ADVERSARIAL IV REGRESSION FOR DEMYSTIFYING CAUSAL FEATURES ON ADVERSARIAL EXAMPLES

Anonymous authors
Paper under double-blind review

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 (Apruzzese et al., 2019; Wang et al., 2019; Sagduyu et al., 2019).

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 (Goodfellow et al., 2015), aberration of statistical fluctuations (Szegedy et al., 2014; Shafahi et al., 2018), and phenomenon induced from frequency information (Yin et al., 2019a). Recently, several works (Ilyas et al., 2019; Kim et al., 2021) 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 (Engstrom et al., 2019a) on underlying causes of adversarial examples, albeit comprehensive endeavors (Tsipras et al., 2019; Hendrycks & Dietterich, 2019). Moreover, the earlier analyses have focused on learning associations between adversarial examples and target labels, namely adversarial training (Madry et al., 2018; Zhang et al., 2019; Wang et al., 2020; Wu et al., 2020; Rade & Moosavi-Dezfooli, 2022) in canonical supervised learning. Such approaches easily induce spurious correlation (*i.e.,* statistical bias) in the learned associations, thereby leading to robustness degradation. This is because they do not learn inherent causal relation between adversarial examples and their target labels, but learn naïve associations under the existence of unknown confounders (*e.g.,* excessive linearity, statistical fluctuations, frequency information, and non-robust features). In order to truly understand where the adversarial vulnerability comes from and deduce true adversarial causality, we need to employ an intervention-oriented approach (*i.e.,* causal inference) that brings in possibly estimating causal relations for the given data population, thereby providing an unbiased environment dissociated from the unknown confounder.

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 (Newey & Powell, 2003; Darolles et al., 2011; Chen & Pouzo, 2012) and provides an unbiased environment for unknown confounder that raises the endogeneity of causal inference (Reiersøl, 1945). In IV regression, the instrument is utilized to eliminate a backdoor path derived from unknown confounders by separating exogenous portions of treatments, for which IV needs to satisfy three valid conditions: independent of the outcome error (*Unconfoundedness*), and not directly affect outcomes (*Exclusion Restriction*) but only affect outcomes through a connection of treatments (*Relevance*).

Once regarding data generating process (DGP) (Phillips & Hansen, 1990) for causal inference as illustrated 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 environment ($Z \rightarrow T \rightarrow Y$). Because IV regression can perform causal inference although unknown confounders remain, it is a suitable causal approach to uncover adversarial origins. Bringing such DGP into adversarial settings, the aforementioned controversial perspectives (*e.g.,* excessive linearity, statistical fluctuations, frequency information, and non-robust features) of adversarial origins can be regarded as possible candidates of unknown confounders U. In most observational studies, everything is endogenous in practice so that we cannot explicitly specify such confounders and conduct full control of them.

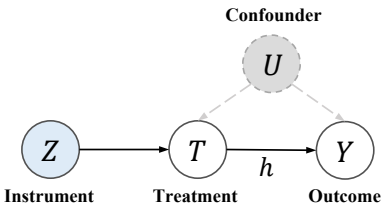


Figure 1: Data generating process (DGP) with IV. By deploying Z, it can estimate causal relation between treatment T and outcome Y under exogenous condition for unknown confounder U.

Accordingly, unknown confounders U in adversarial settings easily induce ambiguous interpretation for the adversarial origin producing spurious correlation between adversarial examples and their target labels, and consequently lead to degradation of adversarial robustness. In order to uncover the adversarial causality, we first need to intervene on the intermediate feature representation 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. Here, once we find causality-related feature representations on adversarial examples, then we name them as *causal features* Y that can encourage robustness of predicting the target labels despite the existence of adversarial perturbation harming model prediction.

In this paper, we propose *adversarial instrumental variable (IV) regression* to identify causal features on adversarial examples with respect to the causal relation of adversarial prediction. Our approach builds an unbiased environment for the unknown confounder U in adversarial settings and estimates inherent causal features on adversarial examples by employing generalized method of moments (GMM) (Hansen, 1982) which is the most flexible estimation for non-parametric IV regression. Similar to the nature of adversarial learning (Goodfellow et al., 2014; Arjovsky et al., 2017), we deploy a zero-sum optimization game (Lewis & Syrgkanis, 2018; Dikkala et al., 2020) 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 (Mahendran & Vedaldi, 2015; Olah et al., 2017), 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 into defense networks for improving adversarial robustness. We demonstrate the effectiveness of inoculating *CAusal FEatures (CAFE)* into the networks with comprehensive experimental results.

2 RELATED WORKS

In the long history of causal inference, there have been a variety of works (Garcia-Retamero & Hoffrage, 2006; Kim & LoSavio, 2009; Hagmayer & Witteman, 2017) to discover how the causal knowledge affects decision-making process. Among various causal approaches, especially in economics, IV regression (Reiersøl, 1945) 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 for an unbiased environment with the confounders.

Earlier works of IV regression (Angrist et al., 1996; Angrist & Pischke, 2008) have limited the relation for causal variables by formalizing it with linear function, which is known as 2SLS estimator (Wooldridge, 2010). With progressive developments of machine learning methods, researchers and data scientists desire to deploy them for non-parametric learning (Newey & Powell, 2003; Darolles et al., 2011; Chen & Pouzo, 2012; Chen & Christensen, 2018) and want to overcome the linear constraints in the functional relation between the variables. As extensions of 2SLS, DeepIV (Hartford et al., 2017), KernelIV (Singh et al., 2019), and Dual IV (Muandet et al., 2020b) 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) (Lewis & Syrgkanis, 2018; Bennett et al., 2019; Dikkala et al., 2020) has been cleverly proposed as 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, Goodfellow et al. (2015) have argued that the excessive linearity in the networks' hyperplane can induce adversarial vulnerability. Several works (Szegedy et al., 2014; Shafahi et al., 2018) have theoretically analyzed such origin as a consequence of statistical fluctuation of data population, or the behavior of frequency information in the inputs Yin et al. (2019a). Recently, the existence of non-robust features in DNNs (Ilyas et al., 2019; Kim et al., 2021) is contemplated as a major cause of adversarial examples, but it still remains inexplicable (Engstrom et al., 2019a).

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 (Lewis & Syrgkanis, 2018; Bennett et al., 2019; Dikkala et al., 2020). 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 getting the causal features to have the ability to overcome all trials and tribulations considered as various types of adversarial perturbation. In the end, we present to inoculate *CAusal FEatures (CAFE)* into defense networks and verify they improve adversarial robustness through empirical evidence in the sense that causal features withstand the unseen perturbations.

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 introducing the way of identifying the 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) (Chamberlain, 1987; Ai & Chen, 2003) 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 (Pearl, 2009). 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 (Chen & Pouzo, 2012) 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 (Nizalova & Murtazashvili, 2016) 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 with GMM

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: $\mathbb{E}_T[h(T) \mid Z] = \int_{t \in T} h(t) \mathrm{d}\mathbb{P}(T = t \mid Z)$, where $\mathbb{P}$ indicates a conditional density measure. In fact, two-stage least squares (2SLS) (Angrist et al., 1996; Angrist & Pischke, 2008; Wooldridge, 2010) 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 (Peters et al., 2017). Later, Hartford et al. (2017) and Singh et al. (2019) 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 (Bennett et al., 2019; Muandet et al., 2020b; Dikkala et al., 2020; Zhang et al., 2020) 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* (Angrist & Pischke, 2008) according to Vapnik's principle (de Mello & Ponti, 2018): *"do not solve a more general problem as an intermediate step"*.

