FLiP: Towards Comprehensive and Reliable Evaluation of Federated Prompt Learning

Dongping Liao[1], Xitong Gao[2,3], Chengzhong Xu[1]

[1]University of Macau, [2]Shenzhen Institutes of Advanced Technology,
[3]Shenzhen University of Advanced Technology

Background

Federated Learning

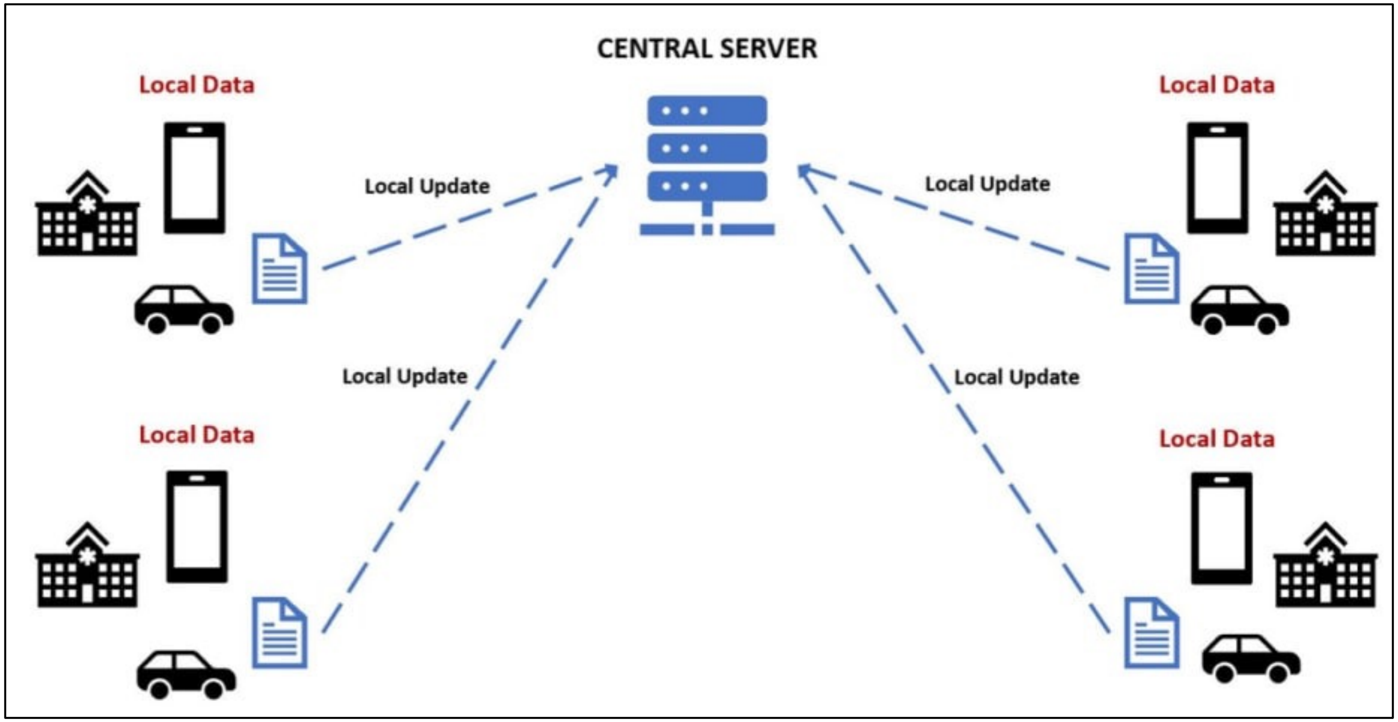


Figure: A conceptual illustration of a federated learning system.

Background

Limitations of Current Federated Prompt Learning Evaluation Approaches

- A **systematic understanding** of their underlying mechanisms and **principled guidelines** for deploying these techniques in different FL scenarios remain absent.
- Existing **inconsistent experimental protocols, limited evaluation scenarios**, and the lack of the proper assessment of **centralized PL methods** have obscured the essence of these algorithms.

