Treatment Learning Causal Transformer for Noisy Image Classification

Chao-Han Huck Yang*,1, Danny I-Te Hung*,2, Yi-Chieh Liu*,1, Pin- Yu Chen[3]
[1]Georgia Institute of Technology, [2]Columbia University, [3]IBM Research AI
{huckiyang, yliu3233}@gatech.edu; ih2320@columbia.edu; pin-yu.chen@ibm.com

Abstract

Current top-notch deep learning (DL) based vision models are primarily based on exploring and exploiting the inherent correlations between training data samples and their associated labels. However, a known practical challenge is their degraded performance against "noisy" data, induced by different circumstances such as spurious correlations, irrelevant contexts, domain shift, and adversarial attacks. In this work, we incorporate this binary information of "existence of noise" as treatment into image classification tasks to improve prediction accuracy by jointly estimating their treatment effects. Motivated from causal variational inference, we propose a transformer-based architecture, Treatment Learning Causal Transformer (TLT), that uses a latent generative model to estimate robust feature representations from current observational input for noise image classification. Depending on the estimated noise level (modeled as a binary treatment factor), TLT assigns the corresponding inference network trained by the designed causal loss for prediction. We also create new noisy image datasets incorporating a wide range of noise factors (e.g., object masking, style transfer, and adversarial perturbation) for performance benchmarking. The superior performance of TLT in noisy image classification is further validated by several refutation evaluation metrics. As a by-product, TLT also improves visual salience methods for perceiving noisy images.

1. Introduction

Although deep neural networks (DNNs) have surpassed human-level "accuracy" in many image recognition tasks [22, 27, 70, 88], current DNNs still implicitly rely on the *assumption* [59] on the existence of a strong correlation between training and testing data. Moreover, increasing evidence and concerns [4, 39] show that using the correlation association for prediction can be problematic against *noisy* images [90], such as pose-shifting of identical objects [4] or imperceptible perturbation [18, 41, 48]. In practice, real-world image classification often involves rich, noisy, and even chaotic contexts, intensifying the demand for generalization in the wild.

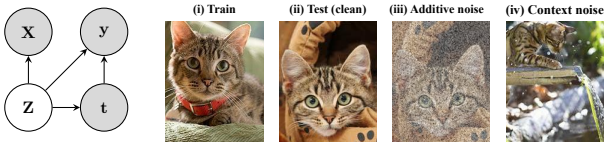


Figure 1: *(a)* An example of deployed causal graphical model (CGM), where $\mathbf{Z}$ denotes unobservable confounder variable (e.g., the concept of "cat"), $\mathbf{X}$ denotes a noisy observation of confounder (e.g., an image can still be recognized as a cat), $\mathbf{y}$ denotes outcome (e.g., a label), and $\mathbf{t}$ denotes the information of a binary treatment (e.g., the existence of extra semantic patterns or additive noise; thus, it is equal to 0 or 1), which is **observable** during **training** and **unobservable** during **testing** time. *(b)* Images with "cat" labels, where (i) and (ii) share the same context of "indoor"; (iii) shows a noisy setup of (ii) undergoing additive Gaussian perturbation; (iv) shows another setup of introducing extra noisy semantic patterns (e.g., "waterside") in NICO [23] noisy images dataset.

To address machine perception against noisy images, we are inspired by how human performs visual recognition. Learning processes of human are often mixed with logic inference (e.g., a symbolic definition from books) and representation learning (e.g., an experience of viewing a visual pattern). One prominent difference between current DNNs and human recognition systems is the capability in causal inference. Mathematically, causal learning [56, 61] is a statistical inference model that infers beliefs or probabilities under uncertain conditions, which aims to identify latent variables (called "confounders") that influence both intervention and outcome. The unobserved confounders may be abstract in a cognitive-level (e.g., concepts) but could be observed via their noisy view in the real-world (e.g., objects). For instance, as shown in Fig. 1 *(a)*, confounder learning aims to model a prediction process by finding a representation (e.g., "cat") and avoiding relying on irrelevant patterns (e.g., "waterside"). Intuitively, with causal modeling and

*The authors have equal contribution. TLT Github: https://github.com/huckiyang/treatment-causal-transformer

confounder inference, correct prediction can be made on noisy inputs, where the generative estimation process, such as causal effect variational autoencoder (CEVAE) [44], affects multiple covariates for predicting data profiles. In this work, we aim to incorporate the effects of causal confounder learning to image classification, as motivated by cognitive psychology for causal learning. Specifically, we use the attention mechanism for noise-resilience inference from patterns. We design a novel sequence-to-sequence learning model, **Treatment Learning Causal Transformer (TLT)**, which leverages upon the conditional query-based attention and the inference power from a variational causal inference model. Our TLT tackles noisy image classification by jointly learning to a generative model of $\mathbf{Z}$ and estimating the effects from the treatment information ($\mathbf{t}$), as illustrated in Fig. 1 (a). This model consists of unobservable confounder variables $\mathbf{Z}$ corresponding to the ground-truth but inaccessible information (e.g., the ontological concept [84] of a label), input data $\mathbf{X}$ from a noisy view of $\mathbf{Z}$ (e.g., images), a treatment [60] information $\mathbf{t}$ given $\mathbf{X}$ and $\mathbf{Z}$ (e.g., secondary information as visual patterns and additive noise without directly affecting our understanding the concept of "cat"), and a classification label $\mathbf{y}$ from the unobservable confounder. Built upon this causal graphical model, our contributions are:

- A transformer architecture (TLT) for noisy image classification are presented, which is based on a treatment estimation architecture and a causal variational generative model with competitive classification performance against noisy image.
- We further curated a new noisy images datasets, Causal Pairs (CPS), to study generalization under different artificial noise settings for general and medical images.
- We use formal statistical refutations tests to validate the causal effect of TLT, and show that TLT can improve visual saliency methods on noisy images.

2. Related Work

Noisy Image Classification. Prior works on noisy images classification have highlighted the importance of using generative models [54] to ameliorate the negative learning effects from noisy data. Xiao *et al.* [90] leverage a conditional generative model [79] to capture the relations among images and noise types from online shopping systems. Direct learning from noisy data is another approach by using statistical sampling [21, 37] and active learning [15] for performance enhancement. Meanwhile, new noisy images dataset and evaluation metrics [23] on context independence have been proposed, such as Strike simulator [4] for synthesizing pose-shifting images and NICO [23, 40, 95] as the open-access noisy image dataset. NICO further aims to highlight the importance of incorporating a statistical inference (e.g., causal model) for improved image classification with large-scale noisy context-patterns (e.g., an image shows "cat in waterside" but given a single label of "cat"). However, different from context-wise noise in NICO, modeling sizeable artificial noise in images is crucial yet remains unexplored. In this work, we create a new image dataset containing various artificial noise and use the NICO [23] with a generative causal model for performance benchmarking.

