Visual Prompting via Image Inpainting

NeurIPS 2022

Amir Bar* [1,2] , Yossi Gandelsman* [1] , Trevor Darrell[1] , Amir Globerson[2] , Alexei A. Efros[1]

1 UC Berkeley 2 Tel Aviv University

Outline

- Authorship
- Background
- Method
- Experiments
- Summary

Background

Prompting

- Pretrain-finetune
  - Finetuning the pretrained model with downstream data & labels
- Pretrain-prompt-predict
  - No longer need finetuning
  - Using prompting to predict
    - E.g. Sentiment prediction. “I love this moive.”
    - Prompt: “I love this moive. Overall, it was a [Z] moive.” We predict sentiment by [Z]

Background

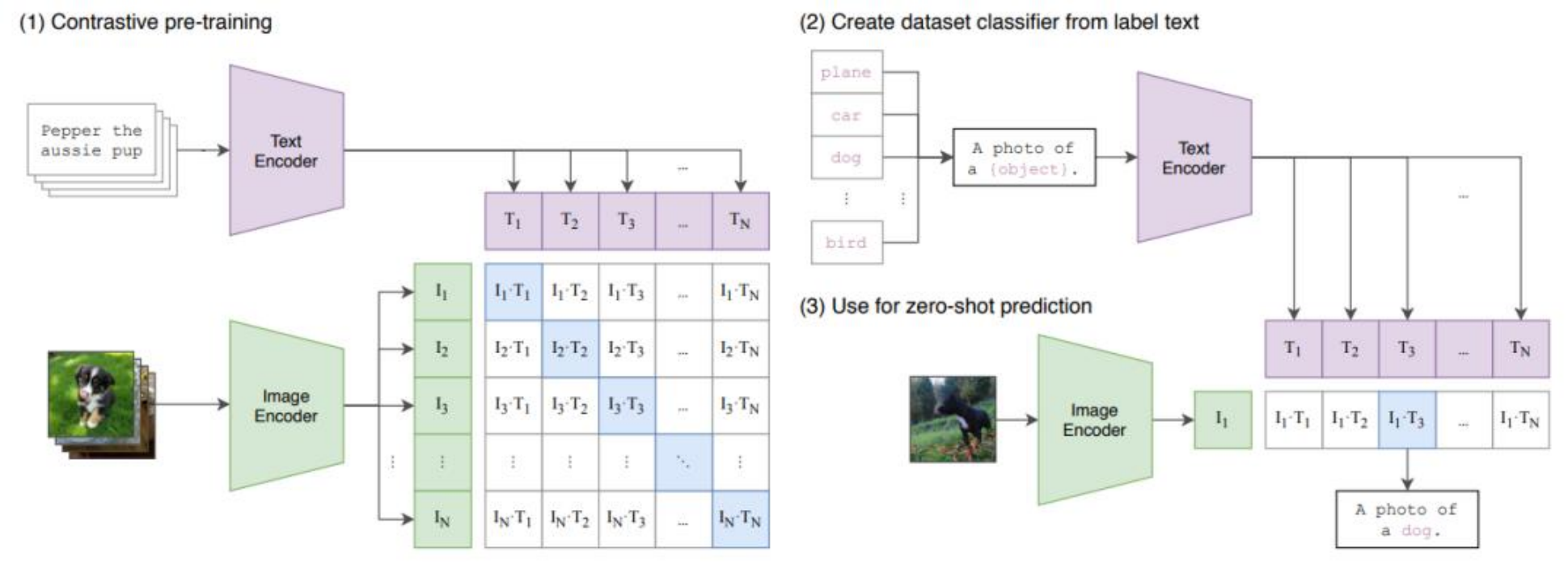
Prompting

- Learned prompt
  - Human-designed prefix is time-consuming/not robust
  - Learning a prefix with few-shot data
    - “I love this moive. Overall, it was a [Z] moive.”
    - “I love this moive.[P][Z].” [P] is tuned on the downstream data.