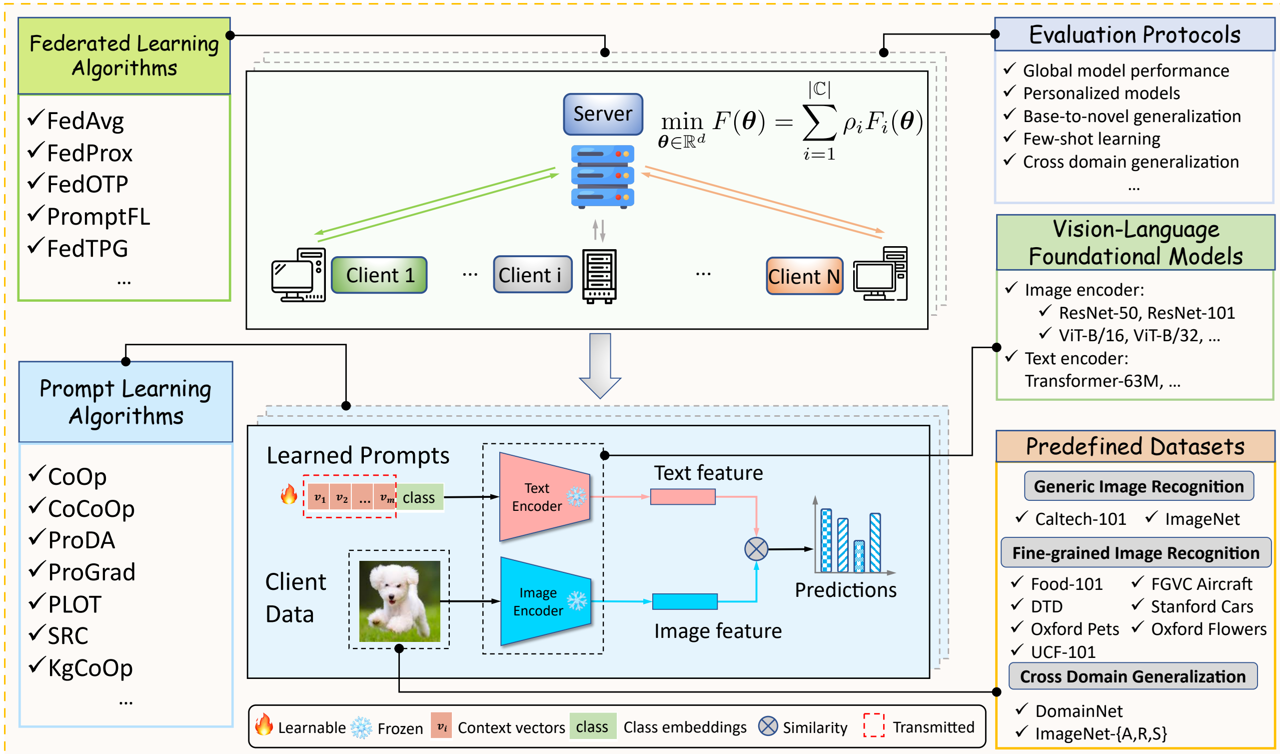
To close above gaps, we introduce a **comprehensive benchmark**, named FLiP, to achieve standardized FPL evaluation.

Research Questions

- **Q1**: How effective are **global and personalized prompt models** obtained by FPL methods?
- **Q2**: How do **various real-world data distribution shifts** impact FPL?
- **Q3**: What are **best practices** for deploying these techniques in different FL scenarios?
- **Q4**: What can we learn about **cost-effectiveness** from different FPL methods?
- **Q5**: What are the **remaining challenges in FPL?**

Benchmark Design

The System Design of the FLiP Benchmark



Core Features

- Decoupled FL and PL modules
- Standardized data interfaces
- Various VL foundational models
- Extensible open-source codebase

Benchmark Statistics

- 3 evaluation protocols
- 6 FPL scenarios
- 12 open datasets
- 13 PL and FPL methods

