Causal reasoning in typical computer vision tasks

Kexuan Zhang[§], Qiyu Sun[§], Chaoqiang Zhao[§], Yang Tang[§*]

[§]Key Laboratory of Advanced Control and Optimization for Chemical Process, Ministry of Education,
East China University of Science and Technology, Shanghai 200237, China

Abstract—Deep learning has revolutionized the field of artificial intelligence. Based on the statistical correlations uncovered by deep learning-based methods, computer vision has contributed to tremendous growth in areas like autonomous driving and robotics. Despite being the basis of deep learning, such correlation is not stable and is susceptible to uncontrolled factors. In the absence of the guidance of prior knowledge, statistical correlations can easily turn into spurious correlations and cause confounders. As a result, researchers are now trying to enhance deep learning methods with causal theory. Causal theory models the intrinsic causal structure unaffected by data bias and is effective in avoiding spurious correlations. This paper aims to comprehensively review the existing causal methods in typical vision and vision-language tasks such as semantic segmentation, object detection, and image captioning. The advantages of causality and the approaches for building causal paradigms will be summarized. Future roadmaps are also proposed, including facilitating the development of causal theory and its application in other complex scenes and systems.

Index Terms—causal reasoning, computer vision tasks, vision-language tasks, semantic segmentation, object detection.

I. INTRODUCTION

Deep learning techniques have improved our understanding of the world, and deep learning-based computer vision methods enable high-performance intelligent perception of our surroundings [1]. Areas such as autonomous cars [2, 3], unmanned aircraft [4], and robotics [5, 6] are developing rapidly with technological innovations. To guarantee performance across one or even more domains, training strategies like attention mechanisms [7, 8], pre-training mechanisms [9, 10], and generic large models [11] have been proposed. Despite their great performance, the basis of these deep learning-based methods is to learn the statistical correlation. However, statistical correlation learns regular knowledge based on the final presentation of the data, which lacks the guidance of prior knowledge. In contrast, causality focuses on the mechanisms of the data generation process and the causal structure of specific tasks. Causality implies the relationship between two variables in which the cause variable directly affects the effect variable [12]. Though both statistical correlation and causality are data-driven [13], the former uses the consistency of data trends as the basis for determining relationships, while causality is the reflection of inherent characteristics and structures both inside and between variables. As there is not always a causal relationship between variables that change in a consistent trend, statistical correlation-based methods tend to

use mistakenly non-causal relationships, also called spurious correlations, as the basis for network inference. In particular, when the data is multidimensional and heterogeneous, the complexity of the relationships may further amplify the impact of spurious correlations. Simpson's paradox proposed by Blyth *et al.* [14] well confirms the flaws of statistical correlation-based methods. It confirms that the direction of an association at the population-level may be reversed within the subgroups comprising that population. As stated in [15], the relation between coffee consumption and neuroticism is positive in each individual, but those individuals who drink more coffee are generally less neurotic. For each individual, the correlation between coffee consumption and neuroticism is positive, but in the population the correlation is negative. Such paradox is worth noticing as different levels of interpretation lead to different results for the same data. It follows that neither individual-level nor population-level statistical correlations can fully characterize the relationship between coffee consumption and neuroticism. However, the causality-based methods will specify the prior knowledge concerning the causal structures and derive the correct causal chains at the specific level of interpretation. As a result, the causality-based methods are more logical and effective compared to the statistical correlation-based ones.

Typical deep learning-based computer vision tasks can be summarized as follows: Given an image X, the goal is to build a network to predict its label Y correctly [16]. A statistical model fitted with a suitable objective function is often used to estimate the conditional probability distribution $P(Y|X)$. However, only under the independent identical distribution ($\mathcal{I}.\mathcal{I}.\mathcal{D}.$) hypothesis can the learned conditional probability distribution $P(Y|X)$ be applied appropriately from the training set to the testing set. It requires that new prediction samples stay consistent with the distribution of the training set. To minimize the impact of domain differences between two sets, deep learning-based methods such as domain adaptation [17] and generalization [18] have been proposed. However, the problem of generalizability cannot be fundamentally solved through these methods, as the bias caused by domain differences is not fundamentally eliminated. Additionally, since deep learning-based methods benefit from many stacked layers and parameters for approximating high-dimensional complex functions, it is hard to interpret the learned models and parameters [16]. In conclusion, existing statistical correlation-based methods overly rely on the given

data and analyze the problems through approximating high-dimensional functions rather than mechanical mining. There are areas for improvement in both generalizability and interpretability.

Due to its mechanical strengths, causality has gained significant attention recently and has rooted its developments across several fields, such as statistics [19, 20], economics [21, 22], epidemiology [23, 24], and computer science [25, 26]. Basic causal methods can be divided into two main aspects, causal discovery and causal inference [27]. Based on the causal structures learned by causal discovery, causal inference leverages those relationships for further analysis. An example is proposed to highlight the advantages of causality-based methods. Fig. 1 compares the statistical correlation-based and the causality-based methods for the same image classification task. Given the input images of the sheep and the corresponding labels in Fig. 1 (a), the model is trained for correct identification. The visualizations of the learned features are shown in Fig. 1 (b). Both statistical correlation-based and causality-based learning mechanisms are specifically shown in Fig. 1 (c). Since the sheep and the grassland frequently coexist in the training data, the statistical correlation-based process tends to regard the grassland characteristics as the basis for the labels due to their similar distribution. In contrast, the causality-based methods focus more on the objective cause chain of sheep and prefer wool features as the basis. Different learning mechanisms lead to different classification results, especially when faced with rare samples, as shown in Fig. 1 (d). When given an image of a sheep standing in the snow, the statistical correlation-based model fails to assign the correct label because no grassland features are found. Conversely, by focusing on the wool features, the causality-based model can accurately categorize the sheep based on the causal features. As a result, the causality-based methods rely not only on the consistency of data trends but also on the underlying processes that generate the causal structure between variables [28]. They are more robust towards an unknown environment, and the learning process tends to be more interpretable.

This review focuses on the implications of causal theory for vision and vision-language tasks, including classification, detection, segmentation, visual recognition, image captioning, and visual question answering. There exist some surveys about causal theory [13, 27, 29, 30]. Kaddour *et al.* [29] group existing causal machine learning work into five categories and comprehensively compare existing approaches. Gao *et al.* [13] focus on the application of causal reasoning to recommendation systems. Li *et al.* [30] outline the advantages of causal theory in industrial applications. Chen *et al.* [27] divide the causal discovery tasks into three types according to the variable paradigm: definite tasks, semidefinite tasks, and undefinite tasks. Unlike these existing causality-based reviews, this work compares the causal structures of different approaches in vision and vision-language tasks and summarizes the correspondence between the commonly used causal structures and the corresponding concerns. Furthermore, the article outlines the benefits of the causal approaches regarding accuracy, generalizability, and interpretability. Section II introduces the basic concepts and terminology of causality, including structural causal models, causal interventions, backdoor adjustments, front-door adjustments, and counterfactual. Section III proposes a systematic and structural review of the specific vision and vision-language tasks. It analyzes the corresponding causal structure and identifies the confounders. Five typical causal structures are summarized for comparison and methodology overviews. Section IV lists the future roadmaps for better development of causal theory and a broad application of causal theory. Section V concludes this review.

II. PRELIMINARIES

In this section, the basic causal methods are introduced, including causal discovery and causal inference. Causal discovery aims to build causal relationships with the structural causal models (SCMs), while causal inference is used to estimate the causal effect.

A. Causal discovery

1) Causal structure: A causal structure is a directed acyclic graph (DAG) in which each node corresponds to a specific variable, and each edge indicates a direct functional relationship between the linked variables [12]. If there is a directed edge pointing from Y to X, X is called the child variable, while Y is the corresponding parent variable. If a variable has its parent variable in the causal structure, it is endogenous; otherwise, it is exogenous. To better describe the concepts involved, the causal structure depicting the process of image generation is shown in Fig. 2 as an example.

The graph has four nodes involved, where X represents the given image, Y represents the corresponding label, C represents the content, and D represents the domain involved. The edge $D \rightarrow X \leftarrow C$ describes image generation process in that the image includes both the content and the domain information. The edge $X \rightarrow Y$ indicates the ultimate goal of a vision task: finding a suitable label for the given image. The edge $C \rightarrow Y$ means that the content in the image determines the label. Since there are two directed edges pointing to X: $D \rightarrow X$ and $C \rightarrow X$, D and C are the parent variables of X. Also, D is exogenous because it has no parent variables in the graph. A path between two variables is a sequence of edges connecting them. The path between X and Y in Fig. 2 can be either $X \rightarrow Y$ or $X \leftarrow C \rightarrow Y$.

2) Structural causal models: Based on the causal structure, a structural causal model can specify how each variable is influenced by its parent variables by effect functions. Given a set of variables $X_1, X_2..., X_n$ indicating the nodes in the causal structure, they can be written as the outcome of their parent variables and the effect function [12].