Background

Prompting in Vision

- CLIP



Background

Prompting in Vision

- CoOp: Using learned prompt

Caltech101

Prompt	Accuracy
a [CLASS].	82.68
a photo of [CLASS].	80.81
a photo of a [CLASS].	86.29
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**91.83**

(a)

Flowers102

Prompt	Accuracy
a photo of a [CLASS].	60.86
a flower photo of a [CLASS].	65.81
a photo of a [CLASS], a type of flower.	66.14
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**94.51**

(b)

Describable Textures (DTD)

Prompt	Accuracy
a photo of a [CLASS].	39.83
a photo of a [CLASS] texture.	40.25
[CLASS] texture.	42.32
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**63.58**

(c)

EuroSAT

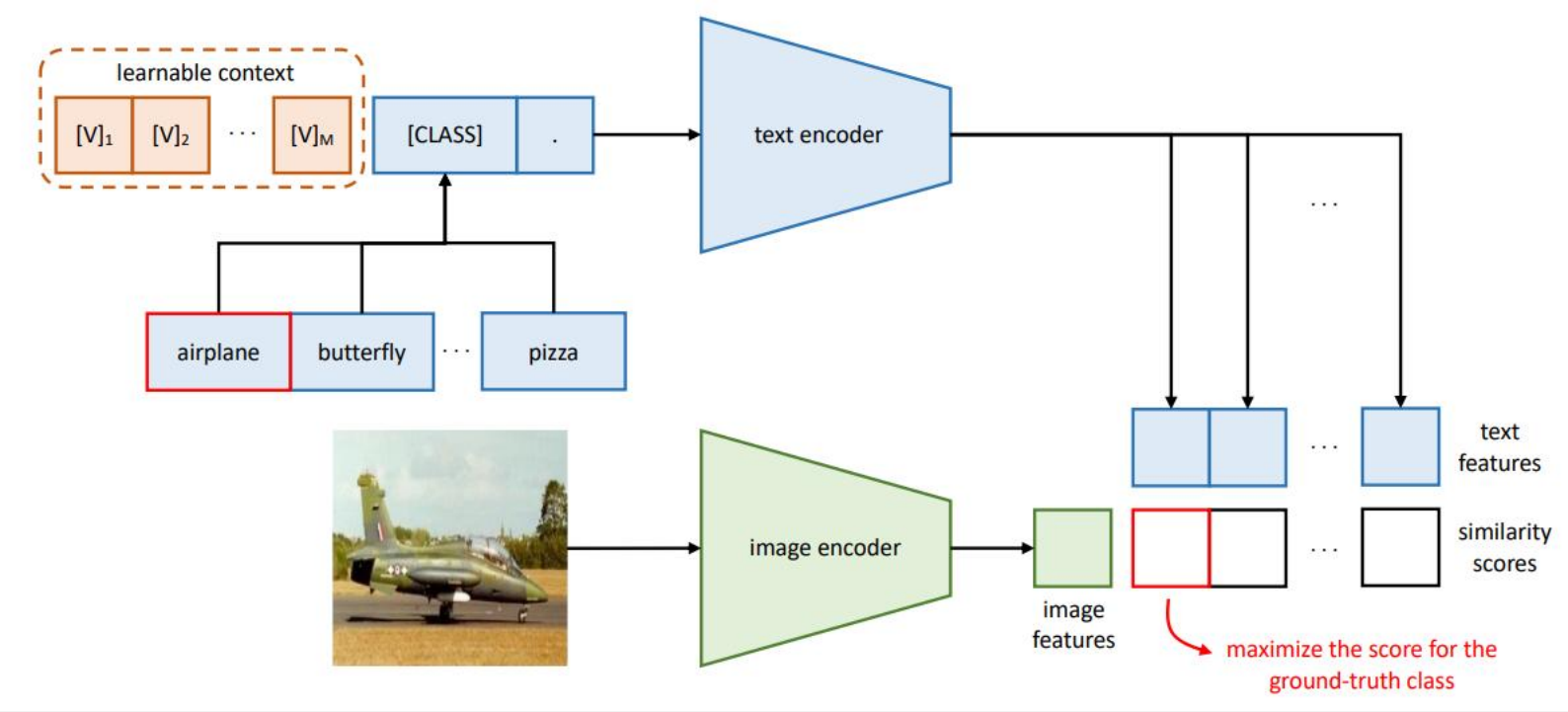
Prompt	Accuracy
a photo of a [CLASS].	24.17
a satellite photo of [CLASS].	37.46
a centered satellite photo of [CLASS].	37.56
$[V]_1$ $[V]_2$ ... $[V]_M$ [CLASS].	**83.53**