Benchmark Results

Global Model Performance

Global α_g	Caltech	DTD	Aircraft	Food	Cars	Flowers	Pets	UCF	Avg.	#
ZS-CLIP	86.0	41.7	16.6	77.9	55.5	65.3	85.7	61.5	61.3	-
PromptFL	91.5±0.5	57.6±1.3	22.8±0.4	79.2±0.1	62.0±0.4	**84.0**±1.7	**89.4**±0.5	70.1±0.8	69.6	-
FedOTP	91.8±0.1	58.0±0.8	21.9±0.4	78.7±0.1	62.8±0.2	83.3±0.6	89.1±0.1	69.4±0.6	69.4	3
FedTPG	90.2±0.1	56.8±1.0	19.0±1.2	79.3±0.2	60.7±0.2	78.0±1.6	89.0±0.6	68.3±0.2	67.6	1
***f*-CoCoOp**	91.7±0.3	54.7±1.0	17.9±4.5	79.3±0.1	60.7±0.5	76.4±0.7	89.1±0.1	68.0±0.9	67.2	2
***f*-PLOT**	91.6±0.3	**58.3**±1.4	21.7±0.5	78.3±0.2	60.7±0.8	83.4±0.6	88.9±0.5	69.7±0.4	69.1	2
***f*-ProDA**	91.6±0.3	57.2±1.1	**23.1**±0.7	79.1±0.2	62.3±0.5	**84.0**±0.8	89.3±0.4	70.2±1.1	69.6	**5**
***f*-ProGrad**	90.7±0.2	57.1±1.0	21.7±0.3	**79.5**±0.1	60.5±0.7	83.4±0.4	89.1±0.2	70.3±0.3	69.1	2
***f*-PromptSRC**	**92.0**±0.8	57.8±0.3	21.2±0.4	78.6±0.4	62.4±0.2	83.6±0.1	89.2±0.7	70.3±1.0	69.4	4
***f*-KgCoOp**	91.8±0.2	58.2±0.8	23.0±0.1	79.4±0.2	61.7±0.7	83.9±0.5	**89.4**±0.2	**70.4**±0.7	**69.7**	**5**

PromptFL, combining CoOp and FedAvg, serves as a simple yet effective baseline and is competitive on fine-grained datasets like Oxford-Pets and Flowers.

Benchmark Results

Personalized Model Performance

Personal α_p	Caltech	DTD	Aircraft	Food	Cars	Flowers	Pets	UCF	Avg.	#
ZS-CLIP	86.0	41.7	16.6	77.9	55.5	65.3	85.7	61.5	61.3	-
PromptFL	91.5±0.4	69.5±4.1	33.8±0.1	**82.1**±0.4	67.7±0.6	**89.7**±0.2	**89.9**±0.6	77.5±1.5	75.2	0
FedOTP	**91.9**±0.4	**73.8**±1.4	**36.1**±0.5	82.0±0.7	**68.1**±1.7	89.6±0.3	89.5±1.2	**80.7**±1.0	**76.5**	5
FedTPG	89.9±0.4	66.8±0.6	31.6±0.3	81.9±0.3	66.2±0.2	86.0±0.4	88.9±0.8	78.0±0.7	73.7	1
FedPGP	91.8±0.4	68.0±0.5	35.2±0.4	81.0±0.2	66.7±0.6	84.1±1.6	87.4±0.3	77.8±0.5	74.0	3
PromptFolio	91.6±0.3	70.2±0.4	34.6±0.5	82.0±0.3	67.4±0.2	89.2±0.5	89.4±0.2	79.2±0.6	75.5	4
DP-FPL	90.2±0.4	65.2±0.5	28.5±0.5	80.1±0.6	64.5±1.4	78.6±0.6	82.4±0.8	72.2±0.7	70.2	0
***f*-CoCoOp**	91.8±0.4	70.3±3.0	34.0±1.7	81.8±0.6	67.4±0.7	86.4±1.9	89.5±0.9	77.1±0.9	74.8	3
***f*-PLOT**	91.7±0.4	71.3±3.1	34.0±0.8	81.4±1.2	67.9±1.1	89.3±0.5	88.6±0.3	79.6±1.8	75.5	5
***f*-ProDA**	91.7±0.9	69.7±2.6	34.7±0.6	**82.1**±1.3	67.8±1.3	89.3±0.6	**89.9**±0.5	77.9±1.5	75.4	**6**
***f*-ProGrad**	91.7±0.6	69.2±0.3	32.9±0.7	81.6±0.8	67.0±1.0	88.8±0.9	89.5±0.7	77.0±1.5	74.7	1
***f*-PromptSRC**	91.7±0.4	69.3±1.3	32.4±2.3	81.8±1.1	67.6±1.1	89.2±1.7	89.3±1.7	78.2±1.1	74.9	2
***f*-KgCoOp**	91.6±0.3	68.6±2.7	31.3±0.5	81.4±0.7	67.1±1.4	88.9±0.9	**89.9**±0.3	76.9±0.9	74.5	1

FedOTP generally outperforms other personalized methods, highlighting the efficacy of distribution alignment in adapting to personalized data.

