Cornell University

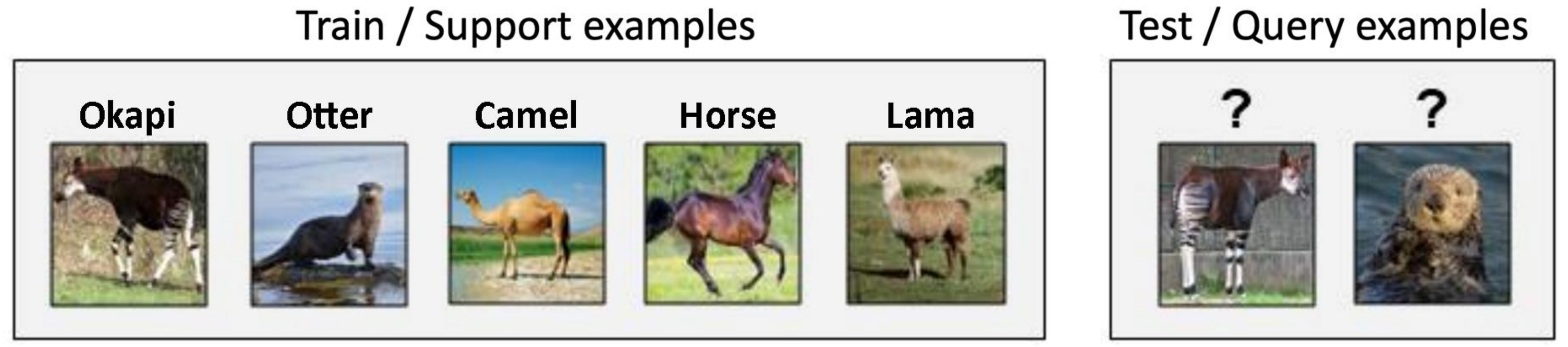
Few-shot Classification with Feature Map Reconstruction Networks

Davis Wertheimer*, Luming Tang*, Bharath Hariharan (*equal contribution)

CVPR VIRTUAL JUNE 19-25

Few-Shot Classification:

Generalize to novel classes using *very few* images per novel class



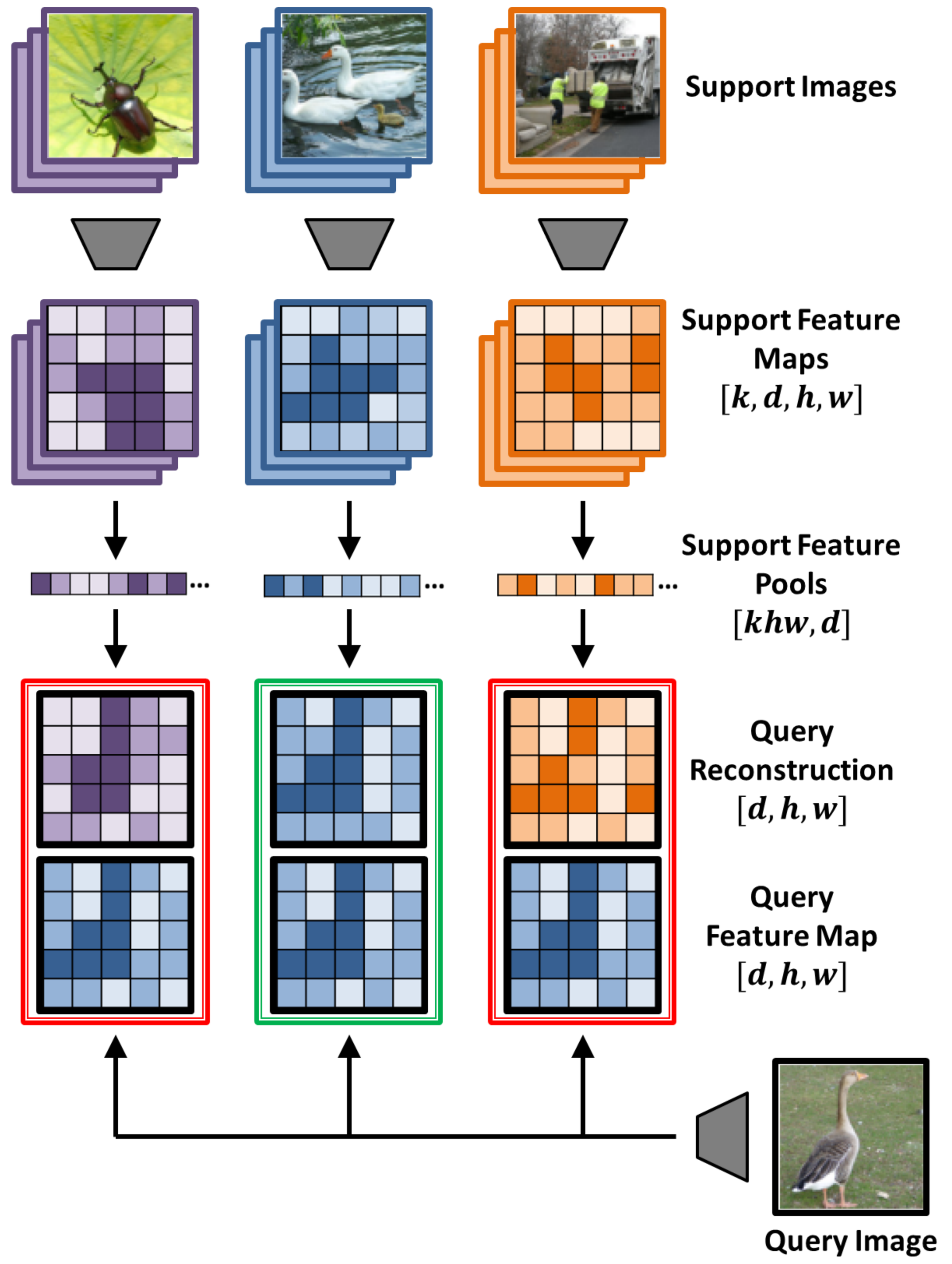
Abstract:

Reformulate few-shot classification as latent space reconstruction

- Approximate query features as weighted sums of support features; assign class based on reconstruction quality
- Simple, closed-formed ridge regression is more performant and efficient than previous approaches
- Across-the-board superiority on four fine-grained datasets and the cross-domain setup
- Competitive results on mini-ImageNet and tiered-ImageNet with minimal bells and whistles

Our Method:

1. Extract feature maps from support images
2. Concatenate into class feature pools
3. Reconstruct query feature map as a weighted sum over each pool
4. Same-class images with semantic overlap should provide the best reconstruction
5. Predict based on reconstruction quality



Calculating Reconstructions:

Given k support features for class y and b query features for image x in d dimensions, i.e.,

$$S \in \mathbb{R}^{k \times d} \qquad Q \in \mathbb{R}^{b \times d} \qquad W \in \mathbb{R}^{b \times k}$$

We wish to find $\bar{W}$ such that:

$$\bar{W}S \approx Q$$

Solve with ridge regression and learned regularization:

$$\bar{W} = \underset{W}{\operatorname{argmin}} \|WS - Q\|^2 + \lambda \|W\|^2$$

The closed-form solution is well-known:

$$\bar{W} = Q(S^T S + \lambda I)^{-1} S^T$$

Prediction probability is proportional to the reconstruction quality:

$$p(y|x) \propto \exp(-\|\bar{W}S - Q\|^2)$$

Comparison to Related Prior Work:

Prior work also compares feature maps via weighted-sum reconstruction, but there are problems:

- **CrossTransformer (CTX)** refines category predictions via attention:
  - Constrained: positive-valued weight matrix, rows sum to 1
  - Reprojection involves many additional parameters
  - Time/memory quadratic in the number of features
- **DeepEMD** compares feature pools via transport cost:
  - Constrained: positive-valued weight matrix, rows *and columns* sum to 1
  - Transport cost requires *very* slow iterative solvers, i.e., *image pairs only*

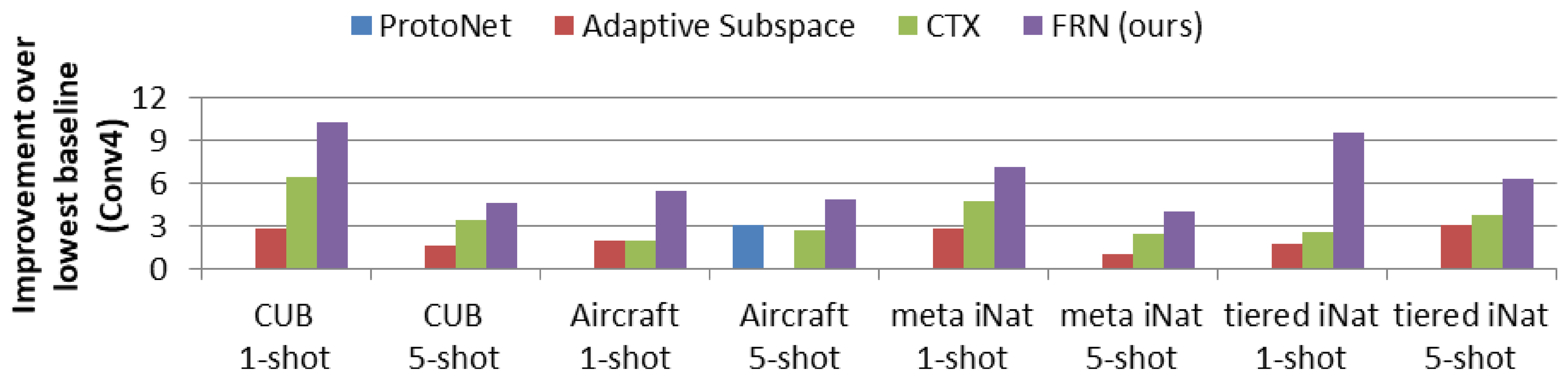
FRN avoids these issues, being expressive, lightweight, and scalable

