THE UNIVERSITY of ADELAIDE

Australian Institute for Machine Learning

CLAP: Isolating Content from Style through Contrastive Learning with Augmented Prompts

Yichao Cai (✉), Yuhang Liu, Zhen Zhang, and Javen Qinfeng Shi

Australian Institute for Machine Learning (AIML), The University of Adelaide, SA 5000, Australia

EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO 2024

Motivation

CLIP - Trained on massive image-text data with a symmetric InfoNCE loss.

Pros
- ✓ Mitigated modality gap
- ✓ High zero-shot performance
- ✓ Great generalization ability

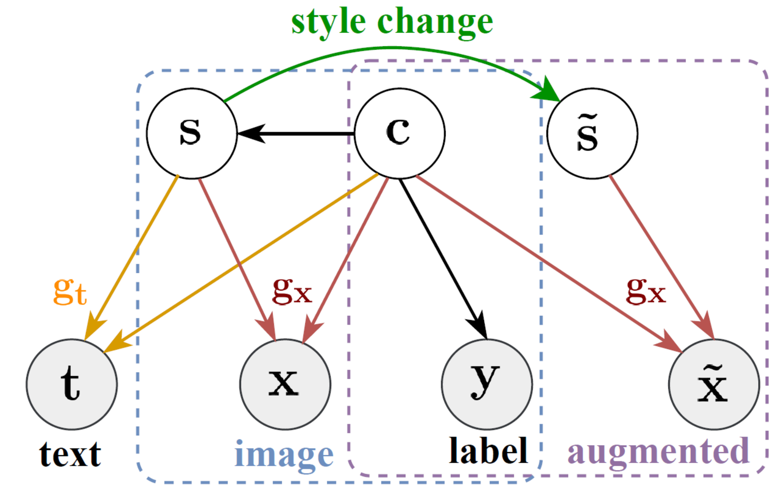
Cons
- Zero-shot sensitive to prompts
- Few-shot performance degrades
- Vulnerable to adversarial attacks

Intuition - There exists spurious correlations in CLIP features, i.e., style-related information is erroneously used to predict class labels.

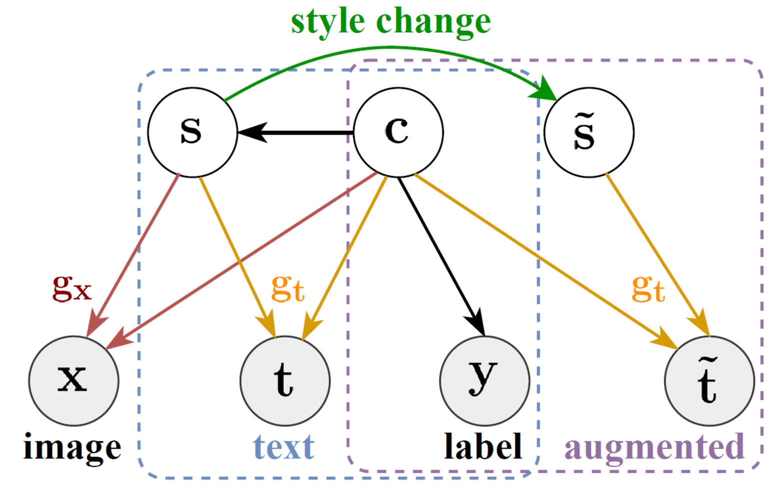
Contribution

- We propose a contrastive learning method to disentangle content and style in pretrained CLIP-like models.
- Our disentangled network, trained on either the image or text encoder, can be seamlessly applied to both modalities.
- Leveraging the high semantic structure of text data, we introduce **CLAP** (Contrastive Learning with Augmented Prompts) to isolate content features within the pre-trained CLIP feature space.

