Meta-Learning of Structured Representation by Proximal Mapping

Mao Li, Yingyi Ma, Xinhua Zhang
University of Illinois at Chicago

Motivation

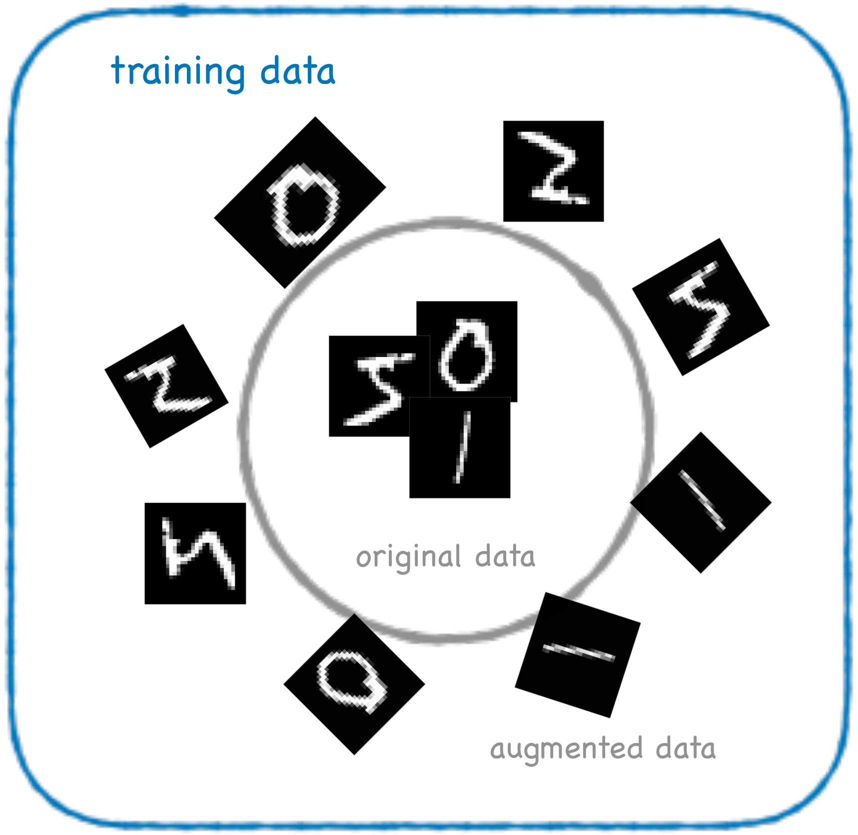
Goal of meta-learning: Extract **prior structures** from a set of **tasks** that allows efficient learning of **new tasks**.

Examples of structural regularities:

- Instance level
  - Input layers: transformation beyond group-based diffeomorphism
  - Within layers: sparsity, disentanglement, spatial invariance, structured gradient accounting for data covariance, manifold smoothness
  - Between layers: equvariance, contractivity, robustness under dropout and adversarial perturbations of preceding nodes
- Batch/Dataset level
  - multi-view, multi-modality, multi-domain
  - diversity, fairness, privacy, causal structure

