智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

引入因果发现学习的跨领域知识泛化方法

李珊珊, 赵清杰, 朱文龙, 阮锦佳, 于铁军, 马少辉, 孙保胜

引用本文:
李珊珊, 赵清杰, 朱文龙, 等. 引入因果发现学习的跨领域知识泛化方法[J]. 智能系统学报, 2025, 20(4): 1033-1045.
LI Shanshan, ZHAO Qingjie, ZHU Wenlong, et al. Cross-domain knowledge generalization method introducing causal discovery learning[J]. *CAAI Transactions on Intelligent Systems*, 2025, 20(4): 1033-1045.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202501005

DOI: 10.11992/tis.202501005
网络出版地址: https://link.cnki.net/urlid/23.1538.tp.20250529.1513.002

引入因果发现学习的跨领域知识泛化方法

李珊珊[1,2], 赵清杰[2], 朱文龙[1], 阮锦佳[3], 于铁军[1], 马少辉[1], 孙保胜[1]
（1. 北京京航计算通讯研究所, 北京 100074; 2. 北京理工大学 计算机学院, 北京 100081; 3. 交通运输部水运科学研究院, 北京 100088）

摘　要: 领域泛化是将多个已知领域的知识泛化到未知目标领域的技术。然而, 现有领域泛化模型在提取图像特征时, 容易受高维噪声的影响, 导致提取的图像特征与标签之间无法建立稳定的因果关系。因此, 受跨域不变因果机制的启发, 本文通过引入因果发现学习技术, 提高跨域知识泛化的准确性。提取图像的低维潜在特征并对其进行变分推理, 保留图像基本信息的同时实现特征变量相互独立; 通过重构潜在特征变量与类别标签之间的因果有向无环图 (directed acyclic graphs, DAG), 发现与类别标签有稳定因果结构的潜在特征变量; 引入反事实对比正则化模块, 利用数据生成过程中的反事实方差和不变性进行因果推断, 生成因果不变表示。为验证本文方法, 在 DomainBed 框架下的 5 个数据集和 SWAD 框架下的 4 个数据集上进行了测试。实验表明, 与现有的领域泛化方法相比, 本文方法在性能和适应性方面有较大提高。

关键词: 迁移学习; 领域泛化; 图像分类; 因果关系; 因果表示学习; 变分推理; 因果发现; 反事实对比
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2025)04−1033−13

中文引用格式: 李珊珊, 赵清杰, 朱文龙, 等. 引入因果发现学习的跨领域知识泛化方法 [J]. 智能系统学报, 2025, 20(4): 1033–1045.
英文引用格式: LI Shanshan, ZHAO Qingjie, ZHU Wenlong, et al. Cross-domain knowledge generalization method introducing causal discovery learning[J]. CAAI transactions on intelligent systems, 2025, 20(4): 1033–1045.

Cross-domain knowledge generalization method introducing causal discovery learning

LI Shanshan[1,2], ZHAO Qingjie[2], ZHU Wenlong[1], RUAN Jinjia[3], YU Tiejun[1], MA Shaohui[1], SUN Baosheng[1]
(1. Beijing Jinghang Research Institute of Computing and Communication, Beijing 100074, China; 2. School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China; 3. China Waterborne Transport Research Institute, Beijing 100088, China)

Abstract: Domain generalization aims to generalize knowledge from multiple known domains to unknown target domains. However, existing models are easily affected by high-dimensional noise when extracting image features, which causes the unstable relationship between the extracted image features and labels. Thus, inspired by the cross-domain invariant causal mechanism, we propose a cross-domain knowledge generalization method introducing causal discovery learning. Specifically, we extract the low-dimensional latent features of the image to retain the basic information of the image. Meanwhile, we perform variational inference on the low-dimensional latent features to achieve mutual independence of latent feature variables. We reconstruct the causal directed acyclic graphs (DAG) between latent feature variables and category labels to discover the latent feature variables that have stable causal structures with category labels. We introduce a counterfactual contrastive regularization term, which exploits counterfactual variance and invariance during data generation to make causal inference and generate causal invariant representations. To verify the proposed method, we conducted tests on five datasets under the DomainBed framework and four datasets under the SWAD framework. Experiments show that compared with existing methods, our domain generalization model has greater improvements in performance and adaptability.

Keywords: transfer learning; domain generalization; image classification; causality; causal representation learning; variational inference; causal discovery; counterfactual contrastive

深度学习 (deep learning, DL)[1-3] 在包括图像分类、目标检测和语义分割等计算机视觉任务上取得了巨大成功。然而, 它们的成功仍然存在潜在

收稿日期: 2025−01−07.　**网络出版日期:** 2025−05−30.
基金项目: 交通运输部水运科学研究院博士科技创新项目 (132415).
通信作者: 赵清杰. E-mail: zhaoqj@bit.edu.cn.

的风险，即面对变化的视觉领域时深度模型十分脆弱。因为深度模型大多是基于理想的假设：训练数据和测试数据是从相同的数据分布中获得。在面对与训练数据分布不同的数据时，模型性能会下降。例如，在晴天拍摄的鸟类图像上训练的目标检测模型，应用到不同天气或不同光照条件下的图像时表现较差。因此，训练一个可以推广到训练数据以外的深度模型是很有必要的。针对这一问题，领域泛化 (domain generalization, DG)[4-5] 技术应运而生，其通过特定算法从一个或几个训练集训练任务模型，无需微调和模型更新，可直接应用到新的测试数据集。

领域泛化是迁移学习 (transfer learning, TL)[6-7] 的一个重要分支。在领域泛化模型的训练阶段，目标领域的数据分布是完全未知的，这就导致领域泛化任务更具有挑战性。大多数领域泛化模型试图从训练域学习跨域不变的特征表示来训练预测模型。然而，最近有研究表明，仅仅学习跨域不变特征是不够的，因为领域泛化问题本质是因果表示学习问题。如果想实现较好的泛化效果，则需要探究潜在特征和标签之间的因果机制。

因果机制不能用布尔逻辑或概率推理的语言来完全描述，它需要额外干预。一些研究者引入因果发现进行因果结构学习，利用观测数据的统计特性恢复因果机制。基于此，本文提出了一种基于因果发现的领域泛化方法，其通过因果发现和反事实样本正则化实现因果结构和因果表示联合优化，从而产生可识别、可解释以及因果相关的特征表示。具体研究内容如下：首先，本文提出了一个特征重构−变分解耦模块，其利用卷积神经网络和变分推理保证了潜在特征中各个子变量的独立性，为后续因果结构学习提供了数据质量的保障。利用前面学习到的潜在特征，建立各维潜在特征变量与类别标签之间的因果有向无环图；通过因果有向无环图的优化，保证了因果特征变量与类别标签的稳定联系，从而发现潜在特征中的不变因果结构，为图像分类和后续方法提供了可信度更高的编码信息。最后，使用反事实对比正则化进行因果推断提取跨域不变因果表示。

1　本文方法

1.1　问题定义与预备知识

为了形式化地描述领域泛化问题，本文符号定义如表 1 所示。其中，$\boldsymbol{X}_i \in \mathcal{X} \subset \mathbf{R}^d$ 和 $Y_i \in \mathcal{Y} \subset \mathbf{R}^d$ 分别表示输入样本和类别标签。各个领域的联合分布不同：$P_{XY}^1 \neq P_{XY}^2 \neq \cdots \neq P_{XY}^S$，但标签空间完全一致：$\mathcal{Y}^1 = \mathcal{Y}^2 = \cdots = \mathcal{Y}^S$。

表 1　主要符号定义

Table 1　Meanings of main symbols

符号	描述
$\mathcal{X}$、$\mathcal{Y}$和$\mathcal{Z}$	输入空间，标签空间和潜在特征空间
$D=\{D_1,D_2,\cdots,D_S\}$	S个领域D_1, D_2,⋯, D_S组成的集合D
$\boldsymbol{X}$和Y	输入样本和类别标签
$P_{XY}^1,P_{XY}^2,\cdots,P_{XY}^S$	每个域输入样本与类别标签的联合分布
$\boldsymbol{Z}$	输入样本在潜在空间中对应的潜在特征
$\boldsymbol{x}$和y	输入样本各维变量和对应的类别标签变量
z	潜在特征$\boldsymbol{Z}$中的各维潜在特征变量
d和m	类别标签的维数和潜在特征的维数

本文引入因果发现学习，旨在发现潜在特征与标签之间的因果关系，进而揭示两者之间的因果结构，下面详细阐述了图像数据的生成过程。假设图像生成过程的因果结构模型 (structural causal model, SCM) 如图 1 所示，定义 $\boldsymbol{X}$ 为生成的图像，$\boldsymbol{Z}$ 为深度潜在特征，Y_{true} 为图像真实类别，D 为域标签。图像 $\boldsymbol{X}$ 是由一组潜在特征变量 $\boldsymbol{Z}=\{z_1,z_2,\cdots,z_n\} \subset \mathbf{R}^n$ 生成，但并不是所有维度的潜在特征变量 z 都会与 Y_{true} 产生联系。其中一部分潜在特征变量 z 与 Y_{true} 产生联系且不受 D 等因素的影响，称为因果变量，这部分是现有深度模型中能产生正确预测标签的深度特征。另一部分受 D 的影响，称为非因果变量，这部分特征变量在深度模型预测任务中往往会导致错误的结论。