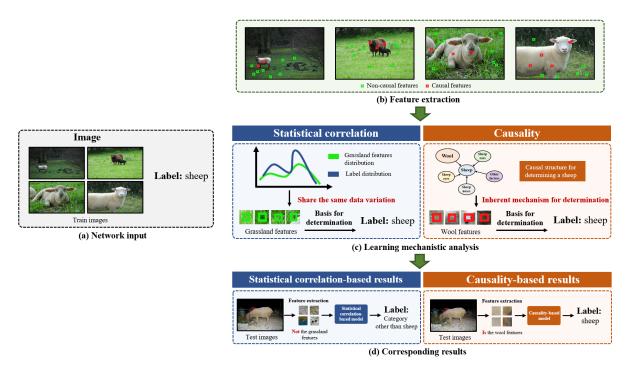$$X_i = f_i(PA_i, U_i), i = 1, ..., n, \tag{1}$$

Fig. 1. Comparison between statistical correlation-based methods and causality-based methods. Given the images and the corresponding labels, the image classification network performs feature extraction, makes an analytical decision on the feature, and finally outputs the corresponding labels. The figure above compares the two learning mechanisms: statistical correlation-based analysis and causality-based analysis, as shown in (c). The input in (a) is visualized by feature extraction in (b). The green boxes indicate the captured non-causal features, while the red boxes indicate the captured causal features. Due to different learning mechanisms, the statistical correlation-based method fails to infer the correct label. In contrast, the causality-based method can get the correct label, as shown in (d). When faced with the learned object in an unknown environment, the correlation-based method tends to be misled by data bias. Contrastly, the causality-based method focuses only on the causal factors associated with the object and is not disturbed by data variation.
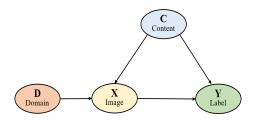

Fig. 2. A general causal structure for image generation, where X represents the given image, Y represents the learned label, C represents the corresponding content, and D represents the domain involved.

where PA_i indicates the parent variables of X_i, U_i are the unobserved background variables including the noise, and f_i indicates the functional relationship. Taking the relationship of variables in Fig. 2 as an example, the variable X is the outcome of its parent variables and the effect functions together as follows:

$$X = f(D, C, U), \qquad (2)$$

where f is the effect function, U indicates the background variable in this functional relationship. The purpose of the causal inference to be presented later is to characterize f and estimate the effects of interventions or counterfactuals.

B. Causal inference

Given a causal structure, causal intervention and counterfactuals can be conducted. The former aims to change the structure by specifying variable data to artificially alter the relationship between variables. In contrast, the latter aims to estimate the variables in a situation opposite to reality. Two typical intervention methods: back-door adjustment and front-door adjustment will also be discussed.

1) Causal intervention: Causal intervention refers to the deliberate manipulation of one or more variables in a system to observe the effect on another variable. The intervention on variable X is formulated with *do*-calculus, $do(X = x)$, which blocks the effect of the parent variables on X and assigns X to x. The causal relationships can be discovered through causal intervention with the modified graphical model

and the corresponding manipulated conditional probabilities. The example in Fig. 2 can be taken for a better illustration. A causal intervention on the domain can be proposed to observe the effect of domain variables on the corresponding labels. When keeping other variables constant, $P(Y|do(D))$ can denote the outcome of Y when changing domain knowledge. Due to the complexity of the relationship between variables, it is difficult to directly determine the causal chain between the cause and the effect without being interrupted by the spurious correlation. Therefore, interventions are needed to ensure independence between variables when exploring the simple direct causal association. The basic connection structures proposed by Rebane *et al.* [31] between three variables X, Y, and C are shown in Fig. 3 (a), (b), and (c). Different ways of intervention are used towards different structures owing to their unique characteristics. The path $X \to C \to Y$ in Fig. 3 (a) is a chain junction where X affects Y via the mediator C. In vision tasks, the feature can be learned from the given image, and the label is made referring to the learned feature. It is easy to find that an intervention on C can easily block the path between X and Y. The path $X \leftarrow C \to Y$ in Fig. 3 (b) is called a confounding junction, where C affects both X and Y and is called a confounder. In vision tasks, the context can affect both the images and the labels, adding a spurious correlation to the actual causal chain between images and labels. Under these circumstances, interventions should be taken on C to block the path. The path $X \to C \leftarrow Y$ in Fig. 3 (c) is called a collider, where both X and Y decide C. In vision tasks, the image can be generated by both the content and domain information. When the value of C is unknown, X and Y are independent. Once the value of C is accessible, X is in relation to Y. Therefore, the value of C cannot be intervened to block the path between X and Y.

In conclusion, causal intervention guarantees the independence between variables and eliminates the effect brought by potential confounders. It is flexible and is decided by the specific structure. Two typical ways of causal interventions will be described below in detail.

2) Back-door adjustment: Back-door adjustment is one of the de-confounding techniques used in the causal intervention [12]. Any path from X to Y that starts with an arrow pointing into X is a back-door path. Assume that there are three variables, X, Y, and C; the corresponding structure is shown in Fig. 3 (d). Since the back-door path $X \leftarrow C \to Y$ is the confounding junction structure, the way to block the path is to manipulate C through an intervention [32] if the value of C is available. The specific manipulation is to stratify C and calculate the average causal effect at each stratum. The corresponding back-door adjustment is shown in Fig. 3 (e), and the formulation is as follows:

$$P(Y|do(X)) = \sum_{c} P(Y|X,c)P(c). \tag{3}$$

With the back-door adjustment, the probability of variable Y conditional on a given variable X, which indicates the

causal relationship can be obtained by calculating the sum of the conditional probabilities $P(Y|X,c)$ and the stratified probability of the confounder C.

3) Front-door adjustment: If C is unavailable, the manipulation cannot be conducted and the back-door adjustment is useless. Under such circumstances, the front-door adjustment is responsible for guiding the discovery of causality [12]. The observed mediator variable M is introduced to block out the relationship between X and Y, as shown in Fig. 3 (f). In the presence of M, the effect of X on M and the effect of M on Y can be calculated, respectively.

To infer the effect of X on M as $P(M = m|do(X))$, the path $X \leftarrow C \to Y \to M$ is blocked out [32]. With the existence of a collider, the equation can be expressed as follows:

$$P(M = m|do(X)) = P(M = m|X). \tag{4}$$

To infer the effect of M on Y as $P(Y = y|do(M))$, the path $W \leftarrow X \leftarrow C \to Y$ should be blocked out. As C is not available, X should be controlled to block the path, and the corresponding equation can be expressed as below:

$$P(Y = y|do(M)) = \sum_{x} P(Y|C,x)P(x). \tag{5}$$

In conclusion, the front-door adjustment can be described as:

$$\begin{aligned} P(Y|do(X)) &= \sum_{m} P(M = m|do(X))P(Y = y|do(M)) \\ &= \sum_{m} P(M = m|X) \sum_{x} P(Y|M = m,x)P(x). \end{aligned} \tag{6}$$

With the front-door adjustment, the probability of variable Y being conditional on a given variable X can be obtained by introducing a mediator variable M. In conclusion, both the back-door adjustment and the front-door adjustment can estimate the causal effect. When choosing the de-confounding techniques, the correlations between variables should be justified, and the characteristics of the confounders should be figured out. If the confounder is available, a back-door adjustment is appropriate; otherwise, the front-door adjustment is the better choice.

4) Counterfactual: Counterfactual is another way of making causal inference. It is the opposite of the factual and is often used to estimate the difference between the variables of interest and their observed values in the real world [13]. Using the error terms [12] through comparison, an intervention on the variables of interest is used to predict the outcome in the counterfactual world. For example, some counterfactual features can be generated randomly using noise to measure the effectiveness of the learned image features on the final label determination. The error terms of the label performance can act as a criterion for effectiveness.

