Demystifying Causal Features on Adversarial Examples and Causal Inoculation for Robust Network by Adversarial Instrumental Variable Regression

Junho Kim,* Byung-Kwan Lee,* Yong Man Ro[†]

Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea

{arkimjh, leebk, ymro}@kaist.ac.kr

Abstract

The origin of adversarial examples is still inexplicable in research fields, and it arouses arguments from various viewpoints, albeit comprehensive investigations. In this paper, we propose a way of delving into the unexpected vulnerability in adversarially trained networks from a causal perspective, namely adversarial instrumental variable (IV) regression. By deploying it, we estimate the causal relation of adversarial prediction under an unbiased environment dissociated from unknown confounders. Our approach aims to demystify inherent causal features on adversarial examples by leveraging a zero-sum optimization game between a casual feature estimator (i.e., hypothesis model) and worst-case counterfactuals (i.e., test function) disturbing to find causal features. Through extensive analyses, we demonstrate that the estimated causal features are highly related to the correct prediction for adversarial robustness, and the counterfactuals exhibit extreme features significantly deviating from the correct prediction. In addition, we present how to effectively inoculate CAusal FEatures (CAFE) into defense networks for improving adversarial robustness.

1. Introduction

Adversarial examples, which are indistinguishable to human observers but maliciously fooling Deep Neural Networks (DNNs), have drawn great attention in research fields due to their security threats used to compromise machine learning systems. In real-world environments, such potential risks evoke weak reliability of the decision-making process for DNNs and pose a question of adopting DNNs in safety-critical areas [4, 58, 66].

To understand the origin of adversarial examples, seminal works have widely investigated the adversarial vulnerability through numerous viewpoints such as excessive linearity in a hyperplane [26], aberration of statistical fluctuations [59, 63], and phenomenon induced from frequency
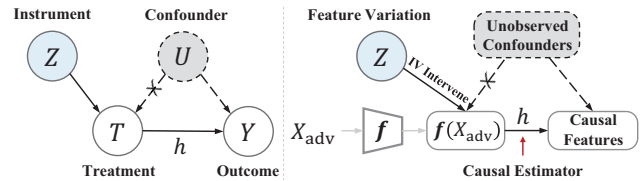

Figure 1. Data generating process (DGP) with IV. By deploying Z, it can estimate causal relation between treatment T and outcome Y under exogenous condition for unknown confounders U.

information [73]. Recently, several works [34, 35, 40] have revealed the existence and pervasiveness of robust and non-robust features in adversarially trained networks and pointed out that the non-robust features on adversarial examples can provoke unexpected misclassifications.

Nonetheless, there still exists a lack of common consensus [22] on underlying causes of adversarial examples, albeit comprehensive endeavors [32, 64]. It is because that the earlier works have focused on analyzing associations between adversarial examples and target labels in the learning scheme of adversarial training [42, 54, 67, 72, 77], which is canonical supervised learning. Such analyses easily induce spurious correlation (*i.e.,* statistical bias) in the learned associations, thereby cannot interpret the genuine origin of adversarial vulnerability under the existence of possibly biased viewpoints (*e.g.,* excessive linearity, statistical fluctuations, frequency information, and non-robust features). In order to explicate where the adversarial vulnerability comes from in a causal perspective and deduce true adversarial causality, we need to employ an intervention-oriented approach (*i.e.,* causal inference) that brings in estimating causal relations beyond analyzing merely associations for the given data population of adversarial examples.

One of the efficient tools for causal inference is instrumental variable (IV) regression when randomized controlled trials (A/B experiments) or full controls of unknown confounders are not feasible options. It is a popular approach used to identify causality in econometrics [13, 16, 47], and it provides an unbiased environment from unknown confounders that raise the endogeneity of causal inference [55]. In IV regression, the instrument is utilized

---

to eliminate a backdoor path derived from unknown confounders by separating exogenous portions of treatments. For better understanding, we can instantiate a case of finding causal relations [9] between education T and earnings Y as illustrated in Fig. 1. Solely measuring correlation between the two variables does not imply causation, since there may exist unknown confounders U (*e.g.,* individual ability, family background, etc.). Ideally, conditioning on U is the best way to identify causal relation, but it is impossible to control the unobserved variables. David Card [9] has considered IV as the college proximity Z, which is directly linked with education T but intuitively not related with earnings Y. By assigning exogenous portion to Z, it can provide an unbiased environment dissociated from U for identifying true causal relation between T and Y.

Specifically, once regarding data generating process (DGP) [53] for causal inference as in Fig. 1, the existence of unknown confounders U could create spurious correlation generating a backdoor path that hinders causal estimator h (*i.e.,* hypothesis model) from estimating causality between treatment T and outcome Y ($T \leftarrow U \rightarrow Y$). By adopting an instrument Z, we can acquire the estimand of true causality from h in an unbiased state ($Z \rightarrow T \rightarrow Y$). Bringing such DGP into adversarial settings, the aforementioned controversial perspectives (*e.g.,* excessive linearity, statistical fluctuations, frequency information, and non-robust features) can be regarded as possible candidates of unknown confounders U to reveal adversarial origins. In most observational studies, everything is endogenous in practice so that we cannot explicitly specify all confounders and conduct full controls of them in adversarial settings. Accordingly, we introduce IV regression as a powerful causal approach to uncover adversarial origins, due to its capability of causal inference although unknown confounders remain.

Here, unknown confounders U in adversarial settings easily induce ambiguous interpretation for the adversarial origin producing spurious correlation between adversarial examples and their target labels. In order to uncover the adversarial causality, we first need to intervene on the intermediate feature representation derived from a network f and focus on what truly affects adversarial robustness irrespective of unknown confounders U, instead of model prediction. To do that, we define the instrument Z as feature variation in the feature space of DNNs between adversarial examples and natural examples, where the variation Z is originated from the adversarial perturbation in the image domain such that Z derives adversarial features T for the given natural features. Note that regarding Z as instrument is reasonable choice, since the feature variation alone does not serve as relevant information for adversarial prediction without natural features. Next, once we find causality-related feature representations on adversarial examples, then we name them as *causal features* Y that can

encourage robustness of predicting target labels despite the existence of adversarial perturbation as in Fig. 1.

In this paper, we propose *adversarial instrumental variable (IV) regression* to identify causal features on adversarial examples concerning the causal relation of adversarial prediction. Our approach builds an unbiased environment for unknown confounders U in adversarial settings and estimates inherent causal features on adversarial examples by employing generalized method of moments (GMM) [28] which is a flexible estimation for non-parametric IV regression. Similar to the nature of adversarial learning [5, 25], we deploy a zero-sum optimization game [20, 41] between a hypothesis model and test function, where the former tries to unveil causal relation between treatment and outcome, while the latter disturbs the hypothesis model from estimating the relation. In adversarial settings, we regard the hypothesis model as a causal feature estimator which extracts causal features in the adversarial features to be highly related to the correct prediction for the adversarial robustness, while the test function makes worst-case counterfactuals (*i.e.,* extreme features) compelling the estimand of causal features to significantly deviate from correct prediction. Consequently, it can further strengthen the hypothesis model to demystify causal features on adversarial examples.