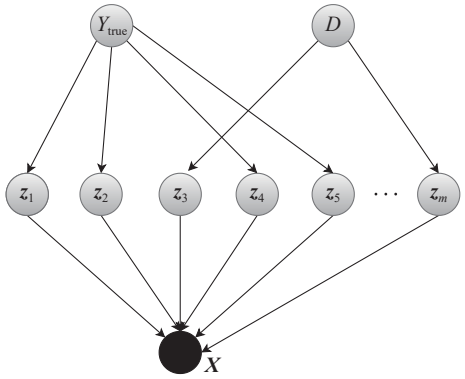

图 1　图像生成过程的因果结构模型

Fig. 1　Structural causal model of image generation process

下面对图像生成过程形式化。

假设 1　图像生成过程[8]。假设图像 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$是由一组潜在特征变量$\boldsymbol{Z}=\{z_1,z_2,\cdots,z_n\}$通过单映射生成函数生成的：

$$h(z)+\boldsymbol{\eta}=\boldsymbol{x}$$

式中：$\boldsymbol{\eta}$为噪声项，$h \in H$是图像生成函数。

因为潜在变量 $\boldsymbol{Z}$ 和类别标签 Y 之间存在因果

关系, 为了将图像生成关系和因果关系区分开来, 假设 2 为形式化因果假设。

假设 2　因果假设[9]。给定一个有向无环图 (directed acyclic graphs, DAG), 它是由一个元组组成 $(\boldsymbol{V}, \boldsymbol{E}, f)$, $\boldsymbol{V}=\boldsymbol{Z}\cup Y=\{\boldsymbol{v}_1,\boldsymbol{v}_2\cdots,\boldsymbol{v}_{d+m}\}$是图中所有节点的集合, $\boldsymbol{E}$ 是因果有向无环图中的边集合, $f=\{f_1,f_2,\cdots,f_d\}$是一组对应每一个类别标签 $Y=\{y_1, y_2,\cdots,y_d\}$的非线性因果函数。根据假设定义:

1) 因果方向, $y_i\to z_j \Rightarrow f_j(y_i)+\varepsilon_j=z_j$;

2) 非因果方向, $y_i\leftarrow z_j \Rightarrow f_j(z_j)+\varepsilon_i=y_i$;

3) Y 和 $\boldsymbol{Z}$ 不会产生混淆;

4) 类别标签 y_i 是相互独立的。

因果函数是非线性相加的模型, 保证了因果关系的可辨识性。为了简单起见, 假设 y_i 的因果关系表示 z_j 是一个一维标量。

本文的目标是发现潜在特征 $\boldsymbol{Z}$ 和类别标签 Y 之间的因果结构, $\boldsymbol{Z}$ 受图像生成过程 $h(\boldsymbol{z})+\boldsymbol{\eta}=\boldsymbol{x}$ 的影响, 可以将这个问题表述为

$$\arg\min_{f}\sum_{i=1}^{d+m}\ell(\boldsymbol{v}_i-f_i(\mathrm{Pa}(\boldsymbol{v}_i)))$$

$$\text{s.t. } h(\boldsymbol{Z})+\boldsymbol{\eta}=\boldsymbol{X},\boldsymbol{v}_i\in\boldsymbol{V},f_i\in f$$

式中: $\ell(\cdot)$表示损失函数, $\mathrm{Pa}(\cdot)$表示 DAG 中某节点的父节点。

相对于传统的因果发现方法, 本文提出的方法面临 3 个挑战。

1) 如何学习图像生成函数 h 和对应的潜在特征表示 $\boldsymbol{Z}$, 在提高表示判别性的同时符合因果发现? 以前的方法通常采用条件变分编码器 (variantional auto-encoder, VAE), 通过近似后验 $p(\boldsymbol{Z}|Y,\boldsymbol{X})$ 来推断 h 和 $\boldsymbol{Z}$。然而, 这些方法通常会丢失一些输入 $\boldsymbol{X}$ 的空间信息。

2) 如何学习由非结构化数据和结构化变量的表示组成的 DAG? 传统的因果 DAG 学习方法由于其过高的维度而难以直接应用于图像特征。

3) 近期虽有反事实方差的相关研究, 但尚未深入探讨其在因果发现框架下的应用。因此, 第 3 个挑战是如何确保表征学习过程中的反事实方差和不变性适用于因果发现任务。

为了应对这些挑战, 本文提出了引入因果发现学习的跨领域知识泛化模型, 如图 2 所示。它主要包含 3 个组件: 特征重构−变分解耦模块, 因果结构学习模块和反事实对比正则化模块。具体来说, 首先通过特征重构−变分解耦模块学习图像的低维嵌入和对低维潜在特征进行变分推理, 来保留图像的基本信息并实现特征向量相互独立。然后通过因果结构学习模块重构潜在特征变量与类别标签之间的因果有向无环图, 来发现与类别标签有稳定因果结构的潜在特征变量。最后, 引入反事实对比正则化项, 利用数据生成过程中的反事实方差和不变性来进行因果推断, 学习领域不变的因果表示。

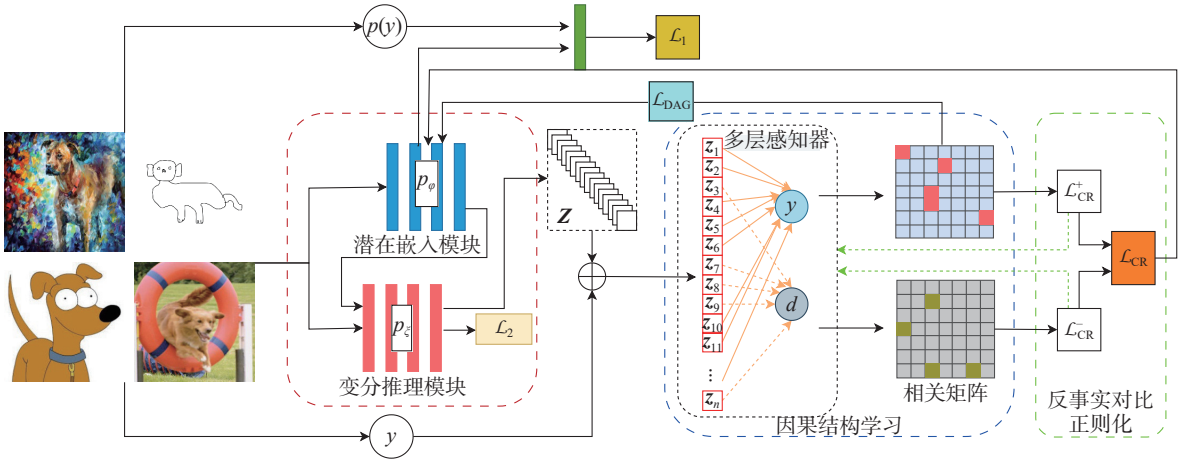

图 2　引入因果发现学习的跨领域知识泛化模型

Fig. 2　Cross-domain knowledge generalization model introducing causal discovery learning

1.2　特征重构−变分解耦模块

本文考虑多域数据的异构结构, 提出了一种双模块表示学习模型。具体来说, 第 1 个模块利用卷积神经网络学习输入的低维嵌入特征 $\boldsymbol{Z}_0$, 它保留了 $\boldsymbol{X}$ 的基本信息; 第 2 个模块基于变分编码器, 在给定 $\boldsymbol{Z}_0$ 和 Y 的情况下, 引入摊销变分估计解耦表示 $\boldsymbol{Z}$。因此, 给定一个样本 $\boldsymbol{X}$ 和其独立副本 $\tilde{\boldsymbol{X}}$, 得到条件生成模型:

$$p_\theta(\boldsymbol{X},\tilde{\boldsymbol{X}},\boldsymbol{Z},\boldsymbol{Z}_0)=p_\varphi(\tilde{\boldsymbol{X}},\boldsymbol{Z}_0)\,p_\xi(\boldsymbol{X},\boldsymbol{Z}|Y)=\underbrace{p_\varphi(\boldsymbol{Z}_0)\,p_\varphi(\tilde{\boldsymbol{X}}|\boldsymbol{Z}_0)}_{\text{Latent Embeddings}}\underbrace{p_\xi(\boldsymbol{X}|\boldsymbol{Z},Y)\,p_\xi(\boldsymbol{Z}|Y)}_{\text{Variational Inference}}\tag{1}$$

式中 θ、φ、ξ 表示模型参数。为了简单起见, $\boldsymbol{X}$ 和

$\tilde{\boldsymbol{X}}$都表示为$\boldsymbol{X}$。式 (1) 中，$p_\varphi(\boldsymbol{X},\boldsymbol{Z}_0)=p_\varphi(\boldsymbol{Z}_0)p_\varphi(\boldsymbol{X}|\boldsymbol{Z}_0)$表示潜在嵌入模块；$p_\xi(\boldsymbol{X},\boldsymbol{Z}|Y)=p_\xi(\boldsymbol{X}|\boldsymbol{Z},Y)p_\xi(\boldsymbol{Z}|Y)$表示变分推理模块，本文用$q_\theta(\boldsymbol{Z}|\boldsymbol{Z}_0,Y)$近似$p_\xi(\boldsymbol{Z}|\boldsymbol{X},Y)$。