5) Potential outcome models: The potential outcome models are first proposed by Neyman*et al.* [33] to estimate the causal effect of a treatment variable on an outcome variable without requiring the causal graph. Ideally, the difference between two potential outcomes can be regarded as the causal
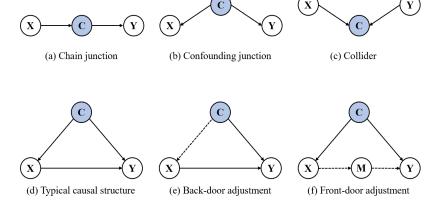
Fig. 3. The three graphs (a), (b), and (c) in the first row show three common connection structures between X, C, and Y. They are the chain junction, the confounding junction, and the collider, respectively. Different interventions are required for different types of connections. Graph (d) in the second row is the most common causal structure, while graphs (e) and (f) are two typical causal intervention methods used for the structure in (d). (e) indicates the back-door adjustment and (f) indicates the front-door adjustment. In each graph, X represents a set of input variables, Y represents a set of output variables, and C represents the confounder that causes the spurious correlation. The dashed line represents the block-out path and the blue nodes represent the confounder.

effect of the treatment on the outcome. For example, given binary treatments $T = 0/1$, the individual treatment effect (ITE) for an individual i is defined as $Y_1^i - Y_0^i$ [13]. However, only one of these results can be observed at a time, either Y_1^i or Y_0^i. As a result, the average treatment effect (ATE) is proposed as an extension for the individual treatment effect to measure the overall average. The formulation is as follows:

$$ATE = \mathbb{E}[Y_1^i - Y_0^i] = \frac{1}{N} \sum_{i=1}^{N} (Y_1^i - Y_0^i), \qquad (7)$$

where $i = \{1, 2, ..., N\}$ represents each individual number in the population.

III. CAUSALITY IN DIFFERENT TASKS

A. Methodology summary

The process of methods based on causal theory includes building the structural causal models through causal discovery and choosing the proper way of causal inference. In the structure, nodes indicate the important variables or variables that are hidden but have an effect on the important ones, while the edges indicate inner causal relationships. Combining different causal structures and different concerns, suitable causal reasoning methods are selected to eliminate spurious correlations. Table I summarizes the causality-based methods both in vision and vision-language tasks. In addition to the task names, the causal structures and the concerns are also listed to compare different methods.

1) Typical causal structures: The causal structures covered in this paper can be summarized into five ones, as shown in Fig. 4. Apart from the simplest structure (I) in Fig. 4, the other structures introduce intermediate variables to further elucidate the relationship.

Fig. 4 (I) describes the relationships between the input, the output, and a confounder that introduces the unwanted spurious correlation. If C is available, a back-door adjustment is often used to cut off the link between X and C to analyze the actual effect of X on Y as follows:

$$P(Y|do(X)) = \sum_c P(Y|X, c)P(c). \qquad (8)$$

Fig. 4 (II) introduces an intermediate variable M between the input and the output. In vision and vision-language tasks, M is a more specific representation of the mapping between inputs and outputs. It is also introduced to avoid the manipulation of C when it cannot be obtained directly in the front-door adjustment.

$$P(Y|do(X)) = \sum_m P(M = m|X) \sum_x P(Y|M = m, x)P(x). \qquad (9)$$

Fig. 4 (III) introduces a new variable M between the confounder and the output, which often indicates a high-dimensional feature of the effect of the confounder on the output. If the confounder C is available, the intervention method will also change.

$$P(Y|do(X)) = \sum_c P(Y|X, M, c)P(M|X, c)P(c). \qquad (10)$$

Otherwise, a mediator is needed to be introduced between X and Y, as shown in Eq. 9 for front-door intervention.

The intermediate variable M can also be placed between the confounder and the input, as shown in Fig. 4 (IV). In this case, conditioning on the intermediate variable M is equivalent to cutting off the spurious correlation.

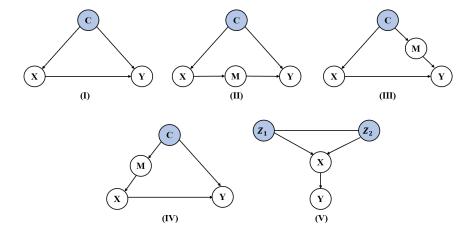$$P(Y|do(X)) = \sum_c P(Y|X, c)P(c, M = m). \qquad (11)$$

Fig. 4. These are the common structures involved in causality-based tasks, in which each colored part indicates confounders.

Different from the previous structures, the structure in Fig. 4 (V) focuses on the effect of different co-occurring confounders on the input. De-correlation of the co-occurring variables or elimination of the effect of these variables are desired in this situation. The potential outcome models proposed by Neyman *et al.* [33] to quantitatively extrapolate results under different conditions often used.

2) Methods for different concerns: The main concerns in introducing causal theory for performance enhancement can be divided into three aspects: accuracy, generalizability and interpretability. These three aspects have their priorities but overlap with each other.

When enhancing model interpretability from a causal perspective, the approach focuses on the intrinsic causal relationship between variables and develops further inference through this objective stable correlation. Methods that use causal theory for accuracy enhancement [34–36] often focus on the task in one given domain and rebuild the structure involving the hidden confounders. They analyze the limitations of existing statistical correlation-based methods at a holistic level and identify the causes of the confounders. With the structural causal model built on prior knowledge, such methods strive to cut off the spurious correlation and inference with the right causal chain. In these methods, confounders are often the specific variables in the inference process of the task. Methods that aim at improving generalizability [37, 38] often focus on the differences of one task between multiple domains and aim to achieve stable results against the domain gap. The confounders in these methods are often the domain and the corresponding context information. For generalizability, such methods aim to eliminate the effects of the domain and learn the domain-invariant features for subsequent inference. Methods that focus on interpretability [39, 40] often expect the model to rely on the correct visual region when making decisions. Instead of making interventions under the guidance of causal structure, such methods often use the potential outcome models and the counterfactual to estimate the impact of the variables as the basis for determining network behavior. Generative models are often used in these methods to eliminate variables with no causal relationship.

When enhancing generalizability from the causal perspective, the spurious correlation brought by the context can be eliminated. In this case, accuracy can also be enhanced since the noise in the context will affect the exploration of causal relationships even in a domain. Guided by prior knowledge, the structural causal model makes the method inherently interpretable.

These three concerns are used to delineate the existing articles in the following section as methods for accuracy enhancement, methods to improve generalizability, and methods to promote interpretability. Since existing work does not fully encompass these three concerns, some tasks will lack concerns. Nevertheless, the application of causal theory for performance enhancement is meaningful in every task.

B. Causal reasoning in vision tasks

This section will cover four tasks: classification, detection, segmentation, and visual recognition. The methods for each task will be classified according to the different concerns: accuracy, generalizability, and interpretability. The causal structure for analysis and the causal inference methods will also be discussed. The approaches that focus on accuracy often construct causal structures, analyze confounders in specific problems, and design effective causal inference methods based on the proposed structure. The approaches that focus on generalizability aim to achieve stable performance between different domains. They learn stable causal representations through methods such as causal inference and reweighting. The approaches that focus on interpretability try to explore the reasons for network decisions by generating new samples to identify the basis for network decisions.

1) Classification: Classification is a fundamental problem in computer vision, which tries to analyze the correlation

TABLE I
THIS IS A SUMMARY OF CAUSAL THEORY-BASED VISION TASKS (VISION-LANGUAGE INCLUDED). FOR CONVENIENCE, ABBREVIATIONS ARE USED TO DENOTE INDIVIDUAL TASKS: CLS FOR CLASSIFICATION, DET FOR OBJECT DETECTION, $SemSeg$ FOR SEMANTIC SEGMENTATION, $WSSS$ FOR WEAKLY SUPERVISED SEMANTIC SEGMENTATION, $MedSeg$ FOR MEDICAL IMAGE SEMANTIC SEGMENTATION, $VisRecog$ FOR TARGET RECOGNITION, $ImgCap$ FOR IMAGE CAPTIONING, VQA FOR VISUAL QUESTION ANSWERING, $3DRecon$ FOR 3D RECONSTRUCTION, $3DPose$ FOR 3D POSE ESTIMATION, AND $ObjNav$ FOR OBJECT NAVIGATION. THE STRUCTURES IN THE LAST COLUMN CORRESPOND TO THE STRUCTURES IN FIG. 4.