Through extensive analyses, we corroborate that the estimated causal features on adversarial examples are highly related to correct prediction for adversarial robustness, and the test function represents the worst-case counterfactuals on adversarial examples. By utilizing feature visualization [43, 50], we interpret the causal features on adversarial examples in a human-recognizable way. Furthermore, we introduce an inversion of the estimated causal features to handle them on the possible feature bound and present a way of efficiently injecting these *CAusal FEatures (CAFE)* into defense networks for improving adversarial robustness.

2. Related Work

In the long history of causal inference, there have been a variety of works [24, 27, 36] to discover how the causal knowledge affects decision-making process. Among various causal approaches, especially in economics, IV regression [55] provides a way of identifying the causal relation between the treatment and outcome of interests despite the existence of unknown confounders, where IV makes the exogenous condition of treatments thus provides an unbiased environment for the causal inference.

Earlier works of IV regression [2, 3] have limited the relation for causal variables by formalizing it with linear function, which is known as 2SLS estimator [71]. With progressive developments of machine learning methods, researchers and data scientists desire to deploy them for non-parametric learning [12, 13, 16, 47] and want to overcome the linear constraints in the functional relation among the vari-

ables. As extensions of 2SLS, DeepIV [29], KernelIV [61], and Dual IV [45] have combined DNNs as non-parametric estimator and proposed effective ways of exploiting them to perform IV regression. More recently, generalized method of moments (GMM) [7, 20, 41] has been cleverly proposed a solution for dealing with the non-parametric hypothesis model on the high-dimensional treatments through a zero-sum optimization, thereby successfully achieving the non-parametric IV regression.

In parallel with the various causal approaches utilizing IV, uncovering the origin of adversarial examples is one of the open research problems that arouse controversial issues. In the beginning, [26] have argued that the excessive linearity in the networks' hyperplane can induce adversarial vulnerability. Several works [59, 63] have theoretically analyzed such origin as a consequence of statistical fluctuation of data population, or the behavior of frequency information in the inputs [73]. Recently, the existence of non-robust features in DNNs [34, 35] is contemplated as a major cause of adversarial examples, but it still remains inexplicable [22].

Motivated by IV regression, we propose a way of estimating inherent causal features in adversarial features easily provoking the vulnerability of DNNs. To do that, we deploy the zero-sum optimization based on GMM between a hypothesis model and test function [7, 20, 41]. Here, we assign the role of causal feature estimator to hypothesis model and that of generating worst-case counterfactuals to test function disturbing to find causal features. This strategy results in learning causal features to overcome all trials and tribulations regarded as various types of adversarial perturbation.

3. Adversarial IV Regression

Our major goal is estimating inherent causal features on adversarial examples highly related to the correct prediction for adversarial robustness by deploying IV regression. Before identifying causal features, we first specify problem setup of IV regression and revisit non-parametric IV regression with generalized method of moments (GMM).

Problem Setup. We start from conditional moment restriction (CMR) [1, 11] bringing in an asymptotically efficient estimation with IV, which reduces spurious correlation (*i.e.,* statistical bias) between treatment T and outcome of interest Y caused by unknown confounders U [51] (see their relationships in Fig. 1). Here, the formulation of CMR can be written with a hypothesis model h, so-called a causal estimator on the hypothesis space $\mathcal{H}$ as follows:

$$\mathbb{E}_T[\psi_T(h) \mid Z] = \mathbf{0}, \tag{1}$$

where $\psi_T : \mathcal{H} \to \mathbb{R}^d$ denotes a generalized residual function [13] on treatment T, such that it represents $\psi_T(h) = Y - h(T)$ considered as an outcome error for regression task. Note that $\mathbf{0} \in \mathbb{R}^d$ describes zero vector and d indicates

the dimension for the outcome of interest Y, and it is also equal to that for the output vector of the hypothesis model h. The treatment is controlled for being exogenous [49] by the instrument. In addition, for the given instrument Z, minimizing the magnitude of the generalized residual function ψ implies asymptotically restricting the hypothesis model h not to deviate from Y, thereby eliminating the internal spurious correlation on h from the backdoor path induced by confounders U.

3.1. Revisiting Non-parametric IV regression

Once we find a hypothesis model h satisfying CMR with instrument Z, we can perform IV regression to endeavor causal inference using h under the following formulation: $\mathbb{E}_T[h(T) \mid Z] = \int_{t \in T} h(t) d\mathbb{P}(T = t \mid Z)$, where $\mathbb{P}$ indicates a conditional density measure. In fact, two-stage least squares (2SLS) [2, 3, 71] is a well-known solver to expand IV regression, but it cannot be directly applied to more complex model such as non-linear model, since 2SLS is designed to work on linear hypothesis model [52]. Later, [29] and [61] have introduced a generalized 2SLS for non-linear model by using a conditional mean embedding and a mixture of Gaussian, respectively. Nonetheless, they still raise an ill-posed problem yielding biased estimates [7,20,45,78] with the non-parametric hypothesis model h on the high dimensional treatment T, such as DNNs. It stems from the curse nature of two-stage methods, known as *forbidden regression* [3] according to Vapnik's principle [17]: *"do not solve a more general problem as an intermediate step"*.

To address it, recent studies [7, 20, 41] have employed generalized method of moments (GMM) to develop IV regression and achieved successful one-stage regression alleviating biased estimates. Once we choose a moment to represent a generic outcome error with respect to the hypothesis model and its counterfactuals, GMM uses the moment to deliver infinite moment restrictions to the hypothesis model, beyond the simple constraint of CMR. Expanding Eq. (1), the formulation of GMM can be written with a moment, denoted by $m : \mathcal{H} \times \mathcal{G} \to \mathbb{R}$ as follows (see Appendix A):

$$m(h, g) = \mathbb{E}_{Z,T}[\psi_T(h) \cdot g(Z)]$$
$$= \mathbb{E}_Z[\underbrace{\mathbb{E}_T[\psi_T(h) \mid Z]}_{\text{CMR}} \cdot g(Z)] = 0, \tag{2}$$

where the operator $\cdot$ specifies inner product, and $g \in \mathcal{G}$ denotes test function that plays a role in generating infinite moment restrictions on test function space $\mathcal{G}$, such that its output has the dimension of $\mathbb{R}^d$. The infinite number of test functions expressed by arbitrary vector-valued functions $\{g_1, g_2, \cdots\} \in \mathcal{G}$ cues potential moment restrictions (*i.e.,* empirical counterfactuals) [8] violating Eq. (2). In other words, they make it easy to capture the worst part of IV which easily stimulates the biased estimates for hypothesis model h, thereby helping to obtain more genuine causal

12304

relation from h by considering all of the possible counterfactual cases g for generalization.