Causal Learning for Computer Vision. Many efforts [13, 14, 34, 62] have leveraged upon causal learning to better understand and interpret toward vision recognition tasks. Lopez-Paz *et al.* [43] propose utilizing DNNs to discover the causation between image class labels for addressing the importance of this direct causal relationship affecting model performance and context grounding. Incorporating causal analysis and regularization showed improved performance in generative adversarial models such as Causal-GANs [5, 32]. However, infusing causal modeling and inference to DNN-based image recognition systems is still an open challenge. For instance, in previous works [43, 92], researchers focus on modeling a direct causal model (DCM) [60] for visual learning. The DCMs treat a visual pattern (e.g., texture) as a causal visual-representation (e.g., patterns of the "cat") and barely incorporate additional label information (e.g., context) or apply noise as a **treatment** in causal analysis. In recent works, causal modeling also show promising results in a large-scale computer vision task, such scene graph [81] generation, visual and language learning [1, 2, 64], and semantic segmentation [94]. The work of Chalupkaet al. [10] is closer to our work by deploying interventional experiments to target causal relationships in the labeling process. However, modeling the aforementioned treatment effects and designing efficient learning models are still not fully explored [59].

Causal Inference by Autoencoder. Recently, classical causal inference tasks, such as regression modeling [8], risk estimation [59], and causal discovery [50], have been incorporated with deep generative models [69] and attained state-of-the-art performance [44, 76]. These generative models often use an encoder-decoder architecture to improve both logic inference and features extracted from a large-scale dataset with noisy observations. TARNet [76] is one foundational DNN model incorporating causal inference loss from a causal graphical model (CGM) and feature reconstruction loss jointly for linear regression, showing better results compared with variational inference models [31]. Inspired by the CGM of TARNet [76], causal-effect variational autoencoder (CEVAE) was proposed in [44, 91] for regression tasks, which draws a connection between causal inference with proxy variables and latent space learning for approximating the hidden and unobservable confounder by the potential outcome model from Rubin's causal inference framework [29, 73]. Our proposed causal model in TLT shares a similar CGM with CEVAE but has a different train-

Table 1: *Causal hierarchy* [58]: questions at level i can only be answered if information from the same or higher level is available.

Level	Activity	PGM	Example
(𝕀) Association	Observing	$P(y\|x)$	ResNet [22]
(𝕀𝕀𝕀) Intervention	Intervening	$P(y\|do(x), z)$	TLT (ours)

ing objective, probabilistic encoding, and specific design for visual recognition, such as the use of attention mechanism.

3. TLT: Treatment Learning Transformer

3.1. Modeling under Causal Hierarchy Theorem

To model a general image classification problem with causal inference, we introduce Pearl's *causal hierarchy Theorem* [7, 58, 77] as shown in Tab. 1, with a non-causal classification model and a causal inference model. Non-causal model is in level (𝕀) of causal hierarchy, which associates the $outcome$ (prediction) to the input directly by $P(y|x)$ from supervised model such as ResNet [22]. Non-causal model could be unsupervised by using approximate inference such as variational encoder-decoder [6] with two parameterized networks, Θ and Φ. The association-level (non-causal) setup in the causal hierarchy can solve visual learning tasks at level (𝕀), such as non-noisy image classification.

For noisy image classification, we argue that the problem setup is elevated to level (𝕀𝕀𝕀) of the causal hierarchy, requiring the capability of confounder learning and the do-$calculus$ [59] (refer to causal inference foundations **supplement** A). We first make a formal definition on a pair of i^{th} query (x_i, y_i) including a noisy image input (x_i) and its associated label (y_i). Suppose for every noisy image, there exists a clean but inaccessible image $(\tilde{x}_i)$ and treatment information (t_i), where the intervened observation is modeled as $P(x_i) = P(do(\tilde{x}_i)) \equiv P(\tilde{x}_i|t_i)$, and t_i encodes full information of the intervention through the do-operator notation $do(\cdot)$. The corresponding confounder z_i follows $P(z_i) = P(\tilde{x}_i, t_i, \tilde{z}_i)$, where $\tilde{z}_i$ is the unobservable part (e.g., undiscovered species of "cat" but belong to its ontological definition) of the confounder. To make a prediction (y_i) of a noisy input of (x_i), we could have the intervened view of the question by:

$$P(y_i|x_i) = P(y_i|do(\tilde{x}_i), z_i) = P(y_i|\tilde{x}_i, t_i, z_i) \quad (1)$$

with do-operator in level (𝕀𝕀𝕀) of the causal hierarchy. Based on the causal hierarchy, we could use the model with the proxy variables (z_i, t_i) in the higher level (𝕀𝕀𝕀) to answer the question in equal or lower level. Next, we introduce our training objective using an encoder-decoder architecture to reparameterize the aforementioned proxy variables for causal learning.

3.2. Training Objective of TLT

We build our TLT model based on the foundational framework of conditional variational encoder-decoder (CVED) [6, 31], which learns a variational latent representation z_i from data x_i and conditional information (e.g., label y_i) for reconstruction or recognition. To effectively learn visual causal pattern recognition, our TLT model uses variational inference to approximate the complex non-linear relationships involving: the pair probability $(p(x_i, z_i))$, the treatment likelihood $P(t_i)$, the model outcome $p(y_i)$, and the joint distribution $p(z_i, x_i, t_i, y_i)$. Specifically, we propose to characterize the causal graphical model in Fig. 1 *(a)* as a latent variable model parameterized by a DNN encoder-decoder as shown in Fig. 6 (in Appendix A). Note that TLT uses an advanced decoding method $p(a_i) = F_T(H_x, H_z \sim P(x_i))$ for approximating $p(z_i)$ from $p(x_i)$ based on the attention (F_T) from transformer [87], which will be detailed in Sec. 3.3.