To address it, recent studies (Lewis & Syrgkanis, 2018; Bennett et al., 2019; Dikkala et al., 2020) 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) (Blundell et al., 2001) 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 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 (Dikkala et al., 2020; Zhang et al., 2020; Muandet et al., 2020a) 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:

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

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 Causal Features on Adversarial Examples

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

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

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 $X_\epsilon = X + \epsilon$. To validate our

IV setup, Appendix B describes its justification on three conditions: *Unconfoundedness*, *Exclusion Restriction*, and *Relevance* in details. 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 this 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'. With them, we re-formulate GMM with counterfactual treatment to fit adversarial IV regression, which can be written as (Note that h and g consist of a simple CNN structure):

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

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, which can be written as follows:

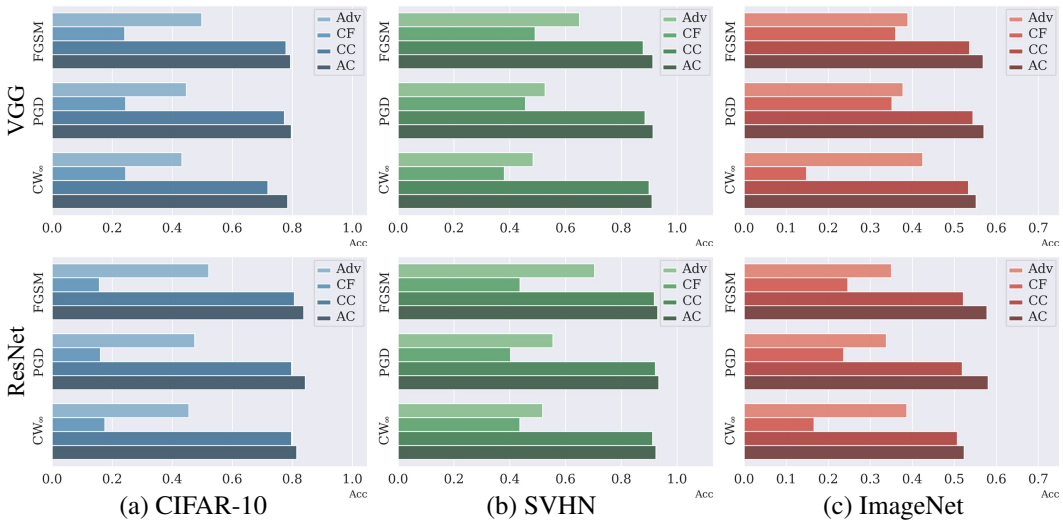
$$\min_{h\in\mathcal{H}} \max_{g\in\mathcal{G}} \mathbb{E}_Z[\psi^{\Omega}_{T'|Z}(h) \cdot (\Omega \circ g)(Z)] = \mathbb{E}_Z[\{G_{\log} - (\Omega \circ h)(T')\} \cdot (\Omega \circ g)(Z)], \tag{6}$$

where $\psi^{\Omega}_{T'|Z}(h)$ 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 $G^{(k)}_{\log} = 0$ when it is $G^{(k)} = 1$ and has $G^{(k)}_{\log} = -\infty$ when it is $G^{(k)} = 0$. To implement it, we just ignore the element $G^{(k)}_{\log} = -\infty$ and use another only.

So far, we construct GMM based on AMR in Eq. (6), namely *AMR-GMM*, to behave adversarial IV regression, but there is absence of regularizing the test function. This results in the existence of a generalization gap bringing in bad effect to causal inference (see Appendix D). To become a rich test function, previous works (Lewis & Syrgkanis, 2018; Bennett et al., 2019; Dikkala et al., 2020; Wang et al., 2021) have employed *Rademacher complexity* (Bartlett & Mendelson, 2002; Koltchinskii & Panchenko, 2002; Yin et al., 2019b) 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 (Wan et al., 2013; Zhai & Wang, 2018; Du & Lee, 2018; Wei & Ma, 2019). Following Appendix C, we build a final objective of AMR-GMM for adversarial IV regression with rich test function as follows:

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

Appendix D delineates how rich test function affects causal inference by conducting ablation studies. In addition, More details for AMR-GMM algorithm are attached in Appendix E.



(a) CIFAR-10 (b) SVHN (c) ImageNet

Figure 2: Adversarial robustness of Adv, CF, CC, AC on VGG-16 and ResNet-18 under three attack modes: FGSM(Goodfellow et al., 2015), PGD (Madry et al., 2018), CW_∞ (Carlini & Wagner, 2017) for CIFAR-10 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2011), ImageNet (Deng et al., 2009).

4 Analyzing Properties of Causal Features in AMR-GMM

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. Here, average treatment effects (ATE) (Holland, 1986), used for conventional validation of causal approach, is replaced with adversarial robustness of the conjunctions. Beyond it, we delve into semantic information of the above feature conjunctions using feature visualization (Olah et al., 2017). Note that, all feature representations are treated at the last convolutional layer of DNNs f as in (Kim et al., 2021), since it mostly contains the high-level object concepts and has the unexpected vulnerability for adversarial perturbation due to high-order interactions (Deng et al., 2022).

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 of 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 has theoretical evidence by nature of AMR-GMM to overcome various types of adversarial perturbation. Note that, our IV setup posits homogeneity assumption (Heckman et al., 2006), a more general version than monotonicity assumption (Angrist et al., 1996), 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

Table 1: Measuring distance metric (unit: m) of KL divergence $\mathcal{D}_{\mathrm{KL}}$ based on Eq. (8), between the prediction of causal features and *causal inversion*, *natural input*, and *adversarial example* of which perturbation budget varies on small and large dataset. The elements below δ_{causal} denote maximum magnitude of causal perturbation budget chosen by a heuristic search to estimate causal features.

Network	CIFAR-10				SVHN				Tiny-ImageNet			
	δ_{causal}	Inversion	Natural	Adversary	δ_{causal}	Inversion	Natural	Adversary	δ_{causal}	Inversion	Natural	Adversary
VGG	8/255	**6.3**	55.5	586.0	4/255	**4.7**	25.6	1011.3	1/255	**35.8**	83.4	800.4
ResNet	4/255	**2.2**	16.5	549.5	1/255	**2.1**	11.6	768.4	.5/255	**35.1**	80.8	762.5
WRN	2/255	**1.2**	7.2	671.8	1/255	**1.6**	6.0	937.9	.5/255	**33.4**	57.5	1062.1

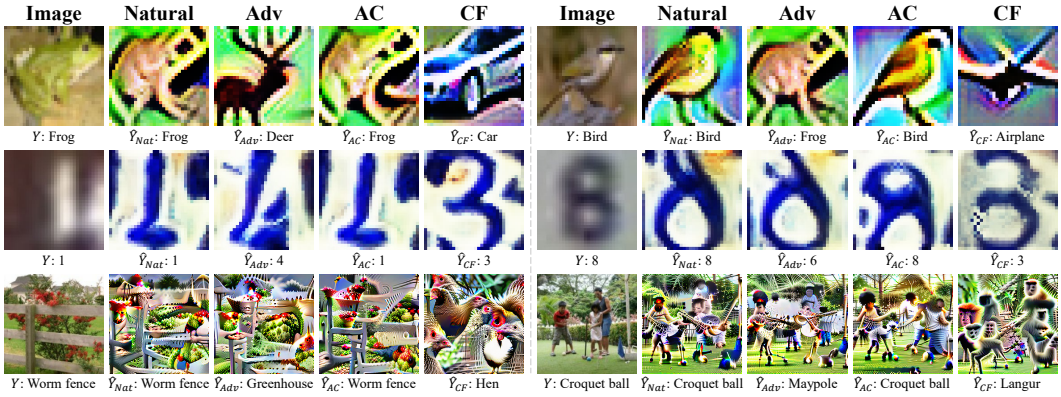


Figure 3: Feature visualization results of representing natural features, Adv, AC, and CF. From the top row, CIFAR-10, SVHN, and ImageNet are sequentially used for the feature visual interpretation.