(d)

Background

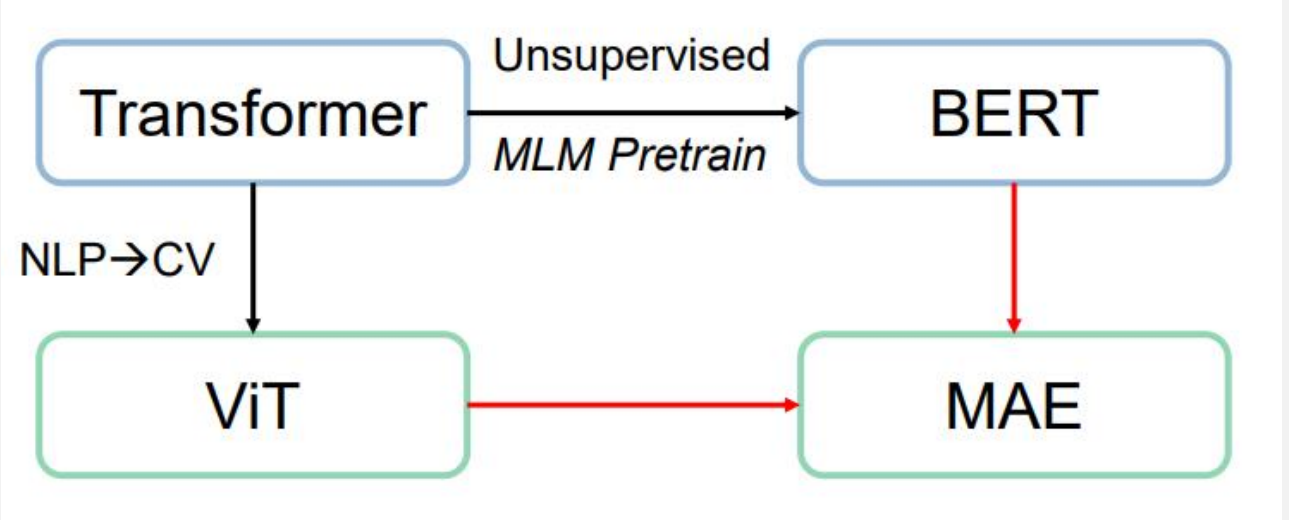
Prompting in Vision

- CoOp: Using learned prompt



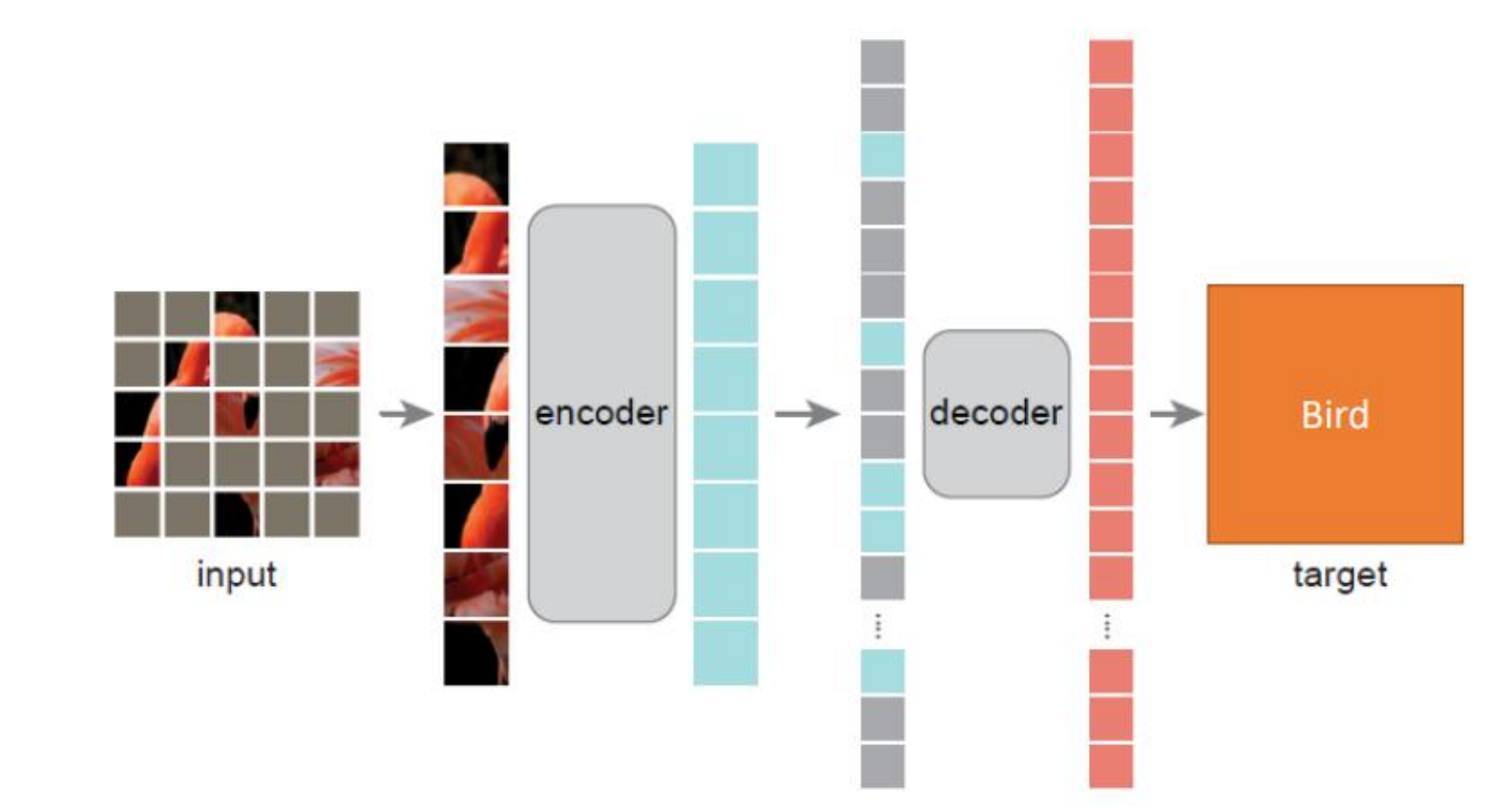
Background

MAE