First, we assume the observations factorize conditioned on the latent variables and use an *general* inference network (encoder) which follows a factorization of the true posterior. For the model network (decoder), instead of conditioning on observations, we approximate the latent variables z. For vision tasks, x_i corresponds to a noisy input image indexed by i, $t_i \in \{0, 1\}$ corresponds to the treatment assignment, y_i corresponds to the outcome and z_i corresponds to the latent hidden confounder. Note that general formation of an approximation of individual outcome (δ_i) is modeling by $\delta_i = t_i \cdot y_i + (1 - t_i) \cdot y_i$ as potential outcome model [25, 29] with its foundation over the causal inference. Next, each of the corresponding factors is described as:

$$p(z_i) = \prod_{z \in z_i} \mathcal{N}(z|0, 1); \;\; p(x_i|z_i) = \prod_{x \in x_i} p(x|z_i);$$
$$p(t_i|z_i) = Bern(\sigma(f_1(z_i)));$$
$$p(y_i|z_i, t_i) = \sigma(t_i f_2(z_i) + (1 - t_i) f_3(z_i))$$

with $\mathcal{N}(\mu, \sigma^2)$ denoting a Gaussian distribution with mean μ and variance σ^2, $p(x|z_i)$ being an appropriate probability distribution, $\sigma(.)$ being a logistic function, and $Bern(.)$ denotes the probability of success of a Bernoulli random variable. Each of the $f_k(.)$ function is an Adaptive Average Pooling plus Linear layer parameterized by its own parameters θ_k for $k = \{1, 2, 3\}$. Here y_i is tailored for categorical classification problems, but our formulation can be naturally extended to different tasks. For example, one can simply remove the final $\sigma(.)$ layer of $p(y_i|z_i, t_i)$ for regression tasks.

Our TLT inference network (encoder), as illustrated in Fig. 2, aims to learn meaningful causal representations in the latent space. As we can see from Fig. 1 *(a)*, the true posterior over $z \in \mathbf{Z}$ depends on $x \in \mathbf{X}$, t, and y. We are required to know the treatment assignment t along with its outcome y prior to inferring the distribution over z. Therefore, unlike variational encoders, which simply passes the feature

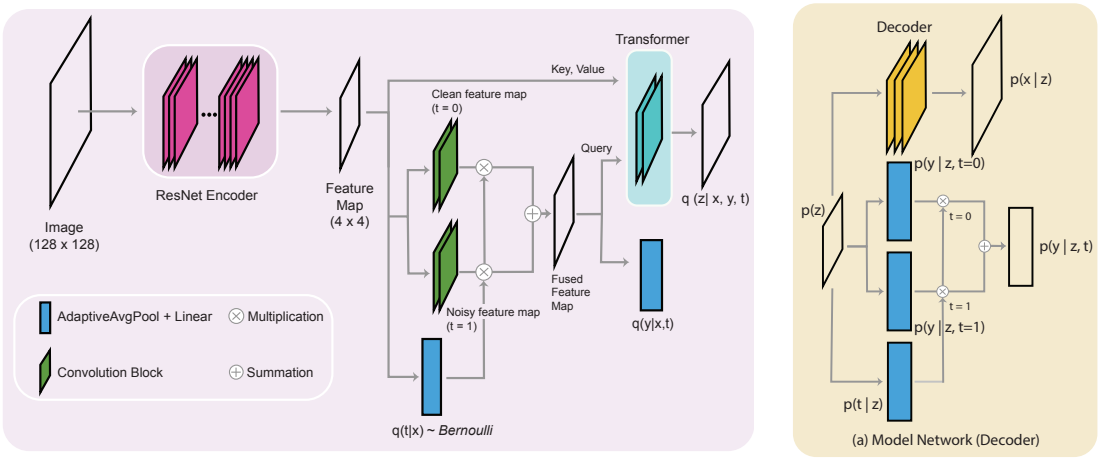


Figure 2: The encoder (inference network) structure of our proposed causal transformer. We leverage bilinear fusion (BF) for $q(z|x, y, t)$ instead of concatenation [44], and decoding conditional queries $H_z \sim q(y|x, t)$ and encoding features $H_x \sim p(x)$ as keys and values to conduct attention. Decoder is shown as Fig. 2 (a) with potential outcome modeling [29, 73] from $p(z)$.

map directly to latent space (the top path in our encoder), the feature map extracted from a residual block is provided to the other switching (the lower and middle paths in our encoder), which provides posterior estimates of treatment t_i and outcome y_i. The switching mechanism (binary selection based on the treatment information of $t_i = 0$ or 1) and its alternative loss training have been widely used in TARNet [76] and CEVAE [44] with theoretical and empirical justification. We employ the distribution by the switching mechanism:

$$\begin{aligned} q(t_i|x_i) &= Bern(\sigma(g_1(x_i)));\\ q(y_i|x_i, t_i) &= \sigma(t_i g_2(x_i) + (1 - t_i)g_3(x_i)), \end{aligned} \tag{2}$$

with each g_k being a neural network approximating $q(t_i|x_i)$ or $q(y_i|x_i, t_i)$. They introduce auxiliary distributions that help us predict t_i and y_i for new samples. To optimize these two distributions, we add an auxiliary objective to our overall model training objective over N data samples:

$$\mathcal{L}_{aux} = \sum_{i=1}^{N} (\log q(t_i = t_i^*|x_i^*) + \log q(y_i = y_i^*|x_i^*, t_i^*)), \tag{3}$$

where x_i^*, t_i^* and y_i^* are the observed values in training set. Since the true posterior over z depends on x, t and y, finally we employ the posterior approximation below:

$$q(z_i|x_i, y_i, t_i) = \prod_{z_i} \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2) \tag{4}$$

$$\boldsymbol{\mu}_i = t_i \boldsymbol{\mu}_{t=1,\mathrm{i}} + (1 - t_i)\boldsymbol{\mu}_{t=0,\mathrm{i}}, \quad \boldsymbol{\sigma}_i^2 = t_i \boldsymbol{\sigma}_{t=1,\mathrm{i}}^2 + (1 - t_i)\boldsymbol{\sigma}_{t=0,\mathrm{i}}^2$$

$$\begin{aligned} \boldsymbol{\mu}_{t=0,i} &= g_4 \circ g_0(x_i, y_i), \quad \boldsymbol{\sigma}_{t=0,i}^2 = \sigma(g_5 \circ g_0(x_i, y_i))\\ \boldsymbol{\mu}_{t=1,i} &= g_6 \circ g_0(x_i, y_i), \quad \boldsymbol{\sigma}_{t=1,i}^2 = \sigma(g_7 \circ g_0(x_i, y_i)) \end{aligned}$$

where g_k again denotes neural network approximation, and $g_0(x_i, y_i)$ is a shared, bilinear-fusioned representation of x, t and y. More specifically, we multiply the feature map with approximated posterior $q(y_i|x_i, t_i)$ without logistic function σ to get $g_0(x_i, y_i)$. Finally, we can have the overall training objective for the inference and model networks. The variational lower bound of TLT to be optimized is given by:

$$\begin{aligned} \mathcal{L}_{TLT} = \ \mathcal{L}_{aux} + \sum_{i=1}^{N} \mathbb{E}_{q(z_i|x_i,t_i,y_i)} [\log p(x_i, t_i|z_i) \\ + \log p(y_i|t_i, z_i) + \log p(z_i) - \log q(z_i|x_i, t_i, y_i)]. \end{aligned} \tag{5}$$

As shown in Fig. 6 (in Appendix A), we could model $q(t|x) \doteq p(t)$ to access the treatment information directly for training to guide one corresponding sub-network in Fig. 2; for testing, $q(t|x)$ could be inferred by a given input x without knowing treatment information from an unsupervised perspective.