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 (Carlini & Wagner, 2017) 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 Visualization in Feature Space

We have reached the causal features in adversarial examples and analyzed their robustness. After that, our next question is *"Can the causal features have semantic information for target objects?"*. Recent works (Engstrom et al., 2019b; Kim et al., 2021) have investigated to figure out the semantic meaning of feature representation in adversarial settings. Following the recent works, we also utilize the feature visualization method (Mahendran & Vedaldi, 2015; Olah et al., 2017; Nguyen et al., 2019) and visualize the feature conjunctions on the input domain to interpret them 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 the 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 the target objects. For instance, as in ImageNet (Deng et al., 2009) 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 the red flowers in the original images have changed into the red cockscomb and patterns of hen feather, in addition, people either have changed into the distinct characteristics of langur, which accelerates the disorientation of feature representation to the counterfactuals. Contrastively, the visualization of AC displays a prominent exhibition and semantic consistency for the target objects, where we can recognize their semantic information by themselves and explicable to human observers. By investigating visual interpretations, we reveal that the 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 have verified semantic meaning of causal features immanent in high-dimensional space despite the existence of the counterfactuals. Next, we explain how to efficiently implant the causal features into various defense networks for improving adversarial robustness.

Table 2: Measuring 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 CIFAR-10, SVHN, Tiny-ImageNet under six attack modes: FGSM, PGD, CW_∞, AP, DLR, AA.

	Method	CIFAR-10							SVHN							Tiny-ImageNet						
		Natural	FGSM	PGD	CW_∞	AP	DLR	AA	Natural	FGSM	PGD	CW_∞	AP	DLR	AA	Natural	FGSM	PGD	CW_∞	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**
	$\Delta(\%)$	*-0.1*	*4.8*	*7.1*	*3.7*	*7.5*	*3.8*	*4.9*	*-0.4*	*3.4*	*6.1*	*2.2*	*6.8*	*2.3*	*1.9*	*-1.2*	*3.0*	*6.4*	*5.2*	*7.8*	*5.6*	*6.9*
	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**
	$\Delta(\%)$	*-3.1*	*2.2*	*4.8*	*1.8*	*5.8*	*2.3*	*2.3*	*-1.8*	*2.1*	*4.6*	*1.9*	*9.3*	*2.9*	*5.0*	*-1.3*	*2.2*	*5.2*	*3.6*	*4.6*	*3.7*	*4.2*
	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**
	$\Delta(\%)$	*-1.8*	*3.4*	*5.1*	*1.2*	*5.6*	*1.6*	*1.9*	*-1.4*	*1.4*	*5.9*	*3.3*	*9.2*	*2.7*	*4.6*	*-0.2*	*2.4*	*4.0*	*1.8*	*4.3*	*2.5*	*3.1*
	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**
	$\Delta(\%)$	*-0.8*	*5.8*	*6.8*	*1.7*	*6.4*	*3.6*	*2.2*	*1.2*	*3.8*	*3.4*	*1.6*	*3.6*	*2.7*	*2.3*	*0.6*	*3.0*	*2.7*	*2.7*	*3.3*	*2.5*	*2.9*
	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**
	$\Delta(\%)$	*-2.3*	*5.0*	*6.4*	*1.5*	*6.6*	*2.2*	*1.8*	*0.3*	*2.3*	*3.3*	*1.4*	*4.2*	*2.4*	*2.5*	*-0.8*	*3.9*	*4.7*	*3.1*	*5.0*	*2.4*	*3.5*
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**
	$\Delta(\%)$	*0.7*	*7.1*	*9.1*	*6.3*	*9.4*	*6.5*	*7.4*	*-0.3*	*4.5*	*6.4*	*4.8*	*7.8*	*3.2*	*5.0*	*-1.5*	*4.6*	*6.2*	*6.2*	*7.2*	*6.1*	*7.6*
	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**
	$\Delta(\%)$	*-2.8*	*2.8*	*3.4*	*2.2*	*4.2*	*2.1*	*1.4*	*-2.0*	*1.4*	*3.4*	*2.9*	*6.9*	*3.2*	*5.0*	*-3.7*	*4.0*	*8.3*	*8.0*	*9.3*	*7.8*	*9.3*
	MART	**83.5**	56.1	50.1	47.1	48.3	47.0	45.5	**93.7**	74.2	58.3	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**
	$\Delta(\%)$	*-1.7*	*2.1*	*3.6*	*2.1*	*3.9*	*2.1*	*1.6*	*-1.6*	*1.0*	*4.7*	*3.3*	*7.7*	*2.1*	*3.8*	*-2.1*	*4.6*	*7.3*	*5.7*	*7.7*	*5.3*	*5.8*
	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**
	$\Delta(\%)$	*0.3*	*4.5*	*5.0*	*2.9*	*4.8*	*2.4*	*1.9*	*1.3*	*4.2*	*1.9*	*1.4*	*1.8*	*2.1*	*1.4*	*0.8*	*2.7*	*2.3*	*3.3*	*3.1*	*3.4*	*4.0*
	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**
	$\Delta(\%)$	*0.1*	*3.5*	*4.6*	*2.1*	*3.9*	*2.2*	*2.4*	*0.3*	*2.6*	*2.5*	*1.7*	*3.9*	*1.3*	*1.7*	*-1.2*	*3.2*	*3.8*	*4.0*	*5.4*	*5.7*	*6.0*
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**
	$\Delta(\%)$	*1.7*	*7.4*	*9.3*	*7.4*	*10.2*	*7.7*	*8.6*	*-0.3*	*5.4*	*4.3*	*3.2*	*4.1*	*2.3*	*2.5*	*-0.5*	*4.4*	*7.1*	*5.4*	*6.9*	*5.4*	*6.4*
	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**
	$\Delta(\%)$	*-3.0*	*2.6*	*4.5*	*2.4*	*5.3*	*2.6*	*2.3*	*-1.6*	*2.2*	*5.0*	*3.4*	*9.5*	*8.6*	*6.1*	*-0.8*	*3.7*	*6.7*	*6.2*	*8.1*	*6.4*	*7.8*
	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**
	$\Delta(\%)$	*-1.0*	*2.2*	*3.8*	*2.8*	*4.0*	*2.0*	*2.7*	*-1.3*	*2.0*	*6.7*	*3.3*	*8.4*	*1.8*	*3.7*	*-0.5*	*4.5*	*7.3*	*6.0*	*6.3*	*5.9*	*6.3*
	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**
	$\Delta(\%)$	*1.1*	*4.5*	*4.1*	*2.1*	*3.4*	*2.2*	*2.1*	*1.1*	*4.9*	*3.1*	*2.9*	*2.8*	*2.9*	*2.0*	*-0.9*	*3.1*	*4.5*	*4.1*	*5.2*	*4.6*	*4.2*
	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**
	$\Delta(\%)$	*-0.9*	*3.3*	*4.0*	*2.1*	*4.0*	*2.1*	*1.9*	*0.5*	*4.3*	*3.0*	*2.2*	*2.1*	*2.0*	*1.9*	*-1.1*	*3.0*	*5.1*	*4.3*	*6.4*	*4.7*	*4.8*

5 INOCULATING CAUSAL FEATURES INTO DEFENSE NETWORKS

Generating Inversion of Causal Features. To eliminate spurious correlation of networks derived from the adversary, the simplest approach that we can come up with is utilizing the hypothesis model to enhance the 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 the 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 by distance metric of KL divergence $\mathcal{D}_{\text{KL}}$ as:

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

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 $\|\delta\|_\infty \leq \gamma$. Table 1 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. It implies how well the generated causal inversion represents causal features on feasible bound so that networks themselves enable to exhibit the causal features. In

other words, it does not harm causal prediction much according to the smaller statistical distance in Table 1, thus we employ it to effectively inject causal features into the defense networks without direct aid of hypothesis model. 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:

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

where $\mathcal{L}_{\text{Defense}}$ specifies a pre-defined loss 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 the 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 the backdoor path-reduced features dissociated from unknown confounders. More details for training algorithm of CAFE are attached in Appendix F. Next, we validate the effectiveness of CAFE for the robustness by comparing it with the defense baselines.