Paper	Task	Year	Concerns	Causal Instrument	Confounder	Structure
[37]	CLS	2017	generalizability	Causal Inference	Context	(I)
[39]	CLS	2019	Interpretability	Potential outcome models	Confounded concept	(V)
[34]	CLS	2020	Accuracy	Back-door adjustment	Momentum in the optimizer	(III)
[41]	CLS	2020	Accuracy	Back-door adjustment	Pre-trained knowledge	(III)
[35]	CLS	2021	generalizability	Potential outcome models	Old knowledge	(I)
[42]	CLS	2021	generalizability	Causal inference	Domain	(I)
[38]	CLS	2021	generalizability	Causal inference	Context	(I)
[43]	CLS	2021	generalizability	Causal inference	Dataset bias	(I)
[44]	CLS	2021	generalizability	Causal inference	Unobserved semantic feature	(I)
[45]	CLS	2022	generalizability	Front-door adjustment	Domain	(II)
[46]	CLS	2022	generalizability	Causal inference	Causal factors	(I)
[47]	CLS	2022	generalizability	Causal inference	Unobserved latent variable	(I)
[48]	CLS	2022	generalizability	Potential outcome models	Domain	(I)
[49]	CLS	2023	Accuracy	Potential outcome models	Unobservable variable	(I)
[50]	CLS	2023	Accuracy	Back-door adjustment	Task identifier	(I)
[51]	CLS	2023	generalizability	Counterfactual	Semantic concept	(I)
[36]	DET	2021	Accuracy	Back-door adjustment	Context	(III)
[52]	DET	2021	generalizability	Causal inference	Scene	(I)
[53]	DET	2022	Accuracy	Back-door adjustment	Contrast distribution	(II)
[54]	DET	2022	Accuracy	Back-door adjustment	Knowledge of pre-trained model	(III)
[55]	DET	2023	generalizability	Causal inference	Non-causal factors	(V)
[56]	$SemSeg$	2021	Accuracy	Counterfactual	Real feature	(I)
[57]	$SemSeg$	2022	Accuracy	Back-door adjustment	Context	(I)
[58]	$WSSS$	2020	Accuracy	Back-door adjustment	Context	(III)
[59]	$WSSS$	2022	Accuracy	Front-door adjustment	Class-specific characteristics	(II)
[60]	$WSSS$	2022	Accuracy	Causal inference	Context	(I)
[61]	$MedSeg$	2022	generalizability	Causal inference	Domain shift	(V)
[62]	$MedSeg$	2022	generalizability	Causal inference	Acquisition process	(III)
[63]	$VisRecog$	2021	Accuracy	Back-door adjustment	Dataset bias	(IV)
[64]	$VisRecog$	2022	Accuracy	Back-door adjustment	Context	(III)
[65]	$VisRecog$	2021	generalizability	Back-door adjustment	Context	(II)
[66]	$VisRecog$	2021	generalizability	Causal inference	Unobserved confounder	(IV)
[67]	$VisRecog$	2022	generalizability	Back-door adjustment	Concept features	(III)
[68]	$ImgCap$	2021	Accuracy	Front-door adjustment	Pre-training dataset	(I)&(II)
[69]	$ImgCap$	2021	Accuracy	Causal inference	Context	(II)
[70]	$ImgCap$	2022	Accuracy	Back-door adjustment	Visual and linguistic features	(II)&(IV)
[71]	VQA	2021	Accuracy	Counterfactual	Language bias	(I)
[72]	VQA	2020	generalizability	Counterfactual	Spurious correlation and bias	(I)
[73]	VQA	2020	generalizability	Back-door adjustment	Dataset biases	(I)
[40]	VQA	2020	Interpretability	Counterfactual	Language bias	(I)
[74]	VQA	2022	generalizability	Causal inference	Causal scene	(V)
[75]	VQA	2023	Accuracy	Causal inference	Domain	(V)
[76]	VQA	2023	Accuracy	Back-door and Front-door adjustment	unobserved confounder	(IV)
[77]	$3DRecon$	2022	Accuracy	Causal inference	Image	(I)
[78]	$3DPose$	2021	generalizability	Causal inference	Content	(I)
[79]	$ObjNav$	2023	Accuracy	Potential outcome models	Observation	(I)

between the images and the corresponding labels. Existing methods [80–82] mainly focus on the statistical correlations between the features and the labels, which will be heavily influenced by spurious correlations in the presence of noise [49] or inconsistent distributions [48]. Therefore, it is important to introduce causal theory into the classification task to learn the causal relationships between the images and the labels.

Methods for accuracy enhancement: The long-tail effect unbalances the impact of different classes on momentum and seriously affects classification accuracy. Tang *et al.* [34] assign the SGD momentum as the confounder and use the back-door adjustment on the structure, including the feature, the SGD momentum, the projection head, and the predicted label in Fig. 4 (III). Such *do*-operator removes the bad confounder bias while keeping the good mediator bias, eliminating the negative effect of SGD momentum. It is a paradox that a stronger pre-trained model may enlarge the dissimilarity between the support set and the query set in few-shot learning, thus affecting the classification accuracy. Yue *et al.* [41] point out that the pre-trained knowledge is the confounder that causes spurious correlations between the sample features and class labels in the support set. They introduce the transformed representation of sample features and use the back-door adjustment in the structure in Fig. 4 (III) to cut off the effect of pre-trained knowledge on the feature representation. The model's misperception of the imposed noise in the image affects the classification accuracy. Qiu *et al.* [50] find the new type of bias as task-induced bias and use the back-door adjustment to the structure in Fig. 4 (I) between image, label and task identifier to transforms biased features into unbiased ones. Yang *et al.* [49] summarized the effects of noise as unobserved confounders. Guided by the structure in Fig. 4 (I), they use the generative model to generate unobserved confounders for estimation and assess causal effects in noisy image classification tasks through treatment estimation. The learning of the robust representation is proposed against any unexpected noise.

Methods to improve generalizability: In practical image classification tasks, the assumption of independent identical distributions is unrealistic. In an ever-changing environment, there exist stable variables that remain constant and unstable variables that often change. The instabilities, in turn, create biases between domains, introduce spurious variables, and affect generalizability. Concerning the need for multi-domain generalization, methods are proposed with different structures. Lv *et al.* [46] build a structural causal model containing causal factors, non-causal factors, raw inputs, and the category label based on Fig. 4 (I). They use a causal intervention module, a causal factorization module, and an adversarial mask module for a more robust representation learning. With the same structure, a causal regularizer is proposed by Shen *et al.* [37] to balance the confounder distributions for each treatment feature by reweighting the samples. Through reweighting, the confounder distributions

can be balanced to correct for bias from the non-random treatment assignments. By adopting a different structure, as shown in Fig. 4 (II), Miao *et al.* [45] use the front-door adjustment due to the specific causal structure between seven variables: the domain, the object, the prior knowledge, the category factor, the domain identity, the unseen images, and the labels. They transfer unseen images to taught knowledge which are the features of seen images and cut off excess causal paths to calculate the causal effect. Chen *et al.* [51] propose a new learning paradigm, namely simulate-analyze-reduce to infer the causes of domain shift between the auxiliary and source domains during training. Through counterfactual inference, they reduce the effect of domain shift between semantic concept, image and category in Fig. 4 (I).

Methods to promote interpretability: It is essential to explain the classification decision drivers of the neural network, however, existing correlation-based explanation methods fail to consider the impact of the confounders and are easily affected by misleading information. Goyal *et al.* [39] propose the conditional generative model for the generation of counterfactuals and quantitatively measure the concept by causal concept effect method. With the structure shown in Fig. 4 (V), they model the relationships between the high-level concepts, the images, and the classifier output.

2) Detection: The goal of object detection is to determine where objects are located in a given image and to which category each object belongs. Different from the classification task, the detection task requires not only an accurate classification but also the precise location of the target in the image. Existing methods [83–85] have made significant achievements in practical applications concerning autonomous driving, but their effectiveness is heavily compromised by biases from complex scenarios. The introduction of causal theory into the detection tasks allows for accurate categorization by learning stable target properties on the one hand and precise localization by sorting out the causal interactions between the target and the environment on the other.

Methods for accuracy enhancement: Existing object detection methods tend to focus on the statistical correlation between instances, bounding boxes, and labels, ignoring spurious correlations introduced by contextual bias, which in turn results in decreased accuracy. Huang *et al.* [36] formulate the causalities in object detection tasks with the structure in Fig. 4 (III) and assign the context as the confounder. Through back-door adjustment, the non-causal but positive correlation between pixels and labels can be avoided. In unsupervised salient object detection tasks, biases caused by semantic contrast and image distribution heavily introduce spurious bias, thus limiting the improvement of performance in accuracy. To eliminate both the contrast distribution bias and the spatial distribution bias, Lin *et al.* [53] identify the visual contrast distribution as the confounder which misleads

the model training towards data-rich visual contrast clusters. They use the causal inference in the structure shown in Fig. 4 (IV) to make each visual contrast cluster contribute fairly and propose an image-level weighting strategy to eliminate the spatial bias. The application of knowledge distillation improves the models ability to capture semantic information in few-shot object detection. However, the empirical error of the teacher model degenerates the student models prediction of the target labels, interfering with the students prediction accuracy on downstream tasks. Li *et al.* [54] designate classification knowledge for specific domains as the confounder and use the back-door adjustment in the structure in Fig. 4 (IV) to remove the effect of the semantic knowledge and re-merge the remaining variables as new knowledge distillation targets.

Methods to improve generalizability: Automated driving technology requires extremely high levels of safety and robustness assurance. Perception modules trained on neural networks are often used for target detection. However, modules with better training results often fail to guarantee performance in unknown scenarios due to deviations between the target and training domains. Resnick *et al.* [52] collect real-world data through an automated driving simulator and make causal interventions on the data to discriminate factors detrimental to safe driving. In turn, the source of the domain variation problem is resolved in advance, and the deviation is eliminated. Xu *et al.* [55] propose to remove non-causal factors from common features by multi-view adversarial training on source domains to remove the non-causal factors and purify the domain-invariant features. They clearly clarify the relationships among causal factors, noncausal factors, domain specific feature and domain common feature from a causal perspective.

