Automatic Visual Instrumental Variable Learning for Confounding-Resistant Domain Generalization

Fuyuan Cao[1, 2], Shichang Qiao[1] *, Kui Yu[3], Jiye Liang[1]

[1]Shanxi University [2]Shanxi Taihang Laboratory [3]Hefei University of Technology

NEURAL INFORMATION PROCESSING SYSTEMS

Automatic Visual Instrumental Variable Learning for Confounding-Resistant Domain Generalization

Motivation

Domain generalization methods based on predefined instrumental variables (IVs) often violate core assumptions, reducing their effectiveness and limiting generalization. We observe that certain intrinsic visual attributes, such as color and texture, often satisfy the conditions of valid IVs relative to causal factors like shape and contour. By extracting representations of these attributes as approximate IVs, models can effectively eliminate confounding effects and improve domain generalization.

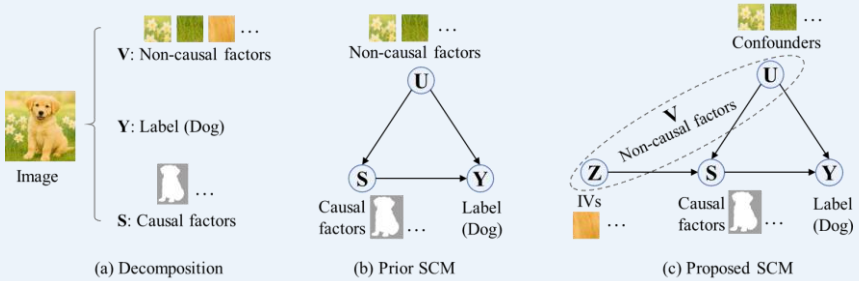

Figure 1: Comparison of proposed and prior SCM.

Contributions

- We propose VIV-DG, a novel approach that automatically learns visual instrumental variables to effectively mitigate the effects of both observed and unobserved confounders, resulting in improved domain generalization.
- We define the novel concept of visual instrumental variables and develop a learner that automatically learns valid ones, effectively overcoming the severe bias caused by violations of IV conditions in predefined approaches.
- Extensive experiments on multiple real-world benchmarks verify the effectiveness and advantages of VIV-DG, demonstrating improved generalization ability.

NEURAL INFORMATION PROCESSING SYSTEMS

Automatic Visual Instrumental Variable Learning for Confounding-Resistant Domain Generalization

Method

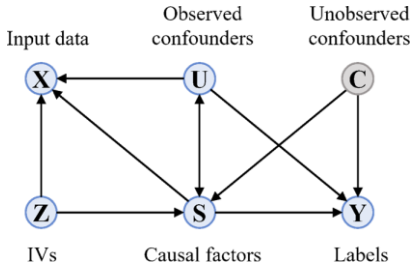

Figure 2: Proposed SCM for DG with IVs.

Definition 1 (Visual instrumental variable). *Suppose the visual space contains a triplet (X, Y, S), where X denotes a visual object (e.g., an image), Y is the label for a downstream task, and S represents a specific causal factor in the visual object such that $S \to Y$. A visual attribute Z is defined as visual instrumental variable (Visual IV) if it satisfies the following three conditions:*

(i) Relevance: Z is significantly associated with the causal factor S, i.e., $Z \not\perp\!\!\!\perp S$;

(ii) Independence: Z is independent of the confounders U, i.e., $Z \perp\!\!\!\perp U$;

(iii) Exclusion: Z affects Y only through its influence on the causal factor S, i.e., $Z \perp\!\!\!\perp Y \mid S$.