Existing Approaches

- Data Augmentation

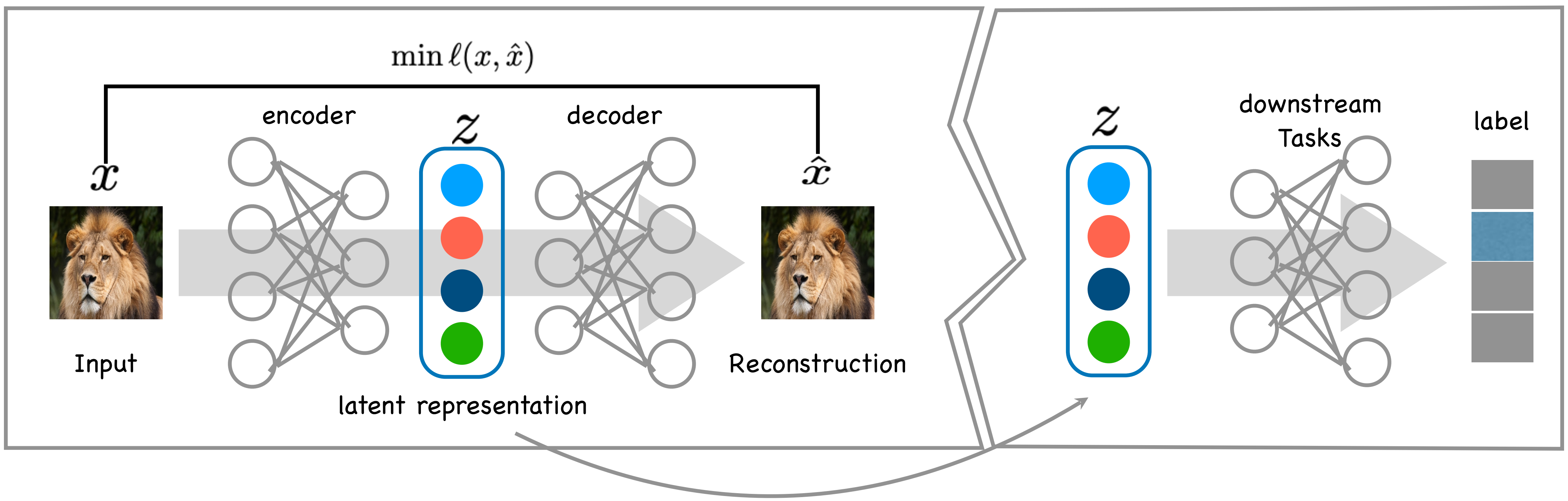


√ boost prediction performance

× unclear the improvement is due to the learned representation or due to a better classifier.

Existing Approaches

- Auto-encoder



√ learned the most salient features

× usually used as an initialization for subsequent supervised task

× not amendable to end-to-end learning

Our goal: learn representations that explicitly encode structural priors in an end-to-end fashion.

Existing Approaches

- Regularization

$$\min_f \ \text{Empirical\_Risk}(f) + R(f)$$

√ simple and efficient