Validating CAFE for Causal Inoculation. We conduct exhaustive experiments on three datasets and three networks to verify generalization in various conditions. For datasets, we take low-dimensional datasets: CIFAR-10 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2011), and a high-dimensional dataset: Tiny-ImageNet (Le & Yang, 2015). To train the three datasets, we adopt standard networks: VGG-16 (Simonyan & Zisserman, 2015), ResNet-18 (He et al., 2016), and an advanced large network: WideResNet-34-10 (Zagoruyko & Komodakis, 2016).

Adversarial Attacks. We use perturbation budget $8/255$ for CIFAR-10, SVHN and $4/255$ for Tiny-ImageNet with two standard attacks: FGSM (Goodfellow et al., 2015), PGD (Madry et al., 2018), and four strong attacks: CW_∞ (Carlini & Wagner, 2017), and AP (Auto-PGD: step size-free), DLR (Auto-DLR: shift and scaling invariant), AA (Auto-Attack: parameter-free) introduced by Croce & Hein (2020). 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_∞ uses gradient clamping for l_∞ with CW objective (Carlini & Wagner, 2017) on $\kappa = 0$ in 100 iterations.

Adversarial Defenses. We adopt a standard defense baseline: ADV (Madry et al., 2018) and four strong defense baselines: TRADES (Zhang et al., 2019), MART (Wang et al., 2020), AWP (Wu et al., 2020), HELP (Rade & Moosavi-Dezfooli, 2022). We generate adversarial examples using PGD (Madry et al., 2018) on perturbation budget $8/255$ where we set 10 steps and 0.0072 step size in training. Especially, adversarially training Tiny-ImageNet is a computational burden, so we employ fast adversarial training (Wong et al., 2020) with FGSM on the budget $4/255$ and its 1.25 times step size. For all training, we use SGD (Robbins & Monro, 1951) with a learning rate of 0.1 scheduled by Cyclic (Smith, 2017) in 120 epochs (Rice et al., 2020; Wong et al., 2020).

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. Subsequently, 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). Table 2 demonstrates CAFE boosts the five defense baselines and outperforms them even on the large network: WideResNet-34-10 and the large dataset: Tiny-ImageNet, so that we verify injecting causal features works well in all networks. Appendix G shows ablation studies for CAFE without causal inversion to identify where the effectiveness comes from.

6 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 through feature visualization. Furthermore, we introduce causal inversion to handle causal features on possible feature bound of network and propose causal inoculation to implant *CAusal FEatures (CAFE)* into defense networks for improving adversarial robustness.

REFERENCES

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

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

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.

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, pp. 1–18. IEEE, 2019.

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

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.

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

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.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pp. 39–57. IEEE Computer Society, 2017.

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

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.

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

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

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

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

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

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*, pp. 248–255. Ieee, 2009.

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.

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, pp. 1329–1338. PMLR, 10–15 Jul 2018.

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*, 2019a.

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*, 2019b.

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

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.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

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

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

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

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

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.

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

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

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.

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.

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.

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.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

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

Greg Lewis and Vasilis Syrgkanis. Adversarial generalized method of moments. *arXiv preprint arXiv:1803.07164*, 2018.

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.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Conference on Computer Vision and Pattern Recognition*, pp. 5188–5196, 2015.

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

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

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

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

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

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.

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

Judea Pearl. *Causality*. Cambridge university press, 2009.

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

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.

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.

Olav Reiersøl. *Confluence analysis by means of instrumental sets of variables*. PhD thesis, Almqvist & Wiksell, 1945.

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

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 400–407, 1951.

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*, pp. 1–9, 2019.

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

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

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

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

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.

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.

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

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. ISSN 0743-7315.

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.

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.

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.

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

Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

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

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, 2019a.

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, pp. 7085–7094. PMLR, 2019b.

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

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

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, pp. 7472–7482, 09–15 Jun 2019.

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

A EXPANDING GMM FROM CONDITIONAL MOMENT RESTRICTION

In order to show what connectivity exists between GMM and moment restriction, we start from Equation (1) of conditional moment restriction (CMR) in our manuscript as:

$$\mathbb{E}_T[\psi_T(h) \mid Z] = \int_{t \in T} \psi_T(h) \mathrm{d}\mathbb{P}(T = t \mid Z) = \mathbf{0}, \tag{10}$$

To generate infinite moments by the test function g regarding numerous case of instrumental variables, GMM selects a moment with the hypothesis model and test function, which can be written as:

$$m(h, g) = \mathbb{E}_{Z,T}[\psi_T(h) \cdot g(Z)]. \tag{11}$$

Here, this moment can be expressed with CMR, then it satisfies zero as follows:

$$\begin{aligned}
\mathbb{E}_{Z,T}[\psi_T(h) \cdot g(Z)] &= \int_{z \in Z, t \in T} \psi_t(h) \cdot g(z) \mathrm{d}\mathbb{P}(Z = z, T = t), \\
&= \int_{z \in Z, t \in T} \psi_t(h) \mathrm{d}\mathbb{P}(T = t \mid Z = z) \cdot g(z) \mathrm{d}\mathbb{P}(Z = z) \\
&= \int_{z \in Z} \mathbb{E}_T[\psi_T(h) \mid Z = z] \cdot g(z) \mathrm{d}\mathbb{P}(Z = z) \\
&= \mathbb{E}_Z[\underbrace{\mathbb{E}_T[\psi_T(h) \mid Z]}_{\text{conditional moment}} \cdot g(Z)] \\
&= \mathbb{E}_Z[\mathbf{0} \cdot g(Z)] \qquad (\because \text{Eq. (10)}) \\
&= 0.
\end{aligned} \tag{12}$$

From this proof, we can infer that once GMM achieves a reduction of moment magnitude, then it successfully expands CMR to perform infinite moment restriction, where the intractable infinite number of moments generated by g is replaced with (infinite-dimensional) non-parametric test function g such as DNNs.

B VALIDITY OF OUR 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 (Hartford et al., 2017; Muandet et al., 2020b): independent of the outcome error such that $\psi \perp Z$ (*Unconfoundedness*) where ψ denotes outcome error, and do not 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 (Madry et al., 2018; Zhang et al., 2019; Wang et al., 2020; Wu et al., 2020; Rade & Moosavi-Dezfooli, 2022) have proposed adversarial training that learns DNNs f with adversarial examples inducing feature variation we consider as IV to earn and improve adversarial robustness. In other words, when we see them in perspective of IV regression, we can regard them as the efforts satisfying conditional moment restriction (CMR) of DNNs f given feature variation Z. In causal perspective, the formulation of adversarial training can be written with CMR as follows:

$$\min_{f}\{\mathbb{E}_Z[\underbrace{\mathbb{E}_T[G - f_{l+}(T) \mid Z]}_{\text{CMR}}]\}^2, \tag{13}$$

where adversarial features T, their onehot vector-valued target labels G, and f_{l+} describes subsequent network from l^{th} intermediate layer to model prediction. Aligned with our causal viewpoints, the first row in Table 3 below shows the existence of adversarial robustness with adversarial features T. Therefore, we can say that our IV (i.e., feature variation) on adversarially trained model satisfies valid condition of *Unconfoundedness* so that IV is independent of the outcome error.