Theorem 1 (Learnability of Visual IVs). *Let (X, S, U, Y) be four random variables, where X denotes observed images, S denotes the causal factors affecting Y, U denotes the confounders, and Y denotes the downstream labels. Assume that $\mathcal{H}_Z = \{h_\omega : X \to Z\}$ is a sufficiently expressive family of mappings (e.g., one that contains all smooth bijections on a latent subspace). Consider the objective*

$$\mathcal{L}(\omega) = -\alpha_1 I(h_\omega(X); S) + \alpha_2 I(h_\omega(X); U) + \alpha_3 I(h_\omega(X); Y \mid S), \tag{23}$$

with $\alpha_1, \alpha_2, \alpha_3 > 0$. Then any global minimizer ω^* yields

$$Z^* = h_{\omega^*}(X), \tag{24}$$

which is learnable as defined in the learnability definition (Definition 2) in the Appendix.

NEURAL INFORMATION PROCESSING SYSTEMS

Automatic Visual Instrumental Variable Learning for Confounding-Resistant Domain Generalization

Method

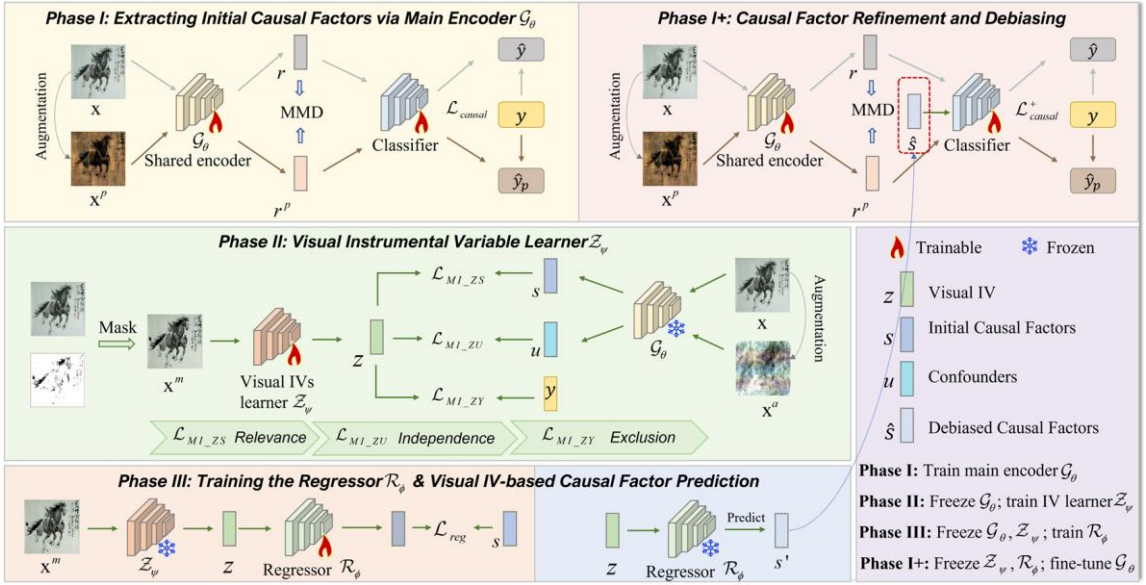

Figure 3: Overview of the VIV-DG framework, which consists of multiple alternately optimized phases: (I) Initial causal factor extraction, (II) Visual IV learning under IV constraints, (III) Causal prediction robust to confounding, and (I+) Causal factor refinement and debiasing.

NEURAL INFORMATION PROCESSING SYSTEMS

Automatic Visual Instrumental Variable Learning for Confounding-Resistant Domain Generalization