Model	No hard constraints on reconstruction	Very few new learned parameters	Scales well with number of features
CTX	✗	✗	✗
DeepEMD	✗	✓	✗
FRN (ours)	✓	✓	✓

Results:

Superior fine-grained performance, competitive results on ImageNet derivatives with minimal bells and whistles

- Consistent improvements over reimplemented baselines in a range of fine-grained settings:



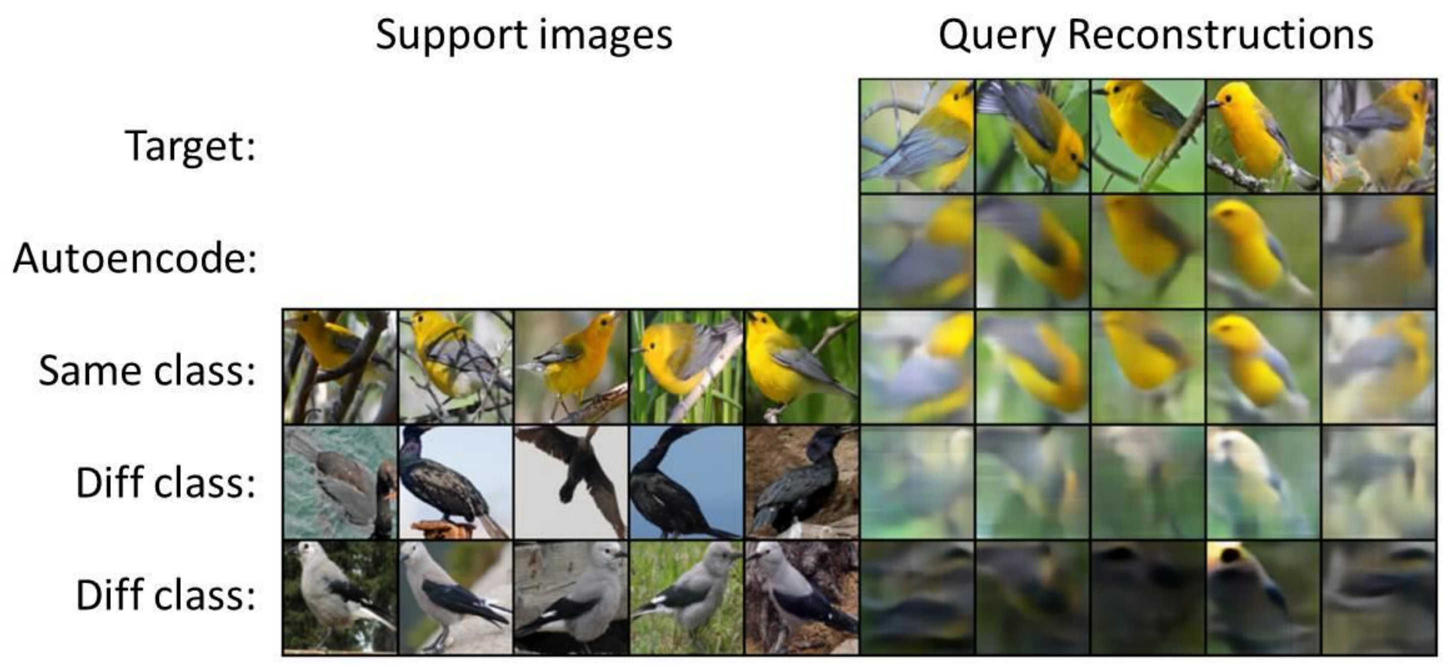
- Large improvements over prior work in fine-grained settings, competitive results in general settings:

Benchmark	Model	Backbone	1-shot	5-shot
CUB	DeepEMD	ResNet-12	75.65±0.83	88.69±0.50
	FRN (ours)	ResNet-12	**83.16±0.19**	**92.59±0.10**
mini-ImageNet →CUB	Centroid Alignment	ResNet-18	46.85±0.75	70.37±1.02
	FRN (ours)	ResNet-12	**53.39±0.21**	**75.16±0.17**
mini-ImageNet	FEAT	ResNet-12	**66.78±0.20**	82.05±0.14
	FRN (ours)	ResNet-12	66.45±0.19	**82.83±0.13**
tiered-ImageNet	DeepEMD	ResNet-12	71.16±0.87	86.03±0.58
	FRN (ours)	ResNet-12	**72.06±0.22**	**86.89±0.14**

- *Much* faster than DeepEMD:

Model	1-shot Latency (ms)	5-shot Latency (ms)
DeepEMD	178	18,292
FRN (ours)	**63**	**79**

- Image decoders verify that same-class reconstructions really are better!



Code is available here: https://github.com/Tsingularity/FRN