For *Exclusion Restriction*, feature variation Z itself cannot solely serve as enlightening information to model prediction in the absence of 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 Table 3 demonstrates that feature variation Z cannot be helpful representation to DNNs. Thereby, our IV is not encouraged to be correlated directly with the outcome, so our IV setup 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 adversarial features T. 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 there is a highly related connection between them as described in the last row of Table 3. Therefore, our IV satisfies valid condition of *Relevance*.

Table 3: Empirical validation for three conditions of our IV with VGG-16, ResNet-18, WRN-34-10 for CIFAR-10, SVHN, Tiny-ImageNet. The first row indicates model performance (%) of adversarial robustness by propagating adversarial features T with subsequent network f_{l+}. The second row shows model performance (%) of residual feature representation itself by propagating feature variation Z with subsequent network f_{l+}. The last row represents Pearson correlation coefficient: $\rho = \text{Cov}(Z,T)/\sigma_Z\sigma_T$ in the range of $-1 \leq \rho \leq 1$.

Network	VGG			ResNet			WRN		
Dataset	CIFAR-10	SVHN	Tiny-ImageNet	CIFAR-10	SVHN	Tiny-ImageNet	CIFAR-10	SVHN	Tiny-ImageNet
$f_{l+}(T)$	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

C Realizing Generalization Gap

We employ *Rademacher Complexity*, *Taylor Expansion*, and *Identity mapping* to realize the generalization gap. By using them, we elaborate the equation to feasibly calculate the generalization gap.

C.1 Rademacher Complexity

Although we construct GMM based on AMR in Eq. (6), namely *AMR-GMM*, to behave adversarial instrumental variable regression, there happens generalization gap between ideal $m(h^*, g^*)$ and empirical moments $m(h^*, g)$ for test functions due to the absence of regularizing the direction for learning a test function. To become a rich test function (Lewis & Syrgkanis, 2018; Bennett et al., 2019; Dikkala et al., 2020), therefore, we introduce *Rademacher complexity* (Bartlett & Mendelson, 2002; Koltchinskii & Panchenko, 2002; Yin et al., 2019b) that provides tight generalization bounds as a strong theoretical foundation for a family of functions, which can be written on AMR-GMM as follows:

$$\phi(g^*, g) = \underbrace{m(h^*, g^*)}_{\text{ideal}} - \underbrace{m(h^*, g)}_{\text{empirical}} \leq 2b\mathcal{R}(\mathcal{G}) + \mathcal{O}(\sqrt{n^{-1}\log\delta^{-1}}), \tag{14}$$

where b describes the upper bound of empirical moments $m(h^*, g)$, and $\mathcal{R}(\mathcal{G})$ denotes Radamecher complexity for all test functions $\forall g \in \mathcal{G}$ with probability greater than $1 - \delta$ for any $\delta \in (0, 1)$. In addition, h^* and g^* represent the most accurate estimator on each $\mathcal{H}$, $\mathcal{G}$. Besides, n indicates total number of training data samples, so that big-$\mathcal{O}$ term converges to 0 as n is large enough. Here, so as to close the gap in Eq. 14, we should tractably calculate and minimize Rademacher complexity measuring tight upper bound of generalization gap $\phi(g^*, g)$, such that $\sup_{g \in \mathcal{G}} |\phi(g^*, g)| = 2b\mathcal{R}(\mathcal{G})$.

C.2 Taylor Expansion

When we make use of AMR-GMM for adversarial instrumental variable regression, there happens generalization gap between ideal $m(h^*, g^*)$ and empirical moments $m(h^*, g)$ for test functions due to the absence of regularizing the direction for learning a test function. Here, the generalization gap can be written as follows:

$$\phi(g^*, g) = m(h^*, g^*) - m(h^*, g), \tag{15}$$

where we suppose the empirical moment has sufficiently converged generalized residual function $\psi^{\Omega}_{T'|Z}(h^*)$ to a small constant value from the best estimator h^*, which can be written as:

$$m(h^*, g) = \mathbb{E}_Z[\psi^{\Omega}_{T'|Z}(h^*) \cdot (\Omega \circ g)(Z)]. \tag{16}$$

Note that, the generalized residual function $\psi^{\Omega}_{T'|Z}(h^*)$ of the ideal moments $m(h^*, g^*)$ is either a small constant value. From their assumption of moments, we can indicate the generalization gap of Eq. (16) with simple subtraction terms with inner product on the small constant of the generalized residual function, which can be written as follows:

$$\phi(g^*, g) = \mathbb{E}_Z[\psi^{\Omega}_{T'|Z}(h^*) \cdot \{(\Omega \circ g^*)(Z) - (\Omega \circ g)(Z)\}]. \tag{17}$$

In this spot, we unpack the log-likelihood function Ω by using *Taylor Expansion* that it satisfies:

$$\Omega(\omega + \Delta\omega) \approx \Omega(\omega) + \Omega'(\omega) \otimes \Delta\omega, \tag{18}$$

with a vector-valued function $\Omega : \mathbb{R}^{H \times W \times C} \to \mathbb{R}^K$ (class number K) and its derivation function $\Omega' : \mathbb{R}^{H \times W \times C} \to \mathbb{R}^{K \times HWC}$. In addition, the operator $\otimes$ denotes dimension squeeze (*i.e.,* vectorize) and multiplication due to its tensor dimension of $\Delta\omega \in \mathbb{R}^{H \times W \times C}$ such that it satisfies $a \otimes b := a \times \text{Vec}(b)$. Then, Taylor Expansion of the log-likelihood function Ω in Eq. (18) can be also applied to a simple setup $\omega = \mathbf{0}$ for the following equation:

$$\Omega(\mathbf{0} + \Delta\omega) \approx \Omega(\omega = \mathbf{0}) + \Omega'(\omega = \mathbf{0}) \otimes \Delta\omega. \tag{19}$$

Eventually, the generalization gap can be possibly approximated by the following equation:

$$\begin{aligned}
\phi(g^*, g) &= \mathbb{E}_Z[\psi^{\Omega}_{T'|Z}(h^*) \cdot \{(\Omega \circ g^*)(Z) - (\Omega \circ g)(Z)\}] \\
&= \mathbb{E}_Z[\psi^{\Omega}_{T'|Z}(h^*) \cdot \{\underbrace{\Omega(\omega = \mathbf{0}) + \Omega'(\omega = \mathbf{0}) \otimes g^*(Z)}_{(\Omega \circ g^*)(Z)} - \underbrace{(\Omega(\omega = \mathbf{0}) + \Omega'(\omega = \mathbf{0}) \otimes g(Z))}_{(\Omega \circ g)(Z)}\}] \\
&= \mathbb{E}_Z[\psi^{\Omega}_{T'|Z}(h^*) \cdot \{\Omega'(\omega = \mathbf{0}) \otimes (g^*(Z) - g(Z))\}].
\end{aligned} \tag{20}$$

C.3 IDENTITY MAPPING

However, once we directly compute this equation, we will take a striking computational burden from the repeated procedure of tensor derivation Ω' and its dimension squeeze and multiplication $\otimes$. To be specific, computing the generalization gap in Eq. (20) naïvely induces a computational complexity $\mathcal{O}(K^2H^2W^2C^2)$, at least, per one iteration.