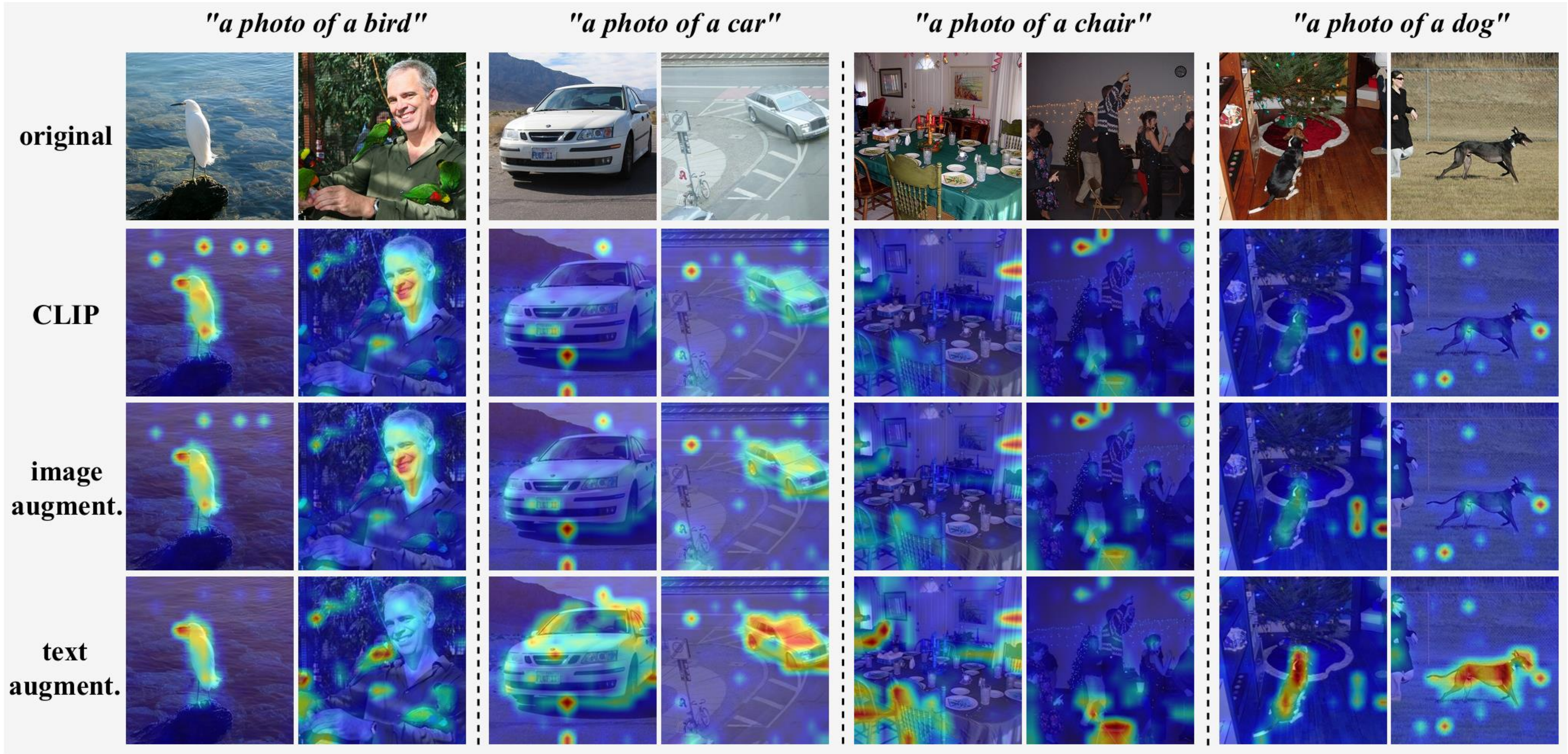
Causal Generative Models of Vision-Language Data



