

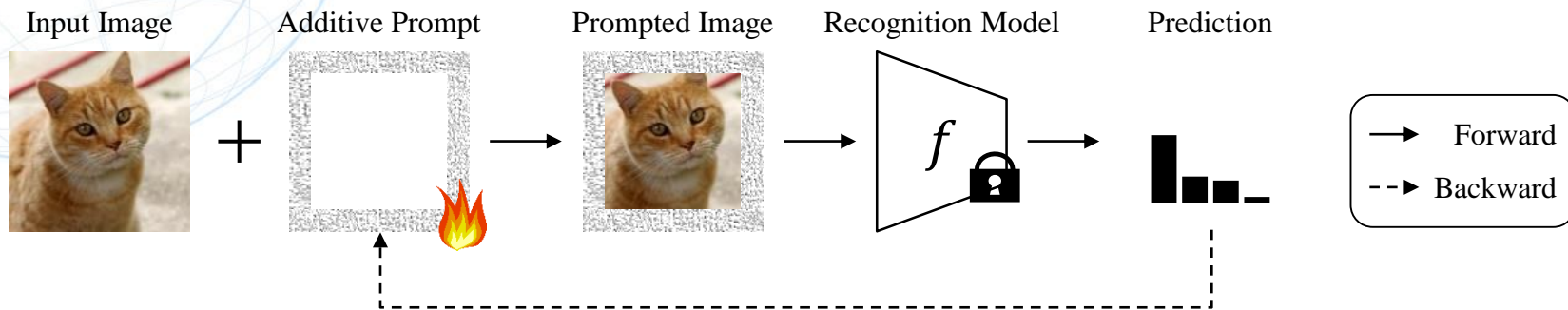
Enhancing Visual Prompting through Expanded Transformation Space and Overfitting Mitigation

NTT

Shohei Enomoto

Visual Prompting (VP)

- Achieving task adaptation without model updates by training parameters of image transformation functions (e.g., noise addition).



- ✓ Parameter-Efficient Fine-Tuning (PEFT).
- ✓ Work on black-box models (e.g., APIs).

The Problems of VP

■ Performance Gap:

- VP often achieves lower accuracy compared to Full Fine-Tuning.



■ Our Core Analysis - Two Critical Limitations:

- Increasing parameters theoretically reduces approximation error but does not improve accuracy.
- **Limited Expressivity:** Simple additive noise has inherent limitations in its expressive power.
- **Overfitting Tendency:** Increasing parameters reduces test accuracy.

Average accuracy on the 12 image classification datasets.

	Full FT	VP
Average Accuracy	87.37	78.39

Padding-type noise vs. Image-sized noise.

Image		
Params	23,280	50,176
Train Acc	97.53 \pm 0.14	97.35 \pm 0.18
Test Acc	93.65 \pm 0.04	89.52 \pm 0.12

No gain

Overfit

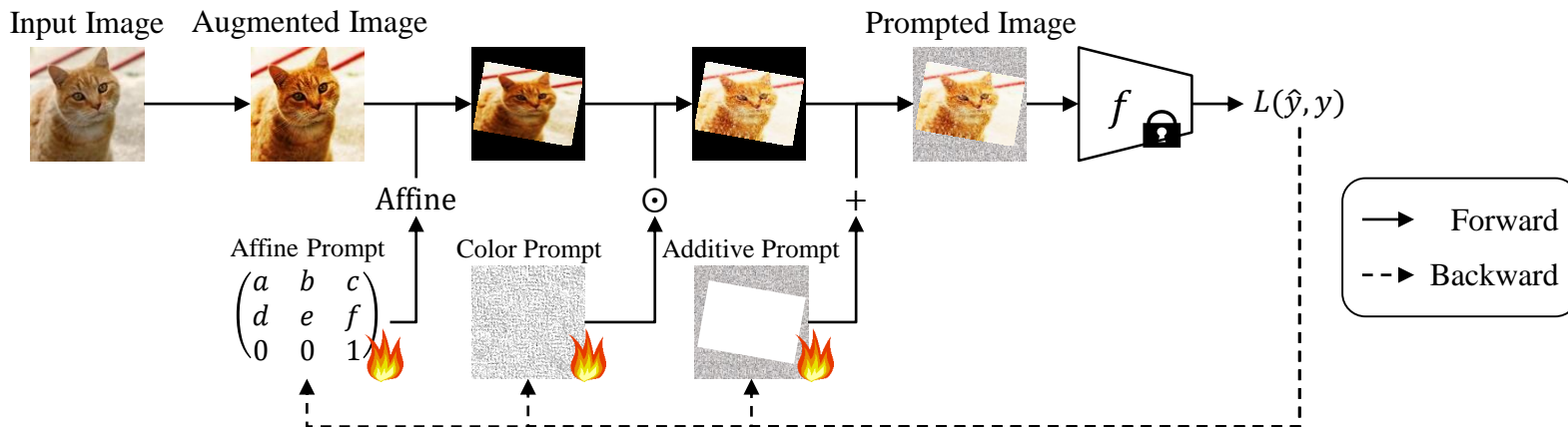
Our Solution: Affine, Color, and Additive VP (ACAVP) **NTT**

1. Expanded Transformation Space

- Using representative and differentiable image transformations.
- Add **affine** and **color** transformations in addition to **additive** transformation.

2. Overfitting Mitigation

- Experimenting with various overfitting mitigation techniques.
- Data augmentation, especially **TrivialAugment**, is effective.



Key Results

■ SOTA Accuracy

- Achieved the highest average accuracy across 12 datasets, outperforming all VP baselines and surpassing Linear Probing (LP).

Method	ZS	VP	EVP	AutoVP	ACAVP	LP	FT
Average	53.12	78.39	78.86	<u>81.54</u>	83.18	82.59	87.37

■ Superior Robustness

- Significantly outperforms baselines on corrupted datasets.

Method	CIFAR10-C	CIFAR100-C
ZS	70.9	42.59
VP	76.65 \pm 0.04	50.52 \pm 0.12
EVP	80.97 \pm 0.14	55.17 \pm 0.16
AutoVP	79.72 \pm 0.28	56.34 \pm 0.15
ACAVP	83.98 \pm 0.20	58.68 \pm 0.29

■ Excellent Transferability

- When trained on InstructBLIP, ACAVP achieves the highest accuracy on the unseen BLIP2 model.

Model	InstructBLIP		BLIP2	
Method	CIFAR10	CIFAR100	CIFAR10	CIFAR100
ZS	88.11	82.41	58.41	60.65
EVP	98.40	85.05	83.20	58.39
TVP	98.78	83.10	85.15	62.80
ACAVP	98.85	88.72	85.32	66.67

Conclusion

- We propose **ACAVP**, expanding the VP transformation space with **Affine** and **Color** transformations.
- Identified **overfitting** as a critical problem in VP and showed **TrivialAugment** is a universal and effective solution.
- ACAVP achieves **SOTA accuracy**, superior **robustness** and **transferability**.