Therefore, we should practically compute the generalization gap and get its fast convergence. Here, *localized Rademacher* enables the operator $\otimes$ and the two weighted factors (ψ^{Ω}, Ω') for $g^*(Z)-g(Z)$ to be ignored in computing the generalization gap, and it allows the generalization gap to be uniform bound, such that

$$|\phi(g^*, g)| \approx |\mathbb{E}_Z[g^*(Z) - g(Z)]|, \tag{21}$$

where its complexity is even $\mathcal{O}(1)$ to our satisfaction. Then, we use an elementary algebraic trick with identity mapping $\mathcal{I}$ to approximate tight upper bound of the generalization gap by triangle inequality for its feasible computation within reach as follows:

$$\begin{aligned} |\phi(g^*, g)| = |m(h^*, g^*) - m(h^*, g)| &= |\underbrace{m(h^*, g^*) - m(h^*, \mathcal{I})}_{\phi(g^*, \mathcal{I})} + \underbrace{m(h^*, \mathcal{I}) - m(h^*, g)}_{\phi(\mathcal{I}, g)}| \\ &\leq |\phi(g^*, \mathcal{I})| + |\phi(\mathcal{I}, g)| \approx |\mathbb{E}_Z[g^*(Z) - Z]| + |\mathbb{E}_Z[Z - g(Z)]|, \end{aligned} \tag{22}$$

where $|\phi(g^*, \mathcal{I})|$ in the upper bound is constant value with respect to g. Once we subtract $|\phi(g^*, \mathcal{I})|$ to the above inequality, we can get the supremum value of $|\phi(g^*, g)| - |\phi(g^*, \mathcal{I})|$, as follows:

$$\sup_{g \in \mathcal{G}} |\phi(g^*, g)| - |\phi(g^*, \mathcal{I})| \leq \sup_{g \in \mathcal{G}} |\phi(\mathcal{I}, g)|. \tag{23}$$

Here, we suppose that the absence of regularizing test function forges a significant difference between a feature variation (*i.e.,* instrument) Z and its counterfactuals (*i.e.,* test function) $g(Z)$. This postulation implies that the output of test function strays from the possible feature bound and the infimum value $\inf_{g \in \mathcal{G}} |\phi(\mathcal{I}, g)| \approx |\mathbb{E}_Z[Z - g(Z)]|$ becomes large enough, thus we can realign Eq. (23) with total range of $|\phi(\mathcal{I}, g)|$, which can be written as follows:

$$\sup_{g \in \mathcal{G}} |\phi(g^*, g)| - |\phi(g^*, \mathcal{I})| \leq \inf_{g \in \mathcal{G}} |\phi(\mathcal{I}, g)| \leq |\phi(\mathcal{I}, g)| \leq \sup_{g \in \mathcal{G}} |\phi(\mathcal{I}, g)|. \tag{24}$$

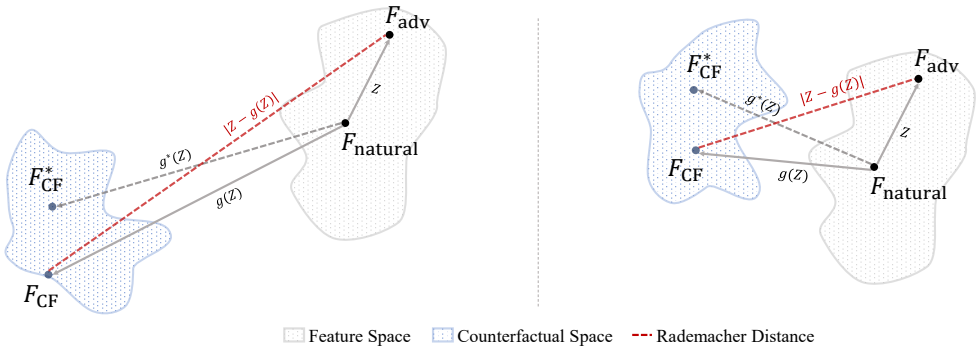
From this inequality, we can show the existence of the triangle inequality $\sup_{g \in \mathcal{G}} |\phi(g^*, g)| \leq |\phi(g^*, \mathcal{I})| + \inf_{g \in \mathcal{G}} |\phi(\mathcal{I}, g)|$ described in our manuscript. In addition, as our manuscript has already explained the connection between the generalization gap and Rademacher complexity, such that $\sup_{g \in \mathcal{G}} |\phi(g^*, g)| = 2b\mathcal{R}(\mathcal{G})$, we eventually get an indirect method to reduce Rademacher complexity once minimizing $|\phi(\mathcal{I}, g)|$ efforlessly. Then, we practically optimize the squared $|\phi(\mathcal{I}, g)|^2$, namely *localized Rademacher regularizer*, together with the main objective of AMR-GMM to maintain a low generalization gap for getting rich test function, which can be written as follows:

$$\min_{h \in \mathcal{H}} \max_{g \in \mathcal{G}} \mathbb{E}_Z[\psi^{\Omega}_{T'|Z}(h) \cdot (\Omega \circ g)(Z)] - |\mathbb{E}_Z[Z - g(Z)]|^2. \tag{25}$$

This ensures the successful achievement of AMR-GMM where the output of test function does not deviate appreciably from the feature variation Z, so that it enables to find out the worst-case counterfactuals within adversarial feature bound. Appendix D describes the triangle inequality clearly with figure and delineates how sufficiently rich test function works, through the lens of empirical evidence by conducting AMR-GMM without the localized Rademacher regularizer.

D RICH TEST FUNCTION BY RADEMACHER COMPLEXITY

In Appendix C, we have verified the realization of the generalization gap on the triangle inequality $\sup_{g\in\mathcal{G}} |\phi(g^*, g)| \leq |\phi(g^*, \mathcal{I})| + \inf_{g\in\mathcal{G}} |\phi(\mathcal{I}, g)|$. To clearly understand it, we then transform representation domain of the triangle inequality to feature and counterfactual space as below figure.

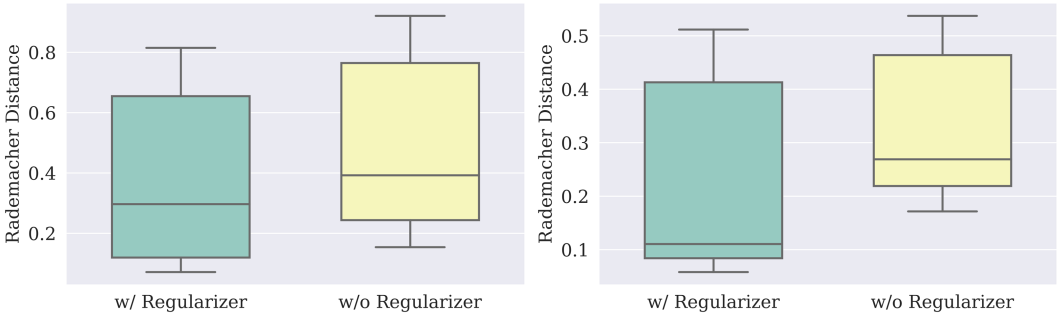


(a) AMR-GMM w/o Regularizer (b) AMR-GMM w/ Regularizer

Figure 4: Representing feature space, counterfactual space, and their space interval (*Rademacher Distance*) according to whether localized Rademacher regularizer is applied in AMR-GMM.