× contention of weights between regularizer and supervised performance

Proposed Method

Morph a representation **z** towards a structured one by proximal mapping:

promote desired structure

$$z \quad \mapsto \quad \mathrm{argmin}_{x \in C} \; \frac{\lambda}{2} \|x - z\|^2 + L(x)$$

z: mini-batch or single-example

a mini-batch $\Longleftrightarrow$ a task in meta-learning

proximal mapping $\Longleftrightarrow$ task-specific base learner

Embed the proximal mapping as a layer into deep networks

Advantages

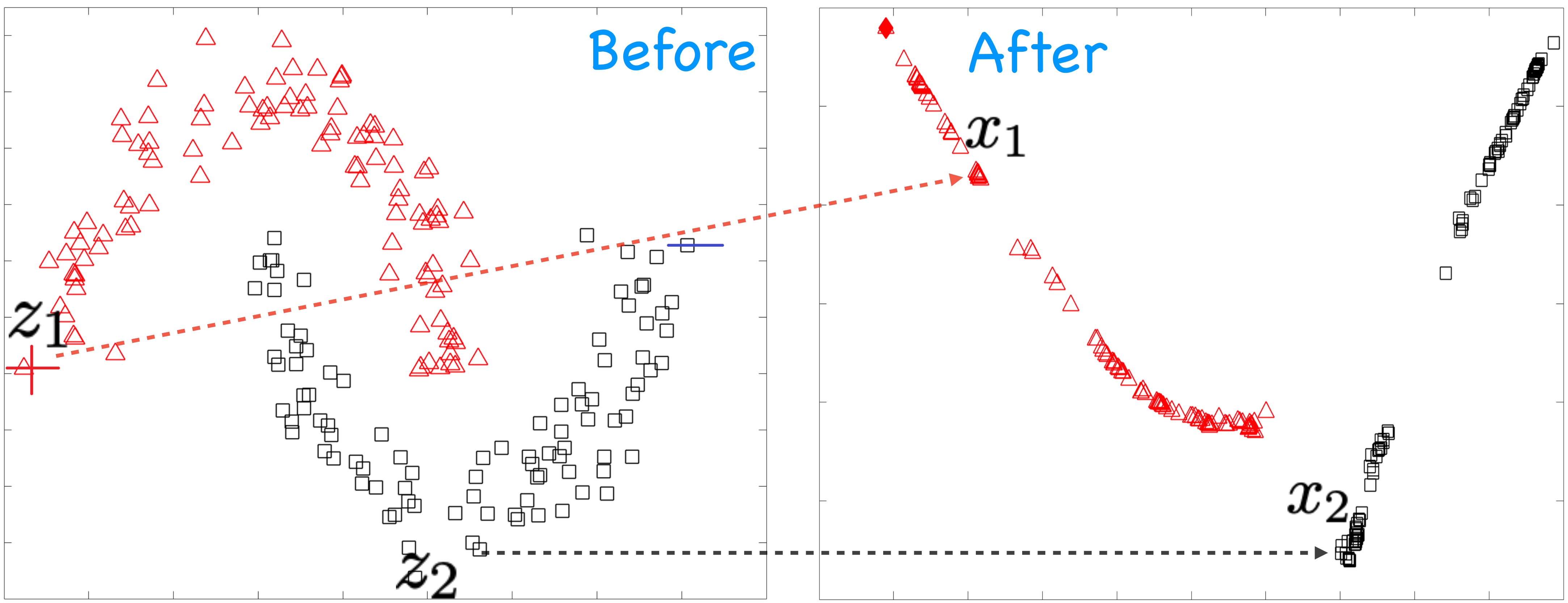
+ decoupling the regularization and supervised learning
+ extend meta-learning to unsupervised base learners