However, it has an analogue limitation that we cannot deal with infinite moments because we only handle observable finite number of test functions. Hence, recent studies construct maximum moment restriction [20, 44, 78] to efficiently tackle the infinite moments by focusing only on the extreme part of IV, denoted as $\sup_{g \in \mathcal{G}} m(h, g)$ in a closed-form expression. By doing so, we can concurrently minimize the moments for the hypothesis model to fully satisfy the worst-case generalization performance over test functions. Thereby, GMM can be re-written with min-max optimization thought of as a zero-sum game between the hypothesis model h and test function g:

$$\min_{h \in \mathcal{H}} \sup_{g \in \mathcal{G}} m(h, g) \approx \min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} \mathbb{E}_{Z,T}[\psi_T(h) \cdot g(Z)], \quad (3)$$

where the infinite number of test functions can be replaced with the non-parametric test function in the form of DNNs. Next, we bridge GMM of Eq. (3) to adversarial settings and unveil the adversarial origin by establishing adversarial IV regression with maximum moment restriction.

3.2. Demystifying Adversarial Causal Features

To demystify inherent causal features on adversarial examples, we first define feature variation Z as the instrument, with adversarially trained DNNs denoted by f as follows:

$$Z = f_l(X_\epsilon) - f_l(X) = F_{\text{adv}} - F_{\text{natural}}, \quad (4)$$

where f_l outputs a feature representation in l^{th} intermediate layer, X represents natural inputs, and X_ϵ indicates adversarial examples with adversarial perturbation ϵ such that $X_\epsilon = X + \epsilon$. In the sense that we have a desire to uncover how adversarial features F_{adv} truly estimate causal features Y which are outcomes of our interests, we set the treatment to $T = F_{\text{adv}}$ and set counterfactual treatment with a test function to $T_{\text{CF}} = F_{\text{natural}} + g(Z)$.

Note that, if we naïvely apply test function g to adversarial features T to make counterfactual treatment T_{CF} such that $T_{\text{CF}} = g(T)$, then the outputs (*i.e.,* causal features) of hypothesis model $h(T_{\text{CF}})$ may not be possibly acquired features considering feature bound of DNNs f. In other words, if we do not keep natural features in estimating causal features, then the estimated causal features will be too exclusive features from natural ones. This results in non-applicable features considered as an imaginary feature we cannot handle, since the estimated causal features are significantly manipulated ones only in a specific intermediate layer of DNNs. Thus, we set counterfactual treatment to $T_{\text{CF}} = F_{\text{natural}} + g(Z)$. This is because above formation can preserve natural features, where we first subtract natural features from counterfactual treatment such that $T' = T_{\text{CF}} - F_{\text{natural}} = g(Z)$ and add the output Y' of

hypothesis model to natural features for recovering causal features such that $Y = Y' + F_{\text{natural}} = h(T') + F_{\text{natural}}$. In brief, we intentionally translate causal features and counterfactual treatment not to deviate from possible feature bound.

Now, we newly define *Adversarial Moment Restriction (AMR)* including the counterfactuals computed by the test function for adversarial examples, as follows: $\mathbb{E}_{T'}[\psi_{T'}(h) \mid Z] = \mathbf{0}$. Here, the generalized residual function $\psi_{T'|Z}(h) = Y' - h(T')$ in adversarial settings deploys the translated causal features Y'. Bring them together, we re-formulate GMM with counterfactual treatment to fit adversarial IV regression, which can be written as:

$$\min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} \mathbb{E}_Z[\underbrace{\mathbb{E}_{T'}[\psi_{T'}(h) \mid Z]}_{\text{AMR}} g(Z)] = \mathbb{E}_Z[\psi_{T'|Z}(h)g(Z)],$$
$$(5)$$

where it satisfies $\mathbb{E}_{T'}[\psi_{T'}(h) \mid Z] = \psi_{T'|Z}(h)$ because Z corresponds to only one translated counterfactual treatment $T' = g(Z)$. Here, we cannot directly compute the generalized residual function $\psi_{T'|Z}(h) = Y' - h(T')$ in AMR, since there are no observable labels for the translated causal features Y' on high-dimensional feature space. Instead, we make use of onehot vector-valued target label $G \in \mathbb{R}^K$ (K: class number) corresponding to the natural input X in classification task. To utilize it, we alter the domain of computing GMM from feature space to log-likelihood space of model prediction by using the log-likelihood function: $\Omega(\omega) = \log f_{l+}(F_{\text{natural}} + \omega)$, where f_{l+} describes the subsequent network returning classification probability after l^{th} intermediate layer. Accordingly, the meaning of our causal inference is further refined to find inherent causal features of correctly predicting target labels even under worst-case counterfactuals. To realize it, Eq. (5) is modified with moments projected to the log-likelihood space as follows:

$$\min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} \mathbb{E}_Z[\psi_{T'|Z}^{\Omega}(h) \cdot (\Omega \circ g)(Z)]$$
$$= \mathbb{E}_Z[\{G_{\log} - (\Omega \circ h)(T')\} \cdot (\Omega \circ g)(Z)], \quad (6)$$

where $\psi_{T'|Z}^{\Omega}(h)$ indicates the generalized residual function on the log-likelihood space, the operator $\circ$ symbolizes function composition, and $G_{\log}$ is log-target label such that satisfies $G_{\log} = \log G$. Each element ($k = 1, 2, \cdots, K$) of log-target label has $G_{\log}^{(k)} = 0$ when it is $G^{(k)} = 1$ and has $G_{\log}^{(k)} = -\infty$ when it is $G^{(k)} = 0$. To implement it, we just ignore the element $G_{\log}^{(k)} = -\infty$ and use another only.

So far, we construct GMM based on AMR in Eq. (6), namely *AMR-GMM*, to behave adversarial IV regression. However, there is absence of explicitly regularizing the test function, thus there happens generalization gap between ideal and empirical moments (see Appendix B). Thereby, it violates possible feature bounds of the test function and brings in imbalanced predictions on causal inference (see Fig. 4). To become a rich test function, previ-

ous works [7, 20, 41, 68] have employed *Rademacher complexity* [6, 37, 74] that provides tight generalization bounds for a family of functions. It has a strong theoretical foundation to control a generalization gap, thus it is related to various regularizers used in DNNs such as weight decay, Lasso, Dropout, and Lipschitz [21, 65, 69, 76]. In AMR-GMM, it plays a role in enabling the test functions to find out the worst-case counterfactuals within adversarial feature bound. Following Appendix B, we build a final objective of AMR-GMM with rich test function as follows:

$$\min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} \mathbb{E}_Z[\psi_{T'|Z}^{\Omega}(h) \cdot (\Omega \circ g)(Z)] - |\mathbb{E}_Z[Z - g(Z)]|^2. \quad (7)$$

Please see details of AMR-GMM algorithm in Appendix D. For the proof of its convergence, we borrow the concept of Lewis *et al.* [41] with *Set Identification* [14] and *Lipschitz* [6] to finding ϵ-equilibrium of the zero-sum game.