(a) Image augmentation



(b) Text augmentation

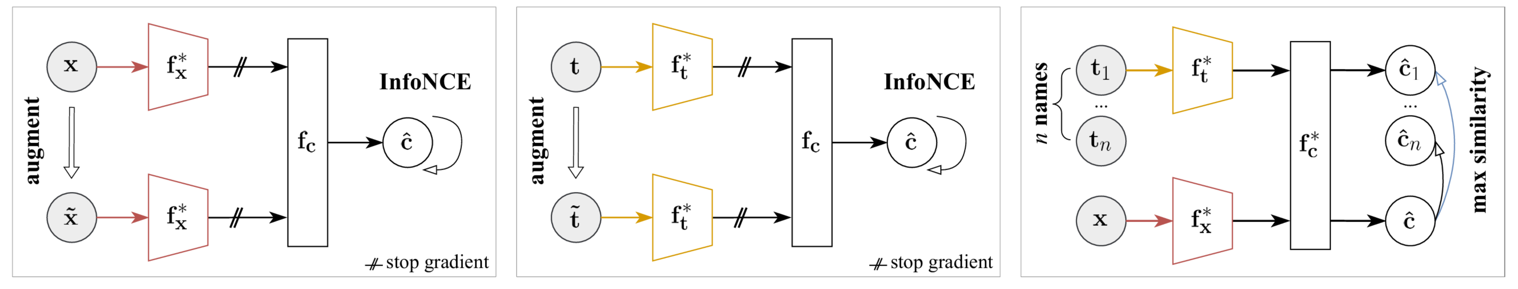


CAM Visualization: Comparative visualization of features from CLIP, Image augmentation (Im.Aug), and Text augmentation (CLAP) for zero-shot inference.

Insight – Image $\mathbf{x}$ and text $\mathbf{t}$, derived from a unified latent space with content $\mathbf{c}$ and style $\mathbf{s}$, follow distinct deterministic processes, $\mathbf{g}_\mathbf{x}$ and $\mathbf{g}_\mathbf{t}$. The class label $\mathbf{y}$ is determined solely by the latent content. (a) Soft interventions on style variables generate augmented images $\tilde{\mathbf{x}}$. (b) Similar interventions produce augmented text $\tilde{\mathbf{t}}$ due to the shared latent space.

Isolating Content through Data Augmentation

Framework



(a) Image augmentation (Im.Aug) (b)Text augmentation (CLAP) (c)Zero-shot inference

Innovation - Contrastive learning with data augmentation in one modality benefits both, with text data being more amendable for style changes due to its semantic structure. The trained adapting network can be seamlessly applied in both modality for zero-shot inference.

Contrastive Learning with Augmented Images (Im.Aug)

Template Prompts *"an [**Art Style**] [**Image Type**] of an [**Object Size**] [**Object Color**] [**Class**]"*

Art Style	Image Type	Object Size	Object Color
realistic, impressionistic	painting, cartoon, inforgraph, sketch, photograph, clipart, mosaic art, sculpture	large, small, normal sized	yellow, green, black, blue, multicolored, orange, red, white, brown, purple

Text-to-image diffusion model

Synthetic Images

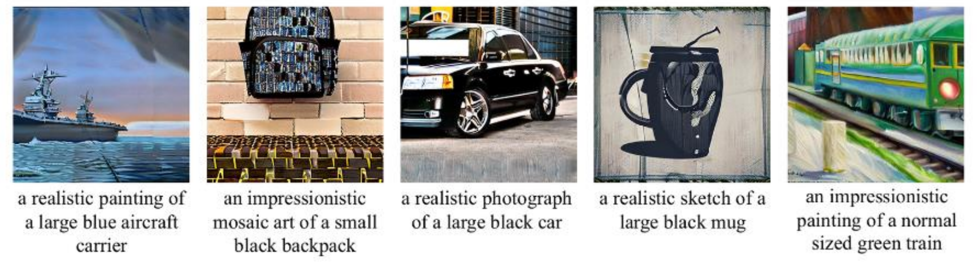


Image Augmentation Random Cropping + Color distortion

Contrastive Learning with Augmented Prompts (CLAP)

Template Prompts *"an [**Art Style**] [**Image Type**] of an [**Object Size**] [**Object Color**] [**Class**]"*
e.g., "a realistic paiting of a large red car"