1.2.1　潜在嵌入模块

潜在嵌入模块是为了学习低维的嵌入$\boldsymbol{Z}_0$，以保留足够多的$\boldsymbol{X}$包含的信息。这个问题可以表示为生成模型$p_\varphi(\boldsymbol{X},\boldsymbol{Z}_0)=p_\varphi(\boldsymbol{Z}_0)p_\varphi(\boldsymbol{X}|\boldsymbol{Z}_0)$，用卷积神经网络$\log p_\varphi(\boldsymbol{X})$实现：

$$\log p_\varphi(\boldsymbol{X})=\log\int p_\varphi(\boldsymbol{X}|\boldsymbol{Z}_0)p_\varphi(\boldsymbol{Z}_0)\mathrm{d}\boldsymbol{Z}_0\geqslant \mathrm{E}_{q_\theta(\boldsymbol{Z}_0|\boldsymbol{X})}\left[\log p_\varphi(\boldsymbol{X}|\boldsymbol{Z}_0)\right]-\mathrm{KL}\left(p_\varphi(\boldsymbol{Z}_0)\|q_\theta(\boldsymbol{Z}_0|\boldsymbol{X})\right)$$

式中：KL 表示 Kullback-Leibler 散度；不等式表示证据下界 (evidence lower bound, ELBO)。因此，潜在嵌入模块的优化目标可以表示为

$$\mathcal{L}_1=\mathrm{E}_{q_\theta(\boldsymbol{Z}_0|\boldsymbol{X})}\left[\log p_\varphi(\boldsymbol{X}|\boldsymbol{Z}_0)-\mathrm{KL}\left(p_\varphi(\boldsymbol{Z}_0)\|q_\theta(\boldsymbol{Z}_0|\boldsymbol{X})\right)\right]$$

1.2.2　变分推理模块

因为利用卷积神经网络学习到的低维嵌入特征，各个维度的特征之间是相互依赖的。领域泛化模型后续模块搜索因果表示时易受其依赖关系影响，导致泛化效果较差。本模块在给定低维嵌入特征$\boldsymbol{Z}_0$和Y的情况下，通过摊销变分推理学习解耦表示$\boldsymbol{Z}$和生成函数h。根据式 (1)，可以将该问题表述为一个条件生成模型，$p_\xi(\boldsymbol{X},\boldsymbol{Z}|Y)=p_\xi(\boldsymbol{X}|\boldsymbol{Z},Y)p_\xi(\boldsymbol{Z}|Y)$。其中，$\xi$表示参数集合。引入$\boldsymbol{Z}$和参数集合$\theta$，$\log p_\xi(\boldsymbol{X}|Y)$计算公式为

$$\begin{aligned}\log p_\xi(\boldsymbol{X}|Y)&=\log\int p_\xi(\boldsymbol{X}|\boldsymbol{Z},Y)p_\xi(\boldsymbol{Z}|Y)=\\&\log\int p_\xi(\boldsymbol{X}|\boldsymbol{Z})p_\xi(\boldsymbol{Z}|Y)\geqslant\\&\mathrm{E}_{q_\theta(\boldsymbol{Z}|\boldsymbol{Z}_0,Y)}\left[\log\left(\frac{p_\xi(\boldsymbol{X}|\boldsymbol{Z})p_\xi(\boldsymbol{Z}|Y)}{q_\theta(\boldsymbol{Z}|\boldsymbol{Z}_0,Y)}\right)\right]\end{aligned}\tag{2}$$

这里的不等式是对数似然的 ELBO。式 (2) 是条件变分编码器的一种变体，不同之处是其没有使用$q_\theta(\boldsymbol{Z}|\boldsymbol{X},Y)$，而是使用后验$q_\theta(\boldsymbol{Z}|\boldsymbol{Z}_0,Y)$来摊销估计，解决多域数据的异构结构。

根据图 1，先验分布$p_\xi(\boldsymbol{Z}|Y)$和后验分布$q_\theta(\boldsymbol{Z}|\boldsymbol{Z}_0,Y)$可以分解成

$$\begin{aligned}p_\xi(\boldsymbol{Z}|\boldsymbol{Y})&=\prod_{j=1}^{m}p_\xi^{(j)}\left(z_j|y_i\right)\\q_\theta(\boldsymbol{Z}|\boldsymbol{Z}_0,\boldsymbol{Y})&=\prod_{j=1}^{m}q_\theta^{(j)}\left(z_j|z_0,y_i\right)\end{aligned}\tag{3}$$

式中：y_i表示与潜在特征变量z_j有因果关系的结构化变量，$p_\xi^{(j)}$和$q_\theta^{(j)}$分别是p_ξ和q_θ对应的第j个分量。遵循文献 [10] 的设置，本文选用高斯分布作为模型的先验分布：

$$p_\xi^{(j)}(z_j|y_i)\sim\mathcal{N}(\mu_j(y_i),\sigma_j(y_i))\tag{4}$$

因此，根据式 (3) 和 (4)，变分推理模块的优化目标可以表示为

$$\mathcal{L}_2=E_{q_\theta(\boldsymbol{Z}|\boldsymbol{Z}_0,Y)}\log\left(\frac{p_\xi(\boldsymbol{X}|\boldsymbol{Z})\prod_{j=1}^{m}p_\xi^{(j)}\left(z_j|y_i\right)}{\prod_{j=1}^{m}q_\theta^{(j)}\left(z_j|\boldsymbol{Z}_0,y_i\right)}\right)=\mathrm{E}_{q_\theta(\boldsymbol{Z}|\boldsymbol{Z}_0,Y)}\log\left[p_\xi(\boldsymbol{X}|\boldsymbol{Z})-\sum_{j=1}^{m}\mathrm{KL}\left(q_\theta^{(j)}\left(z_j|\boldsymbol{Z}_0,y_i\right)\|p_\xi^{(j)}\left(z_j|y_i\right)\right)\right]$$

1.3　因果结构学习

利用上述模块，得到潜在解耦表示$\boldsymbol{Z}$。然后利用因果结构学习模块恢复由$\boldsymbol{Z}$和Y组成的因果 DAG，DAG 是一种表示因果结构的方法，有向边表示直接的因果关系。因果 DAG 中的每个节点对应每个维度的潜在特征变量$\boldsymbol{z}_i$与标签变量y_i。具体来说，假设$\mathcal{A}(\boldsymbol{W})\in\{0,1\}^{(d+m)\times(d+m)}$为二元矩阵，使得$[\mathcal{A}(\boldsymbol{W})]_{ij}=1\Leftrightarrow w_{ij}\neq 0$，否则为 0。满足前述条件，$\boldsymbol{W}$是有向图$\boldsymbol{G}(\boldsymbol{W})$的邻接矩阵。在本文中，类别标签是一维标量，则$d=1$。除了$\boldsymbol{G}(\boldsymbol{W})$之外，还利用$\boldsymbol{W}=[w_1|w_2|\cdots|w_{m+1}]$定义了线性结构方程模型 (structural equation model, SEM)$\boldsymbol{V}_j=\boldsymbol{w}_j^{\mathrm{T}}\boldsymbol{V}+\boldsymbol{e}_j$，是图$\boldsymbol{G}(\boldsymbol{W})$中的节点，$\boldsymbol{e}=(\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_{m+1})$是随机噪声变量。然后利用最小二乘损失来进行图$\boldsymbol{G}(\boldsymbol{W})$和 SEM 之间的转换$\ell(\boldsymbol{W};\boldsymbol{V})=1/2(m+1)\|\boldsymbol{V}-\boldsymbol{V}\boldsymbol{W}\|_{\mathrm{F}}^2$，F 表示 Frobenius 范数。为了实现矩阵$\boldsymbol{W}$的稀疏性，最小化矩阵$\boldsymbol{W}$的ℓ_1范数$\|\boldsymbol{W}\|_1=\|\mathrm{vec}(\boldsymbol{W})\|_1$，并借鉴文献 [11] 的工作，得到正则化分数函数：

$$\begin{aligned}&\min_{\boldsymbol{W}\in\mathbf{R}^{(m+1)\times(m+1)}}\frac{1}{2(m+1)}\|\boldsymbol{V}-\boldsymbol{V}\boldsymbol{W}\|_{\mathrm{F}}^2+\lambda\|\boldsymbol{W}\|_1\\&\text{s.t. }\boldsymbol{G}(\boldsymbol{W})\in\boldsymbol{D}\end{aligned}$$

式中$\boldsymbol{D}$表示有向无环图的集合。为了对正则化分数函数进行黑盒优化，将组合非循环约束$\boldsymbol{G}(\boldsymbol{W})$替换为单个光滑等式约束$h(\boldsymbol{W})=\mathrm{tr}\left(\mathrm{e}^{\boldsymbol{W}\circ\boldsymbol{W}}\right)-(m+1)$，其中$\circ$表示 Hadamard 乘积。$h(\boldsymbol{W})$用于量化$\boldsymbol{W}$是无环图的程度。当且仅当$h(\boldsymbol{W})=0$，$\boldsymbol{W}$是无环的。因此可以将 DAG 恢复问题转化为拉格朗日局部寻优问题：

$$\begin{aligned}&\min_{\boldsymbol{W}\in\mathbf{R}^{(m+1)\times(m+1)}}\frac{1}{2(m+1)}\|\boldsymbol{V}-\boldsymbol{V}\boldsymbol{W}\|_{\mathrm{F}}^2+\lambda\|\boldsymbol{W}\|_1\\&\text{s.t. }h(\boldsymbol{W})=0\end{aligned}$$