4. Analyzing Properties of Causal Features

In this section, we first notate several conjunctions of feature representation from the result of adversarial IV regression with AMR-GMM as: (i) *Adversarial Feature* (Adv): $F_{\text{natural}} + Z$, (ii) *CounterFactual Feature* (CF): $F_{\text{natural}} + g(Z)$, (iii) *Counterfactual Causal Feature* (CC): $F_{\text{natural}} + (h \circ g)(Z)$, and (iv) *Adversarial Causal Feature* (AC): $F_{\text{natural}} + h(Z)$. By using them, we estimate adversarial robustness computed by classification accuracy for which the above feature conjunctions are propagated through f_{l+}, where standard attacks generate feature variation Z and adversarial features T. Note that, implementation of all feature representations is treated at the last convolutional layer of DNNs f as in [35], since it mostly contains the high-level object concepts and has the unexpected vulnerability for adversarial perturbation due to high-order interactions [18]. Here, average treatment effects (ATE) [33], used for conventional validation of causal approach, is replaced with adversarial robustness of the conjunctions.

4.1. Validating Hypothesis Model and Test Function

After optimizing hypothesis model and test function using AMR-GMM for adversarial IV regression, we can then control endogenous treatment (*i.e.,* adversarial features) and separate exogenous portion from it, namely causal features, in adversarial settings. Here, the hypothesis model finds causal features on adversarial examples, highly related to correct prediction for adversarial robustness even with the adversarial perturbation. On the other hand, the test function generates worst-case counterfactuals to disturb estimating causal features, thereby degrading capability of hypothesis model. These learning strategy enables hypothesis model to estimate inherent causal features overcoming all trials and tribulations from the counterfactuals. Therefore, the findings of the causal features on adversarial examples
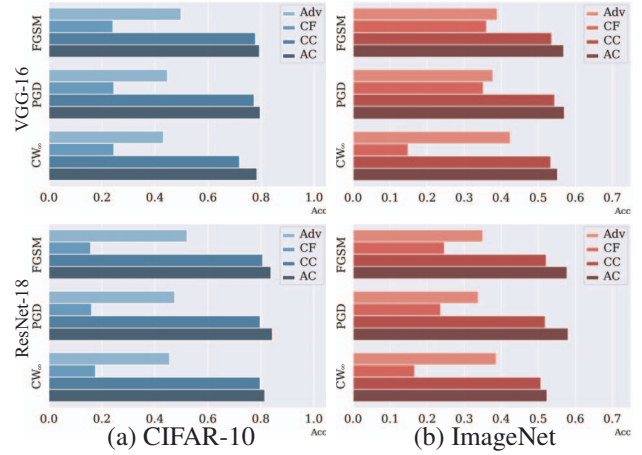

Figure 2. Adversarial robustness of Adv, CF, CC, AC on VGG-16 and ResNet-18 under three attack modes: FGSM [26], PGD [42], CW$_\infty$ [10] for CIFAR-10 [38] and ImageNet [19].

has theoretical evidence by nature of AMR-GMM to overcome various types of adversarial perturbation. Note that, our IV setup posits homogeneity assumption [31], a more general version than monotonicity assumption [2], that adversarial robustness (*i.e.,* average treatment effects) consistently retains high for all data samples despite varying natural features F_{natural} depending on data samples.

As illustrated in Fig. 2, we intensively examine the average treatment effects (*i.e.,* adversarial robustness) for the hypothesis model and test function by measuring classification accuracy of the feature conjunctions (*i.e.,* Adv, CF, CC, AC) for all dataset samples. Here, we observe that the adversarial robustness of CF is inferior to that of CC, AC, and even Adv. Intuitively, it is an obvious result since the test function violating Eq. (7) forces feature representation to be the worst possible condition of extremely deviating from correct prediction. For the prediction results for CC and AC, they show impressive robustness performance than Adv with large margins. Since AC directly leverages the feature variation acquired from adversarial perturbation, they present better adversarial robustness than CC obtained from the test function outputting the worst-case counterfactuals on the feature variation. Intriguingly, we notice that both results from the hypothesis model generally show constant robustness even in a high-confidence adversarial attack [10] fabricating unseen perturbation. Such robustness demonstrates the estimated causal features have ability to overcome various types of adversarial perturbation.

4.2. Interpreting Causal Effects and Visual Results

We have reached the causal features in adversarial examples and analyzed their robustness. After that, our next question is *"Can the causal features per se have semantic information for target objects?"*. Recent works [23, 35] have investigated to figure out the semantic meaning of feature representation in adversarial settings, we also utilize
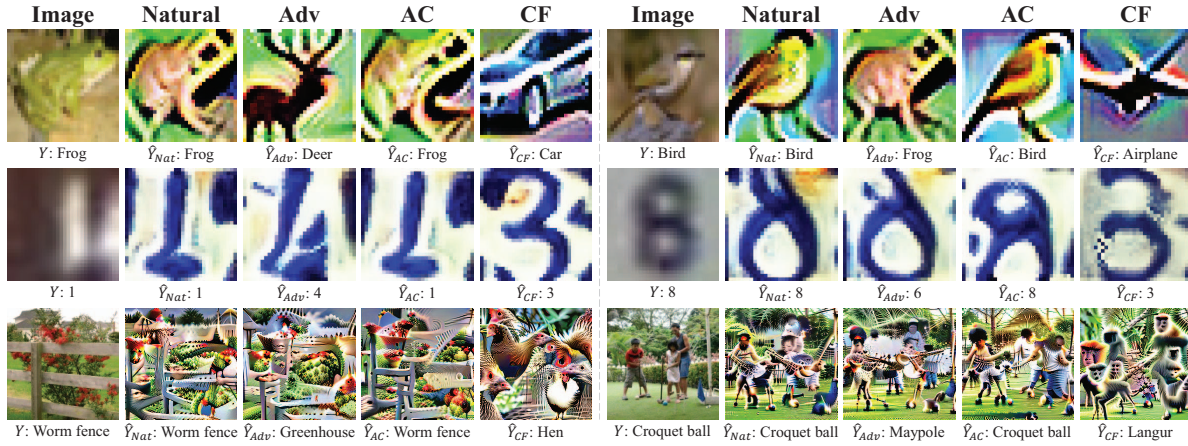
12306

Figure 3. Feature visualization results of representing natural features, Adv, AC, and CF. From the top row, CIFAR-10, SVHN, and ImageNet are sequentially used for the feature visual interpretation.

the feature visualization method [43, 48, 50] on the input domain to interpret the feature conjunctions in a human-recognizable manner. As shown in Fig. 3, we can generally observe that the results of natural features represent semantic meaning of target objects. On the other hand, adversarial features (Adv) compel its feature representation to the orient of adversarially attacked target objects.

