Unsupervised Meta-learning via Few-shot Pseudo-supervised Contrastive Learning

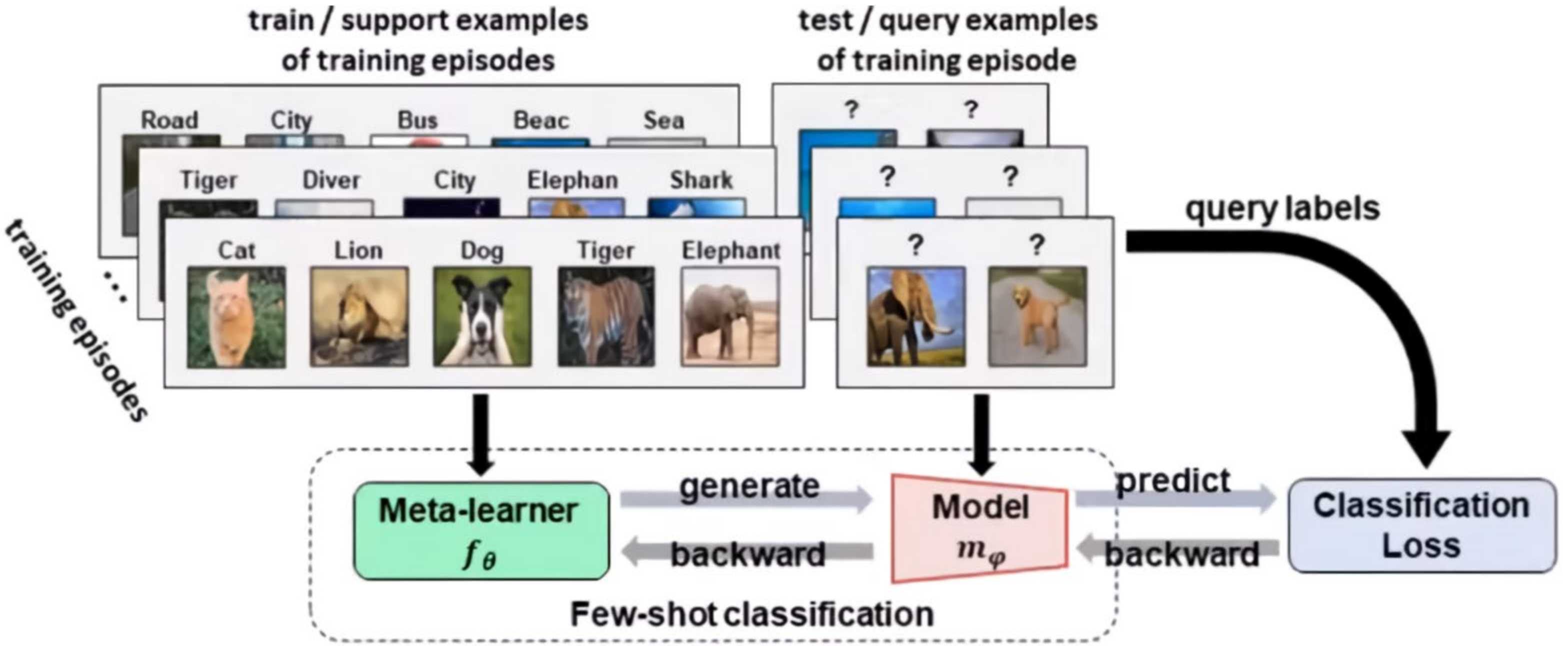
Huiwon Jang[A]*   Hankook Lee[B]*†   Jinwoo Shin[A]

[A]Korea Advanced Institute of Science and Technology (KAIST)
[B]LG AI Research

*Equal contribution
†Work done at KAIST

What is unsupervised meta-learning?