节点$\boldsymbol{v}_i$的优化过程$\boldsymbol{v}_i=f_i(\mathrm{Pa}(\boldsymbol{v}_i))+\varepsilon_i$利用多层感知机 (multi-layer perceptrons, MLP) 实现。重要的是，可以从 MLP 的第一层提取加权邻接矩阵，记为$\boldsymbol{A}^{(1)}$。具体来说，考虑 MLP 具有t个隐藏层，单个激活$\sigma:\mathbf{R}\to\mathbf{R}$和输入$\boldsymbol{V}\in\mathbf{R}^{m+1}$，得到

$$\mathrm{MLP}\left(\boldsymbol{V};\boldsymbol{A}^{(1)}\boldsymbol{A}^{(2)}\cdots\boldsymbol{A}^{(t)}\right)=\sigma\left(\boldsymbol{A}^{(t)}\sigma\left(\cdots\left(\boldsymbol{A}^{(2)}\sigma\left(\boldsymbol{A}^{(1)}(\boldsymbol{V})\right)\right)\right)\right)$$

若$\boldsymbol{A}^{(1)}$的第k列全为0, 则$\mathrm{MLP}\left(\boldsymbol{V};\boldsymbol{A}^{(1)}\boldsymbol{A}^{(2)}\cdots\boldsymbol{A}^{(t)}\right)$与$\boldsymbol{V}$第$i$个坐标$\boldsymbol{v}_i$无关。因此, 如果$\boldsymbol{A}_j^{(1)}$的第$k$列全为0, 则有$\boldsymbol{W}_{kj}=0$。其中, j表示第j层MLP。DAG的学习目标可以形式化为

$$\mathcal{L}_{\mathrm{DAG}}=\frac{1}{m+1}\sum_{i=1}^{m+1}\ell(\boldsymbol{v}_i,\mathrm{MLP}(\boldsymbol{V}))+\alpha_1\left\|\boldsymbol{A}_j^{(1)}\right\|_1+\alpha_2 h(\boldsymbol{W}) \tag{5}$$

$\mathcal{L}_{\mathrm{DAG}}$越小, 意味着学习模型越接近真实的因果DAG。换句话说, 如果存在描述数据生成过程的潜在因果DAG, 则DAG正则化项可以帮助模型基于以因果DAG为特征的不变机制进行预测。将因果DAG学习问题转化为连续优化问题, 使得模型能够联合学习潜在特征$\boldsymbol{Z}$和因果DAG。

1.4　反事实对比正则化

反事实对比正则化模块利用数据生成过程中的一组反事实方差和不变性来实现因果推断, 从而学习因果不变表示。反事实是指如果干预因果模型中的一些变量, 变量的值会如何改变。即在完全一致的现实条件下, 比较不同干预(假设)条件的结果, 这里求的是变量的值而非概率。定义干预为$\mathrm{do}(\boldsymbol{X}=\boldsymbol{X}')$, 即将$\boldsymbol{X}$设为常数$\boldsymbol{X}'$。$p(\mathrm{do}(\boldsymbol{X}=\boldsymbol{X}'))$表示干预的结果分布。本文关注的是干预结构化变量$Y$时, 获得的反事实图像$\boldsymbol{X}$。如果强迫$Y$取值$Y'$, 则定义$\boldsymbol{X}(Y=Y')$为反事实图像。假设有第$k$个样本对$(\boldsymbol{X}^{(k)},Y^{(k)})$和对应的潜在特征表示$\boldsymbol{Z}^{(k)}$。以式(5)中的因果DAG为例, 根据假设2的两种因果结构: 1)$y_i\to z_j$, 2)$z_m\to y_n$, 则有

$$\begin{aligned}\boldsymbol{X}^{(k)}\left(y_i=y_i^{(k)}\right)&\neq\boldsymbol{X}^{(k)}\left(y_i\neq y_i^{(k)}\right)\\ \boldsymbol{X}^{(k)}\left(y_n=y_n^{(k)}\right)&=\boldsymbol{X}^{(k)}\left(y_n\neq y_n^{(k)}\right)\end{aligned} \tag{6}$$

式(6)主要描述了两种因果结构的根本性质:

1) 对于因果方向, 如果y_i改变, $\boldsymbol{X}$也会改变;

2) 对于非因果方向, 如果y_n改变, $\boldsymbol{X}$应该保持不变。

因果表示学习的核心目标是促使模型专注于学习那些与目标标签Y, 具有直接因果关系的表示, 同时尽量减少对与Y无因果关联表示的依赖。为实现这一目标, 提出了以下优化目标, 旨在增强模型捕捉因果特征的能力, 并抑制非因果噪声的干扰, 从而对因果表示学习网络进行算法强化:

$$\begin{aligned}&\arg\max_{p}\\&\left(\sum_{i\in\omega^+}p\left(\hat{\boldsymbol{X}}^{(k)}\left(y_i\neq y_i^{(k)}\right)\neq\boldsymbol{X}^{(k)}|\boldsymbol{X}=\boldsymbol{X}^{(k)}\left(y_i=y_i^{(k)}\right)\right)-\right.\\&\left.\sum_{n\in\omega^-}p\left(\hat{\boldsymbol{X}}^{(k)}\left(y_n\neq y_n^{(k)}\right)\neq\boldsymbol{X}^{(k)}|\boldsymbol{X}=\boldsymbol{X}^{(k)}\left(y_n=y_n^{(k)}\right)\right)\right)\end{aligned} \tag{7}$$

式中: $\omega^+=\left\{y_i|y_i\in\mathrm{Pa}\left(\mathrm{z}_j\right),\mathrm{z}_j\in\boldsymbol{Z},y_i\in Y\right\}$, $\omega^-=\{y_n|y_n\in\mathrm{Ch}(\mathrm{z}_m),\mathrm{z}_m\in\boldsymbol{Z},y_n\in Y\}$, $\hat{\boldsymbol{X}}$表示图像$\boldsymbol{X}$的经验估计。$\mathrm{Ch}(\cdot)$是DAG中节点的子节点集合, $\mathrm{Pa}(\cdot)$是父节点集合。本文方法旨在学习反事实图像的表示, 而非利用因果和反因果方向来区分反事实图像。反事实图像潜在表示的定义见定义1。

定义 1　假设有$(\boldsymbol{X}^{(k)},Y^{(k)})$是第$k$个样本对, $\boldsymbol{Z}$是其对应的潜在表示, 定义$\hat{\boldsymbol{Z}}^{(k)}(y_i\neq y_i^{(k)})$是$\hat{\boldsymbol{X}}^{(k)}(y_i\neq y_i^{(k)})$的反事实图像潜在表示的经验估计, $\hat{\boldsymbol{X}}^{(k)}(y_i\neq y_i^{(k)})=h^{-1}(\hat{\boldsymbol{Z}}^{(k)}(y_i\neq y_i^{(k)}))$。

根据上述定义, 可以将反事实表征的判别问题转化为对比学习任务, 即给定反事实图像表示的经验估计$\hat{\boldsymbol{Z}}^{(k)}(y_i\neq y_i^{(k)})$, 训练一个分类器$C$来区分因果反事实表示$\hat{\boldsymbol{Z}}^{(k)}(y_i\neq y_i^{(k)}),y_i\in\omega^+$和非因果反事实表示$\hat{\boldsymbol{Z}}^{(k)}(y_n\neq y_n^{(k)}),y_n\in\omega^-$。将式(7)中的判别问题形转化为对比学习任务, 提出了反事实表示对比损失:

$$\begin{aligned}\mathcal{L}_{\mathrm{CR}}=\mathrm{E}_{q_\theta}\left[\mathrm{E}_{\omega^+}\left(C\left(\hat{\boldsymbol{Z}}^{(k)}\left(y_i\neq y_i^{(k)}\right)\right)\right)+\right.\\ \left.\mathrm{E}_{\omega^-}\left(1-C\left(\hat{\boldsymbol{Z}}^{(k)}\left(y_n\neq y_n^{(k)}\right)\right)\right)\right]\end{aligned}$$

利用反事实表示对比损失区分因果和非因果表示, 学习领域因果不变表示。

整体优化目标　根据前面3个模块优化目标, 得到整体的优化目标:

$$\mathcal{L}=-\mathcal{L}_1-\mathcal{L}_2+\mathcal{L}_{\mathrm{DAG}}-\mathcal{L}_{\mathrm{CR}}$$

2　实验结果与分析

2.1　实验细节

数据集　本文选用2个通用的领域泛化框架DomainBed[12]和SWAD[13]进行实验来验证方法的有效性。其中DomainBed框架选择Rotated MNIST[14]、VLCS[15]、PACS[16]、TerraIncognita[17]和DomainNet[18]5个数据集, SWAD框架选择VLCS[15]、PACS[16]、TerraIncognita[17]和DomainNet[18]4个数据集。

Rotated MNIST数据集是基于MNIST手写数字数据集的扩展, 但是它在原始图像上进行了旋转变换, 包含10个类别的70 000张图片。每个图片是一个28像素×28像素的灰度图像。根据图像的旋转角度不同, 共分为6个领域: 0°、15°、30°、45°、60°和75°。

VLCS数据集是一个经典的领域泛化数据集, 它由4个数据集的图像组成, 分别是: VOC2007(V)、LabelMe(L)、Caltech(C)和SUN09(S), 每个数据集对应一个领域。包括5种类别(鸟、汽车、椅子、狗和人)的10 729张图片。

PACS 数据集包括 9 991 张图片，共 7 个类别：狗、大象、长颈鹿、吉他、马、房子和人。这些图片来自 4 个差异较大的领域，分别是：艺术 (art-painting, A)、卡通画 (cartoon, C)、照片 (photo, P) 和素描 (sketch, S)。