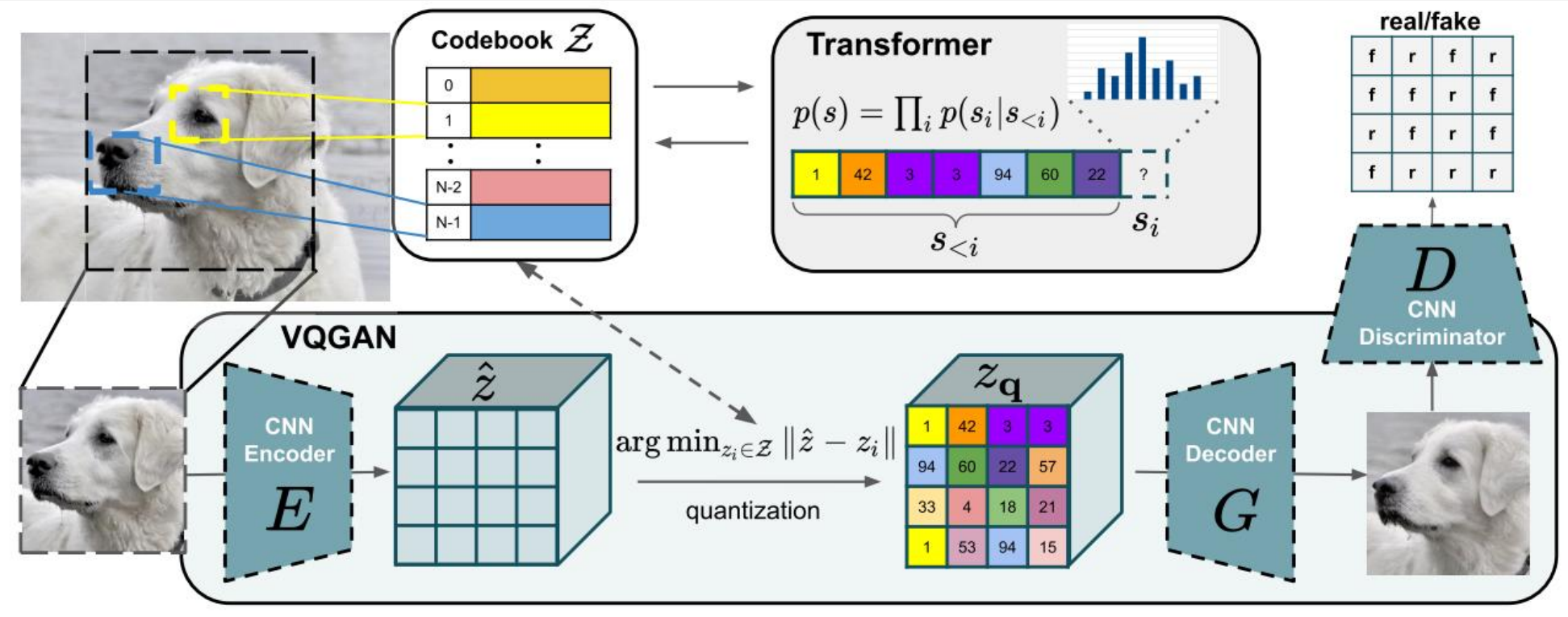
Background

MAE



Background

VQGAN



Method

Motivation

- Prompting in NLP

Je suis désolé	I'm sorry
J'adore la glace	?