From Fig. 4, we can draw three factors $|\phi(g^*, g)|$, $|\phi(g^*, \mathcal{I})|$, $|\phi(\mathcal{I}, g)|$ of the triangle inequality to $|\mathbb{E}_Z[F^*_{CF} - F_{CF}]|$, $|\mathbb{E}_Z[F^*_{CF} - F_{adv}]|$, $|\mathbb{E}_Z[F_{adv} - F_{CF}]|$ and then the inequality for the given instrument can be obviously shown to: $\sup_{F_{CF}|Z} |F^*_{CF} - F_{CF}| \leq |F^*_{CF} - F_{adv}| + \inf_{F_{CF}|Z} |F_{adv} - F_{CF}|$. Therefore, it becomes a more intuitively understandable formulation to explain their relationship.

Here, we newly define $|F_{adv} - F_{CF}| = |Z - g(Z)|$ as Rademacher Distance (red dotted lines) measuring space interval between feature space and its counterfactual space. These red dotted lines are highly related to the localized Rademacher regularizer $|\phi(\mathcal{I}, g)|^2 \approx |\mathbb{E}_Z[Z - g(Z)]|^2$ as explained in Appendix C. Consequently, using this regularizer makes their space interval close compared to not using it, thereby pushing the conterfactual space towards possible feature space.



(a) CIFAR-10 (b) SVHN

Figure 5: Displaying box distribution statistics of Rademacher Distance for all of the test data samples, compared with w/ Regularizer and w/o Regularizer on (a) CIFAR-10 and (b) SVHN for VGG-16.

To validate that the regularizer actually works in practice, we measure Rademacher Distance and display its box distribution as illustrated in Fig. 5. Here, we can apparently observe the existence of the regularization efficiency from narrowed generalization gap. Concretely, both median and average of Rademacher Distance for the regularized test function are larger than the non-regularized one. Next, in order to investigate how rich test function helps causal inference, we examine imbalance 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 observe the ratio 0.5896 of CIFAR-10 with the regularizer, which is an improvement of 47.8% compared to that 0.3989 with non-regularizer. For SVHN, the ratio with the regularizer either shows better balanced ratio 0.7454 than non-regularizer 0.5699. Therefore, we can wind up that rich test function acquired from the localized Rademacher regularizer serves as a key in improving the generalized capacity of causal inference.

E ALGORITHM DETAIL OF AMR-GMM

Algorithm 1 Adversarial Moment Restriction based Generalized Method of Moments (AMR-GMM)

Require: Data Samples $\mathcal{S}$, Pre-trained Network f, Log-likelihood Function Ω

Initialize parameters θ_h and θ_g of hypothesis model h and test function g
for $(X, G) \sim \mathcal{S}$ **do**
  $X_\epsilon \leftarrow \text{Attack}(X, G)$ ▷ PGD Attack
  $F_{\text{adv}} \leftarrow f_l(X_\epsilon)$, $F_{\text{natural}} \leftarrow f_l(X)$ ▷ Adversarial/Natural Features
  $Z \leftarrow F_{\text{adv}} - F_{\text{natural}}$ ▷ Instrumental Variables
  $T' \leftarrow g(Z)$, $G_{\log} \leftarrow \log G$ ▷ Counterfactual Treatment and Log-Target Label
  $\psi^{\Omega}_{T'|Z}(h) \leftarrow G_{\log} - (\Omega \circ h)(T')$ ▷ Generalized Residual Function for AMR
  $\mathcal{L}_{\text{AMR-GMM}}(\theta_h, \theta_g) \leftarrow \psi^{\Omega}_{T'|Z}(h) \cdot (\Omega \circ g)(Z)$ ▷ Main objective of AMR-GMM
  $\mathcal{L}_{\text{Reg}}(\theta_g) \leftarrow |Z - g(Z)|^2$ ▷ Localized Rademacher Regularizer
  $\theta_g \leftarrow \theta_g + \alpha \frac{\partial}{\partial \theta_g}(\mathcal{L}_{\text{AMR-GMM}} - \mathcal{L}_{\text{Reg}})$ ▷ Update θ_g (α: lr) for Maximizing AMR-GMM Loss
  $\theta_h \leftarrow \theta_h - \alpha \frac{\partial}{\partial \theta_h} \mathcal{L}_{\text{AMR-GMM}}$ ▷ Update θ_h (α: lr) for Minimizing AMR-GMM Loss
end for

Both hypothesis model and test function comprise a bundle of the convolutional layers as a simple CNN structure, trained on AdamW with a learning rate of $\alpha = 10^{-4}$ in 10 epochs. For ImageNet described in Fig. 2, we train it with perturbation budget $2/255$ and its 1.25 times step size using fast adversarial training (Wong et al., 2020) based on FGSM and validate its robustness with equal budget. More details are described in our code at supplementary material.

F ALGORITHM DETAIL OF CAFE

Algorithm 2 CAusal FEatures (CAFE)

Require: Data Samples $\mathcal{S}$, Pre-trained Network f and Hypothesis Model h, Defense Loss $\mathcal{L}_{\text{defense}}$

for $(X, G) \sim \mathcal{S}$ **do**
  $X_\epsilon \leftarrow \text{Attack}(X, G)$ ▷ PGD Attack
  $F_{\text{adv}} \leftarrow f_l(X_\epsilon)$, $F_{\text{natural}} \leftarrow f_l(X)$ ▷ Adversarial/Natural Features
  $Z \leftarrow F_{\text{adv}} - F_{\text{natural}}$ ▷ Instrumental Variables
  $F_{\text{AC}} \leftarrow F_{\text{natural}} + h(Z)$ ▷ Calculating Adversarial Causal Features
  $\delta_{\text{causal}} = \arg\min_{\|\delta\|_\infty \leq \gamma} \mathcal{D}_{\text{KL}}\left(f_{l+}(F_{\text{AC}}) \,||\, f(X_\delta)\right)$ ▷ Causal Perturbation
  $X_{\text{causal}} \leftarrow X + \delta_{\text{causal}}$ ▷ Causal Inversion
  $\hat{F}_{\text{AC}} \leftarrow f_l(X_{\text{causal}})$ ▷ Estimated Causal Features
  $\mathcal{L}_{\text{CAFE}}(\theta_f) \leftarrow \mathcal{L}_{\text{Defense}} + \mathcal{D}_{\text{KL}}(f_{l+}(\hat{F}_{\text{AC}}) \,||\, f_{l+}(F_{\text{adv}}))$ ▷ CAFE Loss with parameter θ_f of f
  $\theta_f \leftarrow \theta_f - \alpha \frac{\partial}{\partial \theta_f} \mathcal{L}_{\text{CAFE}}$ ▷ Update θ_f (α: lr) for Minimizing CAFE Loss
end for

As described in line 9, we readily add a causal regularizer $\mathcal{D}_{\text{KL}}$ to pre-defined defense loss $\mathcal{L}_{\text{Defense}}$ and train all networks from scratch to show the true effectiveness of CAFE. Note that, the number of steps for causal inversion is each 10 for CIFAR-10, SVHN and 3 for Tiny-ImageNet (regarding speed). More details are either described in our code at supplementary material.

G ABLATION STUDIES FOR CAFE WITHOUT CAUSAL INVERSION

We experiment ablation studies of CAFE without causal inversion to show the effectiveness of the causal inversion for CAusal FEatures (CAFE). In Algorithm 2, we first remove the procedures of getting the causal inversion and the estimated causal features in line 6-8, and we name it *CAusal FEatures without causal inversion (CAFE†)* of which algorithm is explained in the following Algorithm 3.