Experiments and Results

Table 1: Leave-one-domain-out accuracies on Digits-DG, PACS, and Office-Home

Methods	Digits-DG					PACS (ResNet-18)					Office-Home (ResNet-18)				
	MN	MM	SVHN	SYN	Avg.	A	C	P	S	Avg.	A	C	P	R	Avg.
DeepAll [26]	95.8	58.8	61.7	78.6	73.7	77.6	76.8	95.9	69.5	79.9	57.9	52.7	73.5	74.8	64.7
CCSA [30]	95.2	58.2	65.5	79.1	74.5	-	-	-	-	-	59.9	49.9	74.1	75.7	64.9
JiGen [31]	96.5	61.4	63.7	74.0	73.9	79.4	75.3	96.0	71.4	80.5	53.0	47.5	71.5	72.8	61.2
RSC [32]	-	-	-	-	-	83.4	80.3	96.0	80.9	85.2	58.4	47.9	71.6	74.5	63.1
CrossGrad [33]	96.7	61.1	65.3	80.2	75.8	-	-	-	-	-	58.4	49.4	73.9	75.8	64.4
DDAIG [26]	96.6	64.1	68.6	81.0	77.6	84.2	78.1	95.3	74.7	83.1	59.2	52.3	74.6	76.0	65.5
MatchDG [34]	-	-	-	-	-	81.3	80.7	96.5	79.7	84.6	-	-	-	-	-
L2A-OT [35]	96.7	63.9	68.6	83.2	78.1	83.3	78.2	96.2	73.6	82.8	60.6	50.1	74.8	**77.0**	65.6
CIRL [13]	96.1	**69.9**	76.2	87.7	82.5	86.1	80.6	95.9	82.7	86.3	61.5	55.3	75.1	76.6	67.1
LRDG [36]	-	-	-	-	-	81.9	80.2	95.2	**84.7**	85.5	**61.7**	52.4	73.0	75.9	65.8
FACT [7]	97.9	65.6	72.4	90.3	81.5	85.9	79.4	96.6	80.9	85.7	60.3	54.9	74.5	76.6	66.6
IV-DG [17]	-	-	-	-	-	83.4	78.8	**96.9**	78.7	84.4	60.4	47.7	72.6	76.1	64.2
FAGT [16]	98.3	65.7	70.9	90.4	81.3	**87.5**	80.9	**96.9**	81.9	86.8	60.1	55.0	74.5	75.8	66.4
CDIM [37]	**98.7**	64.0	74.1	**92.9**	82.4	83.6	77.6	95.5	78.2	83.7	-	-	-	-	-
VIV-DG-Lite	97.8	66.8	**77.8**	91.4	**83.5**	86.1	**81.8**	96.6	83.3	87.0	60.8	55.9	**75.8**	76.3	67.2
VIV-DG	97.6	67.1	77.4	91.6	83.4	86.6	81.5	**96.9**	83.9	**87.2**	61.1	**56.3**	75.3	76.8	**67.4**

† The best and second best results are marked in bold and underlined, respectively. Avg. = Average accuracy(%).

Automatic Visual Instrumental Variable Learning for Confounding-Resistant Domain Generalization

Experiments and Results

Table 2: Ablation study on Digits-DG, PACS, and Office-Home datasets

Methods	Setting	Digits-DG					PACS (ResNet-18)					Office-Home (ResNet-18)				
		MN	MM	SV	SY	Avg.	A	C	P	S	Avg.	A	C	P	R	Avg.
VIV-DG	w/o VIV & $\mathcal{R}_\phi$	96.7	62.3	72.8	89.7	80.4	82.7	78.1	93.6	80.5	83.7	58.7	54.6	74.2	76.1	65.9
	w/o $I(Z;S)$	96.8	64.5	74.0	90.6	81.5	85.5	80.4	95.8	83.6	86.3	60.9	54.4	75.6	76.7	66.9
	w/o $I(Z;U)$	96.9	64.1	75.0	90.2	81.6	84.9	79.7	95.9	83.0	85.9	60.9	53.9	75.3	76.6	66.7
	w/o $I(Y;Z\|S)$	96.7	64.6	74.7	90.5	81.6	84.6	79.1	95.3	82.4	85.4	60.7	53.6	75.4	76.5	66.6
	Full Model	97.6	67.1	77.4	91.6	83.4	86.6	81.5	96.9	83.9	87.2	61.1	56.3	75.3	76.8	67.4
VIV-DG-Lite	w/o VIV & $\mathcal{R}_\phi$	96.7	62.3	72.8	89.7	80.4	82.7	78.1	93.6	80.5	83.7	58.7	54.6	74.2	76.1	65.9
	w/o $I(Z;S)$	97.1	64.4	74.9	90.4	81.7	85.6	79.7	95.4	82.9	85.9	60.5	53.9	75.4	75.9	66.4
	w/o $I(Z;U)$	96.9	63.9	74.6	90.2	81.4	84.7	78.4	95.5	82.4	85.3	60.4	53.7	75.6	76.1	66.5
	w/o $I(Y;Z\|S)$	96.5	64.2	74.4	90.1	81.3	84.5	78.2	95.5	81.9	85.0	60.2	53.3	75.2	75.7	66.1
	Full Model	97.8	66.8	77.8	91.4	83.5	86.1	81.8	96.6	83.3	87.0	60.8	55.9	75.8	76.3	67.2