3) Segmentation: Segmentation is equivalent to a pixel-wise classification problem as the most demanding task in terms of accuracy, and the spatial-semantic uncertainty principle is the main challenge [86]. Among the causality-based segmentation tasks, fully supervised and weakly supervised segmentation problems tend to focus on accuracy, while medical segmentation problems are more concerned with generalizability to ensure safety.

Methods for accuracy enhancement: The wrongly explored semantic correlations between word embeddings and visual features in generative models can lead to spurious correlations and compromise performance. To eliminate the bias between visible and invisible classes caused by confounders in zero-sample semantic segmentation, Shen *et al.* [56] adopt the counterfactual theory with the causal structure containing true and false features, word embeddings, and labels. With the same structure as Fig. 4 (I), a causal structure consisting of images, category labels, and contextual information is constructed by Li *et al.* [57]. The confusion bias is eliminated through causal intervention, while a fusion module is designed to fuse original features and causal

contextual relationships for a closer resemblance to the human learning process. For high-quality seed regions in weakly supervised semantic segmentation, Zhang *et al.* [58] construct the structural causal model between pixels, contexts, and labels as shown in Fig. 4 (III) and introduce a back-door adjustment to eliminate the negative impact of contexts on class activation graph generation as shown in Fig. 5. Different from the settings in [58], Wang *et al.* [59] attribute the poor performance of existing methods to a set of class-specific latent confounders in the dataset and analyze the causality between image, image-level tag, pixel-level localization, and a set of class-specific latent confounders, referring to Fig. 4 (II). They use the front-door adjustment to cut off the spurious correlation between the confounder and the images. Focusing on weakly supervised semantic segmentation in medical images, an approach that focuses on both the category-causality chain and the anatomy-causality chain is proposed by Chen *et al.* [60]. They use causal interventions to get the actual cause of category prediction and integrate anatomical constraints for the actual cause of segmentation.

Methods to improve generalizability: Generalizability is a top priority in medical image segmentation due to the personal safety involved. Causal reasoning is frequently used in medical image processing to mitigate domain shifts caused by imaging modalities, scanning protocols, and device manufacturers [61, 62]. A causal structure involving the operating system, the surgical environment, and the segmentation graph is built by Ding *et al.* [61]. They align tool models with image observations by updating the initially incorrect robot kinematic parameters through forward kinematics and differentiable rendering to optimize im- age feature similarity end-to-end. With a more formal causal structure between acquisition, content, feature maps, medical images, and the ground-truth segmentation mask, Ouyang *et al.* [62] propose a simple causality-inspired data augmentation approach to expose a segmentation model to synthesized domain-shifted training examples.

4) Visual recognition: Visual recognition aims to imitate the human visual system as much as possible. Previous works [87–89] have noticed the inconsistency between the training set and the application environment, compensating for the bias effect through data augmentation, re-weighted loss, and normalization. However, such methods cannot design generic methods of eradicating deviations for different scenarios. As a result, the revisiting of the fundamental process of building a visual recognition system is processed with the help of causal theory to distinguish the paradoxical character of the bias.

Methods for accuracy enhancement: Contextual bias misdirects attention to the co-occurrence context rather than the objects, leading to the loss of accuracy. Liu *et al.* [64] build the structure between the object representations, prior context knowledge, image-specific context, and predictions.
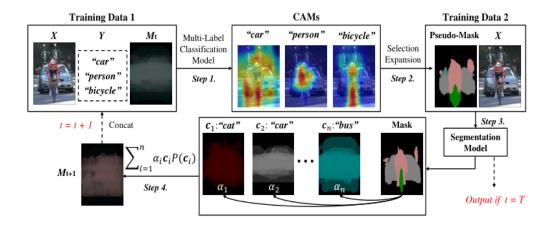
Fig. 5. This is the overview of the proposed Context Adjustment (CONTA) in [58]. In their setting, Y represents a set of given labels indicating the specific classes of the objects, c_i represents the average mask of the ith class images, and M can be viewed as a linear combination of c_i. This weakly supervised semantic segmentation task aims to enhance the quality of the generated class activation maps.
   Since the confounder proposed is unobserved, it uses an iterative procedure to establish $P(c)$ in a back-door adjustment. The confounders are better learned through iteration, resulting in high-quality class activation maps for segmentation.

They propose a novel paradigm with both back-door adjustment and counterfactual inference to conquer the effect of contextual bias with the structure in Fig. 4 (III). Dataset bias always contributes to the biased statistic correlation-based model, resulting in decreased performance. Qin *et al.* [63] argue that such bias misleads the correlation between the input images and the output labels. They build the causal structure between image, label, context, common sense, and bias, referring to the structure in Fig. 4 (IV), and cut off the backdoor path involving the confounder by the back-door adjustment.

Methods to improve generalizability: Since the association between images and labels is not generalizable across domains, the out-of-distribution performance is always poor. Guided by the causal-transportability language [90], Mao *et al.* [67] build a causal structure between the input images, the corresponding labels, and unobserved variables encoding external sources of variation not captured in the images and the labels themselves with the structure in Fig. 4 (I). From a more complex structure in Fig. 4 (II), Mao *et al.* [66] build a causal structure between six variables. Variables include object features, unobserved confounders, background features, images, and labels. Aware that attention mechanisms can no longer be robustly characterized in any confusing environment, Wang *et al.* [65] focus on the deficiency of the attention mechanism when generating robust representations and propose a causal attention module that self-annotates the context confounders in an unsupervised fashion.

C. Causal reasoning in vision-language tasks

This section will cover two tasks: image captioning and visual question answering. The concerns, corresponding causal structure, and causal inference methods will also be discussed.

Different from vision tasks, vision-language tasks aim to combine the information from vision, and language to perform complex tasks that mimic human behavior [91, 92]. The main training steps for visual-language tasks involve encoding images and texts into single-modal embeddings for representation learning, designing an encoder to integrate information from both modalities, and using different aggregation methods. Despite the great success of existing methods [93, 94], the bias between modalities cannot be ignored, as the semantics of images and language differ significantly and are all susceptible to the influence of the environment. These confounding factors confuse the model regarding the causal chains between important information and thus deteriorate performance. In order to avoid bias, it is necessary to reconsider the link between two modalities from a causal viewpoint in addition to learning stable characteristics from each modality separately.

1) Image captioning: Image captioning is a task that aims to decipher automatically the semantic information contained in an image and produce an accurate description of it [95]. Numerous efforts [96, 97] have been made to improve the performance of image captioning systems, but the endogenous language bias is neglected since the existing image captioning models are inclined to build spurious connections between images and the high-frequency concurrent categories. By applying causal theory to this task, it is necessary to learn a stable representation from the images and analyze the causal relationships between images and text for appropriate responses.

Methods for accuracy enhancement: Since the dataset bias is inevitable, the accuracy performance is always affected by spurious associations caused by the bias. Yang *et al.* [68] introduce the semantic structure set as a mediator and use the front-door adjustment in Fig. 4 (II) combined with the

back-door adjustment in Fig. 4 (I) to eliminate the spurious correlation caused by language resources, which is denoted as the confounder. As a result, the dataset bias introduced by the confounder is reduced. Noticing that Yang *et al.* [68] neglect the confounded visual features in the encoder, Chen *et al.* [69] eliminate the spurious correlation between visual features and certain expressions by using the back-door adjustment and estimate the confounder with variational inference. Considering the limitation in [69] that the pre-training dataset is hard to stratify, the confounder is explicitly divided into two classes by Liu *et al.* [70]. In most transformer-based image captioning methods, the bias caused by both visual and linguistic confounders is often overlooked. However, the resulting spurious correlations often compromise the accuracy of the network. Liu *et al.* [70] propose to disentangle the region-based features and use the back-door adjustment to deconfound the causal structure as shown in Fig. 4 (II) and (V), which can effectively eliminate the spurious correlations caused by both visual and linguistic confounders.

2) Visual question answering: Given an image-question pair, visual question answering (VQA) tasks expect the model to answer questions correctly based on the given information. Since the existing models are proven to be fragile to linguistic variations in questions and answers [98], the causal relationships between images and language information are investigated to grasp the meaning of the inquiry.

Methods for accuracy enhancement: Due to the inclusion of both image and language modal data, VQA requires visual analysis, language understanding, and multi-modal reasoning. However, existing models suffer from linguistic bias and do not make correct use of semantic information. Niu *et al.* [71] refer to Fig. 4 (I) to analyze this problem and mitigate the language bias by subtracting the direct language effect from the total causal effect with the counterfactual inference as shown in Fig. 6. Zang *et al.* [75] are concerned about the bias introduced by multi-modal data in VQA tasks. They capture visual features related to the semantics of the question and weaken the influence of local language semantics on the question answering. Liu *et al.* [76] collaboratively disentangle the spurious correlations between vision and language modal through back-door adjustment and front-door adjustment. They capture the fine-grained interactions between visual and linguistic semantics and learn the global semantic-aware visual-linguistic representations adaptively.