Benchmark Results

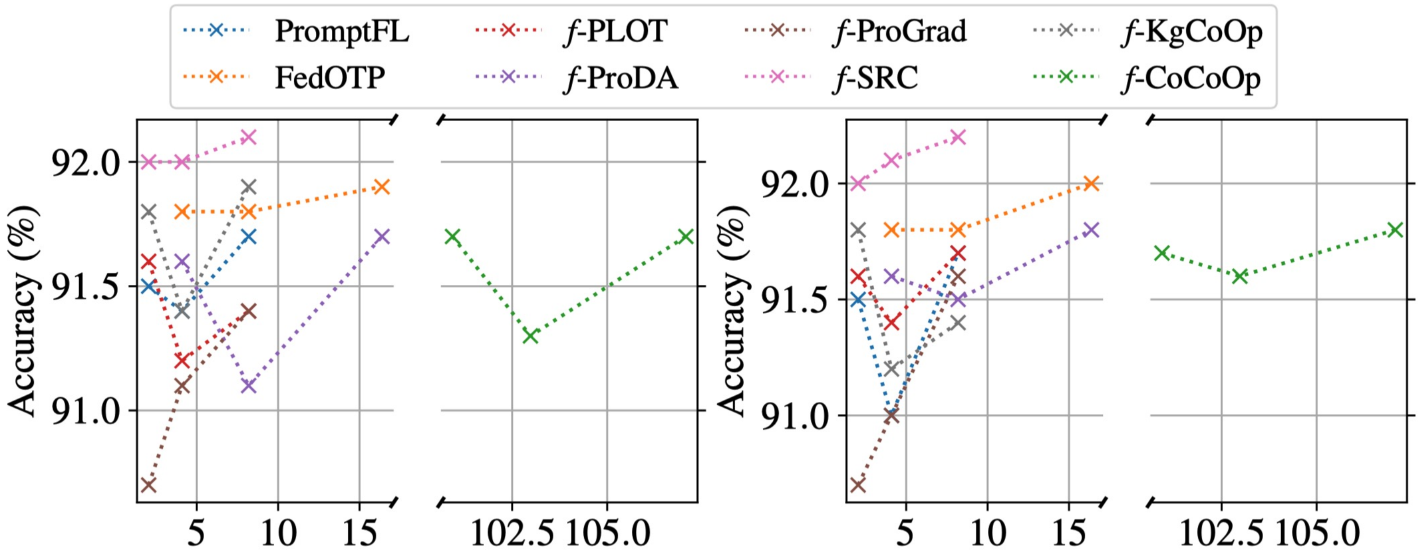
Base-to-Novel Class Generalization

	Caltech			Aircraft			Cars			Flowers			Avg.			
Metric	α_b	α_n	α_h	α_b	α_n	α_h	α_b	α_n	α_h	α_b	α_n	α_h	α_b	α_n	α_h	#
ZS-CLIP	88.2	92.6	90.3	19.6	24.7	21.8	59.5	68.1	63.5	77.2	71.0	73.9	61.1	64.1	62.4	-
PromptFL	92.8±0.8	92.7±0.7	92.6±0.2	20.9±0.4	24.7±0.5	22.6±0.2	63.0±0.7	67.6±0.7	65.2±0.1	79.9±2.9	69.3±0.7	74.2±1.0	64.1	63.8	63.8	-
FedOTP	93.1±0.2	93.7±0.4	<u>93.4</u>±0.1	21.0±0.8	23.7±1.0	22.2±0.8	62.1±0.1	66.0±0.7	64.0±0.4	81.0±0.6	68.7±1.9	74.3±1.2	64.3	63.0	63.5	2
FedTPG	93.6±0.4	90.0±0.6	91.8±0.4	19.5±0.3	22.7±0.5	21.0±0.5	66.4±0.2	67.7±0.2	**67.0**±0.2	75.0±0.3	66.0±0.5	70.2±0.4	63.6	61.6	62.5	2
FedPGP	93.2±0.4	92.7±1.1	92.9±0.5	19.8±1.0	19.2±2.1	19.5±0.8	63.7±1.4	67.8±0.7	65.7±0.9	80.7±1.4	66.8±1.7	73.1±1.2	64.3	61.6	62.8	2
***f*-CoCoOp**	92.7±0.8	93.6±0.6	93.1±0.2	18.0±2.1	17.1±2.4	17.2±2.2	62.8±0.6	66.7±0.3	64.7±0.2	79.4±1.4	70.7±1.8	<u>74.8</u>±0.6	63.2	62.0	62.4	2
***f*-PLOT**	93.4±0.6	93.5±1.0	**93.5**±0.8	19.0±0.8	23.4±0.4	21.0±0.4	62.4±0.5	65.4±1.4	63.8±0.7	78.7±3.0	