TerraIncognita 数据集是抓拍的野生动物数据集，根据拍摄相机所在陷阱的位置不同，共分为 4 个领域：L100、L38、L43、L46，共有 24 788 张图片，分布在 10 个类别：鸟、山猫、猫、土狼、狗、空镜、负鼠、兔子、浣熊、松鼠。

DomainNet 数据集包含约 60 万张图片，分布在 345 个类别和 6 个领域中，包括：素描 (sketch)、剪贴画 (clipart)、绘画 (painting)、快速绘制 (quick-draw)、信息图 (infograph) 和照片 (real)，是迄今为止最大的领域泛化数据集。

网络结构与参数设置　DomainBed 框架和 SWAD 框架均认为领域泛化算法的性能严重依赖其所使用的网络架构以及超参数。因此，2 个框架均提出，使所有模型在同一网络架构以及完全相同的超参数下进行性能对比。Rotated MNIST 数据集选择 MNIST ConvNet 网络作为基础网络模型，PACS、VLCS、TerraIncognita 和 DomainNet 数据集则选择 ResNet-50 网络作为基础网络模型。表 2 给出了各个数据集统一参数设置，本文所有的实验都是在 PyTorch 框架下实现。

测试策略　DomainBed 框架中，根据验证集的选择方式，有 2 种不同的模型测试策略。

1) 训练域验证集。在这种策略下，将每个训练域划分为训练子集和验证子集。然后，将每个训练域的验证子集集合起来，创建一个整体的验证集。最后，选择在整个验证集上精度最大化的模型。

2) 测试域验证集。在这个策略下，验证集是基于测试域中的数据形成的，超参数是基于测试时性能进行调优的。为了避免将问题呈现为领域自适应而不是领域泛化，只有在训练结束时才能访问验证集，提前停止训练是不可行的。在这个场景中，所有基于不同算法的模型都应该经过固定的训练步骤，以便公平地相互比较。

表 2　各个数据集对应的超参数和默认值

Table 2　Hyperparameters and default values for each dataset

数据集	参数名	预设值
PACS/VLCS/ TerraIncognita/ DomainNet	学习率	0.000 05
	批大小	32
	权重衰减	0
	随机失活	0
Rotated MNIST	学习率	0.001
	批大小	64
	权重衰减	0

SWAD 框架计算每个域的域外 (out-of-domain) 分类准确率和整体平均值，即模型在训练域上进行训练和验证，并在不可见的测试域上进行评估。

2.2　实验结果

表 3 提供了 DomainBed 框架下“训练域验证集”测试策略在 7 个数据集上的图像分类准确率和平均准确率，表 4 总结了在“测试域验证集”策略下的实验结果，表 5 提供了 SWAD 框架下 5 个数据集的图像分类准确率与平均值。

表 3　训练域验证集测试策略下 DomainBed 框架实验结果

Table 3　Experimental results on DomainBed framework under “Training-domain validation set” test strategy

算法	RotatedMNIST	VLCS	PACS	TerraIncognita	DomainNet	平均值
IRM[19]	97.7±0.1	78.5±0.5	83.5±0.8	47.6±0.8	33.9±2.8	68.2
ERM[20]	98.0±0.0	77.5±0.4	85.5±0.2	46.1±1.8	40.9±0.1	69.6
RSC[21]	97.6±0.1	77.1±0.5	85.2±0.9	46.6±1.0	38.9±0.5	69.1
SagNet[22]	98.0±0.0	77.8±0.5	**86.3±0.2**	48.6±1.0	40.3±0.1	70.2
AND-mask[23]	97.6±0.1	78.1±0.9	84.4±0.9	44.6±0.3	37.2±0.6	68.4
SAND-mask[24]	97.4±0.1	77.4±0.2	84.6±0.9	42.9±1.7	32.1±0.6	66.9
Fish[25]	98.0±0.0	77.8±0.3	85.5±0.3	45.1±1.3	**42.7±0.2**	69.8
Fishr[26]	97.8±0.0	77.8±0.3	85.5±0.4	47.4±1.6	41.7±0.0	70.0
RIDG[27]	—	77.8±0.4	84.7±0.2	47.8±1.1	41.9±0.3	—
本文方法	**98.1±0.0**	**79.2±0.2**	86.1±0.0	**51.4±1.4**	41.3±0.3	**71.2**

注：“—”表示数据缺失，加粗表示本列最优结果。

表 4　在测试域验证集测试策略下 DomainBed 框架实验结果

Table 4　Experimental results on DomainBed framework under "Test-domain validation set" test strategy

算法	RotatedMNIST	VLCS	PACS	Terralncognita	DomainNet	平均值
IRM[19]	97.5±0.2	76.9±0.6	84.5±1.1	50.5±0.7	28.0±5.1	67.5
ERM[20]	97.8±0.1	77.6±0.3	86.7±0.3	53.0±0.3	41.3±0.1	71.3
RSC[21]	97.6±0.1	77.8±0.6	86.2±0.5	52.1±0.2	38.9±0.6	70.5
SagNet[22]	97.9±0.0	77.6±0.1	86.4±0.4	52.5±0.4	40.8±0.2	71.0
AND-mask[23]	97.5±0.0	76.4±0.4	86.4±0.4	49.8±0.4	37.9±0.6	69.6
SAND-mask[24]	97.4±0.1	76.2±0.5	85.9±0.4	50.2±0.1	32.2±0.6	68.4
Fish[25]	97.9±0.1	77.8±0.6	85.8±0.6	50.8±0.4	43.4±0.3	71.1
Fishr[26]	97.8±0.1	78.2±0.2	86.9±0.2	53.6±0.4	**41.8±0.2**	71.7
本文方法	**98.3±0.0**	**80.4±0.7**	**87.1±0.1**	**53.6±0.5**	41.2±0.1	**72.1**

注: 加粗表示本列最优结果。

表 5　在 SWAD 框架上实验结果

Table 5　Experimental results on SWAD framework

算法	PACS	VLCS	Terralncognita	DomainNet	平均值
ERM[20]	84.2	77.3	47.8	44.0	63.3
SAGM[28]	86.6	80.0	48.8	**45.0**	65.1
DCAug[29]	86.1	78.9	48.7	43.7	64.4
本文方法	**87.9**	**80.5**	**50.4**	44.8	**65.9**

注: 加粗表示本列最优结果。

由表 3 和表 4 可以看出, 本文方法在 2 种测试策略下的整体平均分类准确率达到最优。在表 3"训练域验证集"测试策略下, 本文提出的泛化模型在 DomainBed 框架下的 5 个数据集上取得了 71.2% 的平均准确率, 达到了最好的泛化性能。该结果表明通过恢复图像特征与类别标签之间的因果结构, 可以使模型更稳定地学习可迁移特征, 从而提高模型的泛化性能。本文提出的方法在 5 个数据集上的平均准确率都超过了 ERM (empirical risk minimization) 方法, 进一步验证了挖掘图像特征与类别标签之间的因果机制的重要性。另外, 本文方法在 VLCS 和 TerraIncognita 数据集上达到了最优性能, VLCS 和 TerraIncognita 数据集都是自然环境下拍摄的数据集, 这表明了本文方法可以很好地适应自然环境图像的泛化任务, 这对于领域泛化模型的实际应用有重要意义。

表 4 为在"测试域验证集"测试策略下的实验结果。在相同参数设置下, 本文提出的方法取得了依旧取得了最好的性能, 比排名第二的 Fishr 方法提高了 0.4 百分点, 模型在领域差异较大的数据集上也有相同表现。另外, 与 ERM 方法相比, 本文方法的性能也取得了较大提升, 充分证明了通过因果发现来提取因果不变表示, 可以提高模型的泛化性能。

表 5 给出了在 SWAD 框架的 4 个基准数据集上的域外分类准确率和整体平均值, 本文方法取得了最优性能。在领域差异较大 PACS 数据集上, 本文方法比第二名的 SAGM 方法提高了 1.3 百分点。在自然环境下的 Terralncognita 数据集上, 其同样取得了最高分类准确率, 比 SAGM 提高了 1.6 百分点。值得注意的是, 在 DomainNet 数据集上, 本文方法并未能取得最优的实验效果, 这可能是由于这数据集包含较多实例类别, 其包含 345 类的目标实例。本文方法在构建潜在特征与类别标签之间的 DAG 时, 随着类别数量的增加, DAG 的复杂度也会急剧上升, 导致模型在训练过程中不易有效地学习到各个类别之间的判别性特征, 进而影响了模型的泛化性能。