- **Meta-learning** aims to learn **generalizable knowledge** from **prior experiences**
  - It can solve **unseen**, yet **relevant tasks**



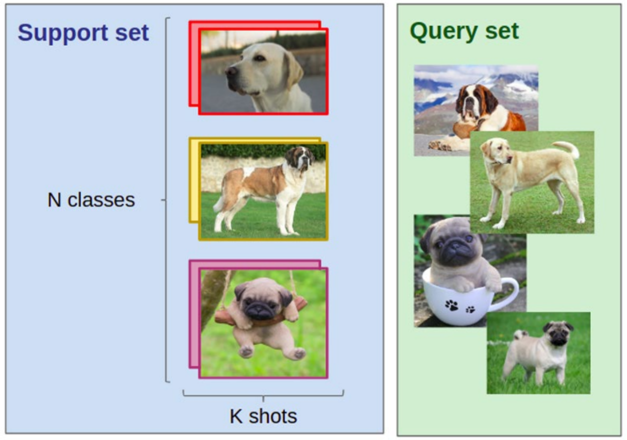
Limitation of meta-learning: <u>Task (episode) construction</u> phase <u>requires a lot of human-annotations</u>

What is unsupervised meta-learning?

- **Meta-learning** aims to learn **generalizable** knowledge from **prior experiences**
- **Unsupervised meta-learning** aims at **meta-learning** from **unlabeled** data

Meta-train (Unlabeled)

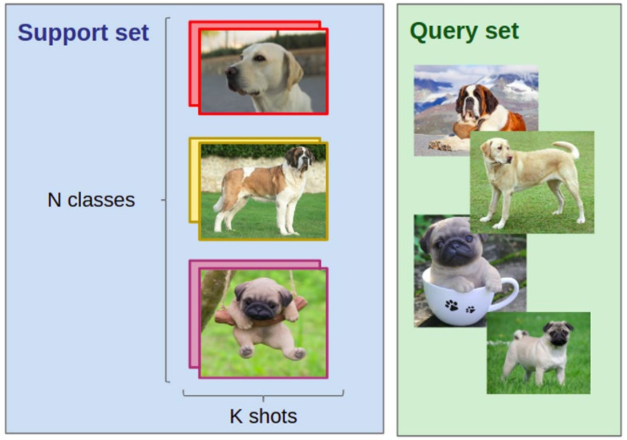
Meta-test (Labeled)



What is unsupervised meta-learning?

- **Meta-learning** aims to learn **generalizable** knowledge from **prior experiences**
- **Unsupervised meta-learning** aims at **meta-learning** from **unlabeled** data
  - **Challenge**: It requires to **construct synthetic tasks** to perform meta-learning without labels