Proposed Method

Morph a representation **z** towards a structured one by proximal mapping:

promote desired structure

$$z \quad \mapsto \quad \operatorname{argmin}_{x \in C} \; \frac{\lambda}{2} \|x - z\|^2 + L(x)$$



L: graph-Laplacian (for smoothness on manifold)

MetaProx for Multi-view Learning

In multiview learning, observations are available as pairs of views: $\{x_i, y_i\}$.

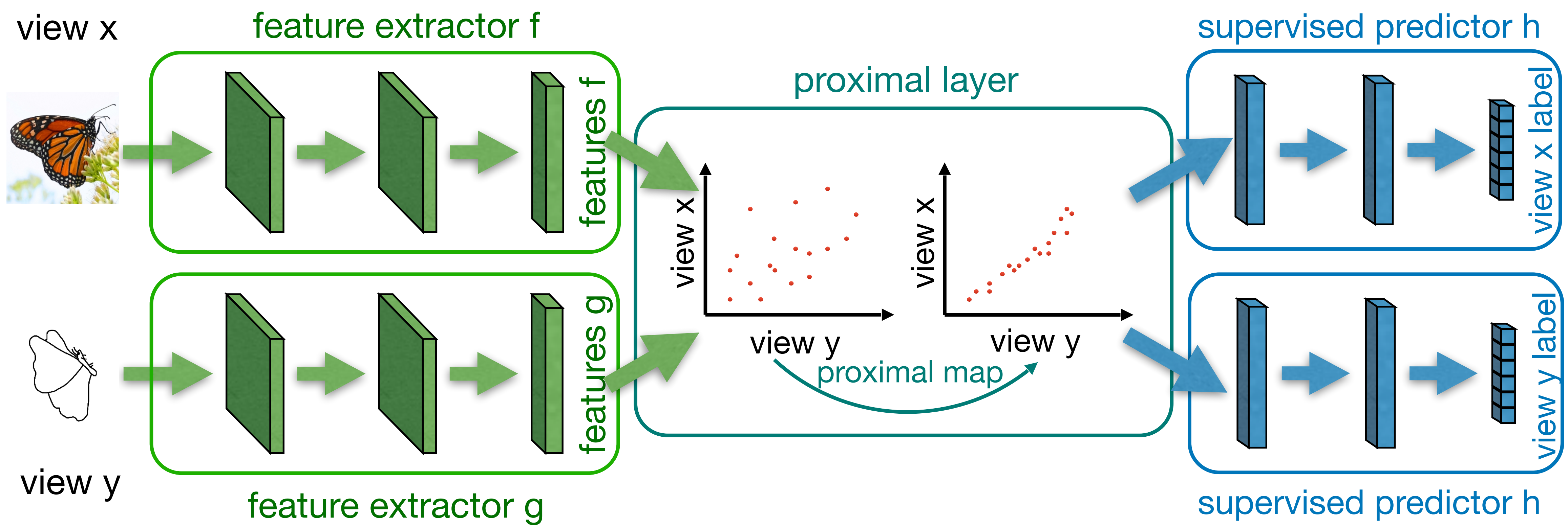
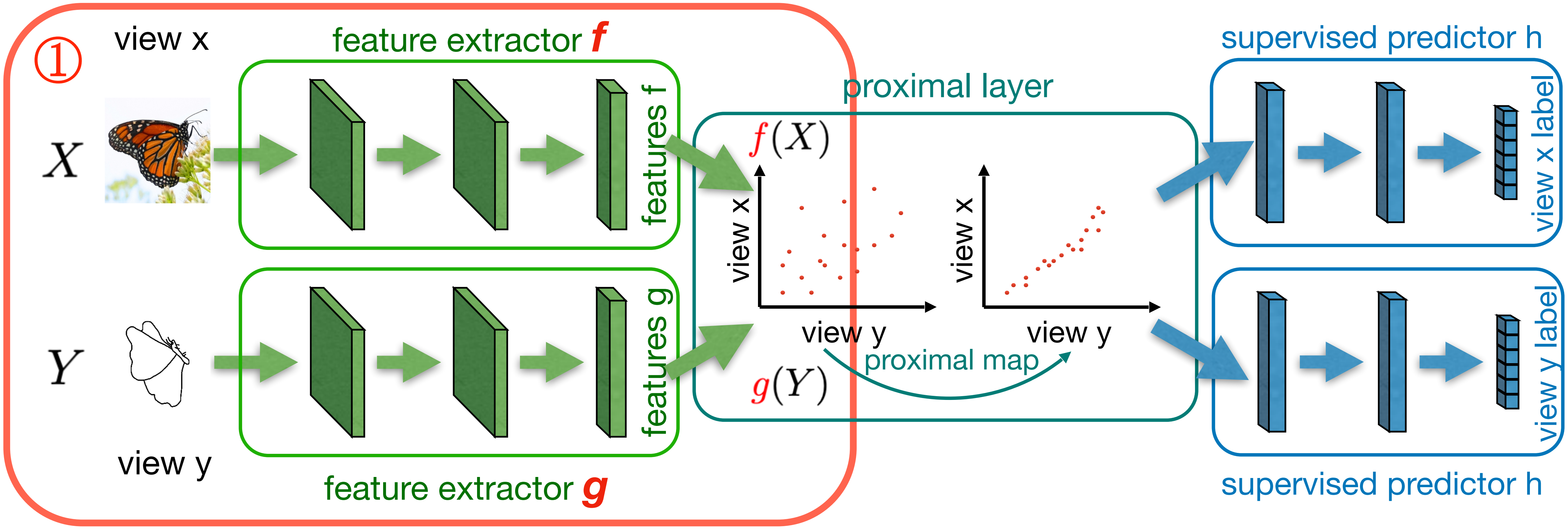