3.3. Attention mechanism of TLT

Attention mechanism is one of the human learning components to capture global dependencies for discovering logical and causal relationships [53] from visual patterns in the cognitive psychology community [11]. Transformer [87] based attention mechanism has, recently, shown its connection from the **sequential** energy update rule to Hopfield networks [67], which stands for a major framework to model human memory. With the intuition on leveraging human-inspired attention upon inference from noisy images, we incorporate a new type of Transformer module for the proposed causal modeling, which explicitly model all pairwise interactions between elements in a sequence. The idea is to learn the causal signal [43] via self-attention setup, where we set the interference signal (H_z) for learning query and image features (H_x) for learning key and value. As shown in Fig 2, we use a feature map with a ResNet$_{34}$ [22] encoder

extracting from input image $p(x_i)$ feeding into keys (K) and value (V) with queries $q(y_i)$ from Eq. (2):

$$Q = \text{unroll}\left(F_Q(H_z \sim q(y_i|x_i, t_i))\right) \quad (6)$$

$$K = \text{unroll}\left(F_K(H_x \sim p(x_i))\right) \quad (7)$$

$$V = \text{unroll}\left(F_V(H_x \sim p(x_i))\right);\ a_i = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (8)$$

where F_Q, F_K, F_V are convolutional neural networks and d_k is dimension of keys. Finally, we model $q(z_i)$ by using $q(t_i|x_i)$ and $p(a_i|x_i)$ with the causal two model extended from Eq. (4) for approximating posterior distribution $p(z_i)$:

$$p(z_i) \leftarrow q(z_i|x_i, a_i, y_i, t_i) = \prod_{z_i} \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2). \quad (9)$$

We also have conducted ablation studies on architecture selection and required parameters with respect to supervised learning [22], attention networks [87], and causal model [76] in **supplement** B to validate our model design of TLT. To sum up, the proposed causal architecture attains the best performance with the same amount of parameters.

4. Evaluating Causal Effects on Noisy Images

In this section, we introduce noisy image datasets and conduct statistical refutation tests on TLT to evaluate its causal effect based on the CGM in Fig. 1 *(a)*. That is, we provide an affirmative answer to whether there **exist causal effects** in the studied noisy image classification tasks.

4.1. Estimate Causal Effects

Estimation of **expected causal effects** is one general approach [44, 58, 59] to evaluate whether a CGM (from a logic hypothesis) is valid on the selected test dataset. The underlying graphical model will undergo a series of randomization tests of graphical connection and sub-set sampling to measure its estimation errors on estimating causal effects. In general, a causal model is reliable with the CGM when exhibiting a lower absolute error on the causal effects. In this work, we use average treatment effects (ATE), as used in prior arts [44], for comprehensive analysis.

Average Treatment Effects (ATEs). In the binary treatment setting [56], for the i-th individual and its associated model outcome y_i considering the treatment effect, the ATE is calculated by:

$$y_i = y_{t_i=0,i}\,(1 - t_i) + y_{t_i=1,i}(t_i), \quad (10)$$

$$ATE = |\mathbb{E}\left[y_i = y_i^*|t_i^* = 1\right] - \mathbb{E}\left[y_i = y_i^*|t_i^* = 0\right]|, \quad (11)$$

where $y_{t_i,i}$ denotes the prediction with estimated treatment $t_i \in \{0, 1\}$. y_i^* and t_i^* are the observations. The ATE is taken over all subjects. From [20], these metrics cannot be properly estimated if there are confounding variables in the system. On the other hand, Pearl [56] introduces the "do-operator" [59] on treatment to study this problem under intervention. The do symbol removes the treatment t from the given mechanism and sets it to a specific value by some external intervention. The notation $P(y|do(t))$ denotes the probability of y with possible interventions on treatment. Following Pearl's back-door adjustment formula [58] and the CGM in Fig. 1, it is proved in [44] that the causal effect for a given binary treatment t, a proxy variable x, an outcome y and a confounding variable z can be evaluated by (similarly for $t = 0$):

$$p(y|x, do(t = 1)) = \int_z p(y|x, t = 1, z)p(z|x)dz \quad (12)$$

To intervene the information of t ($do(t)$), flipping errors [44] with different rates (see **supplement** C) are applied to change the t_i label(s) [55] in our experiments in Section 5.1. The proposed CGM and its associated TLT show resilient ATE estimation under statistical refutations.

Table 2: Applying causal modeling on noisy images classification. **Case 1**: perfect labeled visual pattern (H_{lab}) with secondary patterns (H_{iid}); **Case 2**: original labeled images (H_{ori}) under an additive perturbation (F_{per}).

Treatment	$\tilde{x}$	$do(\tilde{x})$	t=1 or 0
1. Context	H_{lab}	H_{lab}+ H_{iid}	Additional patterns (e.g., "waterside") (1) or not (0)
2. Perturbation	H_{ori}	$F_{per}(H_{ori})$	Artificial noise (e.g., Gaussian) (1) or not (0)

Visual Patterns in the Intervention Level ($\mathbb{III}$). We clarify two common scenarios, noisy context and under perturbation, in the intervention level ($\mathbb{III}$) for noisy image classification. As shown in Tab. 2, the treatment information (t) is binary with an accessible noisy input x and **inaccessible** ontological (clean) representation $\tilde{x}$ from Eq. (1) for visual pattern modeling. Next, we introduce datasets in the regime of the case 1 and 2 for our experiments in this work.