Method

Motivation

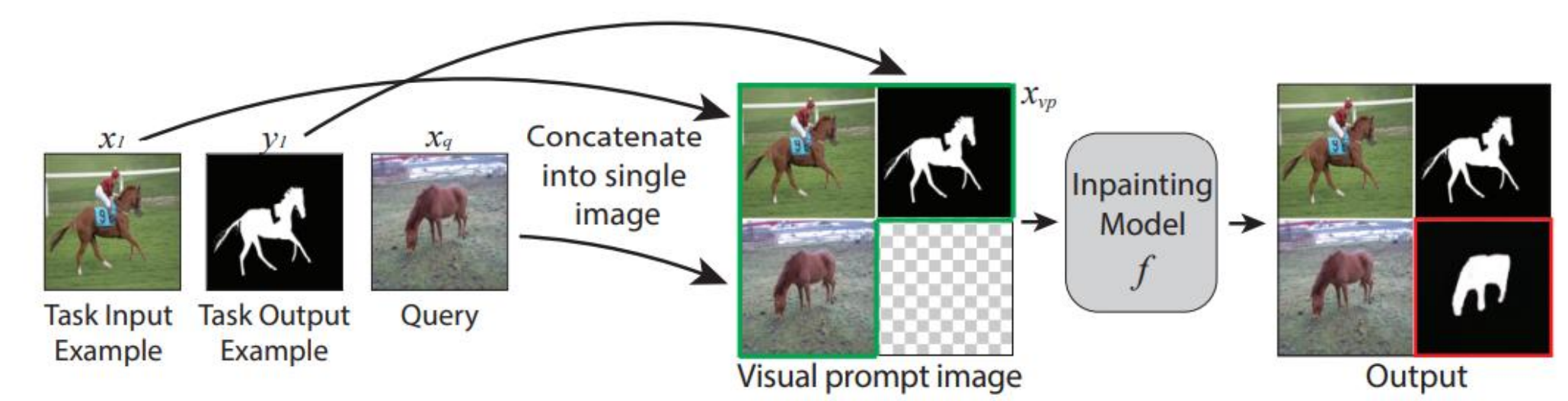
- Prompting in NLP

Je suis désolé	I'm sorry
J'adore la glace	I love ice cream

Method

Overview

- Prompting in NLP→CV



Method

Overview

- Prompting in NLP→CV

Method

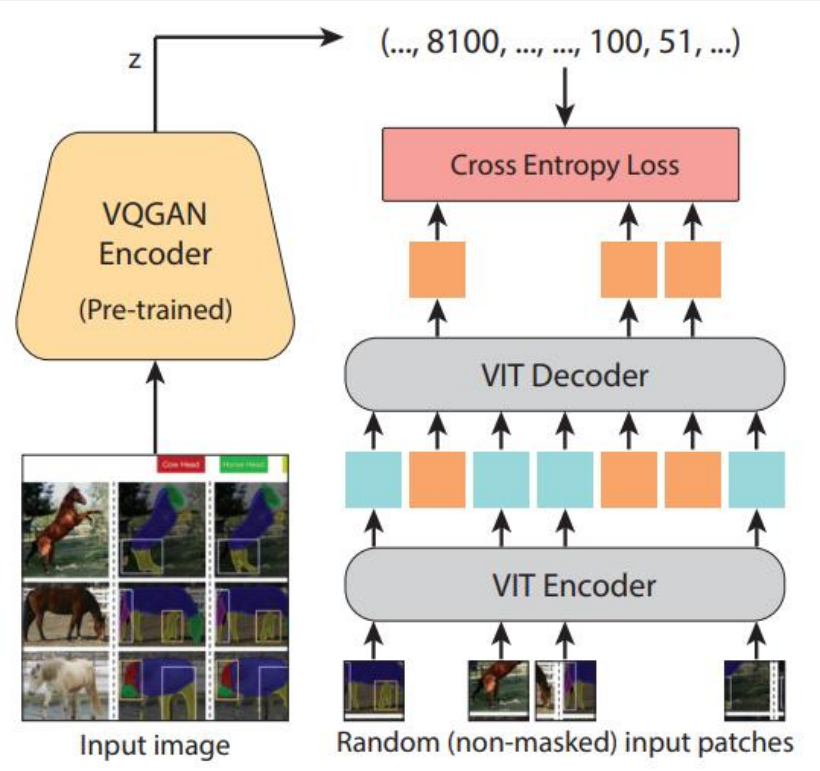
Data

- 88k unlabeled figures from Arxiv [Opensourced]