Figure 1: training framework of MetaProx

MetaProx for Multi-view Learning

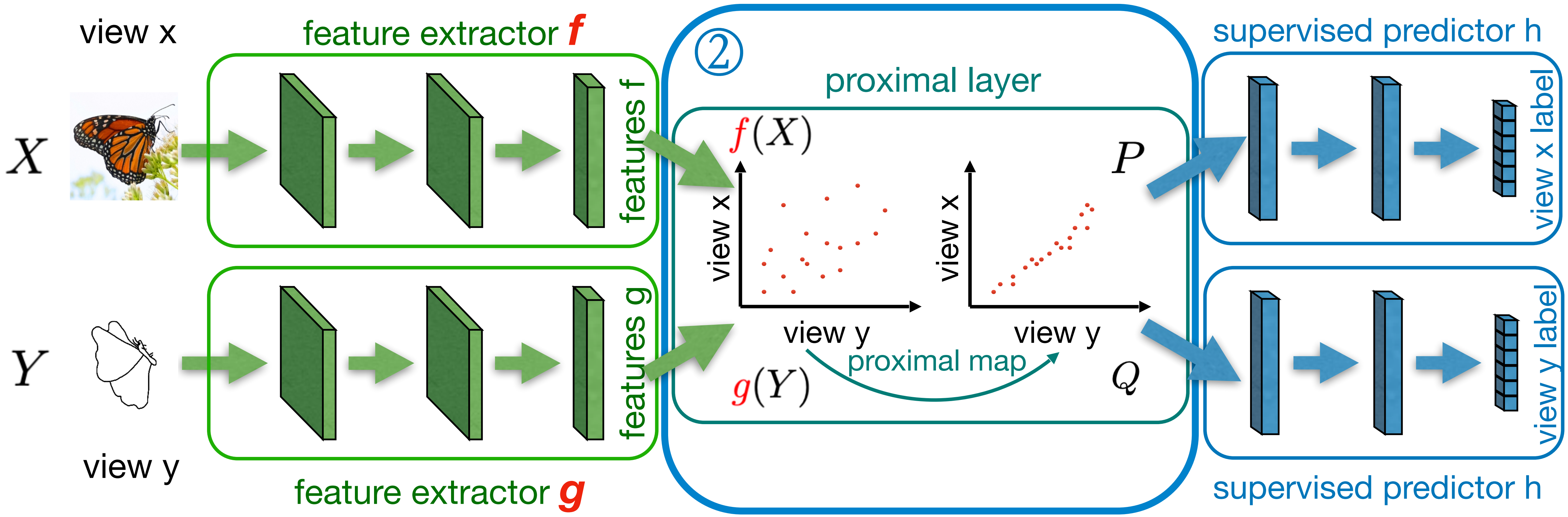


① feature extraction:

$$X \longrightarrow f(X)$$

$$Y \longrightarrow g(Y)$$

MetaProx for Multi-view Learning



② proximal mapping: promote high correlation between two views

$$\arg\min_{P,Q} \frac{1}{2}\|P - f(X)\|^2 + \frac{1}{2}\|Q - g(Y)\|^2 + \mathrm{CCA}(P,Q)$$

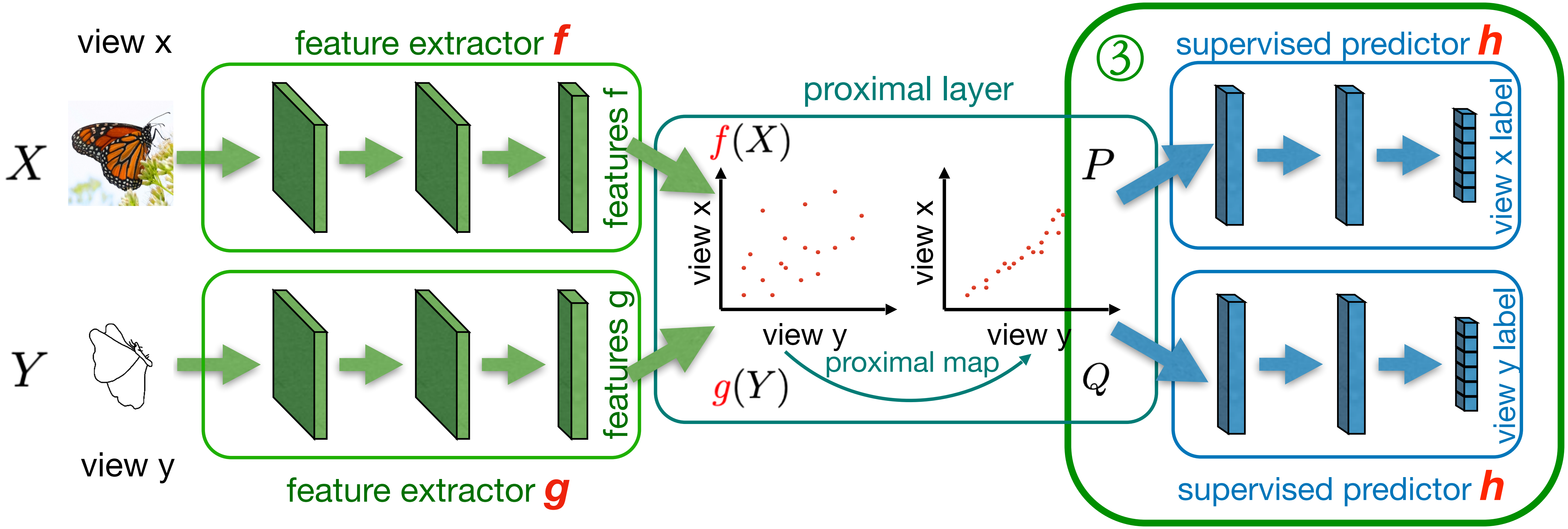
$$\mathrm{CCA}(P,Q) := \min_{U,V} -\mathrm{tr}(U^\top P Q^\top V),$$