Algorithm 3 CAusal FEatures without Causal Inversion (CAFE†)

Require: Data Samples $\mathcal{S}$, Pre-trained Network f and Hypothesis Model h, Defense Loss $\mathcal{L}_{\text{defense}}$

1: **for** $(X, G) \sim \mathcal{S}$ **do**
2: $X_\epsilon \leftarrow \text{Attack}(X, G)$ ▷ PGD Attack
3: $F_{\text{adv}} \leftarrow f_l(X_\epsilon)$, $F_{\text{natural}} \leftarrow f_l(X)$ ▷ Adversarial/Natural Features
4: $Z \leftarrow F_{\text{adv}} - F_{\text{natural}}$ ▷ Instrumental Variables
5: $F_{\text{AC}} \leftarrow F_{\text{natural}} + h(Z)$ ▷ Calculating Adversarial Causal Features
6: $\mathcal{L}_{\text{CAFE}^\dagger} \leftarrow \mathcal{L}_{\text{Defense}} + \mathcal{D}_{\text{KL}}(f_{l+}(F_{\text{AC}}) \,||\, f_{l+}(F_{\text{adv}}))$ ▷ CAFE† Loss
7: $\theta_f \leftarrow \theta_f - \alpha \frac{\partial}{\partial \theta_f} \mathcal{L}_{\text{CAFE}^\dagger}$ ▷ Update θ_f (α: lr)
8: **end for**

Table 4: Measuring adversarial robustness of CAFE† not using causal inversion (Algorithm 3) and comparing the robustness with original CAFE (Algorithm 2) on five defense baselines: ADV, TRADES, MART, AWP, HELP, trained with VGG-16 for CIFAR-10, SVHN, Tiny-ImageNet under six attack modes: FGSM, PGD, CW$_\infty$, AP, DLR, AA.

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
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$_{\text{CAFE}^\dagger}$	**79.5**	50.6	45.0	43.8	43.5	44.1	41.7	**92.0**	64.9	52.0	49.5	47.6	48.7	45.4	**53.7**	25.4	21.8	21.2	20.6	21.1	20.0
ADV$_{\text{CAFE}}$	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**
$\Delta_{\text{CAFE}^\dagger}$(%)	***1.3***	*1.5*	*0.4*	*3.0*	*0.8*	*2.6*	*2.4*	***0.2***	*0.2*	*-0.2*	*1.2*	*-0.8*	*0.5*	*0.4*	***0.9***	*0.6*	*1.7*	*1.2*	*2.2*	*1.5*	*1.8*
Δ_{CAFE}(%)	*-0.1*	***4.8***	***7.1***	***3.7***	***7.5***	***3.8***	***4.9***	*-0.4*	***3.4***	***6.1***	***2.2***	***6.8***	***2.3***	***1.9***	*-1.2*	***3.0***	***6.4***	***5.2***	***7.8***	***5.6***	***6.9***
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$_{\text{CAFE}^\dagger}$	78.2	50.0	45.1	43.5	43.9	43.6	41.9	90.6	64.1	52.8	49.5	48.5	48.8	45.9	**53.5**	25.5	22.1	21.5	20.9	21.5	20.3
TRADES$_{\text{CAFE}}$	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**
$\Delta_{\text{CAFE}^\dagger}$(%)	***-1.7***	*-0.8*	*-1.4*	*0.5*	*-1.0*	*1.6*	*0.4*	***-1.4***	*-3.4*	*-1.5*	*1.0*	*-1.1*	*2.2*	*1.5*	***1.3***	*-1.9*	*-1.6*	*-1.4*	*-3.0*	*-1.4*	*-2.0*
Δ_{CAFE}(%)	*-3.1*	***2.2***	***4.8***	***1.8***	***5.8***	***2.3***	***2.3***	*-1.8*	***2.1***	***4.6***	***1.9***	***9.3***	***2.9***	***5.0***	*-1.3*	***2.2***	***5.2***	***3.6***	***4.6***	***3.7***	***4.2***
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$_{\text{CAFE}^\dagger}$	79.4	51.7	45.8	43.3	44.1	43.7	41.6	92.0	65.8	53.1	49.1	48.2	48.2	44.9	**53.5**	25.4	21.8	21.3	20.7	21.3	20.2
MART$_{\text{CAFE}}$	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**
$\Delta_{\text{CAFE}^\dagger}$(%)	***-0.5***	*-1.3*	*-3.0*	*-0.2*	*-3.2*	*-0.3*	*-0.8*	***-0.6***	*-1.2*	*-2.0*	*2.3*	*-2.8*	*2.4*	*1.1*	***0.7***	*1.5*	*1.7*	*0.9*	*1.7*	*1.5*	*1.5*
Δ_{CAFE}(%)	*-1.8*	***3.4***	***5.1***	***1.2***	***5.6***	***1.6***	***1.9***	*-1.4*	***1.4***	***5.9***	***3.3***	***9.2***	***2.7***	***4.6***	*-0.2*	***2.4***	***4.0***	***1.8***	***4.3***	***2.5***	***3.1***
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$_{\text{CAFE}^\dagger}$	76.3	50.9	47.0	43.8	45.9	44.0	42.4	83.2	58.0	51.8	49.0	49.8	48.7	47.0	52.5	26.5	23.4	22.6	22.4	22.5	21.6
AWP$_{\text{CAFE}}$	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**
$\Delta_{\text{CAFE}^\dagger}$(%)	*-2.2*	*-1.6*	*-2.4*	*0.6*	*-2.7*	*1.5*	*-0.4*	*-8.3*	*-11.4*	*-8.6*	*-2.9*	*-7.9*	*-2.0*	*-3.3*	*-0.2*	*-5.4*	*-8.7*	*-4.0*	*-9.6*	*-4.3*	*-5.3*
Δ_{CAFE}(%)	***-0.8***	***5.8***	***6.8***	***1.7***	***6.4***	***3.6***	***2.2***	***1.2***	***3.8***	***3.4***	***1.6***	***3.6***	***2.7***	***2.3***	***0.6***	***3.0***	***2.7***	***2.7***	***3.3***	***2.5***	***2.9***
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$_{\text{CAFE}^\dagger}$	76.2	51.0	47.2	43.9	46.1	44.2	42.7	87.6	61.7	53.7	49.6	51.3	49.2	47.3	52.9	27.0	24.1	23.0	23.2	23.0	22.1
HELP$_{\text{CAFE}}$	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**
$\Delta_{\text{CAFE}^\dagger}$(%)	***-1.6***	*-1.6*	*-2.2*	*0.0*	*-2.4*	*0.9*	*-0.4*	*-4.0*	*-6.3*	*-5.1*	*-2.4*	*-5.0*	*-2.0*	*-3.0*	***-0.1***	*-4.4*	*-7.0*	*-3.8*	*-7.8*	*-3.2*	*-4.2*
Δ_{CAFE}(%)	*-2.3*	***5.0***	***6.4***	***1.5***	***6.6***	***2.2***	***1.8***	***0.3***	***2.3***	***3.3***	***1.4***	***4.2***	***2.4***	***2.5***	*-0.8*	***3.9***	***4.7***	***3.1***	***5.0***	***2.4***	***3.5***

Table 4 shows that CAFE without causal inversion (CAFE†) cannot further enhance adversarial robustness of networks compared with that of original CAFE with causal inversion, and even CAFE† has mostly worse robustness than its corresponding baselines. Totally, CAFE† produces a negative effect on pre-defined defense loss. This is because adversarial features cannot easily assimilate causal features far from possible feature bound. This is why we introduce a causal inversion that helps to estimate causal features and fit their prediction. We can then enlighten causal inversion has a remarkable effect to elevate robustness in all of the defense networks and conclude the effectiveness of the CAFE comes from the causal inversion.