Method

Network Structure

- Based on MAE-VQGAN



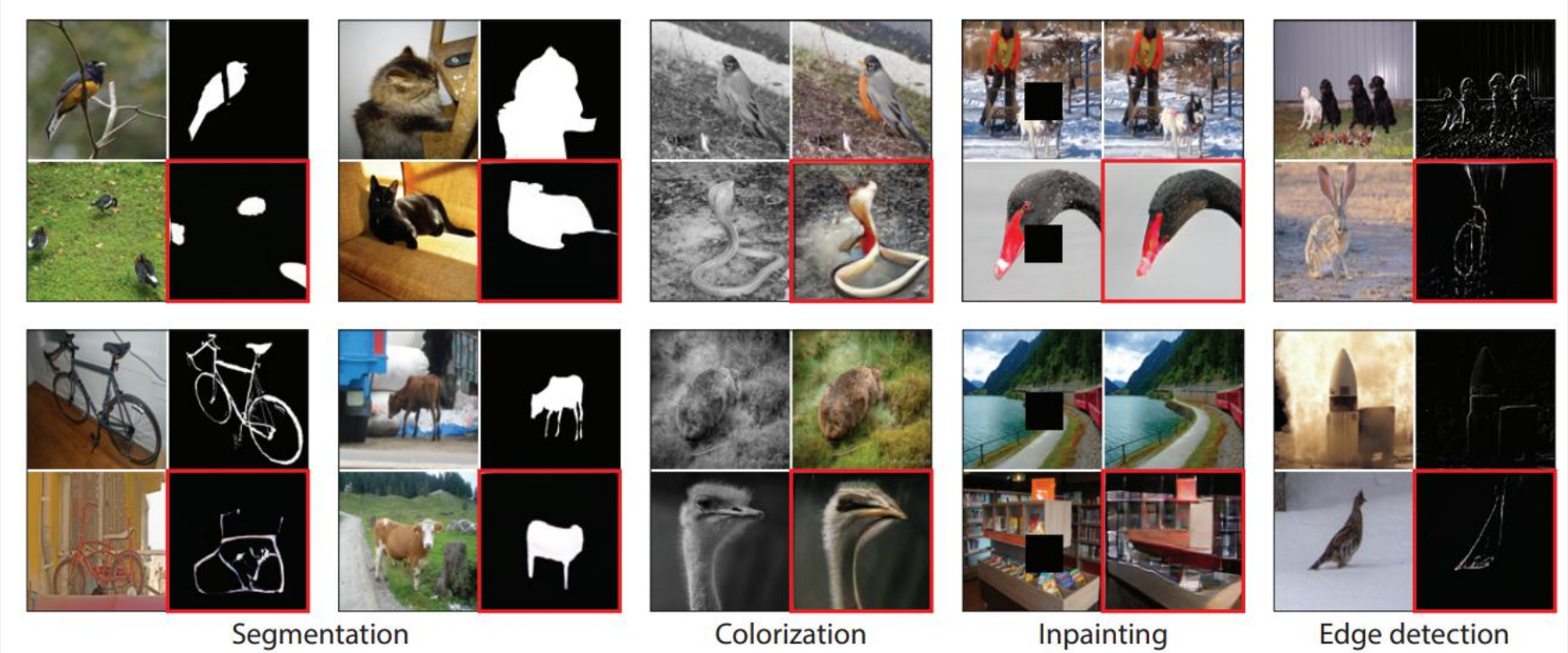
Experiment

CV Tasks

Model	Foreground Segmentation ↑				Single Object Detection ↑				Colorization ↓	
	Split 0	Split 1	Split 2	Split 3	Split 1	Split 2	Split 3	Split 4	MSE	LPIPS
Copy	12.92	17.90	13.52	15.29	12.14	13.50	13.03	12.38	2.63	0.75
BEiT (IN-21k)	0.38	0.93	0.90	0.95	0.24	0.32	0.19	0.10	1.25	0.73
VQGAN (IN-1k)	6.96	10.55	9.59	9.43	5.19	4.99	5.09	5.10	2.44	0.66
MAE (IN-1k)	1.92	6.76	3.85	4.57	1.37	1.98	1.62	1.62	1.13	0.87
MAE-VQGAN (IN-1k)	2.22	7.07	5.48	6.28	3.34	3.21	2.80	2.80	3.31	0.75
BEiT (Figures)	5.38	3.94	3.20	3.29	0.17	0.02	0.14	0.16	0.60	0.70
VQGAN (Figures)	12.56	17.51	14.27	15.06	2.27	2.37	2.48	1.99	1.50	0.56
MAE (Figures)	17.42	25.70	18.64	16.53	5.49	4.98	5.24	5.84	**0.43**	0.55
MAE-VQGAN (Figures)	**27.83**	**30.44**	**26.15**	**24.25**	**24.19**	**25.20**	**25.36**	**25.23**	0.67	**0.40**