Text augmentation Randomly apply the combination of these techniques:

Object Size Deletion **(OSD)**	Object Color Deletion **(OCD)**	Image Type Deletion **(ITD)**	Art Style Deletion **(ASD)**	Swapping Prompt Order **(SPO)**	Inserting Gaussian Noise **(IGN)**
a realistic paiting of a red car	a realistic paiting of a large car	a realistic of a large red car	a paiting of a large red car	a large red car in a realistic painting	[noise] a realistic paiting of a large red car

Experimental Results

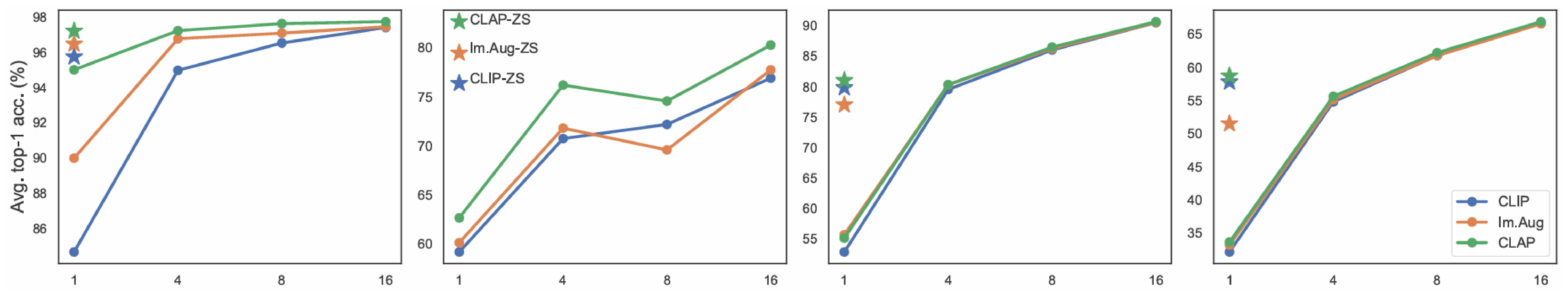
Zero-shot performance

Prompt	Method	Zero-shot performance, avg. top-1 acc. (%) (↑) PACS	VLCS	Off.Home	Dom.Net	Overall
ZS(C)	CLIP	95.7	76.4	79.8	57.8	77.4
	Im.Aug	96.5	79.5	77.0	51.5	76.1
	CLAP	**97.2**	**82.6**	**81.0**	**58.7**	**79.9**
ZS(CP)	CLIP	95.2	82.0	79.5	57.0	78.4
	Im.Aug	96.3	82.9	75.8	50.7	76.4
	CLAP	**97.3**	**83.4**	**80.5**	**58.0**	**79.8**
ZS(PC)	CLIP	96.1	82.4	82.5	57.7	79.7
	Im.Aug	96.5	83.0	78.6	51.6	77.4
	CLAP	**97.2**	**83.4**	**83.0**	**59.0**	**80.6**
ZS(NC)	CLIP	90.8	68.3	71.5	51.0	70.4
	Im.Aug	94.8	73.1	67.5	44.0	69.9
	CLAP	**97.2**	**81.0**	**73.5**	**52.6**	**76.1**