表 6~10 提供了所有数据集的各个子任务的实验结果。

表 6　在 RoatedMINIST 数据集上的实验结果

Table 6　Experimental results on RoatedMINIST dataset

测试策略	算法	0°	15°	30°	45°	60°	75°	平均值
训练域验证集	IRM[19]	95.5±0.1	98.8±0.2	98.7±0.1	98.6±0.1	98.7±0.0	95.9±0.2	97.7
	ERM[20]	95.9±0.1	98.9±0.0	98.8±0.0	98.9±0.0	98.9±0.0	96.4±0.0	98.0
	RSC[21]	94.8±0.5	98.7±0.1	98.8±0.1	98.8±0.0	98.9±0.1	95.9±0.2	97.6
	SagNet[22]	95.9±0.3	98.9±0.1	99.0±0.1	99.1±0.0	99.0±0.1	96.3±0.1	98.0
	AND-mask[23]	95.9±0.4	**99.0±0.1**	98.8±0.1	98.9±0.1	**99.1±0.1**	96.7±0.2	98.1
	SAND-mask[24]	95.5±0.2	99.0±0.0	98.7±0.2	98.8±0.1	98.8±0.0	96.4±0.0	97.9
	Fishr[26]	95.4±0.1	98.6±0.1	98.6±0.1	98.9±0.0	98.8±0.1	95.4±0.3	97.6
	本文方法	**96.0±0.1**	98.9±0.0	**99.1±0.1**	**99.1±0.0**	98.9±0.1	**96.7±0.3**	**98.1**
测试域验证集	IRM[19]	94.9±0.6	98.7±0.2	98.6±0.1	98.6±0.2	98.7±0.1	95.2±0.3	97.5
	ERM[20]	95.3±0.2	98.7±0.1	98.9±0.1	98.7±0.2	98.9±0.0	96.2±0.2	97.8
	RSC[21]	95.4±0.1	98.6±0.1	98.6±0.1	98.9±0.0	98.8±0.1	95.4±0.3	97.6
	SagNet[22]	95.9±0.1	99.0±0.1	98.9±0.1	98.6±0.1	98.8±0.1	96.3±0.1	97.9
	AND-mask[23]	94.9±0.1	98.8±0.1	98.8±0.1	98.7±0.2	98.6±0.2	95.5±0.2	97.5
	SAND-mask[24]	94.7±0.2	98.5±0.2	98.6±0.1	98.6±0.1	98.5±0.1	95.2±0.1	97.4
	Fishr[26]	95.8±0.1	98.3±0.1	98.8±0.1	98.6±0.3	98.7±0.1	96.5±0.1	97.8
	本文方法	**96.3±0.1**	**99.0±0.0**	**99.2±0.1**	**99.2±0.0**	**99.0±0.0**	**96.8±0.3**	**98.3**

注：加粗表示该项最优结果。

表 7　在 VLCS 数据集上的实验结果

Table 7　Experimental results on VLCS dataset

测试策略	算法	C	L	S	V	平均值
训练域验证集	IRM[19]	98.6±0.3	66.0±1.1	69.3±0.9	71.5±1.9	76.3
	ERM[20]	98.0±0.4	62.6±0.9	70.8±1.9	77.5±1.9	77.2
	RSC[21]	97.5±0.6	63.1±1.2	73.0±1.3	76.2±0.5	77.5
	SagNet[22]	97.3±0.4	61.6±0.8	73.4±1.9	**77.6±0.4**	77.5
	AND-mask[23]	97.8±0.4	64.3±1.2	73.5±0.7	76.8±2.6	78.1
	SAND-mask[24]	98.5±0.3	63.6±0.9	70.4±0.8	77.1±0.8	77.4
	Fishr[26]	**98.9±0.3**	64.0±0.5	71.5±0.2	76.8±0.7	77.8
	本文方法	98.3±0.6	**67.8±0.2**	**73.7±0.3**	77.1±0.8	**79.2**
测试域验证集	IRM[19]	97.3±0.2	66.7±0.1	71.0±2.3	72.8±0.4	76.9
	ERM[20]	97.6±0.3	67.9±0.7	70.9±0.2	74.0±0.6	77.6
	RSC[21]	98.0±0.4	67.2±0.3	70.3±1.3	75.6±0.4	77.8
	SagNet[22]	97.4±0.3	66.4±0.4	71.6±0.1	75.0±0.8	77.6
	AND-mask[23]	98.3±0.3	64.5±0.2	69.3±1.3	73.4±1.3	76.4
	SAND-mask[24]	97.6±0.3	64.5±0.6	69.7±0.6	73.0±1.2	76.2
	Fishr[26]	97.6±0.7	67.3±0.5	72.2±0.9	75.7±0.3	78.2
	本文方法	**98.9±0.4**	**69.0±0.6**	**75.8±1.0**	**78.2±1.4**	**80.4**

注：加粗表示该项最优结果。

表 8　在 PACS 数据集上的实验结果

Table 8　Experimental results on PACS dataset

测试策略	算法	A	C	P	S	平均值.
训练域验证集	IRM[19]	84.8 ±1.3	76.4±1.1	96.7±0.6	76.1±1.0	83.5
	ERM[20]	84.7±0.4	80.8±0.6	97.2±0.3	79.3±1.0	85.5
	RSC[21]	85.4±0.8	79.7±1.8	**97.6±0.3**	78.2±1.2	85.2
	SagNet[22]	87.4±1.0	80.7±0.6	97.1±0.1	**80.0±0.4**	**86.3**
	AND-mask[23]	85.3±1.4	79.2±2.0	96.9±0.4	76.2±1.4	84.4
	SAND-mask[24]	85.8±1.7	79.2±0.8	96.3±0.2	76.9±2.0	84.6
	Fishr[26]	**88.4±0.2**	78.7±0.7	97.0±0.1	77.8±2.0	85.5
	本文方法	84.8±1.0	**83.0±0.3**	96.4±0.8	79.9±1.8	86.1
测试域验证集	IRM[19]	84.2±0.9	79.7±1.5	95.9±0.4	78.3±2.1	84.5
	ERM[20]	86.5±1.0	81.3±0.6	96.2±0.3	**82.7±1.1**	86.7
	RSC[21]	86.0±0.7	81.8±0.9	96.8±0.7	80.4±0.5	86.2
	SagNet[22]	87.4±0.5	81.2±1.2	96.3±0.8	80.7±1.1	86.4
	AND-mask[23]	86.4±1.1	80.8±0.9	97.1±0.2	81.3±1.1	86.4
	SAND-mask[24]	86.1±0.6	80.3±1.0	97.1±0.3	80.0±1.3	85.9
	Fishr[26]	**87.9±0.6**	80.8±0.5	**97.9±0.4**	81.1±0.8	86.9
	本文方法	85.7±0.3	**83.2±0.7**	97.3±0.2	82.2±0.2	**87.1**

注: 加粗表示该项最优结果。

表 9　在 TerraIncognita 数据集上的实验结果

Table 9　Experimental results on TerraIncognita dataset

测试策略	算法	L100	L38	L43	L46	平均值
训练域验证集	IRM[19]	54.6±1.3	39.8±1.9	56.2±1.8	39.6±0.8	47.6
	ERM[20]	49.8±4.4	42.1±1.4	56.9±1.8	35.7±3.9	46.1
	RSC[21]	50.2±2.2	39.2±1.4	56.3±1.4	40.8±0.6	46.6
	SagNet[22]	53.0±2.9	43.0±2.5	**57.9±0.6**	40.4±1.3	48.6
	AND-mask[23]	50.0±2.9	40.2±0.8	53.3±0.7	34.8±1.9	44.6
	SAND-mask[24]	45.7±2.9	31.6±4.7	55.1±1.0	39.0±1.8	42.9
	Fishr[26]	50.2±3.9	43.9±0.8	55.7±2.2	39.8±1.0	47.4
	本文方法	**57.3±3.0**	**49.5±0.9**	57.0±1.4	**41.8±0.5**	**51.4**
测试域验证集	IRM[19]	56.5±2.5	49.8±1.5	57.1±2.2	38.6±1.0	50.5
	ERM[20]	59.4±0.9	49.3±0.6	60.1±1.1	43.2±0.5	53.0
	RSC[21]	59.9±1.4	46.7±0.4	57.8±0.5	44.3±0.6	52.1
	SagNet[22]	56.4±1.9	**50.5±2.3**	**59.1±0.5**	44.1±0.6	52.5
	AND-mask[23]	54.7±1.8	48.4±0.5	55.1±0.5	41.3±0.6	49.8
	SAND-mask[24]	56.2±1.8	46.3±0.3	55.8±0.4	42.6±1.2	50.2
	Fishr[26]	60.4±0.9	50.3±0.3	58.8±0.5	**44.9±0.5**	53.6
	本文方法	**64.0±0.8**	49.5±0.9	58.0±0.7	42.7±1.1	**53.6**

注: 加粗表示该项最优结果。

表 10　在 DomainNet 数据集上的实验结果
Table 10　Experimental results on DomainNet dataset

测试策略	算法	clipart	infograph	painting	quickdraw	real	sketch	平均值
训练域验证集	IRM[19]	32.1±13.3	11.0±4.6	26.8±11.3	8.7±2.1	32.7±13.8	28.9±11.9	23.4
	ERM[20]	43.8±1.3	14.8±0.3	38.2±0.6	9.0±0.3	47.0±1.1	39.9±0.6	32.1
	RSC[21]	57.9±0.5	18.5±0.4	46.0±0.1	12.5±0.1	59.5±0.3	49.2±0.1	40.6
	SagNet[22]	58.1±0.3	18.8±0.3	46.7±0.3	12.2±0.4	59.6±0.1	49.8±0.4	40.9
	AND-mask[23]	59.1±0.2	19.1±0.3	45.8±0.7	13.4±0.3	59.6±0.2	50.2±0.4	41.2
	SAND-mask[24]	59.2±0.1	19.7±0.2	46.6±0.3	13.4±0.4	59.8±0.2	50.1±0.6	41.5
	Fishr[26]	58.2±0.5	**20.2±0.2**	**47.7±0.3**	12.7±0.2	**60.3±0.2**	**50.8±0.1**	**41.7**
	本文方法	**59.7±0.4**	19.3±0.0	46.3±0.4	**14.7±1.1**	57.8±0.0	50.2±1.3	41.3
测试域验证集	IRM[19]	32.2±13.3	11.2±4.5	26.8±11.3	8.8±2.2	32.7±13.8	29.0±11.8	23.5
	ERM[20]	40.4±6.6	12.1±2.7	31.4±5.7	9.8±1.2	37.7±9.0	36.7±5.3	28.0
	RSC[21]	57.7±0.3	19.1±0.1	46.3±0.5	13.5±0.4	58.9±0.4	49.5±0.2	40.8
	SagNet[22]	58.6±0.3	19.2±0.2	47.0±0.3	13.2±0.2	59.9±0.3	49.8±0.4	41.3
	AND-mask[23]	**59.3±0.1**	19.6±0.2	46.8±0.2	13.4±0.2	60.1±0.4	50.4±0.3	41.6
	SAND-mask[24]	59.2±0.1	19.9±0.2	47.4±0.2	14.0±0.4	59.8±0.2	50.4±0.4	41.8
	Fishr[26]	58.3±0.5	**20.2±0.2**	**47.9±0.2**	13.6±0.3	**60.5±0.3**	50.5±0.3	**41.8**
	本文方法	59.2±0.1	19.3±0.0	46.0±1.5	**14.6±1.1**	57.6±0.1	**50.5±1.3**	41.2