As aforementioned, the test function distracts treatments to be worst-case counterfactuals, which exacerbates the feature variation from adversarial perturbation. Thereby, the visualization of CF is remarkably shifted to the violated feature representation for target objects. For instance, as in ImageNet [19] examples, we can see that the visualization of CF displays *Hen* and *Langur* features, manipulated from *Worm fence* and *Croquet ball*, respectively. We note that red flowers in original images have changed into red cockscomb and patterns of hen feather, in addition, people either have changed into distinct characteristics of langur, which accelerates the disorientation of feature representation to the worst counterfactuals. Contrastively, the visualization of AC displays a prominent exhibition and semantic consistency for target objects, where we can recognize their semantic information by themselves and explicable to human observers. By investigating visual interpretations, we reveal that feature representations acquired from the hypothesis model and test function both have causally semantic information, and their roles are in line with the theoretical evidence of our causal approach. In brief, we validate semantic meaning of causal features immanent in high-dimensional space despite the counterfactuals.

4.3. Validating Conditions of IV Setup

The instrumental variable needs to satisfy the following three valid conditions in order to successfully achieve non-parametric IV regression based on previous works [29, 45]: independent of the outcome error such that $\psi \perp Z$ (*Unconfoundedness*) where ψ denotes outcome error, and do not

	VGG			ResNet			WRN		
	CIFAR	SVHN	Tiny	CIFAR	SVHN	Tiny	CIFAR	SVHN	Tiny
$f_{l+}(T)$	44.8	52.1	21.5	46.5	55.4	24.2	48.7	56.7	25.5
$f_{l+}(Z)$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ρ	0.9	0.8	0.8	0.9	0.8	0.7	0.9	0.9	0.8

Table 1. Empirical validation for three conditions of our IV setup. $f_{l+}(T)$ and $f_{l+}(Z)$ indicates model performance (%) of adversarial robustness by propagating adversarial features T and feature variation Z with subsequent network, respectively. The last row represents Pearson correlation: $\rho = \mathrm{Cov}(Z, T)/\sigma_Z \sigma_T$.

directly affect outcomes such that $Z \perp Y \mid T, \psi$ (*Exclusion Restriction*) but only affect outcomes through a connection of treatments such that $\mathrm{Cov}(Z, T) \neq 0$ (*Relevance*).

For *Unconfoundedness*, various works [42, 54, 67, 72, 77] have proposed adversarial training robustifying DNNs f with adversarial examples inducing feature variation that we consider as IV to improve robustness. In other words, when we see them in a perspective of IV regression, we can regard them as the efforts satisfying CMR in DNNs f for the given feature variation Z. Aligned with our causal viewpoints, the first row in Tab. 1 shows the existence of adversarial robustness with adversarial features T. Therefore, we can say that our IV (*i.e.,* feature variation) on adversarially trained models satisfies valid condition of *Unconfoundedness*, so that IV is independent of the outcome error.

For *Exclusion Restriction*, feature variation Z itself cannot serve as enlightening information to model prediction without natural features, because only propagating the residual feature representation has no effect to model prediction by the learning nature of DNNs. Empirically, the second row in Tab. 1 demonstrates that Z cannot be helpful representation for prediction. Thereby, our IV is not encouraged to be correlated directly with the outcome, and it satisfies valid condition of *Exclusion Restriction*.

For *Relevance*, when taking a look at the estimation procedure of adversarial feature T such that $T = Z + F_{\text{natural}}$, feature variation Z explicitly has a causal influence on T.

