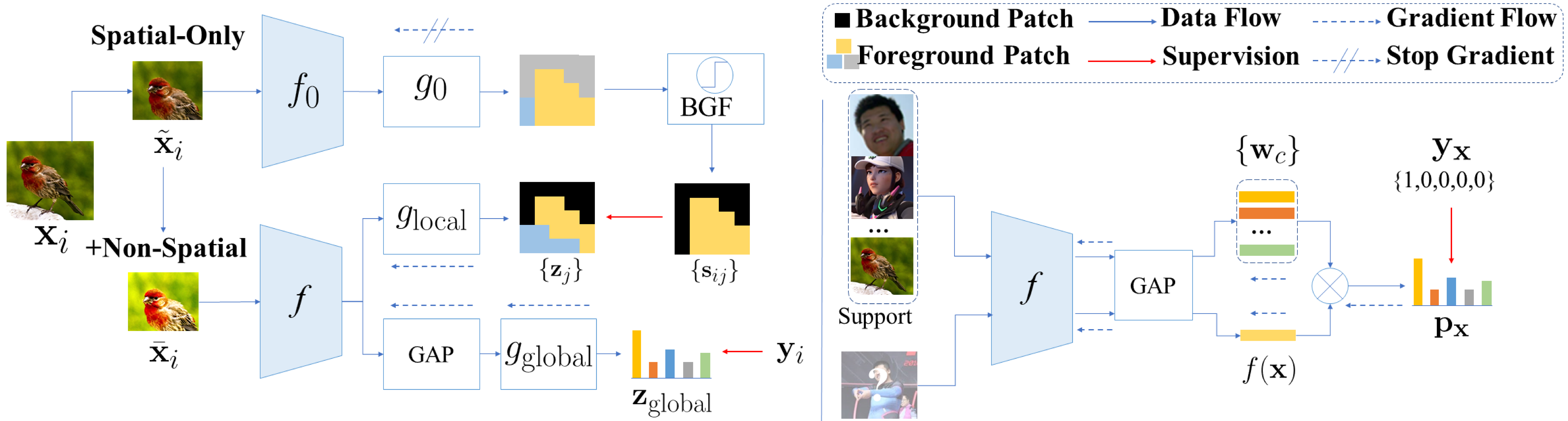
Self-Promoted Supervision for Few-Shot Transformer

Bowen Dong[1], Pan Zhou[2], Shuicheng Yan[2], Wangmeng Zuo[1,3]

[1] Harbin Institute of Technology [2] National University of Singapore [3] Peng Cheng Laboratory



(a) Phase 1: Meta-Training

(b) Phase 2: Meta-Tuning

ViT Type	1-shot w/o SUN	1-shot w/ SUN-M
LT-ViT	43.08±0.38	**59.00±0.44**
Visformer	47.61±0.43	**67.80±0.45**
Swin	54.63±0.45	**64.94±0.46**
NesT	54.57±0.46	**66.54±0.45**

Method	5-way 1-shot	5-way 5-shot
SUN − M	**67.80±0.45**	**83.25±0.30**
SUN − F	**66.90±0.44**	**82.63±0.30**
SUN − D	**69.56±0.44**	**85.38±0.49**

Motivation:

- Our work starts from two questions:
  - Whether ViTs can perform well under few-shot learning setting or not?
  - If not, how to improve the few-shot learning ability of ViTs?

Why Investigating ViTs:

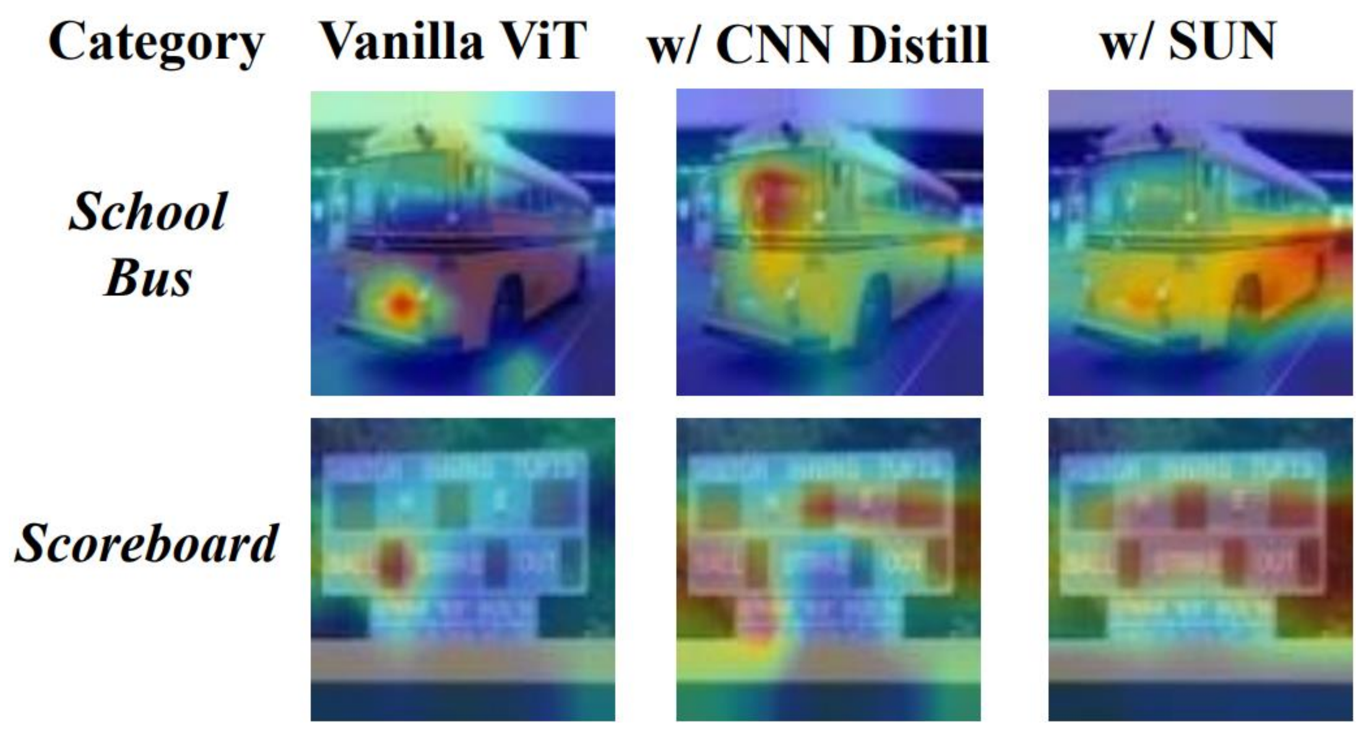
- ViT has three advantages over CNN:
  - **Better performance** than CNN.
  - **Unify** vision and language models.
  - **More highly parallelized** than CNN.
- SUN has proved that ViTs can perform well on few-shot learning scenarios.

Preliminary Study:

- We use Meta-Baseline to evaluate ViTs and ResNet-12, and find that **ViTs perform much worse than CNNs**.
- We reveal that ViTs face **severe overfitting on base classes**, but obtain **worse generalization on novel classes**.

Analysis:

- **The lack of inductive bias leads to the severe performance degeneration.**
  - CNN (alike) inductive bias or local attention benefits to generalization ability
- **Low-quality token dependency learning results in overfitting on base classes.**



Self-promoted sUpervisioN (SUN):

- **Meta-Training:** to learn a meta-learner f with **location-specific supervision**, such that f is able to fast adapt itself to novel classes with a few training data.
  - First optimize a teacher ViT f_g and use f_g to obtain **location-specific** supervision.
  - Optimize meta-learner f via both ground-truth labels and location-specific supervision.
  - Augmented Training via **Spatial-Consistent Augmentation** and **Background Filtration**.
- **Meta-Tuning:** we fine-tune the meta-learner f via training it on multiple "*N-way K-shot*" tasks sampled from base set.
- **Tips:** CNN-based patch embedding, longer training epochs and relative large drop-path rate (*e.g.*, 0.5) benefits to ViTs for SUN training framework.

Experiments:

- We evaluate our SUN on different ViTs, and obtain promising **performance improvement on all ViT feature extractors**.
- We use the same meta-training phase and **adopt various few-shot learning methods** as meta-tuning phase, all achieve good classification accuracy (SUN-D is the best)
- SUN achieves comparable accuracy **against state-of-the-art CNN-based few-shot learning methods** on multiple benchmarks.

Conclusion:

- ViT faces severe accuracy drop on few-shot learning, and reveal that **slow token dependency learning and limited training data** lead to the performance degeneration.
- We propose **self-promoted supervision (SUN)** to generate individual location-specific for few-shot learning.
- The first work to empirically analyze ViTs for few-shot learning, and provides **a simple yet solid baseline** for few-shot classification.