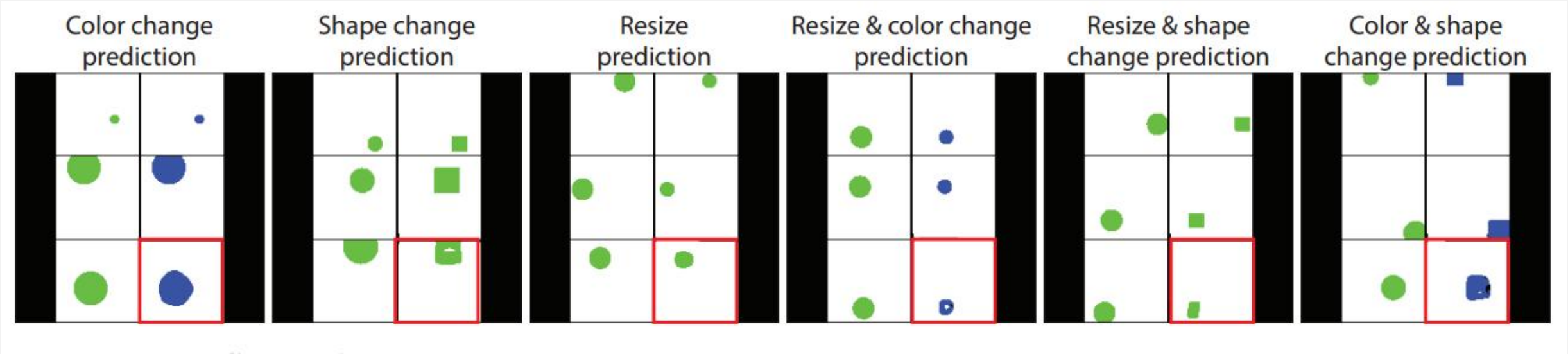
Experiment

CV Tasks



Experiment

Synthetic Data



Experiment

Synthetic Data

	Color	Shape	Size	Color & Shape	Color & Size	Shape & Size
Copy	5.53	6.71	1.17	6.74	1.17	1.86
VQGAN (IN-1k)	0.91	6.51	6.24	2.40	0.70	6.53
BEiT (IN-22k)	15.99	9.08	1.26	7.23	2.84	2.66
MAE (IN-1k)	0.00	2.07	1.20	0.00	0.00	1.56
MAE-VQGAN (IN-1k)	0.13	2.94	3.71	0.00	0.01	3.60
VQGAN (Figures)	6.96	19.11	16.21	7.40	2.24	18.41
BEiT (Figures)	40.92	31.43	7.12	**33.10**	**21.21**	12.98
MAE (Figures)	**70.23**	43.99	34.72	19.30	18.99	**46.02**
MAE-VQGAN (Figures)	40.40	**46.53**	**42.04**	20.41	18.27	40.33

Experiment

Comparing to Finetuning and 1-shot Segmentaion

Pretraining	# Labeled Images	# Shots	Model	Split 0	Split 1	Split 2	Split 3
Unlabeled ImageNet	1	1	Finetune MAE	11.1	13.4	13.0	12.3
	4	4		12.9	15.8	14.3	15.0
	16	16		13.7	16.1	16.8	17.1
Unlabeled Figures	1	1	MAE-VQGAN	32.5	33.8	32.7	27.2
Labeled Pascal 5i (Segmentation masks)	2086 – 5883	1	FWB [36]	51.3	64.5	56.7	52.2