68.3±1.3	73.1±0.6	63.4	62.6	62.8	1
***f*-ProDA**	93.0±0.4	92.7±1.1	92.8±0.4	22.0±0.5	25.1±1.1	<u>23.4</u>±0.6	63.3±0.7	67.7±0.4	65.4±0.2	77.9±0.7	69.6±0.5	73.5±0.1	64.0	63.8	63.8	<u>3</u>
***f*-ProGrad**	93.2±0.4	93.0±0.4	93.1±0.4	21.6±0.5	25.2±1.6	23.3±0.4	64.2±0.6	67.9±0.5	<u>66.0</u>±0.3	80.3±1.1	70.6±0.5	**75.2**±0.2	64.7	64.2	**64.3**	**4**
***f*-SRC**	90.2±0.2	93.0±0.1	91.6±0.1	21.7±1.0	24.9±1.0	23.1±0.8	61.2±0.2	67.1±0.3	64.0±0.1	78.9±1.1	70.2±0.8	74.3±0.7	62.3	63.8	62.8	<u>3</u>
***f*-KgCoOp**	93.5±0.4	93.4±1.0	<u>93.4</u>±0.5	22.3±0.7	25.1±0.5	**23.6**±0.5	63.7±0.3	67.5±0.3	65.6±0.1	79.6±2.7	68.9±1.2	73.9±0.5	64.8	63.7	<u>64.1</u>	<u>3</u>

Regularization prevents overfitting and balances base and novel class metrics, without it (PromptFL) is less effective.

Benchmark Results

Cost-Performance Tradeoff



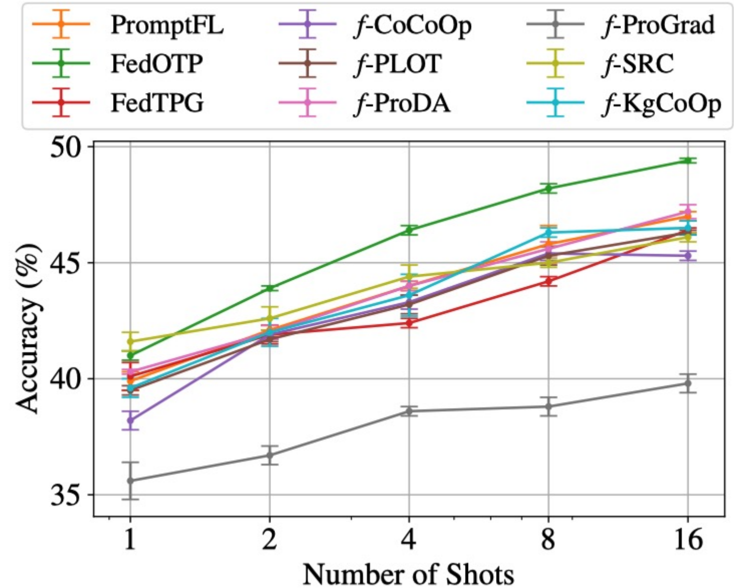
(a) Number of prompts: $\{1, 2, 4\}$.   (b) Number of tokens: $\{4, 8, 16\}$.

Extra learnable parameters (CoCoOp) do not necessarily improve performance but could be detrimental to computational and communication efficiency.