4.2. Case 1: NICO Dataset with Noisy Extra Visual Patterns

NICO [23] is a large-scale and open-access benchmark dataset for noisy image classification, which is motivated by studying non-independent image classification with causal modeling. The NICO dataset labels images with both main concepts (e.g., "cat") and contexts as sub-labels (e.g., "water"). NICO is constructed by two super-classes: "animal" and "vehicle", with 10 classes for "animal" and 9 classes for "vehicle". In total, NICO contains 19 classes, 188 contexts, and 25,000 images. The design intuition of NICO is to provide a causal modeling benchmark for large-scale image classification. The authors evaluate several major image classification dataset (e.g., ImageNet, Pascal, and MS-COCO) and found out the auxiliary context information (treatment) is much random and inaccurate from statistical measurement

for structuring validated causal inference. By selecting different contexts of the concept, testing data distribution can be unknown and different from training data distribution, which can be used to evaluate a causal inference model.

In our experiments, we follow the standard NICO evaluation process [23], where a concept is incorporated with two contexts. We further use **context** as treatment in the intervention level as in **Case 1** of Tab. 2. One context is the attribute of concept ($t = 1$) while another context is the background or scene of a concept ($t = 0$).

4.3. Case 2: Curated Causal Pairs (CPS) Dataset with Additive Artificial Noises

Despite many efforts in providing benchmark datasets for causal inference on non-vision tasks [24, 26, 33, 61], visual causal data collection is relatively limited to bare causal effect evaluation with conditional visual treatments [43]. Motivated by the perturbation-based causation studies testing biological network and the efforts from NICO, we further curate two datasets from public sources, named causal pairs **(CPS)**, by using a diverse set of image perturbation types as treatment (i.e., **Case 2** in Tab. 2). We select two representative datasets, Microsoft COCO [38], and a medicine dataset, Decathlon [78], to create our CPS datasets. Each CPS contains pairs of original and perturbed images, as well as **five** different perturbation types described in Sec. 4.4. Table. 3 summarizes the NICO and our CPS datasets. Next, we introduce how to generate noisy images in CPS.

Table 3: Comparison of noisy image classification datasets: CPS with perturbation as a treatment (see Sec. 4.4) and NICO with noisy context (e.g., "indoor").

Dataset	Treatment (Binary Information)	Numbers	Super-classes	Total classes
CPS (ours)	Receiving artificial noise (or not)	13,752	General / Medical	16
NICO	Existing context-wise pattern (or not)	25,000	Animal / Vehicle	19

Super-class 1: Generating Noisy General Objects. To generate CPS dataset from MS-COCO [38] for general super-class, we selected **six** similar object classes that could possibly result in confusing interpretation and recognition by human psychology studies [51, 68] (e.g., giraffe and elephant, etc.). We conduct a survey with 1,000 volunteers from Amazon mechanical turk [85] and pick the top-3 similarity label pairs. Specifically, we format three different common causal pairs, namely giraffe-elephant (CPS_1) with 3316 images, stop sign-fire hydrant (CPS_2) with 2419 images, and bike-motorcycle (CPS_3) with 4729 images, where the dataset is visualized in Fig. 3 (a).

Super-class 2: Generating Noisy Medical Images. For the medical super-class, we use an identical setting with 2630 training and 658 test CT images for **ten** different types (total classes) of human disease from Decathlon [78], which includes: (1) Liver Tumours; (2) Brain Tumours; (3) Hippocampus; (4) Lung Tumours; (4) Prostate; (5) Cardiac; (6) Pancreas Tumour; (7) Colon Cancer; (8) Hepatic Vessels, and (10) Spleen. More details and visualization (Fig. 3 (b)) about this dataset are given in **supplement** B. From these two super-classes, we randomly selected $\mathbf{50}\%$ of these labeled images and applied visual modifications to generate interventional observations. Each generated image is assigned with a binary treatment indicator vector t_i, where its i-th element denotes the binary treatment label according to the i-th visual modification.

4.4. Visual Perturbation (Treatment) in CPS

We employ five distinct types of image modification methods as independent intervention variables: (i) image scrambling; (ii) neural style transfer; (iii) adversarial example; (iv) object masking, and (v) object-segment background shifting. Below we provide brief descriptions for these visual treatments as illustration in Fig. 3.**Image Scrambling (IS)** [93] algorithms re-align all pixels in an image to different positions to permute an original image into a new image, which is used in privacy-preserved classification [82].

Neural Style Transfer (ST) [16] creates texture effect with perceptual loss [30] and super-resolution along with instance normalization [86].

Adversarial Example (AE) adds input perturbation for prediction evasion. We employ the Fast Gradient Sign Method (FGSM) [19] with a scaled ℓ_∞ perturbation bound of $\epsilon = 0.3$. We also evaluated other attacks including C&W [9] and PGD [47] in **supplement** B.

Object Masking (OM) & Background Refilling (BR): Object masking (OM) was proposed in previous studies [43, 92] for causal learning. We applied OM and another masking methods, background refilling (BR), that duplicates non-object background into the mask segment as treatments.

5. Experiments

5.1. Noisy Image Classification on NICO and CPS

Generative Model Baselines

For a fair comparison, we select two benchmark conditional generative model incorporating both information of label (y) and binary treatment (t): modified conditional VAE [31, 79] (CVAE') and modified CEVAE [44] (CEVAE'), where CVAE' use $p(t, y)$ for concatenation as a conditional inference and CEVAE' follows a similar causal variational inference process [44] without features fusion and conditional queries. Both model are enhanced by ResNet [22] and attention layers with similar parameters (7.1M) with TLT. Noted CEVAE [44] is originally designed and applied only on linear regression tasks but benefited from our causal modeling for noisy image classification.