注: 加粗表示该项最优结果。

RotatedMNIST　表 6 给出了在 RotatedMNIST 数据集上 6 个子任务的实验结果, 可以观察到在“测试域验证集”测试策略下, 本文方法在 6 个子任务上都取得了最高的泛化准确率。

VLCS　VLCS 中的图像都是从真实世界中收集的, 与模拟数据集 (如 RotatedMNIST) 相比, 具有更大的类内方差和领域偏移。从表 7 可以观察到, 各个子任务的准确率相差较大, 这也证明了该数据集领域分布差异较大。本文方法在各个子任务和平均准确率均取得了最优或者次优的泛化性能, 在 L、S 2 个较难的子任务中均取得了最优性能。这充分证明了类别标签对应的图片特征会对模型的性能产生更积极的影响。

PACS　PACS 有艺术绘画、卡通、照片和素描 4 个领域, 由于 4 个领域风格上的巨大差异, 其领域偏移比 VLCS 还要大。但是相比于 VLCS, PACS 图像中的目标占据了所在图像的很大一部分, 并且很好地集中在一起。表 8 实验结果证明, 本文方法同样在 2 个较难的子任务卡通 C 和素描 S 上取得了较高的泛化性能。

TerraIncognita　该数据集由相机在不同地点拍摄的野生动物照片组成。其各个子任务的实验结果如表 9 所示。VLCS 和 TerraIncognita 数据集的图像数据都是自然环境下拍摄的图片, 本文方法都取得了最优性能, 表明本文方法在处理自然环境图像方面具有优势。

DomainNet　该数据集是几个数据集中难度最大的数据集。表 10 提供了在 DomainNet 数据集上的实验结果, 本文方法在快速绘制 (quickdraw) 和素描 (sketch)2 个子任务上取得了最佳性能, 在另外的子任务上也取得了次优的泛化效果。本文方法在面对复杂和多类别领域泛化数据集时, 依旧表现优异, 充分证明学习类别标签与特征之间的因果机制比调整特征更有优势。

消融实验　为了验证本文方法各个模块的实验性能, 对本文提出方法的每个组件进行了消融实验, 以评估本文方法各个部分的有效性。本文共有 3 个模块, 分别是特征重构−变分解耦模块、因果结构学习模块 M_{DAG} 和反事实对比正则化模块 M_{CR}。其中, 特征重构−变分解耦模块又分为 2 个子模块, 潜在嵌入模块 M_1 和变分推理模块 M_2。表 11 提供了在 PACS 数据集上进行实验验证的结果, 可以看出本文提出的每个模块都扮演着重要的角色。具体来说, 潜在嵌入模块 M_1 的表示只应用特征提取网络, Model Ⅰ作为本文方法的基线方法, 其泛化性能较差, 在 Model Ⅱ引入变分推理模块 M_2 之后, 2 种测试策略下的平均准确率都有了显著提高, 表明对潜在嵌入特征进行变分推理使其解耦, 可以提高特征的判别性。Model Ⅲ则只应用了因果发现学习模块 M_{DAG}, 其

平均泛化性能比 Model Ⅰ也有较大提高, 表明学习观测数据的因果关系, 来学习因果不变表示, 对于提高模型的泛化性能有积极作用。Model Ⅴ是在 Model Ⅲ的基础上, 增加了反事实对比正则化模块 M_{CR}, 进一步提高了模型性能。以上实验结果证明了本文方法中各个模块的有效性。

表 11　在 PACS 数据集上的消融实验结果

Table 11　Ablation experimental results on PACS dataset

测试策略	算法	M_1	M_2	M_{DAG}	M_{CR}	A	C	P	S	平均值
训练域验证集	Model I	√	—	—	—	83.5±0.1	78.9±0.5	95.7±0.3	78.2±0.6	84.1
	Model II	√	√	—	—	83.8±0.7	81.5±0.1	96.0±0.2	78.6±1.2	85.0
	Model III	√	—	√	—	83.6±0.2	81.9±1.0	96.0±0.2	78.9±0.8	85.1
	Model IV	√	√	√	—	84.2±1.1	82.2±0.4	96.3±0.4	79.0±0.5	85.4
	Model V	√	—	√	√	84.6±0.8	**83.1±0.2**	96.3±0.1	79.2±0.9	85.8
	本文方法	√	√	√	√	**84.8±1.0**	83.0±0.3	**96.4±0.2**	**79.9±1.8**	**86.1**
测试域验证集	Model I	√	—	—	—	83.9±0.4	78.0±0.6	95.3±0.2	78.3±0.1	83.9
	Model II	√	√	—	—	84.5±0.6	80.2±0.4	96.0±0.2	79.5±0.5	85.1
	Model III	√	—	√	—	84.6±0.1	81.3±0.5	97.2±0.0	81.2±0.8	86.1
	Model IV	√	√	√	—	84.8±0.8	81.8±0.1	97.0±0.4	81.3±0.0	86.2
	Model V	√	—	√	√	85.2±0.3	82.9±0.2	97.1±0.1	82.0±0.7	86.8
	本文方法	√	√	√	√	**85.7±0.3**	**83.2±0.7**	**97.3±0.2**	**82.2±0.2**	**87.1**

注: 加粗表示该项最优结果。

类激活映射　为了对因果结构学习模块获得的 DAG 结构进行可视化, 从而进一步评估因果结构模块的贡献, 本文在 PACS 数据集上的 C、A、S、P 任务上, 进行类激活映射 (class activation mapping, CAM) 实验, 实验结果如图 3 所示。第 1 行是原始输入图像, 第 2 行是 ERM 方法的潜在特征类激活映射结果, 第 3 行是本文方法映射结果。从图中可以看出, 本文方法比 ERM 方法关注更小和更准确的区域, 可以更好地识别和关注不可见域中的相关对象。通过深入研究 DAG 约束下的优化过程, 可以发现因果发现模块主要学习特征之间的层次结构和聚类关系。这些层次化的特征结构使得网络能够更精准地捕捉图像中的因果关系, 从而在 CAM 实验中展现出对目标区域的高度关注。这一结果同时验证了采用因果发现方法学习潜在特征中领域不变表示的有效性和优越性。

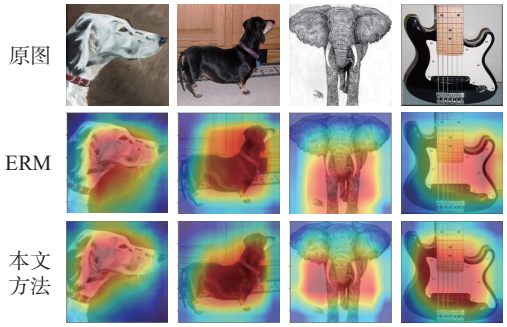

图 3　在 PACS 数据集上的类激活映射

Fig. 3　Class activation mapping on PACS dataset

t-SNE 特征可视化　本节还利用 t-SNE 可视化方法对 ERM 方法和本文方法在 PACS 数据集 A、C、S、P 泛化子任务上学习的深度特征进行 t-SNE 可视化, 实验结果如图 4 所示, 每个点代表一个数据点, 不同颜色对应不同类别。

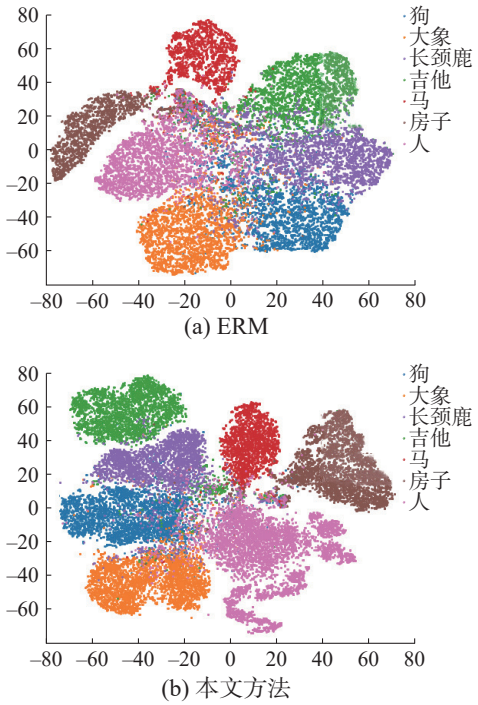

图 4　在 PACS 数据集上的 t-SNE 特征可视化

Fig. 4　t-SNE feature visualization on PACS dataset

从图 4(a) 可以看出, ERM 方法学习到的特征, 不同颜色的散点在特征空间中分布相对随

机，没有明显的聚类或分隔，某些类的决策边界较为模糊，不同类别的特征空间有重叠。这意味着在 ERM 方法下，不同类别的数据在特征空间中没有得到很好的区分。相比之下，本文方法的 t-SNE 可视化特征图中散点分布比 ERM 更集中，某些颜色（即类别）的散点之间出现了更明显的聚类现象。这表明本文方法在特征提取过程中更好地保留了数据的内在结构，使得同类别的数据点更为紧密地聚集在一起，各个类别的边界较为清晰。综上所述，本文方法在图像分类任务上具有更好的性能，证明了因果不变表示在面对差异较大的数据分布时的鲁棒性。

