition tasks.

10. Learning to Decompose and Disentangle Representations for Video Prediction

URL: View paper

Brief Assessment

Decompose Disentangle Video[39] focuses on decomposing video frames into spatial components (individual objects) and disentangling appearance from motion for video prediction tasks. The ORIGINAL paper addresses learning intervention representations in latent space for causal reasoning across distribution shifts, which is a fundamentally different problem domain.

Contribution 2: Multi-objective loss function for learning causal representations

Description: The authors design a training objective combining cross-entropy loss, supervised contrastive loss, and sparsity regularization to enforce the desired properties of Causal Delta Embeddings. This loss function enables learning intervention representations from image pairs without additional supervision.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Invariant causal representation learning for out-of-distribution generalization

URL: View paper

Brief Assessment

Invariant Causal Representation[41] focuses on identifying latent causal variables from observations using exponential family priors and score matching, not on multi-objective loss functions combining cross-entropy, contrastive, and sparsity losses for intervention representations from image pairs.

2. DGCDN: robust acoustic fault diagnosis via domain-generalized causal disentanglement

URL: View paper

Brief Assessment

DGCDN[43] focuses on acoustic fault diagnosis in industrial machinery using domain generalization, not on learning intervention representations from image pairs or causal delta embeddings as in the original paper.

3. Towards robust and adaptive motion forecasting: A causal representation perspective

URL: View paper

Brief Assessment

Robust Motion Forecasting[42] uses a multi-objective loss combining task loss, invariant loss, and style contrastive loss for motion forecasting, not for learning sparse object-invariant causal representations from image pairs as in the original paper.

Contribution 3: Patch-wise extension for multi-object scenes

Description: The authors extend their global CDE model to handle complex multi-object scenes by computing delta embeddings at the patch level and aggregating the top-K patches with largest changes. This architectural extension addresses scenarios where interventions affect only localized image regions.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Semantic-decoupled Spatial Partition Guided Point-supervised Oriented Object Detection

URL: View paper

Brief Assessment

Semantic Decoupled Partition[28] addresses oriented object detection in remote sensing with spatial partitioning for instance segmentation, not causal intervention modeling in multi-object scenes. The technical domains and objectives are fundamentally different.

2. Dynamic-eDiTor: Training-Free Text-Driven 4D Scene Editing with Multimodal Diffusion Transformer

URL: View paper

Brief Assessment

Dynamic eDiTor[31] focuses on 4D scene editing with spatio-temporal propagation across multi-view videos, not on patch-wise models for causal representation learning or interventional scene understanding as in the original paper's multi-object extension.

3. ASIMO: Agent-centric scene representation in multi-object manipulation

URL: View paper

Brief Assessment

ASIMO[23] focuses on scene decomposition for vision-based RL in multi-object manipulation tasks, not on patch-wise delta embeddings for causal representation learning from intervention pairs.

4. A Local-to-Global Approach to Multi-modal Movie Scene Segmentation

URL: View paper

Brief Assessment

Local to Global Segmentation[26] addresses movie scene segmentation using spatial patch features for video understanding, not causal representation learning with interventional image pairs. The patch-wise approach serves different purposes in different domains.

5. Eligen: Entity-level controlled image generation with regional attention

URL: View paper

Brief Assessment

Eligen[22] focuses on entity-level image generation using regional attention mechanisms in diffusion transformers for spatial control, not on causal representation learning or patch-wise delta embeddings for intervention modeling in multi-object scenes.

6. PRISM: Progressive Restoration for Scene Graph-based Image Manipulation

URL: View paper

Brief Assessment

PRISM[30] focuses on progressive image manipulation using scene graphs with a multi-head decoder architecture for object-level detail generation, not on patch-wise delta embeddings for causal intervention representation in multi-object scenes.

7. Comprehensive Visual Question Answering on Point Clouds through Compositional Scene Manipulation

URL: View paper

Brief Assessment

Point Cloud VQA[24] focuses on 3D point cloud scene understanding with compositional scene manipulation for VQA tasks, not patch-wise visual representations for intervention-based causal learning in 2D images.

8. Graph-to-3D: End-to-End Generation and Manipulation of 3D Scenes Using Scene Graphs

URL: View paper

Brief Assessment

Graph to 3D[27] focuses on 3D scene generation from scene graphs using patch-wise features for spatial localization in 3D scenes, not on causal representation learning with interventional image pairs or delta embeddings for multi-object scenes.