Methods to improve generalizability: Recent VQA models are proven to be brittle to linguistic and semantic variations in questions and images. Agarwal *et al.* [72] propose to enhance the robustness through automated semantic image manipulations and tests for consistency in model predictions. A data augmentation strategy that involves adding noise to the input image to improve model performance on challenging counting questions is also introduced. Li *et al.* [74] propose a learning framework to ground the question-critical scene for invariant inference across different scenes.

Methods to promote interpretability: Enhancing the interpretable aspects of the model helps to understand the basis on which the model is used in decision-making. However, existing methods are guided by attention and this statistic-based approach is vulnerable to language bias. Also, there are issues with existing methods such as the need for additional annotations. Chen *et al.* [40] consider the VQA task as a multi-class classification problem and use counterfactual analysis to synthesize the samples, encouraging the model to perceive the difference in questions when changing some critical words and clarify the basis for their decisions.

IV. Future roadmaps

A. Reasonable causal structure

Though causality has shown great potential in improving accuracy, generalizability, and interpretability, all the approaches are premised on the causal structures that are tailored to the specific problems [99]. However, many of the variables involved in the vision tasks are simply ignored and only some important ones are treated as nodes to construct the simplified structure. Different choices of variables can lead to different structures, thus leading to different perspectives on the same task. When solving the problem of weakly supervised semantic segmentation, Zhang *et al.* [58] construct the causal structure between pixels, contexts, and labels, while Chen *et al.* [60] build the structure between images, categories, context confounders, anatomical structures, shapes of segmentation, and pseudo masks. Although addressing the same issues, the two approaches have their focus, one on contextual information and the other on biased category information. While constructing causal structures is key to causal learning, true causality is complex and there is no uniform modeling framework. When analyzing complex vision problems, it is of great importance to model a reasonable, and comprehensive causal structure for a clearer interpretation of the relationships between visually relevant variables.

B. More guidance for multi-modal fusion

Multi-modal fusion integrates different modal information to overcome the restrictions of incomplete information given by a single modality, hence realizing modal information complementarity and improving feature representation [100]. However, data fusion is challenging. Data generation is driven by numerous underlying processes that depend on many inaccessible variables, while the data itself is always heterogeneous and complex. Data fusion aims to allow modalities to communicate entirely and inform each other. As a result, selecting an analytical model that properly depicts the link between modalities and provides a meaningful combination of them is critical. [101]. Causality can uncover causal chains between variables objectively and can further guide analytical model building between multi-modal data to enhance
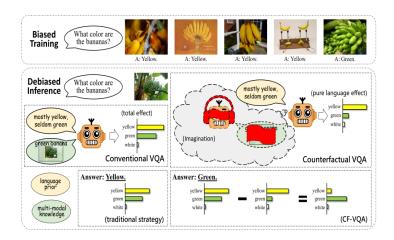
Fig. 6. The biased training section shows the negative impact of language bias. The trained model will focus more on the question than the visual content, which seriously limits the generalization. The debiased strategy is proposed for a better generalization when the distributions are inconsistent across different stages. With this motivation, a counterfactual inference framework is constructed based on language-prior knowledge in [71].

accuracy, generalizability, and interpretability. For example, image data enables the precise visualization of the location of problems, and time-series data offers a thorough history of changes in the pertinent variables. A reliable causal chain is constructed by detecting faults with image data and analyzing the root cause of faults with time-series data, enabling more accurate fault detection and root cause identification. The interpretability of causal theories can also enhance the interactive ability of multi-modal fusion.

C. More applications of causal theory

So far, the applications of causal theory to vision and vision-language tasks have been discussed. Since the causal perspective can explore the relationships between variables at a fundamental level, it is suitable for guiding the study of complex systems and tasks, such as industrial processes [102, 103], electrical systems [104], and Earth systems [105]. Robots must make stable and timely decisions in complex industrial processes in response to the ever-changing environment. It is also important to make the decision-making process of robots transparent to humans. Recently, the task plan explanation methods [106, 107] have promoted human-robot interaction in robotics and enhanced the interpretability of robotic decision-making mechanisms with the help of causality. The generalization techniques also make the deployment of tasks in unknown environments a reality. The power system is a unified whole capable of producing, transmitting, distributing, and consuming electrical energy. Recently, with the improvement of top-level design, the power system is developing towards safety, high efficiency, low carbon, and intelligent integration. However, the strong interactions between systems introduce new challenges in maintaining high supply security, as new factors can affect the overall security of the power system [108]. The causal theory can be used to explain the relationships between power system variables, respond

to power system trends, and assist researchers by providing them with more interpretable analysis. In large-scale complex dynamical systems such as the Earth system, anthropogenic intervention is uncontrolled and against humanitarianism. To find out the mechanisms between such complex systems, the causal theory is the best choice [109]. With an explosion in the availability of large-scale time series data and an increasingly accurate perception of the Earth model, the causal interdependencies of the underlying system can be discovered and quantified to improve the theoretical understanding of the Earth system. Combining time series, satellite remote sensing images, and other multi-source data, the nonlinear dynamical interactions between different climate factors can be analyzed and the effective prevention of tipping points can be predicted. In addition, a deeper understanding of the influence of meteorological parameters on meteorological phenomena can be obtained.

V. CONCLUSION

This review aims to contribute to the development of causality in typical computer vision tasks and provide detailed explanations of existing methods from the perspective of causal structures. The concept and necessity of causality in guiding vision tasks are first discussed. The survey on existing methods of causal reasoning in vision and vision-language tasks is then conducted from different perspectives: accuracy, generalizability, and interpretability. Finally, future roadmaps are suggested to encourage the theory and applications of this promising field.

REFERENCES

[1] Y. Pang, X. Bai, and G. Zhang, "Special focus on deep learning for computer vision," *SCIENCE CHINA Information Sciences*, vol. 62, no. 12, pp. 220 100–, 2019.

[2] D. Yang, K. Jiang, D. Zhao, C. Yu, Z. Cao, S. Xie, Z. Xiao, X. Jiao, S. Wang, and K. Zhang, "Intelligent and connected vehicles: Current status and future perspectives," *SCIENCE CHINA Technological Sciences*, vol. 61, no. 10, pp. 1446–1471, 2018.

[3] D. Li, M. Liu, F. Zhao, and Y. Liu, "Challenges and countermeasures of interaction in autonomous vehicles," *SCIENCE CHINA Information Sciences*, vol. 62, no. 5, pp. 050 201–, 2019.

[4] Y. Wang, H. Song, Q. Li, and H. Zhang, "Research on a full envelop controller for an unmanned ducted-fan helicopter based on switching control theory," *SCIENCE CHINA Technological Sciences*, vol. 62, no. 10, pp. 1837–1844, 2019.

[5] Z. Xie, Q. Zhang, Z. Jiang, and H. Liu, "Robot learning from demonstration for path planning: A review," *SCIENCE CHINA Technological Sciences*, vol. 63, no. 8, pp. 1325–1334, 2020.

[6] Z. Chu, D. Jie, L. Su, J. Cui, and S. Fuchun, "A gecko-inspired adhesive robotic end effector for critical-contact manipulation," *SCIENCE CHINA Information Sciences*, vol. 65, no. 8, pp. 182 203–, 2022.

[7] H. Li, S. Cao, Y. Chen, m. Zhang, and D. Feng, "Tulam: Trajectory-user linking via attention mechanism," *SCIENCE CHINA Information Sciences*, pp. –, 2023. [Online]. Available: http://www.sciengine.com/publisher/ScienceChinaPress/journal/SCIENCECHINAInformationSciences///10.1007/s11432-021-3673-6,

[8] G. Cheng, P. Lai, D. Gao, and J. Han, "Class attention network for image recognition," *SCIENCE CHINA Information Sciences*, vol. 66, no. 3, pp. 132 105–, 2023.

[9] P. Yan, Y. Tan, and Y. Tai, "Repeatable adaptive key-point detection via self-supervised learning," *SCIENCE CHINA Information Sciences*, vol. 65, no. 11, pp. 212 103–, 2022.

[10] Y. Shao, Z. Geng, Y. Liu, J. Dai, H. Yan, F. Yang, Z. Li, H. Bao, and X. Qiu, "Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation," *SCIENCE CHINA Information Sciences*, pp. –, 2022.

[11] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[12] J. Pearl, *Causality*. Cambridge university press, 2009.

[13] C. Gao, Y. Zheng, W. Wang, F. Feng, X. He, and Y. Li, "Causal inference in recommender systems: A survey and future directions," *arXiv preprint arXiv:2208.12397*, 2022.