Method	CIFAR-10							SVHN							Tiny-ImageNet						
	Natural	FGSM	PGD	CW$_\infty$	AP	DLR	AA	Natural	FGSM	PGD	CW$_\infty$	AP	DLR	AA	Natural	FGSM	PGD	CW$_\infty$	AP	DLR	AA
VGG ADV	**78.5**	49.8	44.8	42.6	43.2	42.9	40.7	**91.9**	64.8	52.1	48.9	48.0	48.5	45.2	**53.2**	25.3	21.5	21.0	20.2	20.8	19.6
ADV$_{CAFE}$	78.4	**52.2**	**47.9**	**44.1**	**46.4**	**44.5**	**42.7**	91.5	**67.0**	**55.3**	**50.0**	**51.3**	**49.6**	**46.1**	52.6	**26.0**	**22.8**	**22.1**	**21.8**	**22.0**	**21.0**
TRADES	79.5	50.4	45.7	43.2	44.4	42.9	41.8	**91.9**	66.4	53.6	49.1	49.1	47.7	45.2	**52.8**	25.9	22.5	21.9	21.5	21.8	20.7
TRADES$_{CAFE}$	77.0	**51.6**	**47.9**	**44.0**	**47.0**	**43.9**	**42.7**	90.3	**67.8**	**56.1**	**50.0**	**53.6**	**49.1**	**47.5**	52.1	**26.5**	**23.6**	**22.6**	**22.5**	**22.6**	**21.6**
MART	79.7	52.4	47.2	43.4	45.5	43.8	42.0	**92.6**	66.6	54.2	47.9	49.6	47.1	44.4	**53.1**	25.0	21.5	21.2	20.4	21.0	19.9
MART$_{CAFE}$	78.3	**54.2**	**49.7**	**43.9**	**48.1**	**44.5**	**42.7**	91.3	**67.6**	**57.3**	**49.5**	**54.2**	**48.3**	**46.4**	53.0	**25.6**	**22.3**	**21.6**	**21.3**	**21.5**	**20.5**
AWP	**78.0**	51.7	48.2	43.5	47.2	43.4	42.6	90.8	65.5	56.6	50.4	54.0	49.7	48.6	52.6	28.0	25.7	23.6	24.8	23.5	22.8
AWP$_{CAFE}$	77.4	**54.8**	**51.4**	**44.2**	**50.2**	**44.9**	**43.5**	**91.9**	**67.9**	**58.6**	**51.2**	**55.9**	**51.1**	**49.7**	**52.9**	**28.8**	**26.4**	**24.2**	**25.6**	**24.1**	**23.4**
HELP	77.4	51.8	48.3	43.9	47.3	43.9	42.9	91.2	65.8	56.6	50.9	53.9	50.2	48.8	**53.0**	28.3	25.9	23.9	25.1	23.8	23.1
HELP$_{CAFE}$	75.6	**54.4**	**51.4**	**44.6**	**50.4**	**44.8**	**43.7**	**91.5**	**67.3**	**58.5**	**51.6**	**56.2**	**51.4**	**50.0**	52.6	**29.4**	**27.1**	**24.7**	**26.4**	**24.4**	**23.9**
ResNet ADV	82.0	52.1	46.5	44.8	44.8	44.8	43.0	92.8	70.4	55.4	51.3	50.9	51.0	47.5	**57.2**	27.3	24.2	23.2	22.8	23.2	21.8
ADV$_{CAFE}$	**82.6**	**55.9**	**50.7**	**47.6**	**49.0**	**47.7**	**46.2**	92.5	**73.6**	**58.9**	**53.8**	**54.9**	**52.6**	**49.8**	56.3	**28.6**	**25.7**	**24.7**	**24.4**	**24.6**	**23.5**
TRADES	83.0	55.0	49.8	47.5	48.3	47.3	46.1	93.2	72.8	57.7	52.6	53.0	51.5	48.9	56.5	28.4	25.3	24.4	24.2	24.3	23.2
TRADES$_{CAFE}$	80.7	**56.6**	**51.4**	**48.5**	**50.4**	**48.3**	**46.7**	91.3	**73.9**	**59.6**	**54.1**	**56.7**	**53.2**	**51.3**	54.5	**29.6**	**27.4**	**26.3**	**26.5**	**26.2**	**25.4**
MART	83.5	56.1	50.1	47.1	48.3	47.0	45.5	93.7	74.2	55.3	51.7	53.2	50.8	47.8	**57.1**	27.4	24.2	23.2	22.9	23.2	22.2
MART$_{CAFE}$	82.1	**57.3**	**51.9**	**48.1**	**50.2**	**48.0**	**46.2**	92.2	**74.9**	**61.0**	**53.4**	**57.3**	**51.8**	**49.7**	55.9	**28.6**	**25.9**	**24.6**	**24.7**	**24.5**	**23.5**
AWP	81.2	55.3	51.6	48.0	50.5	47.8	46.9	92.2	71.1	59.8	54.3	56.8	53.6	52.0	56.2	30.5	28.5	26.2	27.6	26.2	25.5
AWP$_{CAFE}$	**81.5**	**57.8**	**54.2**	**49.4**	**52.9**	**49.0**	**47.8**	**93.4**	**74.0**	**60.9**	**55.0**	**57.8**	**54.8**	**52.7**	**56.6**	**31.4**	**29.2**	**27.1**	**28.4**	**27.0**	**26.5**
HELP	80.5	55.8	52.1	48.4	51.1	48.5	47.4	92.6	72.0	59.8	54.4	56.6	53.9	52.0	**56.1**	31.0	28.6	26.3	27.7	26.3	25.7
HELP$_{CAFE}$	**80.6**	**57.8**	**54.5**	**49.4**	**53.1**	**49.5**	**48.5**	**92.9**	**73.9**	**61.3**	**55.3**	**58.8**	**54.6**	**52.8**	55.4	**32.0**	**29.7**	**27.4**	**29.2**	**27.8**	**27.3**
WRN ADV	84.3	54.5	48.7	47.8	47.0	47.9	45.6	**94.0**	71.8	56.7	53.2	51.9	52.8	49.0	**60.9**	29.8	25.5	25.8	24.2	26.0	23.9
ADV$_{CAFE}$	**85.7**	**58.5**	**53.3**	**51.3**	**51.8**	**51.5**	**49.5**	93.7	**75.7**	**59.1**	**54.9**	**54.0**	**54.1**	**50.2**	60.6	**31.1**	**27.3**	**27.2**	**25.8**	**27.4**	**25.4**
TRADES	86.3	57.1	52.1	50.8	50.6	50.7	49.0	93.8	74.0	58.1	53.9	53.0	53.4	49.9	**60.8**	30.5	26.4	26.7	25.0	26.8	24.6
TRADES$_{CAFE}$	83.7	**58.6**	**54.5**	**52.0**	**53.2**	**52.0**	**50.1**	92.4	**75.6**	**61.0**	**55.7**	**58.0**	**58.0**	**53.0**	60.3	**31.7**	**28.2**	**28.3**	**27.0**	**28.5**	**26.5**
MART	86.5	58.5	52.6	50.0	50.7	49.9	48.0	**94.2**	75.0	58.0	53.1	52.8	52.8	48.9	**60.7**	29.9	25.6	25.9	24.0	25.5	23.6
MART$_{CAFE}$	85.7	**59.8**	**54.6**	**51.4**	**52.7**	**50.9**	**49.3**	93.0	**76.5**	**61.9**	**54.9**	**57.2**	**53.8**	**50.7**	60.4	**31.2**	**27.5**	**26.8**	**25.5**	**27.0**	**25.1**
AWP	83.7	58.0	54.7	51.3	53.7	51.2	50.1	93.2	73.4	60.8	55.9	57.5	55.5	53.6	**61.9**	35.5	32.8	31.0	31.6	31.1	29.6
AWP$_{CAFE}$	**84.6**	**60.6**	**56.9**	**52.4**	**55.5**	**52.3**	**51.1**	**94.2**	**76.9**	**62.7**	**57.5**	**59.2**	**57.1**	**54.6**	61.4	**36.6**	**34.2**	**32.3**	**33.2**	**32.5**	**30.8**
HELP	83.8	58.6	54.9	51.6	53.8	51.6	50.3	93.5	73.4	60.8	56.5	57.6	56.1	54.0	**61.8**	35.9	33.0	31.3	31.8	31.3	29.8
HELP$_{CAFE}$	83.1	**60.5**	**57.1**	**52.7**	**56.0**	**52.6**	**51.3**	**94.0**	**76.6**	**62.6**	**57.7**	**58.8**	**57.2**	**55.0**	61.1	**37.0**	**34.7**	**32.6**	**33.8**	**32.8**	**31.2**

Table 2. Comparison of adversarial robustness and improvement from CAFE on five defense baselines: ADV, TRADES, MART, AWP, HELP, trained with VGG-16, ResNet-18, WideResNet-34-10 for three datasets under six attacks: FGSM, PGD, CW$_\infty$, AP, DLR, AA.

This is because, in our IV setup, the treatment T is directly estimated by instrument Z given natural features F_{natural}. By using all data samples, we empirically compute Pearson correlation coefficient to prove existence of highly related connection between them as described in the last row of Tab. 1. Therefore, our IV satisfies *Relevance* condition.

5. Inoculating CAusal FEatures for Robustness

Next, we explain how to efficiently implant the causal features into various defense networks for robust networks. To eliminate spurious correlation of networks derived from the adversary, the simplest approach is utilizing the hypothesis model to enhance robustness. However, there is a realistic obstacle that it works only when we already identify what is natural inputs and their adversarial examples in inference phase. Therefore, it is not feasible approach to directly exploit the hypothesis model to improve robustness.

To address it, we introduce an inversion of causal features (*i.e.,* causal inversion) reflecting those features on input domain. It takes an advantage of well representing causal features within allowable feature bound regarding network parameters of the preceding sub-network f_l for the given adversarial examples. In fact, causal features are manipulated on an intermediate layer by the hypothesis model h, thus they are not guaranteed to be on possible feature bound. The causal inversion then serves as a key in resolving it without harming causal prediction much, and its formulation can be written with causal perturbation using

distance metric of KL divergence $\mathcal{D}_{\text{KL}}$ as:

$$\delta_{\text{causal}} = \underset{\|\delta\|_\infty \leq \gamma}{\arg\min} \, \mathcal{D}_{\text{KL}}\left(f_{l+}(F_{\text{AC}}) \,\|\, f(X_\delta)\right), \quad (8)$$

where F_{AC} indicates adversarial causal features distilled by hypothesis model h, and δ_{causal} denotes causal perturbation to represent causal inversion X_{causal} such that $X_{\text{causal}} = X + \delta_{\text{causal}}$. Note that, so as not to damage the information of natural input during generating the causal inversion X_{causal}, we constraint the perturbation δ to l_∞ within γ-ball, as known as perturbation budget, to be human-imperceptible one such that $\|\delta\|_\infty \leq \gamma$. Appendix C shows the statistical distance away from confidence score for model prediction of causal features, compared with that of causal inversion, natural input, and adversarial examples. As long as being capable of handling causal features using the causal inversion such that $\hat{F}_{\text{AC}} = f_l(X_{\text{causal}})$, we can now develop how to inoculate *CAusal FEatures (CAFE)* to defense networks as a form of empirical risk minimization (ERM) with small population of perturbation ϵ, as follows:

$$\min_{f \in \mathcal{F}} \mathbb{E}_{\mathcal{S}} \left[\max_{\|\epsilon\|_\infty \leq \gamma} \mathcal{L}_{\text{Defense}} + \mathcal{D}_{\text{KL}}(f_{l+}(\hat{F}_{\text{AC}}) \,\|\, f_{l+}(F_{\text{adv}})) \right],$$
$$(9)$$

where $\mathcal{L}_{\text{Defense}}$ specifies a pre-defined loss such as [42, 54, 67, 72, 77] for achieving a defense network f on network parameter space $\mathcal{F}$, and $\mathcal{S}$ denotes data samples such that $(X, G) \sim \mathcal{S}$. The rest term represents a causal regularizer

serving as *causal inoculation* to make adversarial features F_{adv} assimilate causal features F_{AC}. Specifically, while $\mathcal{L}_{\text{Defense}}$ robustifies network parameters against adversarial examples, the regularizer helps to hold adversarial features not to stretch out from the possible bound of causal features, thereby providing networks to backdoor path-reduced features dissociated from unknown confounders. More details for training algorithm of CAFE are attached in Appendix E.

6. Experiments

6.1. Implementation and Experimental Details

We conduct exhaustive experiments on three datasets and three networks to verify generalization in various conditions. For datasets, we take CIFAR-10 [38], SVHN [46], and Tiny-ImageNet [39]. To train the three datasets, we adopt standard networks: VGG-16 [60], ResNet-18 [30], and an advanced large network: WideResNet-34-10 [75].

For attacks, we use perturbation budget $8/255$ for CIFAR-10, SVHN and $4/255$ for Tiny-ImageNet with two standard attacks: FGSM [26], PGD [42], and four strong attacks: CW$_\infty$ [10], and AP (Auto-PGD: step size-free), DLR (Auto-DLR: shift and scaling invariant), AA (Auto-Attack: parameter-free) introduced by [15]. PGD, AP, DLR have 30 steps with random starts where PGD has step sizes 0.0023 and 0.0011 respectively, and AP, DLR have momentum coefficient $\rho = 0.75$. CW$_\infty$ uses gradient clamping for l_∞ with CW objective [10] on $\kappa = 0$ in 100 iterations. For defenses, we adopt a standard defense baseline: ADV [42] and four strong defense baselines: TRADES [77], MART [67], AWP [72], HELP [54]. We generate adversarial examples using PGD [42] on perturbation budget $8/255$ where we set 10 steps and 0.0072 step size in training. Especially, adversarially training for Tiny-ImageNet is a computational burden, so we employ fast adversarial training [70] with FGSM on the budget $4/255$ and its 1.25 times step size. For all training, we use SGD [57] with 0.9 momentum and learning rate of 0.1 scheduled by Cyclic [62] in 120 epochs [56, 70].

6.2. Comparing Adversarial Robustness

We align the above five defense baselines with our experiment setup to fairly validate adversarial robustness. From Eq. (8), we first acquire causal inversion to straightly deal with causal features. Next, we employ the causal inversion to carry out causal inoculation to all networks by adding the causal regularizer to the pre-defined loss of the defense baselines from scratch, as described in Eq. (9). Tab. 2 demonstrates CAFE boosts the five defense baselines and outperforms them even on the large network and large dataset, so that we verify injecting causal features works well in all networks. Appendix F shows ablation study for CAFE without causal inversion to identify where the effectiveness comes from, and it additionally demonstrates the power of test function in CAFE and the stability of CAFE.

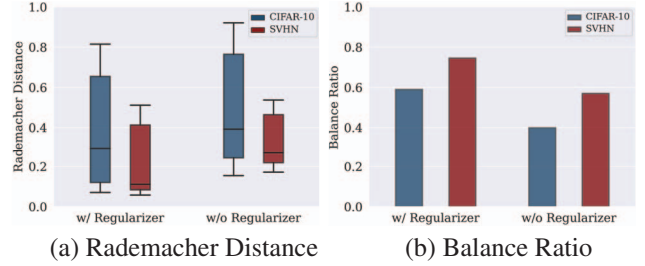(a) Rademacher Distance   (b) Balance Ratio

Figure 4. Displaying box distribution statistics of Rademacher distance and balance ratio for prediction results, compared with w/ Regularizer and w/o Regularizer on two datasets for VGG-16.

6.3. Ablation Studies on Rich Test Function

To validate that the regularizer truely works in practice, we measure Rademacher Distance and display its box distribution as illustrated in Fig. 4 (a). Here, we can apparently observe the existence of the regularization efficiency through narrowed generalization gap. Concretely, both median and average of Rademacher Distance for the regularized test function are smaller than the non-regularized one. Next, in order to investigate how rich test function helps causal inference, we examine balance ratio of prediction results for the hypothesis model, which is calculated as # of minimum predicted classes divided by # of maximum predicted classes. If the counterfactual space deviates from possible feature bound much, the attainable space that hypothesis model can reach is only restricted areas. Hence, the hypothesis model may predict biased prediction results for the target objects. As our expectation, we can observe the ratio with the regularizer is largely improved than non-regularizer for both datasets as in Fig. 4 (b). Consequently, we can summarize that rich test function acquired from the localized Rademacher regularizer serves as a key in improving the generalized capacity of causal inference.

7. Conclusion

In this paper, we build AMR-GMM to develop adversarial IV regression that effectively demystifies causal features on adversarial examples in order to uncover inexplicable adversarial origin through a causal perspective. By exhaustive analyses, we delve into causal relation of adversarial prediction using hypothesis model and test function, where we identify their semantic information in a human-recognizable way. Further, we introduce causal inversion to handle causal features on possible feature bound of network and propose causal inoculation to implant *CAusal FEatures (CAFE)* into defenses for improving adversarial robustness.