$$\text{s.t } U^\top P P^\top U = I$$

$$V^\top Q Q^\top V = I$$

$$u_i^\top P Q^\top v_j = 0, \forall i \neq j \text{ from } 1 \text{ to } k.$$

MetaProx for Multi-view Learning

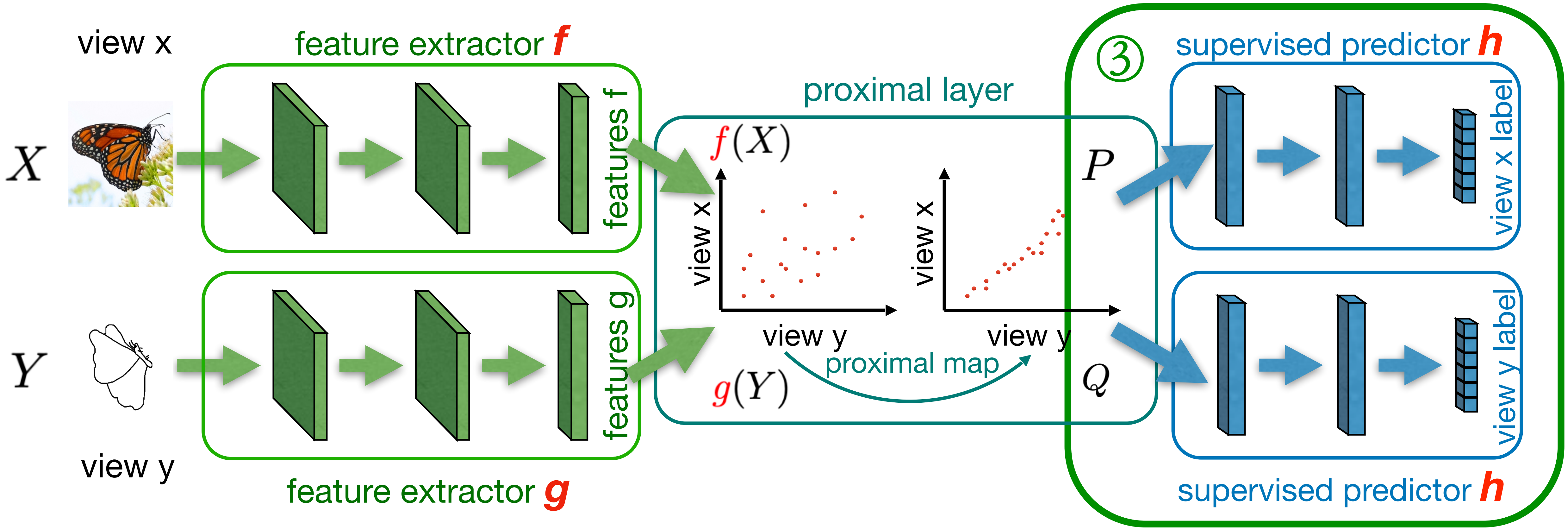


③ supervised task

$$\min_{f,g,h} \; loss \left(h \left(\begin{array}{rl} \arg\min\limits_{P,Q} & \frac{1}{2}\|P - f(X)\|^2 \\ + & \frac{1}{2}\|Q - g(Y)\|^2 \\ + & \mathrm{CCA}(P,Q) \end{array} \right), \text{ground true label} \right)$$

h: supervised predictor

MetaProx for Multi-view Learning



③ supervised task

$$\min_{f,g,h} \; loss\left(h\left(\begin{array}{rl} \arg\min\limits_{P,Q} & \frac{1}{2}\|P - f(X)\|^2 \\ + & \frac{1}{2}\|Q - g(Y)\|^2 \\ + & \mathrm{CCA}(P,Q) \end{array} \right), \text{ground true label} \right)$$