3　结束语

本文提出了一种引入因果发现学习的跨领域知识泛化方法，旨在解决现有领域泛化模型提取的图像特征与标签之间无法建立稳定的因果关系，进而导致泛化性能较差的问题。针对这一问题，首先引入了摊销变分推理方法，解耦各个维度深度图像特征。然后使用因果结构学习来寻找深度特征与标签之间的稳定因果结构，消除不同训练域之间分布差异的影响，提高模型的泛化性能。同时，本文采用反事实对比正则化方法，来区分因果和非因果特征，改进因果发现学习，进一步提高模型的鲁棒性。最后，在 DomainBed 框架和 SWAD 框架下进行实验，在网络结构及预设参数完全一致的情况下，本文提出的方法在 2 种领域泛化框架下的整体表现均优于其他对比算法。尤其是在自然环境数据集中的性能达到了最优，这充分说明了本文提出的方法具有良好的跨域分类能力以及实用性。未来的工作中，将尝试把本文方法应用到目标检测或者语义分割等其他计算机视觉任务中，提高模型的泛化能力。

参考文献：

[1] 张凯, 杨朋澄, 彭开香, 等. 基于深度置信网络的多模态过程故障评估方法及应用[J]. 自动化学报, 2024, 50(1): 89−102.
ZHANG Kai, YANG Pengcheng, PENG Kaixiang, et al. A deep belief network-based fault evaluation method for multimode processes and its applications[J]. Acta automatica sinica, 2024, 50(1): 89−102.

[2] 张汝波, 蔺庆龙, 张天一. 基于深度学习的图像篡改检测方法综述[J]. 智能系统学报, 2025, 20(2): 283−304.
ZHANG Rubo, LIN Qinglong, ZHANG Tianyi. A review of image tampering detection methods based on deep learning[J]. CAAI transactions on intelligent systems, 2025, 20(2): 283−304.

[3] 丁贵广, 陈辉, 王澳, 等. 视觉深度学习模型压缩加速综述[J]. 智能系统学报, 2024, 19(5): 1072−1081.
DING Guiguang, CHEN Hui, WANG Ao, et al. Review of model compression and acceleration for visual deep learning[J]. CAAI transactions on intelligent systems, 2024, 19(5): 1072−1081.

[4] LI Shanshan, ZHAO Qingjie, ZHANG Changchun, et al. Deep discriminative causal domain generalization[J]. Information sciences, 2023, 645: 119335.

[5] 邵海东, 肖一鸣, 颜深, 等. 域泛化机械故障诊断研究进程与展望[J]. 中国科学: 技术科学, 2025, 55(1): 14−32.
SHAO Haidong, XIAO Yiming, YAN Shen, et al. Progress and prospects of domain generalization mechanical fault diagnosis research[J]. Scientia sinica (technologica), 2025, 55(1): 14−32.

[6] 崔腾, 张海军, 代伟. 基于分布共识的联邦增量迁移学习[J]. 计算机学报, 2024, 47(4): 821−841.
CUI Teng, ZHANG Haijun, DAI Wei. Federated incremental transfer learning based on distributed consensus [J]. Chinese journal of computers, 2024, 47(4): 821−841.

[7] 李鑫尧, 李晶晶, 朱磊, 等. 资源受限的大模型高效迁移学习算法研究综述[J]. 计算机学报, 2024, 47(11): 2491−2521.
LI Xinyao, LI Jingjing, ZHU Lei, et al. Efficient transfer learning of large models with limited resources: a survey[J]. Chinese journal of computers, 2024, 47(11): 2491−2521.

[8] MAHAJAN D, TOPLE S, SHARMA A. Domain generalization using causal matching[C]//Proceedings of International Conference on Machine Learning. [S.l.]: OpenReview.net, 2021: 7313−7324.

[9] MAO Haiyi, LIU Hongfu, DOU J, et al. Towards cross-modal causal structure and representation learning[C]// Machine Learning for Health. New Orleans: PMLR, 2022: 120−140.

[10] ZHAO Shengjia, SONG Jiaming, ERMON S. Learning hierarchical features from deep generative models[C]// Proceedings of the International Conference on Machine Learning. Sydney: OpenReview.net, 2017: 4091−4099.

[11] ZHENG Xun, DAN Chen, ARAGAM B, et al. Learning sparse nonparametric dags[C]//Proceedings of the International Conference on Artificial Intelligence and Statistics. [S.l.]: PMLR, 2020: 3414−3425.

[12] GULRAJANI I, LOPEZ-PAZ D. In search of lost domain generalization[C]//Proceedings of International Conference on Learning Representations. [S.l.]: PMLR, 2020: 1−29.

[13] CHA J, CHUN S, LEE K, et al. SWAD: domain general-

ization by seeking flat minima[EB/OL]. (2021−02−17) [2025−01−07]. https://arxiv.org/abs/2102.08604.

[14] GHIFARY M, KLEIJN W B, ZHANG Mengjie, et al. Domain generalization for object recognition with multi-task autoencoders[C]//2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 2551−2559.

[15] FANG Chen, XU Ye, ROCKMORE D N. Unbiased metric learning: on the utilization of multiple datasets and web images for softening bias[C]//2013 IEEE International Conference on Computer Vision. Sydney: IEEE, 2013: 1657−1664.

[16] LI Da, YANG Yongxin, SONG Yizhe, et al. Deeper, broader and artier domain generalization[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 5543−5551.

[17] BEERY S, VAN HORN G, PERONA P. Recognition in terra incognita[C]//Computer Vision–ECCV 2018. Cham: Springer International Publishing, 2018: 472−489.

[18] PENG Xingchao, BAI Qinxun, XIA Xide, et al. Moment matching for multi-source domain adaptation[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 1406−1415.

[19] ARJOVSKY M, BOTTOU L, GULRAJANI I, et al. Invariant risk minimization[EB/OL]. (2019−07−05) [2025−01−07]. https://arxiv.org/abs/1907.02893.

[20] VAPNIK V N. An overview of statistical learning theory [J]. IEEE transactions on neural networks, 1999, 10(5): 988−999.

[21] HUANG Zeyi, WANG Haohan, XING E P, et al. Self-challenging improves cross-domain generalization[C]// Computer Vision–ECCV 2020. Cham: Springer International Publishing, 2020: 124−140.

[22] NAM H, LEE H, PARK J, et al. Reducing domain gap by reducing style bias[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 8686−8695.

[23] PARASCANDOLO G, NEITZ A, ORVIETO A, et al. Learning explanations that are hard to vary[C]//Proceedings of International Conference on Learning Representations. [S.l.]: PMLR, 2021: 1−24.

[24] SHAHTALEBI S, GAGNON-AUDET J C, LALEH T, et al. SAND-mask: an enhanced gradient masking strategy for the discovery of invariances in domain generalization [EB/OL]. (2021−06−04) [2025−01−07]. https://arxiv.org/pdf/2106.02266.

[25] SHI Yuge, SEELY J, TORR P, et al. Gradient matching for domain generalization[C]//Proceedings of International Conference on Learning Representations. [S.l.]: OpenReview.net, 2022: 1−28.

[26] RAME A, DANCETTE C, CORD M. Fishr: invariant gradient variances for out-of-distribution generalization [C]//Proceedings of International Conference on Machine Learning. Baltimore: OpenReview.net, 2022: 18347−18377.

[27] CHEN Liang, ZHANG Yong, SONG Yibing, et al. Domain generalization via rationale invariance[C]//2023 IEEE/CVF International Conference on Computer Vision. Paris: IEEE, 2023: 1751−1760.

[28] WANG Pengfei, ZHANG Zhaoxiang, LEI Zhen, et al. Sharpness-aware gradient matching for domain generalization[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023: 3769−3778.

[29] AMINBEIDOKHTI M, PEÑA F A G, MEDEIROS H R, et al. Domain generalization by rejecting extreme augmentations[C]//2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 2024: 2204−2214.

作者简介：

李珊珊，工程师，博士，主要研究方向为图像智能信息处理、迁移学习和智能算法评测。发表学术论文 10 余篇。E-mail：liss0033@163.com。

赵清杰，教授，博士生导师，主要研究方向机器视觉和智能体系统。主持国家自然科学基金项目、国家重点研发计划项目等 30 余项。获发明专利授权 30 余项，发表学术论文 200 余篇，出版专著 6 部。E-mail：zhaoqj@bit.edu.cn。

朱文龙，高级工程师，主要研究方向为计算机视觉算法评测和智能评测系统。E-mail：78664659@qq.com。

[责任编辑：丁钰]