9. Remote Sensing Scene Classification via Multi-Branch Local Attention Network

URL: View paper

Brief Assessment

Multi Branch Attention[25] focuses on attention mechanisms for remote sensing scene classification, not on patch-wise causal delta embeddings or intervention-based representation learning for multi-object scenes.

10. Moving object detection in complex scene using spatiotemporal structured-sparse RPCA

URL: View paper

Brief Assessment

Spatiotemporal Sparse RPCA[29] addresses moving object detection using superpixel-based spatial regularization in RPCA, not patch-wise delta embeddings for causal intervention representation in multi-object scenes.

Appendix: Text Similarity Detection

Textual similarity detection checked 24 papers and found 3 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Causal Triplet: An Open Challenge for Intervention-centric Causal Representation Learning

Detected in: Core Task (sibling)

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] Learning Robust Intervention Representations with Delta Embeddings View paper
- [1] Clothes-invariant feature learning by causal intervention for clothes-changing person re-identification View paper
- [2] Emoedit: Evoking emotions through image manipulation View paper
- [3] Discovering causal signals in images View paper
- [4] Weakly supervised causal representation learning View paper
- [5] Intra-and Inter-Image Causal Intervention for Robust Semantic Segmentation in Remote-Sensing Images View paper
- [6] Benchmarking algorithmic bias in face recognition: An experimental approach using synthetic faces and human evaluation View paper
- [7] Benchmarking counterfactual image generation View paper
- [8] Not Just Pretty Pictures: Text-to-Image Generators Enable Interpretable Interventions for Robust Representations View paper
- [9] Counterfactual Generative Modeling with Variational Causal Inference View paper
- [10] CABIN: Debiasing Vision-Language Models Using Backdoor Adjustments View paper
- [11] Robust non-rigid registration through agent-based action learning View paper
- [12] Deformable image registration based on similarity-steered CNN regression View paper
- [13] From Causal Pairs to Causal Graphs View paper
- [14] Learning High-Order Features for Fine-Grained Visual Categorization with Causal Inference View paper
- [15] Self-images play a causal role in social phobia View paper
- [16] Causal inference and the evolution of opposite neurons View paper
- [17] The nature of phonological processing and its causal role in the acquisition of reading skills. View paper
- [18] Spatial-temporal Causal Inference for Partial Image-to-video Adaptation View paper
- [19] How to Understand" Support"? An Implicit-enhanced Causal Inference Approach for Weakly-supervised Phrase Grounding View paper
- [20] Toward Robust Partial-Image Based Template Matching Techniques for MRI-Guided Interventions. View paper
- [21] Causal Triplet: An Open Challenge for Intervention-centric Causal Representation Learning View paper
- [22] Eligen: Entity-level controlled image generation with regional attention View paper
- [23] ASIMO: Agent-centric scene representation in multi-object manipulation View paper
- [24] Comprehensive Visual Question Answering on Point Clouds through Compositional Scene Manipulation View paper
- [25] Remote Sensing Scene Classification via Multi-Branch Local Attention Network View paper
- [26] A Local-to-Global Approach to Multi-modal Movie Scene Segmentation View paper
- [27] Graph-to-3D: End-to-End Generation and Manipulation of 3D Scenes Using Scene Graphs View paper
- [28] Semantic-decoupled Spatial Partition Guided Point-supervised Oriented Object Detection View paper
- [29] Moving object detection in complex scene using spatiotemporal structured-sparse RPCA View paper
- [30] PRISM: Progressive Restoration for Scene Graph-based Image Manipulation View paper
- [31] Dynamic-eDiTor: Training-Free Text-Driven 4D Scene Editing with Multimodal Diffusion Transformer View paper
- [32] Linear causal disentanglement via interventions View paper
- [33] Identifiability guarantees for causal disentanglement from soft interventions View paper
- [34] Counterfactual image editing with disentangled causal latent space View paper
- [35] Counterfactual explanations as interventions in latent space View paper

- [36] Drivedreamer: Towards real-world-drive world models for autonomous driving View paper
- [37] Interventional causal representation learning View paper
- [38] Nonparametric identifiability of causal representations from unknown interventions View paper
- [39] Learning to Decompose and Disentangle Representations for Video Prediction View paper
- [40] Universal visual decomposer: Long-horizon manipulation made easy View paper
- [41] Invariant causal representation learning for out-of-distribution generalization View paper
- [42] Towards robust and adaptive motion forecasting: A causal representation perspective View paper
- [43] DGCDN: robust acoustic fault diagnosis via domain-generalized causal disentanglement View paper