Performance on NICO Dataset. We first evaluate models performance trained on NICO dataset. From the reported results in the paper [23, 95], we select the best reported

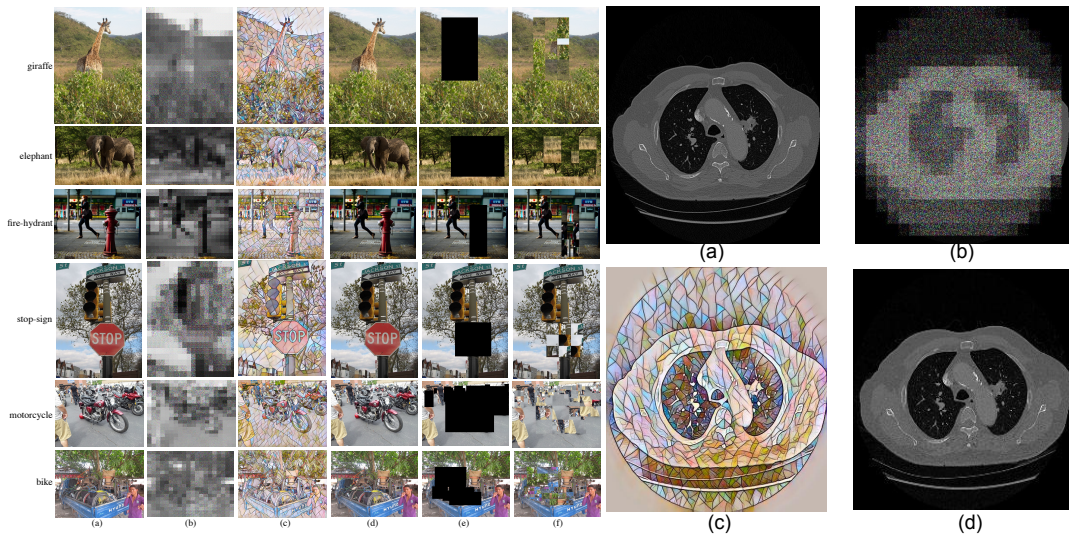


Figure 3: Illustration of our generated **CPS** dataset for noisy image classification. We randomly selected $\mathbf{50}\%$ of labeled images from both datasets and applied visual modifications to generate interventional observations. We selected similar object classes by 1,000 human surveys. *Left*: three causal pairs – giraffe/elephant, fire-hydrant/stop-sign, and motorcycle/bike. From left to right in **CPS**, the visual treatments are: (a) original input image, (b) image scrambling, (c) neural style transfer; (d) adversarial example. We further discuss masking intervention effects used in [43, 92] on general subjects by (e) object masking; and (f) background refilling. *Right*: a demonstration of Lung Tumours in Decathlon of the same format.

Table 4: Perturbation (e.g., texture) effects with classification accuracy (%) on the average of **CPS** images (∼13.7k) for different treatments and their causal effect estimates. Note that TLT (7.39M) has similar parameters compared with CVAE' and CEVAE', which are enhanced by ResNet as discussion in the ablation studies. **n** is for treatment noise level.

	Classification Accuracy (↑)			**Average Treatment Effect** (↑)		
Type of t (with n = 0.05)	CVAE'	CEVAE'	TLT (ours)	CVAE'	CEVAE'	TLT (ours)
Original (without **t**)	83.31 ±0.12	83.31 ±0.23	83.31 ±0.13	0.012	0.018	0.017
Style Transfer (ST)	73.67 ±0.31	74.34 ±0.26	**76.12** ±0.27	0.324	0.343	**0.359**
Image Scrambling (IS)	72.31 ±1.27	76.21 ±0.81	**80.12** ±0.54	0.057	**0.295**	0.288
Adversarial Example (AE)	79.12 ±0.25	81.12 ±0.17	**83.12** ±0.12	0.025	0.027	**0.036**
Object Masking (OM)	70.12 ±0.19	72.73 ±0.21	**74.06** ±0.11	0.179	0.241	**0.253**
Background Refilling (BR)	71.32 ±0.28	72.59 ±0.29	**74.91**±0.17	0.213	0.221	**0.238**

Table 5: Classification accuracy (%) on NICO.

Model	**StableNet** [23]	**CVAE'**	**CEVAE'**	**TLT**
Acc.	59.76 ±1.52	57.23 ±2.12	62.17 ±1.82	**65.98** ±1.74

model, StableNet from [95] with sample weighting, which outperforms six existing competitive models [36, 49, 52, 74, 83, 97] including SagNet [52] and GroupDRO [74] from official report. As shown in Table. 5, generative models with proposed causal modeling attain competitive results on NICO with compositional bias setup, where TLT attains a best performance of **65.98**%. We provide more analysis under different setup of NICO, where TLT remains as the best model in **supplement** C.

Performance on CPS Dataset. In Table 4, we compare TLT with modified CVAE' and modified CEVAE' as baselines[1] trained on **CPS** dataset. The accuracy of TLT in the original image, IS, ST, AE, OM and BR settings are consistently better than CVAE' and CEVAE', with substantially large margins ranging from $\mathbf{1.60}\%$ to $\mathbf{7.81}\%$. CEVAE' and TLT are also shown to have higher causal estimate (CE) than CVAE' in all settings except for ST. Interestingly, ST leads to a higher causal value (from $\mathbf{0.318}$ to $\mathbf{0.354}$) when compared to the other modifications such as IS and AT. This finding accords to the recent studies on DNN's innate bias of using edges and textures for vision task [17]. CEVAE' and TLT having lower value in ST setting could be explained by a more unbiased representation learned by inference network with lower dependency on edges and textures. A benchmark

[1]We have also conducted experiments on the seven algorithms used in the NICO for CPS. However, all the evaluated algorithms perform worse than our selected CPS baselines, possibly due to the challenges of visual perturbation deployed in CPS. The code and weights will open for reference.

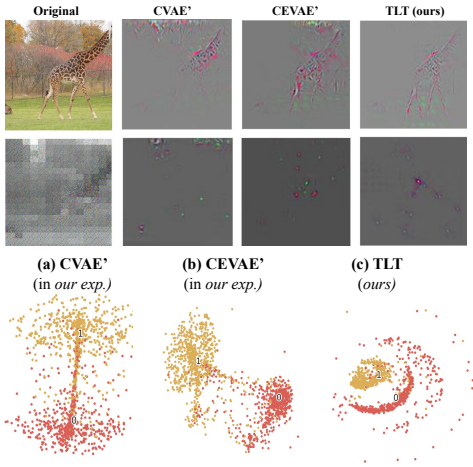


Figure 4: *(a)* With proposed TLT and CPS dataset, neural saliency methods can be extended to visual pattern from inference. Take the top row as an example, using TLT, guided grad-CAM [75] can be more aligned with the concise human-interpretable giraffe patterns instead of forest texture and edges. More correlation analyses between saliency and labels in NICO and CPS are given in supplement C. *(b)* Visualization of learned manifolds of $q(z)$ by tSNE [46], proposed TLT's results largest intra-cluster pair-wise sample distances between additive noise (adversarial) ($t = 1$) and vanilla ($t = 0$) image samples from CPS_1.

visualization of Guided Grad-CAM [75] in Fig. 4 *(a)* validates this hypothesis and highlights the importance of our inference network in gaining robust visual understanding from latent space z as tSNE [46] results Fig. 4 *(6)*. One critical issue for visual intervention is its difficulty in investigating the effect on object mask size [43, 62]. **supplement C** shows a consistent and stable performance of TLT against varying mask sizes.