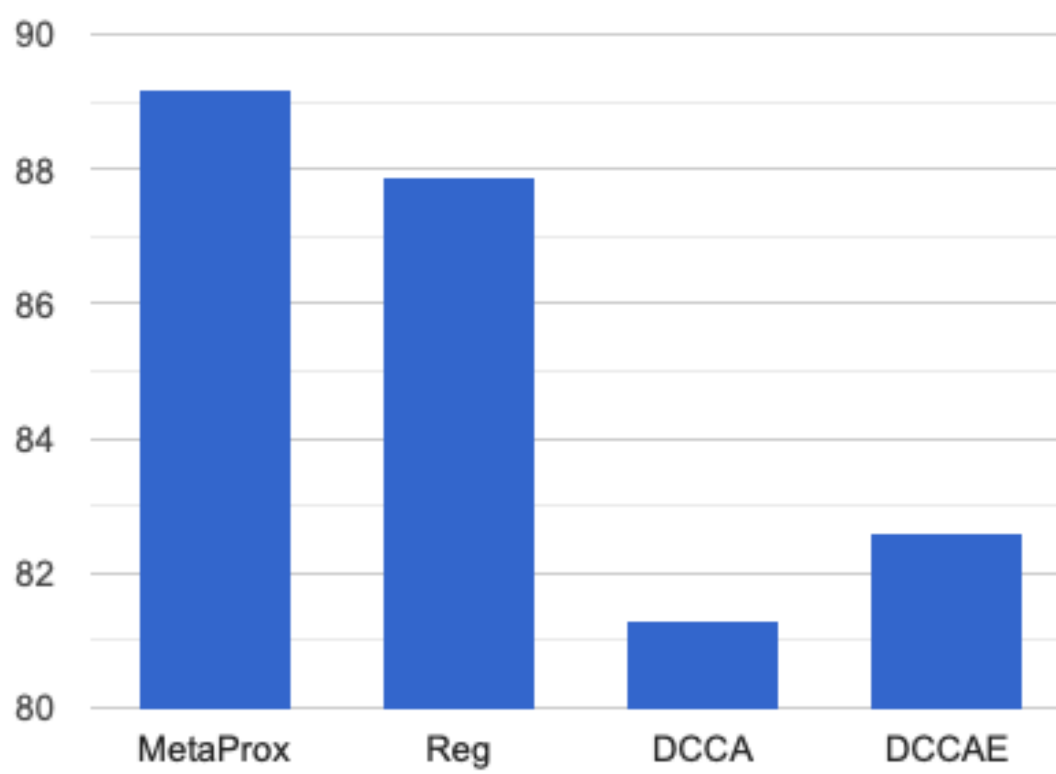
optimize over red variables

Experiment Results

Multi-view image classification

- **Dataset**: a subset of Sketchy (20 classes)

{( , ),'butterfly'; ... ...;( , ),'cat' }



Test accuracy for image classification

Experiment Results

Crosslingual word embedding

– **Dataset**: WS353, SimLex999

– **Metric**: Spearman's correlation between the rankings by model and human

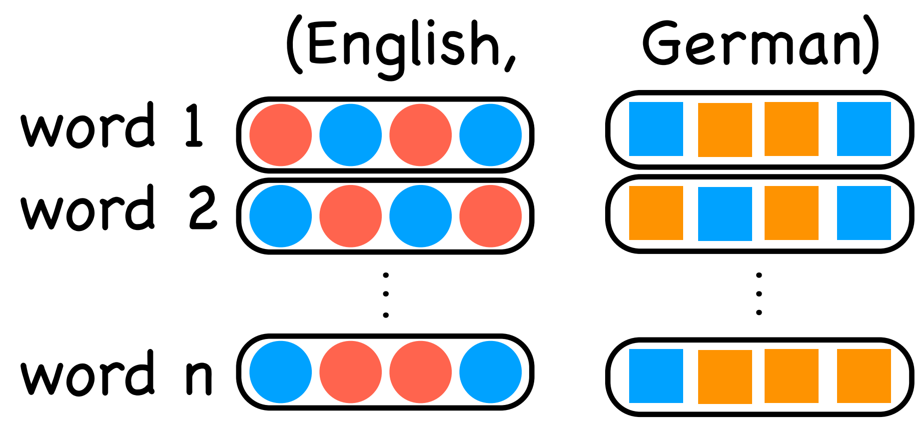


Table 1: Spearman's correlation for word similarities

	WS-353		WS-SIM		WS-REL		SimLex999	
	EN	DE	EN	DE	EN	DE	EN	DE
Baseline	73.35	52.68	77.84	63.34	67.66	44.24	37.15	29.09
linearCCA	73.79	68.45	76.06	73.02	67.01	62.95	37.84	43.34
DCCA	73.86	69.09	**78.69**	74.13	66.57	64.66	38.78	43.29
DCCAE	72.39	**69.67**	75.74	74.65	65.96	64.20	36.72	41.81
MetaProx	**75.38**	69.19	78.28	**75.40**	**70.97**	**66.81**	**39.99**	**44.23**
DEPEMB	-	-	-	-	-	-	35.60	30.60

At the poster:
More details and discussions

Thanks!

MetaProx $\neq$ "Efficient Meta Learning via Minibatch Proximal Update" (NeurIPS 2019)

"Meta-Learning with Implicit Gradients" (NeurIPS 2019)

modeling | optimization