† The results of "w/o IV & $\mathcal{R}_\phi$" are identical for both VIV-DG and VIV-DG-Lite under this setting, and are provided for both methods to facilitate direct comparison.

Experiments and Results

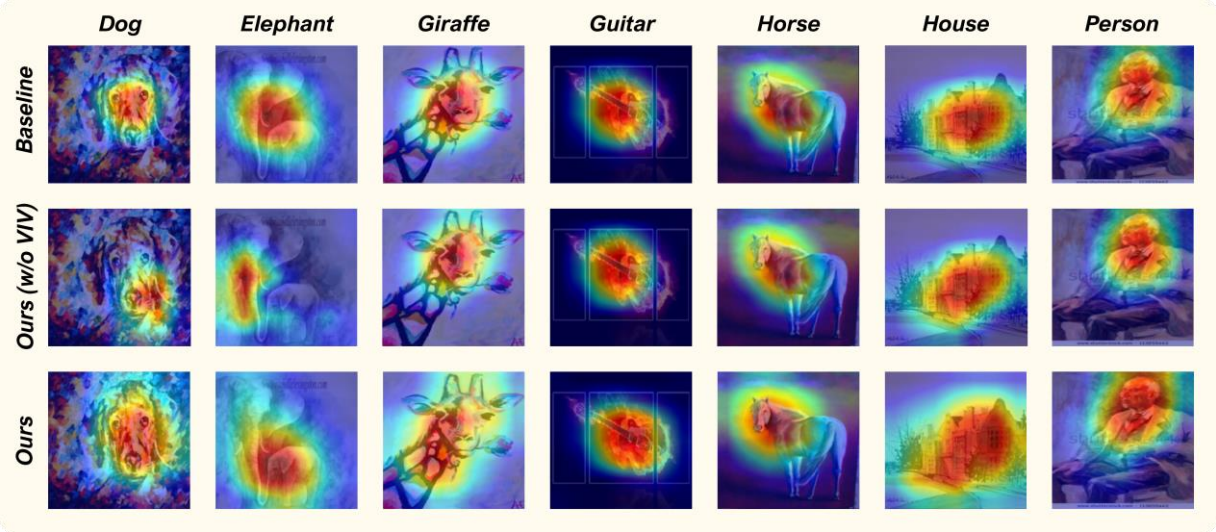

Figure 4: Grad-CAM visualization on the PACS dataset with Art-Painting as the target domain.

Conclusion

We propose a novel confounding-resistant DG approach, termed VIV-DG. We break away from the conventional view that simply categorizes non-causal factors as confounders and observe that certain visual attributes in image data satisfy the conditions of IVs. Building on this insight, We develop a framework that automatically learns valid Visual IVs and mitigates the significant bias arising from violations of IV conditions in predefined IVs. By mitigating confounding effects, including those from unobserved confounders, VIV-DG consistently achieves improved generalization.


NEURAL INFORMATION
PROCESSING SYSTEMS


SHANXI UNIVERSITY
1902


1945

Thank You!!