Benchmark Results

Few-shot Accuracy

$k=$	1	2	4	8	16	Avg.	#
PromptFL	39.9±0.4	42.1±0.2	44.0±0.2	45.8±0.8	47.0±0.2	43.7	-
FedOTP	41.0±0.2	**43.9**±0.1	**46.4**±0.2	**48.2**±0.2	**49.4**±0.1	**45.8**	**5**
FedTPG	40.1±0.6	41.9±0.4	42.4±0.2	44.2±0.4	46.4±0.2	41.8	1
***f*-CoCoOp**	38.2±0.4	41.9±0.2	43.3±0.3	45.4±0.3	45.3±0.2	42.8	0
***f*-PLOT**	39.5±0.2	41.7±0.1	43.2±0.4	45.3±0.4	46.3±0.1	43.2	0
***f*-ProDA**	40.3±0.1	42.0±0.3	44.0±0.2	45.6±0.3	47.2±0.3	43.8	2
***f*-ProGrad**	35.6±0.8	36.7±0.4	38.6±0.2	38.8±0.4	39.8±0.4	37.9	0
***f*-SRC**	**41.6**±0.4	42.6±0.5	44.4±0.5	45.0±0.2	46.1±0.2	43.9	3
***f*-KgCoOp**	39.6±0.4	42.0±0.6	43.6±0.9	46.3±0.2	46.5±0.3	43.6	1



Methods that use multiple prompts (e.g., FedOTP, f-ProDA, f-SRC) perform best in few-shot scenarios, indicating ensemble helps reduce sample selection bias.

Benchmark Results

Feature Shift Heterogeneity

Feature Shift	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Avg.	#
ZS-CLIP	54.8	40.9	48.8	6.0	77.7	49.3	44.6	-
PromptFL	59.6±0.2	45.6±0.3	53.7±0.3	8.9±0.1	79.8±0.1	**54.2**±0.2	48.1±0.1	-
FedOTP	58.4±0.2	45.2±0.1	53.4±0.2	9.0±0.1	79.2±0.1	53.2±0.1	47.7±0.1	1
FedTPG	59.8±0.1	45.8±0.3	53.6±0.2	8.5±0.2	79.9±0.3	54.0±0.3	47.9±0.3	3
***f*-CoCoOp**	**60.0**±0.1	**46.1**±0.2	53.0±0.3	**9.1**±0.2	79.8±0.2	**54.2**±0.2	48.1±0.2	**5**
***f*-PLOT**	58.5±0.3	44.8±0.2	53.0±0.1	9.0±0.4	79.2±0.1	53.3±0.1	47.6±0.1	1
***f*-ProDA**	59.5±0.2	45.6±0.1	**53.8**±0.2	9.0±0.2	79.6±0.1	54.0±0.3	48.0±0.1	3
***f*-ProGrad**	58.8±0.2	44.5±0.2	52.5±0.1	7.5±0.2	80.0±0.1	53.0±0.1	47.3±0.1	1
***f*-SRC**	59.0±0.1	44.6±0.4	52.6±0.1	7.8±0.1	79.7±0.1	52.9±0.1	47.3±0.1	0
***f*-KgCoOp**	59.9±0.1	45.9±0.2	53.6±0.1	8.8±0.1	**80.2**±0.1	54.1±0.1	**48.2**±0.1	4

It remains challenging to counteract the adverse impact of feature shift for all evaluated FPL methods.

Benchmark Results

Cross-domain Generalization

Table 8: Comparing the cross-domain performance of FPL methods. Here, the source domain is ImageNet (IN), and the target domains are ImageNet-A(dversarial), -R(endition), and -S(ketch), respectively denoted as IN-A, IN-R, IN-S.

Cross-domain $\alpha_{\text{IN}\rightarrow\cdot}$	**IN-A**	**IN-R**	**IN-S**	Avg.	#
ZS-CLIP	21.7	56.1	33.4	37.1	-
PromptFL	**24.9**±0.4	58.2±0.3	35.6±0.6	39.6	-
FedOTP	23.8±0.3	58.3±0.8	35.2±0.4	39.1	1
FedTPG	24.5±0.4	58.6±0.6	35.7±0.3	39.6	**2**
***f*-CoCoOp**	24.0±0.6	**59.8**±1.1	**36.0**±1.0	**39.9**	**2**
***f*-PLOT**	23.8±0.7	57.5±0.3	34.6±0.6	38.6	0
***f*-ProDA**	24.7±1.7	58.3±0.6	35.6±0.8	39.5	1
***f*-ProGrad**	23.5±1.1	58.3±0.5	35.5±1.2	39.1	1
***f*-SRC**	24.0±0.7	58.7±0.9	35.1±0.5	39.3	1
***f*-KgCoOp**	24.7±1.2	58.7±1.4	35.9±0.7	39.8	**2**

Test-time image feature injection and self-regularization contribute to improving the robustness against cross domain shift.

Project & Code

Thanks