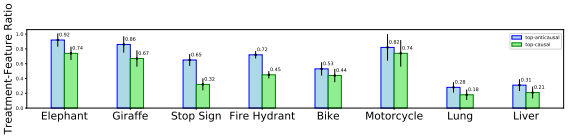
Case Study on the medical super-class: We conduct the same experiments with medical super-class to identify visual clinical features. Both the classification and estimation performance are consistent with general CPS objects, where TLT attains the highest accuracy 88.74% in the original setting and 82.57% in the scrambling setting (e.g., data encryption operation) settings. TLT is most effective in classifying noisy image and more sensible in measuring ATE on adversarial example. We also conduct expert evaluation on the activation saliency of clinical patterns (Fig. 4). Based on their domain knowledge [65, 66, 89], three physicians independently and unanimously give the highest confidence scores on saliency attributes to our method.

Statistical Refutation of Causal Models: To rigorously validate our ATE estimation result, we follow a standard refuting setting [57, 60, 72] with the causal model in Fig. 1 to run three major tests, as reported in **supplement E** and Table S15, which validate our method is robust.

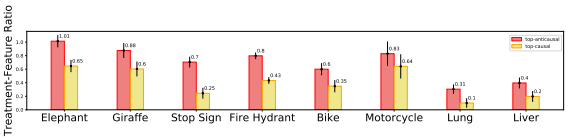
5.2. Neural Causation Coefficient (NCC)

Neural Causation Coefficient (NCC) [43] is a **benchmark causal discovery technique** to validate its significance of causal signal [43, 92] in a deployed experiment.

NCC is used to discover for joint distribution of a pair of related proxy variables that are computed by applying CNNs to the image pixels. Lopez et al. [43] used an augmented NCC network to prove the existence of causal relations in ResNet [22] between object and context in an image, and showed that in object-feature ratio anticausal signal consistently has stronger relation than causal signal.



(a) TFR calculated by feature f_R from ResNet34 [22] as [43].



(b) TFR calculated by feature f_C from our proposed TLT.

Figure 5: Evaluation of causal pairs by treatment feature ratio (TFR) score [43]. The average and standard deviation of TFR associated to the top-1% causal/anticausal feature scores are displayed. The results show the visual perturbation measurement is coherent with the previous study [43].

We reproduce the NCC architecture from [43] and find all the anti-causal scores of COCO is larger than causal score as shown as [43], where the causal signals in the proposed dataset have been validated with extra NCC tests.

6. Conclusion

Motivated by human-inspired attention mechanism and causal hierarchy theorem, in this paper we proposed a novel framework named treatment learning transformer (TLT) for tackling noisy image classification with treatment estimation. In addition to showing significantly improved accuracy of TLT on the NICO dataset with noisy contexts, we also curated a new causal-pair dataset (CPS) based on five different visual image perturbation types for performance benchmarking in general and medical images. We validated the causal effect of TLT through statistical refutation testing on average treatment effects. We also show derived advantages of TLT in terms of improved visual saliency maps and representation learning. Our results suggest promising means and a new neural network architecture for the advancement of research at the intersection of deep learning and visual causal inference. Our supplementary code will be open-resource under Apache License 2.0 to the community.

References

[1] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10044–10054, 2020. 2

[2] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698, 2020. 2

[3] Jan Philipp Albrecht. How the gdpr will change the world. *Eur. Data Prot. L. Rev.*, 2:287, 2016. 13

[4] Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4845–4854, 2019. 1, 2

[5] Mohammad Taha Bahadori, Krzysztof Chalupka, Edward Choi, Robert Chen, Walter F Stewart, and Jimeng Sun. Causal regularization. *arXiv preprint arXiv:1702.02604*, 2017. 2

[6] Hareesh Bahuleyan, Lili Mou, Olga Vechtomova, and Pascal Poupart. Variational attention for sequence-to-sequence models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1672–1682, 2018. 3

[7] E Bareinboim, JD Correa, D Ibeling, and T Icard. On pearl's hierarchy and the foundations of causal inference. *ACM Special Volume in Honor of Judea Pearl (provisional title)*, 2020. 3

[8] Peter Bühlmann, Jonas Peters, Jan Ernest, et al. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014. 2

[9] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14. ACM, 2017. 6, 14, 16

[10] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. *arXiv preprint arXiv:1412.2309*, 2014. 2

[11] Tianwen Chen, Lars Michels, Kaustubh Supekar, John Kochalka, Srikanth Ryali, and Vinod Menon. Role of the anterior insular cortex in integrative causal signaling during multisensory auditory–visual attention. *European Journal of Neuroscience*, 41(2):264–274, 2015. 4

[12] Paul Downing, Jia Liu, and Nancy Kanwisher. Testing cognitive models of visual attention with fmri and meg. *Neuropsychologia*, 39(12):1329–1342, 2001. 17

[13] Amy Fire and Song-Chun Zhu. Using causal induction in humans to learn and infer causality from video. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35, 2013. 2

[14] Amy Fire and Song-Chun Zhu. Learning perceptual causality from video. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(2):23, 2016. 2

[15] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. *arXiv preprint arXiv:1703.02910*, 2017. 2

[16] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 6

[17] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *ICLR*, 2019. 7

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015. 1

[19] Michael Goodfellow and Amanda L Jones. Laceyella. *Bergey's Manual of Systematics of Archaea and Bacteria*, pages 1–4, 2015. 6, 14, 16

[20] Sander Greenland, James M Robins, Judea Pearl, et al. Confounding and collapsibility in causal inference. *Statistical science*, 14(1):29–46, 1999. 5, 20

[21] Jiangfan Han, Ping Luo, and Xiaogang Wang. Deep self-learning from noisy labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5138–5147, 2019. 2

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 3, 4, 5, 6, 8

[23] Yue He, Zheyan Shen, and Peng Cui. Towards non-iid image classification: A dataset and baselines. *Pattern Recognition*, page 107383, 2020. 1, 2, 5, 6, 7, 15, 17

[24] Jennifer L Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. 6

[25] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986. 3

[26] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pages 689–696, 2009. 6

[27] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1

[28] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019. 18

[29] Guido W Imbens and Donald B Rubin. Rubin causal model. In *Microeconometrics*, pages 229–241. Springer, 2010. 2, 3, 4