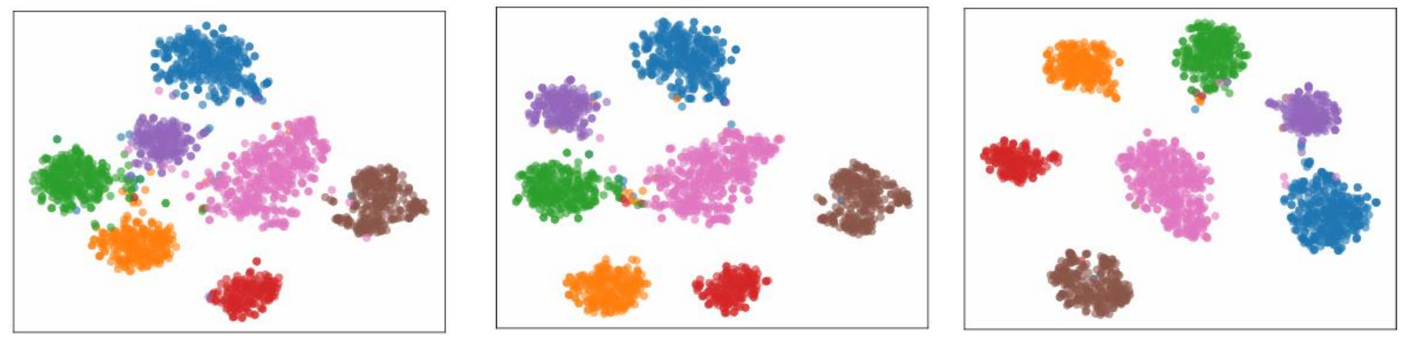
Metric	Method	Performance variance, avg. top-1 acc. (%) (↓) PACS	VLCS	Off.Home	Dom.Net	Overall
R	CLIP	0.9	6.1	3.1	**0.8**	2.7
	Im.Aug	**0.1**	3.6	2.8	0.9	1.9
	CLAP	**0.1**	**0.8**	**2.5**	1.0	**1.1**
δ	CLIP	0.4	2.8	1.4	**0.4**	1.2
	Im.Aug	0.1	1.7	1.2	**0.4**	0.8
	CLAP	**0.0**	**0.4**	**1.1**	**0.4**	**0.5**
$\Delta_{(NC)}$	CLIP	4.9	8.1	8.3	6.8	7.0
	Im.Aug	1.6	6.4	9.5	7.5	6.3
	CLAP	**0.0**	**1.6**	**7.5**	**6.1**	**3.8**

Inference prompts: **ZS(C)** – "[class]" **ZS(CP)** – "a [class] in a photo" **ZS(PC)** – "a photo of a [class]" **ZS(NC)** – "[Gaussian noise][class]"

Few-shot performance



Qualitative analysis with t-SNE visualization



t-SNE comparison of CLIP, Im.Aug, and CLAP features (from left to right) on the PACS dataset, Art Painting domain

Adversarial robustness

Setting	Method	FGSM PACS	FGSM VLCS	FGSM O.H.	FGSM D.N.	PGD-20 PACS	PGD-20 VLCS	PGD-20 O.H.	PGD-20 D.N.	CW-20 PACS	CW-20 VLCS	CW-20 O.H.	CW-20 D.N.	Avg.
ZS(C)	CLIP	86.8	65.6	57.9	22.5	29.1	2.0	10.1	10.7	27.4	1.5	7.4	7.6	29.2
	Im.Aug	88.0	69.6	55.1	37.9	**31.3**	2.1	10.4	9.0	29.4	1.7	7.0	5.8	31.1
	CLAP	**88.7**	**71.9**	**58.5**	**44.2**	30.8	**3.2**	**10.6**	**11.2**	**29.8**	**2.3**	**8.1**	**8.0**	**32.7**
1-shot	CLIP	66.7	45.2	34.3	22.5	34.8	16.0	5.6	11.3	18.9	0.7	4.5	3.2	23.7
	Im.Aug	79.4	47.1	**37.1**	23.5	55.2	16.1	**8.5**	**12.5**	23.2	0.9	**5.1**	3.4	28.0
	CLAP	**89.6**	**52.2**	**37.1**	**23.9**	**73.4**	**21.2**	7.4	**12.5**	**27.0**	**1.1**	5.0	**3.5**	**31.9**

Avg. top-1 acc. (%) under adversarial attacks(↑)

- CLAP successfully refines pretrained CLIP features by effectively isolating content from style, as demonstrated by improved zero-shot and few-shot performance, increased robustness against adversarial attacks, and enhanced qualitative visualizations.
- Exploring the more efficient data augmentation techniques, including the combination of augmentations from both modalities could be valuable, given their complementary strengths.

Scan me!