[14] C. R. Blyth, "On simpson's paradox and the sure-thing principle," *Journal of the American Statistical Association*, vol. 67, no. 338, pp. 364–366, 1972.

[15] D. Borsboom, R. A. Kievit, D. Cervone, and S. B. Hood, *The Two Disciplines of Scientific Psychology, or: The Disunity of Psychology as a Working Hypothesis*. New York, NY: Springer US, 2009, pp. 67–97.

[16] N. Malik and P. V. Singh, "Deep learning in computer vision: Methods, interpretation, causation, and fairness," in *Operations Research & Management Science in the Age of Analytics*. INFORMS, 2019, pp. 73–100.

[17] Q. Sun, C. Zhao, Y. Tang, and F. Qian, "A survey on unsupervised domain adaptation in computer vision tasks," *SCIENTIA SINICA Technologica*, vol. 52, no. 1, pp. 26–54, 2022.

[18] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization in vision: A survey," *arXiv preprint arXiv:2103.02503*, 2021.

[19] R. E. Heidel, "Causality in statistical power: Isomorphic properties of measurement, research design, effect size, and sample size," *Scientifica*, vol. 2016, p. 8920418, 2016.

[20] A. P. Dawid, "Statistical causality from a decision-theoretic perspective," *Annual Review of Statistics and Its Application*, vol. 2, pp. 273–303, 2015.

[21] J. J. Heckman and R. Pinto, "Causality and econometrics," National Bureau of Economic Research, Tech. Rep. 29787, 2022.

[22] J. Geweke, "Inference and causality in economic time series models," *Handbook of econometrics*, vol. 2, pp. 1101–1144, 1984.

[23] M. Kundi, "Causality and the interpretation of epidemiologic evidence," *Environmental Health Perspectives*, vol. 114, no. 7, pp. 969–974, 2006.

[24] H. Ohlsson and K. S. Kendler, "Applying causal inference methods in psychiatric epidemiology: A review," *JAMA psychiatry*, vol. 77, no. 6, pp. 637–644, 2020.

[25] J. F. Hair Jr and M. Sarstedt, "Data, measurement, and causal inferences in machine learning: opportunities and challenges for marketing," *Journal of Marketing Theory and Practice*, vol. 29, no. 1, pp. 65–77, 2021.

[26] M. Prosperi, Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He, S. Rich, M. Wang, I. E. Buchan, and J. Bian, "Causal inference and counterfactual prediction in machine learning for actionable healthcare," *Nature Machine Intelligence*, vol. 2, no. 7, pp. 369–375, 2020.

[27] H. Chen, K. Du, X. Yang, and C. Li, "A review and roadmap of deep learning causal discovery in different variable paradigms," *arXiv preprint arXiv:2209.06367*, 2022.

[28] J. Pearl, "Bayesian networks," 2011.

[29] J. Kaddour, A. Lynch, Q. Liu, M. J. Kusner, and R. Silva, "Causal machine learning: A survey and open problems," *arXiv preprint arXiv:2206.15475*, 2022.

[30] Z. Li, Z. Zhu, X. Guo, S. Zheng, Z. Guo, S. Qiang, and Y. Zhao, "A Survey of Deep Causal Models and Their Industrial Applications," 2023.

[31] G. Rebane and J. Pearl, "The recovery of causal poly-trees from statistical data," *arXiv preprint arXiv:1304.2736*, 2013.

[32] D. C. Castro, I. Walker, and B. Glocker, "Causality matters in medical imaging," *Nature Communications*, vol. 11, no. 1, p. 3673, 2020.

[33] J. Splawa-Neyman, D. M. Dabrowska, and T. P. Speed, "On the application of probability theory to agricultural experiments. essay on principles. section 9." *Statistical Science*, pp. 465–472, 1990.

[34] K. Tang, J. Huang, and H. Zhang, "Long-tailed classification by keeping the good and removing the bad momentum causal effect," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1513–1524, 2020.

[35] X. Hu, K. Tang, C. Miao, X.-S. Hua, and H. Zhang, "Distilling causal effect of data in class-incremental learning," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021, pp. 3957–3966.

[36] W. Huang, M. Jiang, M. Li, B. Meng, J. Ren, S. Zhao, R. Bai, and Y. Yang, "Causal intervention for object detection," in *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2021, pp. 770–774.

[37] Z. Shen, P. Cui, K. Kuang, B. Li, and P. Chen, "On image classification: Correlation vs causality," *arXiv preprint arXiv:1708.06656*, 2017.

[38] C. Liu, X. Sun, J. Wang, H. Tang, T. Li, T. Qin, W. Chen, and T.-Y. Liu, "Learning causal semantic representation for out-of-distribution prediction," *Advances in Neural Information Processing Systems*, vol. 34, pp. 6155–6170, 2021.

[39] Y. Goyal, A. Feder, U. Shalit, and B. Kim, "Explaining classifiers with causal concept effect (cace)," *arXiv preprint arXiv:1907.07165*, 2019.

[40] L. Chen, X. Yan, J. Xiao, H. Zhang, S. Pu, and Y. Zhuang, "Counterfactual samples synthesizing for robust visual question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 800–10 809.

[41] Z. Yue, H. Zhang, Q. Sun, and X.-S. Hua, "Interventional few-shot learning," *Advances in neural information processing systems*, vol. 33, pp. 2734–2746, 2020.

[42] D. Mahajan, S. Tople, and A. Sharma, "Domain generalization using causal matching," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7313–7324.

[43] X. Sun, B. Wu, X. Zheng, C. Liu, W. Chen, T. Qin, and T.-Y. Liu, "Recovering latent causal factor for generalization to distributional shifts," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 846–16 859, 2021.

[44] Z. Yue, Q. Sun, X.-S. Hua, and H. Zhang, "Transporting causal mechanisms for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8599–8608.

[45] Q. Miao, J. Yuan, and K. Kuang, "Domain generalization via contrastive causal learning," *arXiv preprint arXiv:2210.02655*, 2022.

[46] F. Lv, J. Liang, S. Li, B. Zang, C. H. Liu, Z. Wang, and D. Liu, "Causality inspired representation learning for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8046–8056.

[47] X. Wang, M. Saxon, J. Li, H. Zhang, K. Zhang, and W. Y. Wang, "Causal balancing for domain generalization," *arXiv preprint arXiv:2206.05263*, 2022.

[48] Y. Wang, F. Liu, Z. Chen, Y.-C. Wu, J. Hao, G. Chen, and P.-A. Heng, "Contrastive-ace: Domain generalization through alignment of causal mechanisms," *IEEE Transactions on Image Processing*, vol. 32, pp. 235–250, 2022.

[49] C.-H. H. Yang, I.-T. Hung, Y.-C. Liu, and P.-Y. Chen, "Treatment learning causal transformer for noisy image classification," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023, pp. 6139–6150.

[50] B. Qiu, H. Li, H. Wen, H. Qiu, L. Wang, F. Meng, Q. Wu, and L. Pan, "Cafeboost: Causal feature boost to eliminate task-induced bias for class incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 016–16 025.

[51] J. Chen, Z. Gao, X. Wu, and J. Luo, "Meta-causal learning for single domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7683–7692.

[52] C. Resnick, O. Litany, A. Kar, K. Kreis, J. Lucas, K. Cho, and S. Fidler, "Causal bert: Improving object detection by searching for challenging groups," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2972–2981.

[53] X. Lin, Z. Wu, G. Chen, G. Li, and Y. Yu, "A causal debiasing framework for unsupervised salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 1610–1619.

[54] J. Li, Y. Zhang, W. Qiang, L. Si, C. Jiao, X. Hu, C. Zheng, and F. Sun, "Disentangle and remerge: Interventional knowledge distillation for few-shot object detection from a conditional causal perspective," *arXiv preprint arXiv:2208.12681*, 2022.

[55] M. Xu, L. Qin, W. Chen, S. Pu, and L. Zhang, "Multiview adversarial discriminator: Mine the non-causal factors for object detection in unseen domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8103–8112.

[56] F. Shen, J. Liu, and P. Hu, "Conterfactual generative zero-shot semantic segmentation," *arXiv preprint arXiv:2106.06360*, 2021.

[57] W. Li and Z. Li, "Causal-setr: A segmentation transformer variant based on causal intervention," in *Proceedings of the Asian Conference on Computer Vision*,

2022, pp. 756–772.

[58] D. Zhang, H. Zhang, J. Tang, X.-S. Hua, and Q. Sun, "Causal intervention for weakly-supervised semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 655–666, 2020.

[59] Y. Wang, "Causal class activation maps for weakly-supervised semantic segmentation," in *UAI 2022 Workshop on Causal Representation Learning*, 2022.

[60] Z. Chen, Z. Tian, J. Zhu, C. Li, and S. Du, "C-cam: Causal cam for weakly supervised semantic segmentation on medical image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 676–11 685.