		1	CyCTR [59]	67.2	71.1	57.6	59.0

Experiment

Prompting Engineering

Experiment

Prompting Engineering

	Horizontal	Vertical
Black/White	27.17	31.57
Purple/Yellow	23.44	28.47

Experiment

Prompting Ensembling

Prompt Layout	Color	Shape	Size
Horizontal	39.97	46.54	42.01
+ Vertical	41.31	54.71	46.18
+ Vertical w/ Rows Swap	**44.14**	**60.42**	**49.42**

Experiment

Style/content Extrapolation

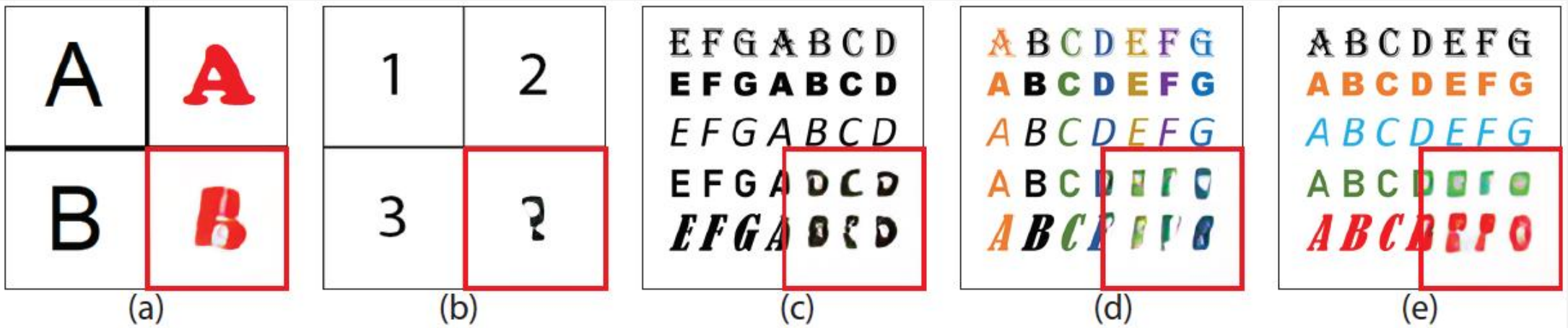


Figure 10: **Style and content extrapolation using MAE-VQGAN**. The model can extrapolate the style of a new content (a), but fails to predict a new content (b). The model struggles to extrapolate new style and content of longer sequences (c-e).

Conclusion

- The dataset is interesting and may be helpful
- Transferring ideas from NLP can benefit CV
- Combining pretraining models (MAE+VQGAN) may be helpful
- Visual prompting can be adapted to many vision tasks
  - especially for some specific scenario