References

[1] Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003. 3

[2] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996. 2, 3, 5

[3] Joshua D Angrist and Jörn-Steffen Pischke. Mostly harmless econometrics. In *Mostly Harmless Econometrics*. Princeton university press, 2008. 2, 3

[4] Giovanni Apruzzese, Michele Colajanni, Luca Ferretti, and Mirco Marchetti. Addressing adversarial attacks against security systems based on machine learning. In *International Conference on Cyber Conflict*, volume 900, pages 1–18. IEEE, 2019. 1

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223. PMLR, 2017. 2

[6] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002. 5

[7] Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. *Advances in Neural Information Processing Systems*, 32, 2019. 3, 5

[8] Richard Blundell, Stephen Bond, and Frank Windmeijer. *Estimation in dynamic panel data models: improving on the performance of the standard GMM estimator*. Emerald Group Publishing Limited, 2001. 3

[9] David Card. Using geographic variation in college proximity to estimate the return to schooling, 1993. 2

[10] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57. IEEE Computer Society, 2017. 5, 8

[11] Gary Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987. 3

[12] Xiaohong Chen and Timothy M Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9(1):39–84, 2018. 2

[13] Xiaohong Chen and Demian Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321, 2012. 1, 2, 3

[14] Victor Chernozhukov, Han Hong, and Elie Tamer. Estimation and confidence regions for parameter sets in econometric models 1. *Econometrica*, 75(5):1243–1284, 2007. 5

[15] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pages 2206–2216. PMLR, 2020. 8

[16] Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011. 1, 2

[17] R Fernandes de Mello and M Antonelli Ponti. Statistical learning theory. *Machine Learning*, 2018. 3

[18] Huiqi Deng, Qihan Ren, Hao Zhang, and Quanshi Zhang. Discovering and explaining the representation bottleneck of dnns. In *International Conference on Learning Representations*, 2022. 5

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 5, 6

[20] Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33:12248–12262, 2020. 2, 3, 4, 5

[21] Simon Du and Jason Lee. On the power of over-parametrization in neural networks with quadratic activation. In *International Conference on Machine Learning*, volume 80, pages 1329–1338. PMLR, 10–15 Jul 2018. 5

[22] Logan Engstrom, Justin Gilmer, Gabriel Goh, Dan Hendrycks, Andrew Ilyas, Aleksander Madry, Reiichiro Nakano, Preetum Nakkiran, Shibani Santurkar, Brandon Tran, Dimitris Tsipras, and Eric Wallace. A discussion of 'adversarial examples are not bugs, they are features'. *Distill*, 2019. 1, 3

[23] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*, 2019. 5

[24] Rocio Garcia-Retamero and Ulrich Hoffrage. How causal knowledge simplifies decision-making. *Minds and Machines*, 16(3):365–380, 2006. 2

[25] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014. 2

[26] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 1, 3, 5, 8

[27] York Hagmayer and Cilia Witteman. Causal knowledge and reasoning in decision making. In *Psychology of Learning and Motivation*, volume 67, pages 95–134. Elsevier, 2017. 2

[28] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982. 2

[29] Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR, 2017. 3, 6

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 8

[31] James J Heckman, Sergio Urzua, and Edward Vytlacil. Understanding instrumental variables in models with essential heterogeneity. *The review of economics and statistics*, 88(3):389–432, 2006. 5

[32] Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 1

[33] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986. 5

[34] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 1, 3

[35] Junho Kim, Byung-Kwan Lee, and Yong Man Ro. Distilling robust and non-robust features in adversarial examples by information bottleneck. In *Advances in Neural Information Processing Systems*, 2021. 1, 3, 5

[36] Nancy S Kim and Stefanie T LoSavio. Causal explanations affect judgments of the need for psychological treatment. *Judgment and Decision Making*, 4(1):82, 2009. 2

[37] Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002. 5

[38] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5, 8

[39] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7:7, 2015. 8

[40] Byung-Kwan Lee, Junho Kim, and Yong Man Ro. Masking adversarial damage: Finding adversarial saliency for robust and sparse network. In *Conference on Computer Vision and Pattern Recognition*, pages 15126–15136, 2022. 1

[41] Greg Lewis and Vasilis Syrgkanis. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164*, 2018. 2, 3, 5

[42] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1, 5, 6, 7, 8

[43] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Conference on Computer Vision and Pattern Recognition*, pages 5188–5196, 2015. 2, 6

[44] Krikamol Muandet, Wittawat Jitkrittum, and Jonas Kübler. Kernel conditional moment test via maximum moment restriction. In *Conference on Uncertainty in Artificial Intelligence*, pages 41–50. PMLR, 2020. 4

[45] Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. Dual instrumental variable regression. *Advances in Neural Information Processing Systems*, 33:2710–2721, 2020. 3, 6

[46] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 8

[47] Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003. 1, 2

[48] Anh Nguyen, Jason Yosinski, and Jeff Clune. Understanding neural networks via feature visualization: A survey. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 55–76. Springer, 2019. 6

[49] Olena Y Nizalova and Irina Murtazashvili. Exogenous treatment and endogenous factors: Vanishing of omitted variable bias on the interaction term. *Journal of Econometric Methods*, 5(1):71–77, 2016. 3

[50] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017. 2, 6

[51] Judea Pearl. *Causality*. Cambridge university press, 2009. 3

[52] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017. 3

[53] Peter CB Phillips and Bruce E Hansen. Statistical inference in instrumental variables regression with i (1) processes. *The Review of Economic Studies*, 57(1):99–125, 1990. 2

[54] Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *International Conference on Learning Representations*, 2022. 1, 6, 7, 8

[55] Olav Reiersøl. *Confluence analysis by means of instrumental sets of variables*. PhD thesis, Almqvist & Wiksell, 1945. 1, 2

[56] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, volume 119, pages 8093–8104, 2020. 8

[57] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951. 8

[58] Y. E. Sagduyu, Y. Shi, and T. Erpek. Iot network security from the perspective of adversarial deep learning. In *International Conference on Sensing, Communication, and Networking*, pages 1–9, 2019. 1

[59] Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? In *International Conference on Learning Representations*, 2018. 1, 3

[60] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 8

[61] Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[62] Leslie N Smith. Cyclical learning rates for training neural networks. In *IEEE Winter Conference on Applications of Computer Vision*, pages 464–472. IEEE, 2017. 8

[63] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. 1, 3

[64] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. 1

[65] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *International Conference on Machine Learning*, pages 1058–1066. PMLR, 2013. 5

[66] Xianmin Wang, Jing Li, Xiaohui Kuang, Yu an Tan, and Jin Li. The security of machine learning in an adversarial setting: A survey. *Journal of Parallel and Distributed Computing*, 130:12 – 23, 2019. 1

[67] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020. 1, 6, 7, 8

[68] Ziyu Wang, Yuhao Zhou, Tongzheng Ren, and Jun Zhu. Scalable quasi-bayesian inference for instrumental variable regression. *Advances in Neural Information Processing Systems*, 34, 2021. 5

[69] Colin Wei and Tengyu Ma. Data-dependent sample complexity of deep neural networks via lipschitz augmentation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 5

[70] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020. 8

[71] Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010. 2, 3

[72] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020. 1, 6, 7, 8

[73] Dong Yin, Raphael Gontijo Lopes, Jon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 3

[74] Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, pages 7085–7094. PMLR, 2019. 5

[75] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, 2016. 8

[76] Ke Zhai and Huan Wang. Adaptive dropout with rademacher complexity regularization. In *International Conference on Learning Representations*, 2018. 5

[77] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, volume 97, pages 7472–7482, 09–15 Jun 2019. 1, 6, 7, 8

[78] Rui Zhang, Masaaki Imaizumi, Bernhard Schölkopf, and Krikamol Muandet. Maximum moment restriction for instrumental variable regression. *arXiv preprint arXiv:2010.07684*, 2020. 3, 4