[30] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 6

[31] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2, 3, 6

[32] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. *arXiv preprint arXiv:1709.02023*, 2017. 2

[33] Robert J LaLonde. Evaluating the econometric evaluations

of training programs with experimental data. *The American economic review*, pages 604–620, 1986. 6

[34] Karel Lebeda, Simon Hadfield, and Richard Bowden. Exploring causal relationships in visual object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3065–3073, 2015. 2

[35] Ute Leonards, Stefan Sunaert, Paul Van Hecke, and Guy A Orban. Attention mechanisms in visual search—an fmri study. *Journal of Cognitive Neuroscience*, 12(Supplement 2):61–75, 2000. 17

[36] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 7

[37] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1910–1918, 2017. 2

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6, 13, 16, 17

[39] Bing Liu, Dong Wang, Xu Yang, Yong Zhou, Rui Yao, Zhiwen Shao, and Jiaqi Zhao. Show, deconfound and tell: Image captioning with causal inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18041–18050, 2022. 1

[40] Jiashuo Liu, Zheyan Shen, Peng Cui, Linjun Zhou, Kun Kuang, Bo Li, and Yishi Lin. Stable adversarial learning under distributional shifts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 2

[41] Ruyang Liu, Hao Liu, Ge Li, Haodi Hou, TingHao Yu, and Tao Yang. Contextual debiasing for visual recognition with causal mechanisms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12755–12765, 2022. 1

[42] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 20

[43] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6979–6987, 2017. 2, 4, 6, 7, 8, 13

[44] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6446–6456, 2017. 2, 4, 5, 6, 13, 16, 20

[45] Steven J Luck, Geoffrey F Woodman, and Edward K Vogel. Event-related potential studies of attention. *Trends in cognitive sciences*, 4(11):432–440, 2000. 17

[46] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 8

[47] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 6, 14, 16

[48] Chengzhi Mao, Kevin Xia, James Wang, Hao Wang, Junfeng Yang, Elias Bareinboim, and Carl Vondrick. Causal transportability for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7521–7531, 2022. 1

[49] Toshihiko Matsuura and Tatsuya Harada. Domain generalization using a mixture of multiple latent domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11749–11756, 2020. 7

[50] Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. Causal discovery with general non-linear relationships using non-linear ica. In *Uncertainty in Artificial Intelligence*, pages 186–195. PMLR, 2020. 2

[51] Gail Musen and Anne Treisman. Implicit and explicit memory for visual patterns. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(1):127, 1990. 6

[52] Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8690–8699, 2021. 7

[53] Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):312–340, 2019. 4

[54] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014. 2

[55] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[56] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995. 1, 5, 13, 17, 20

[57] Judea Pearl. On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 435–443. Morgan Kaufmann Publishers Inc., 1995. 8

[58] Judea Pearl. *Causality*. Cambridge university press, 2009. 3, 5, 13, 17

[59] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019. 1, 2, 3, 5, 13, 17, 19

[60] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 2, 8

[61] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014. 1, 6

[62] Lyndsey C Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Scholkopf, and William T Freeman. Seeing the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2035–2042, 2014. 2, 8

[63] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8571–8580, 2018. 18

[64] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10860–10869, 2020. 2

[65] Pranav Rajpurkar, Jeremy Irvin, Aarti Bagul, Daisy Ding, Tony Duan, Hershel Mehta, Brandon Yang, Kaylie Zhu, Dillon Laird, Robyn L Ball, et al. Mura: Large dataset for abnormality detection in musculoskeletal radiographs. *arXiv preprint arXiv:1712.06957*, 2017. 8

[66] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017. 8

[67] Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020. 4

[68] Stephen K Reed. *Cognition: Theories and applications*. CENGAGE learning, 2012. 6

[69] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on International Conference on Machine Learning-Volume 32*, pages II–1278, 2014. 2

[70] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1

[71] Paul R Rosenbaum and Donald B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218, 1983. 20

[72] Kenneth J Rothman and Sander Greenland. Causation and causal inference in epidemiology. *American journal of public health*, 95(S1):S144–S150, 2005. 8

[73] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974. 2, 4

[74] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019. 7

[75] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626. IEEE, 2017. 8, 14, 16, 17, 18

[76] Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017. 2, 4, 5, 13, 20

[77] Ilya Shpitser and Judea Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9(Sep):1941–1979, 2008. 3

[78] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019. 6, 16

[79] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pages 3483–3491, 2015. 2, 6

[80] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 20

[81] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725, 2020. 2

[82] Michael J Tarr and Heinrich H Bülthoff. Image-based object recognition in man, monkey and machine. *Cognition*, 67(1-2):1–20, 1998. 6

[83] Xavier Thomas, Dhruv Mahajan, Alex Pentland, and Abhimanyu Dubey. Adaptive methods for aggregated domain generalization. *arXiv preprint arXiv:2112.04766*, 2021. 7

[84] Christine Trampusch and Bruno Palier. Between x and y: how process tracing contributes to opening the black box of causality. *New political economy*, 21(5):437–454, 2016. 2

[85] Amazon Mechanical Turk. Amazon mechanical turk. *Retrieved August*, 17:2012, 2012. 6

[86] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 6

[87] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3, 4, 5

[88] Lan Wang and Vishnu Naresh Boddeti. Do learned representations respect causal relationships? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 264–274, 2022. 1

[89] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 3462–3471, 2017. 8

[90] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015. 1, 2

[91] Chao-Han Huck Yang, I Danny, Te Hung, Yi Ouyang, and

Pin-Yu Chen. Causal inference q-network: Toward resilient reinforcement learning. In *Self-Supervision for Reinforcement Learning Workshop-ICLR 2021*, 2021. 2

[92] Chao-Han Huck Yang, Yi-Chieh Liu, Pin-Yu Chen, and Xiaoli Ma. When causal intervention meets image masking and adversarial perturbation for deep neural networks. *arXiv preprint arXiv:1902.03380*, 2019. 2, 6, 7, 8, 13, 14, 17, 18

[93] Guodong Ye. Image scrambling encryption algorithm of pixel bit based on chaos map. *Pattern Recognition Letters*, 31(5):347–354, 2010. 6

[94] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[95] Xingxuan Zhang, Peng Cui, Renzhe Xu, Linjun Zhou, Yue He, and Zheyan Shen. Deep stable learning for out-of-distribution generalization. *arXiv preprint arXiv:2104.07876*, 2021. 2, 6, 7

[96] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 16, 17

[97] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *International Conference on Learning Representations*, 2020. 7