[61] H. Ding, J. Zhang, P. Kazanzides, J. Y. Wu, and M. Unberath, "Carts: Causality-driven robot tool segmentation from vision and kinematics data," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VII.* Springer, 2022, pp. 387–398.

[62] C. Ouyang, C. Chen, S. Li, Z. Li, C. Qin, W. Bai, and D. Rueckert, "Causality-inspired single-source domain generalization for medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 42, no. 4, pp. 1095–1106, 2023.

[63] W. Qin, H. Zhang, R. Hong, E.-P. Lim, and Q. Sun, "Causal interventional training for image recognition," *IEEE Transactions on Multimedia*, vol. 25, pp. 1033–1044, 2023.

[64] R. Liu, H. Liu, G. Li, H. Hou, T. Yu, and T. Yang, "Contextual debiasing for visual recognition with causal mechanisms," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 12 755–12 765.

[65] T. Wang, C. Zhou, Q. Sun, and H. Zhang, "Causal attention for unbiased visual recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3091–3100.

[66] C. Mao, A. Cha, A. Gupta, H. Wang, J. Yang, and C. Vondrick, "Generative interventions for causal learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3947–3956.

[67] C. Mao, K. Xia, J. Wang, H. Wang, J. Yang, E. Bareinboim, and C. Vondrick, "Causal transportability for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7521–7531.

[68] X. Yang, H. Zhang, and J. Cai, "Deconfounded image captioning: A causal retrospect," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[69] W. Chen, J. Tian, C. Fan, H. He, and Y. Jin, "Dependent multi-task learning with causal intervention for image captioning," *arXiv preprint arXiv:2105.08573*, 2021.

[70] B. Liu, D. Wang, X. Yang, Y. Zhou, R. Yao, Z. Shao, and J. Zhao, "Show, deconfound and tell: Image captioning with causal inference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 041–18 050.

[71] Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, and J.-R. Wen, "Counterfactual vqa: A cause-effect look at language bias," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 700–12 710.

[72] V. Agarwal, R. Shetty, and M. 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*, 2020, pp. 9690–9698.

[73] S. Zhang, T. Jiang, T. Wang, K. Kuang, Z. Zhao, J. Zhu, J. Yu, H. Yang, and F. Wu, "Devlbert: Learning deconfounded visio-linguistic representations," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 4373–4382.

[74] Y. Li, X. Wang, J. Xiao, W. Ji, and T.-S. Chua, "Invariant grounding for video question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2928–2937.

[75] C. Zang, H. Wang, M. Pei, and W. Liang, "Discovering the real association: Multimodal causal reasoning in video question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 027–19 036.

[76] Y. Liu, G. Li, and L. Lin, "Cross-modal causal relational reasoning for event-level visual question answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–17, 2023.

[77] W. Liu, Z. Liu, L. Paull, A. Weller, and B. Schölkopf, "Structural causal 3d reconstruction," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I.* Springer, 2022, pp. 140–159.

[78] X. Zhang, Y. Wong, X. Wu, J. Lu, M. Kankanhalli, X. Li, and W. Geng, "Learning causal representation for training cross-domain pose estimator via generative interventions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 270–11 280.

[79] S. Zhang, X. Song, W. Li, Y. Bai, X. Yu, and S. Jiang, "Layout-based causal inference for object navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 792–10 802.

[80] C.-F. R. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 357–366.

[81] Y. Li, C.-Y. Wu, H. Fan, K. Mangalam, B. Xiong,

J. Malik, and C. Feichtenhofer, "Mvitv2: Improved multiscale vision transformers for classification and detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4804–4814.

[82] T. Wang, Y. Gu, G. Gao, and X. Zeng, "A coupled multi-task feature boosting method for remote sensing scene classification," *SCIENCE CHINA Technological Sciences*, vol. 66, no. 3, pp. 663–673, 2023.

[83] N. Zeng, P. Wu, Z. Wang, H. Li, W. Liu, and X. Liu, "A small-sized object detection oriented multi-scale feature fusion approach with application to defect detection," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–14, 2022.

[84] L. Fang, Q. Tang, L. Ouyang, J. Yu, J. Lin, S. Ding, and L. Tang, "Long-tailed object detection of kitchen waste with class-instance balanced detector," *SCIENCE CHINA Technological Sciences*, pp. –, 2023.

[85] X. Xie, G. Cheng, Q. Li, S. Miao, K. Li, and J. Han, "Fewer is more: Efficient object detection in large aerial images," *SCIENCE CHINA Information Sciences*, pp. –, 2023.

[86] Q. Geng, Z. Zhou, and X. Cao, "Survey of recent progress in semantic image segmentation with cnns," *SCIENCE CHINA Information Sciences*, vol. 61, pp. 1–18, 2018.

[87] L. Yuan, Q. Hou, Z. Jiang, J. Feng, and S. Yan, "Volo: Vision outlooker for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 6575–6586, 2023.

[88] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 519–16 529.

[89] X. Wei, S. Xu, H. Chen, L. Xiao, and Y. Peng, "Prototype-based classifier learning for long-tailed visual recognition," *SCIENCE CHINA Information Sciences*, vol. 65, no. 6, pp. 160 105–, 2022.

[90] E. Bareinboim and J. Pearl, "A general algorithm for deciding transportability of experimental results," *Journal of causal Inference*, vol. 1, no. 1, pp. 107–134, 2013.

[91] Y. Du, Z. Liu, J. Li, and W. X. Zhao, "A survey of vision-language pre-trained models," *arXiv preprint arXiv:2202.10936*, 2022.

[92] H. Jin, Y. Cao, T. Wang, X. Xing, and W. Xiaojun, "Recent advances of neural text generation: Core tasks, datasets, models and challenges," *SCIENCE CHINA Technological Sciences*, vol. 63, no. 10, pp. 1990–2010, 2020.

[93] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2019, pp. 3195–3204.

[94] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 13 041–13 049.

[95] J. Wang, Y. Li, Y. Pan, T. Yao, J. Tang, and T. Mei, "Contextual and selective attention networks for image captioning," *SCIENCE CHINA Information Sciences*, vol. 65, no. 12, pp. 1–15, 2022.

[96] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 578–10 587.

[97] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 971–10 980.

[98] A. Ray, K. Sikka, A. Divakaran, S. Lee, and G. Burachas, "Sunny and dark outside?! improving answer consistency in vqa through entailed question generation," *arXiv preprint arXiv:1909.04696*, 2019.

[99] Y. Liu, Y. Wei, H. Yan, G. Li, and L. Lin, "Causal reasoning meets visual representation learning: A prospective study," *Machine Intelligence Research*, pp. 1–27, 2022.

[100] J. Li and Q. Wang, "Multi-modal bioelectrical signal fusion analysis based on different acquisition devices and scene settings: Overview, challenges, and novel orientation," *Information Fusion*, vol. 79, pp. 229–247, 2022.

[101] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[102] J. Fu, Y. Lv, and W. Yu, "Robust adaptive time-varying region tracking control of multi-robot systems," *SCIENCE CHINA Information Sciences*, vol. 66, no. 5, pp. 159 202–, 2023.

[103] Y. Zhang, C. Yang, S. Xu, and Y. Ou, "Obstacle avoidance in human-robot cooperative transportation with force constraint," *SCIENCE CHINA Information Sciences*, vol. 66, no. 1, pp. 119 205–, 2023.

[104] l. Jin, Y. He, C. Zhang, X. Shangguan, L. Jiang, and M. Wu, "Equivalent input disturbance-based load frequency control for smart grid with air conditioning loads," *SCIENCE CHINA Information Sciences*, vol. 65, no. 2, pp. 122 205–, 2022.

[105] X. Chen, Z. Gong, X. Zhao, W. Zhou, and W. Yao, "A machine learning surrogate modeling benchmark for temperature field reconstruction of heat source systems," *SCIENCE CHINA Information Sciences*, vol. 66, no. 5, pp. 152 203–, 2023.

[106] F. Lindner and C. Olz, "Step-by-step task plan expla-

nations beyond causal links," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*.   IEEE, 2022, pp. 45–51.

[107] A. Daruna, D. Das, and S. Chernova, "Explainable knowledge graph embedding: Inference reconciliation for knowledge inferences supporting robot actions," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.   IEEE, 2022, pp. 1008–1015.

[108] R. V. Yohanandhan, R. M. Elavarasan, R. Pugazhendhi, M. Premkumar, L. Mihet-Popa, J. Zhao, and V. Terzija, "A specialized review on outlook of future cyber-physical power system (cpps) testbeds for securing electric power grid," *International Journal of Electrical Power & Energy Systems*, vol. 136, p. 107720, 2022.

[109] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí *et al.*, "Inferring causation from time series in earth system sciences," *Nature communications*, vol. 10, no. 1, p. 2553, 2019.