Meta-train (Unlabeled) **Meta-test (Labeled)**

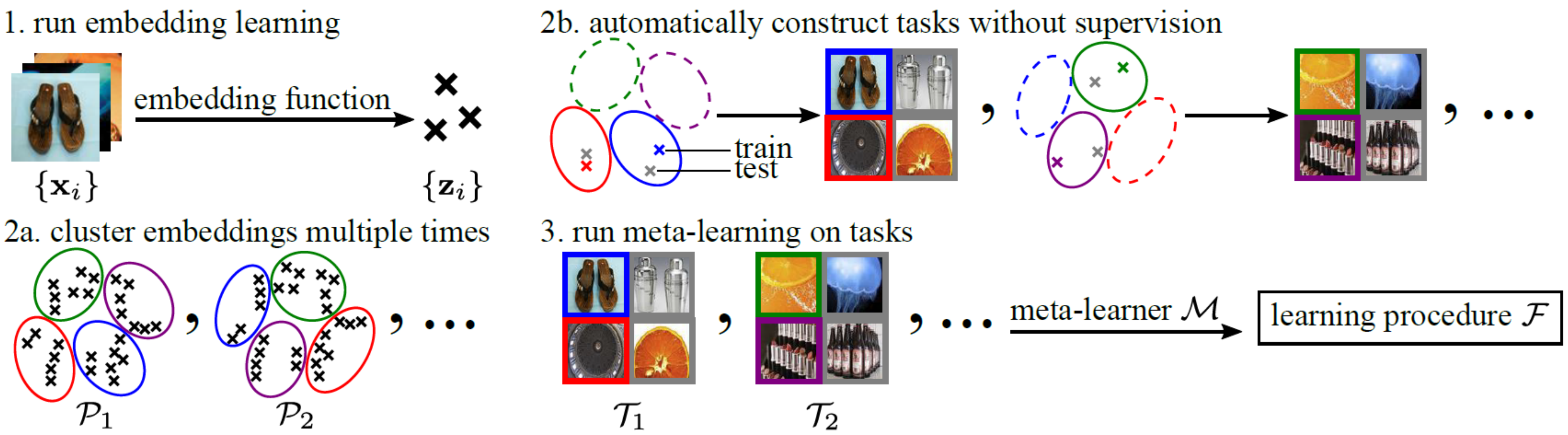


Benefits of unsupervised meta-learning:

- **Take the advantage of meta-learning**: Generalized model across tasks, which adapt to new tasks quickly
- **Mitigate the limitation of meta-learning**: Task construction phase requires a lot of human-annotations

Previous Approaches to Construct Synthetic Tasks

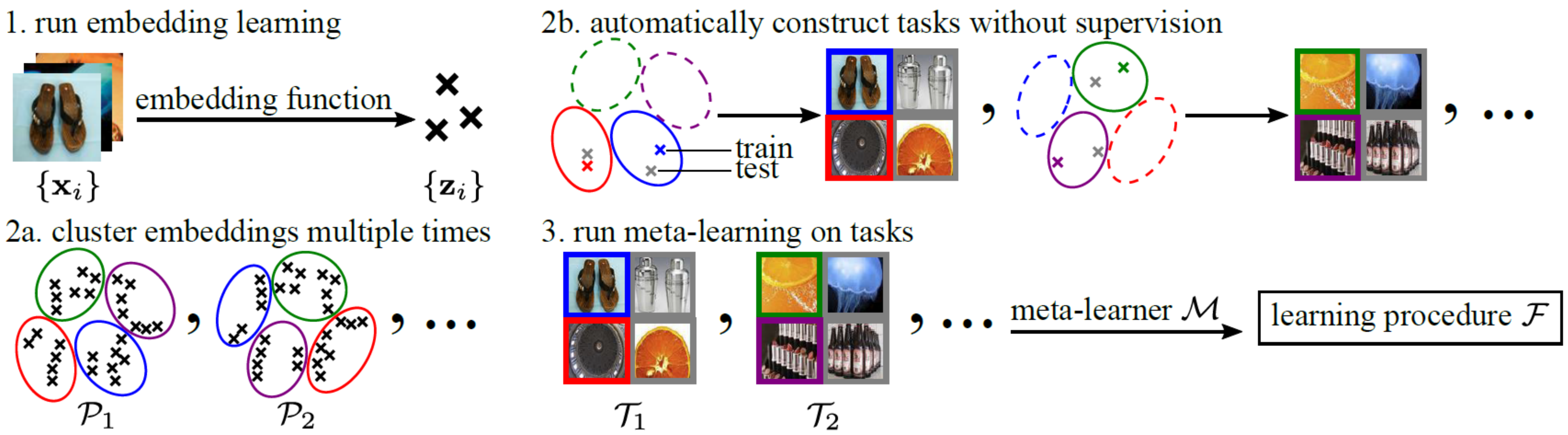
1. Assigning pseudo-labels [1-2]
   - They utilize unsupervised representation or augmentations to assign pseudo-labels
   - **Limitation:** Pseudo-labels are fixed during meta-training, and <u>impossible to correct mislabeled samples</u>



[1] Hsu et al., Unsupervised Learning via Meta-learning, ICLR 2019
[2] Khodadadeh et al., Unsupervised Meta-learning for Few-shot Image Classification, NeurIPS 2019

Previous Approaches to Construct Synthetic Tasks

1. Assigning pseudo-labels [1-2]
   - They utilize unsupervised representation or augmentations to assign pseudo-labels
   - **Limitation:** Pseudo-labels are fixed during meta-training, and <u>impossible to correct mislabeled samples</u>

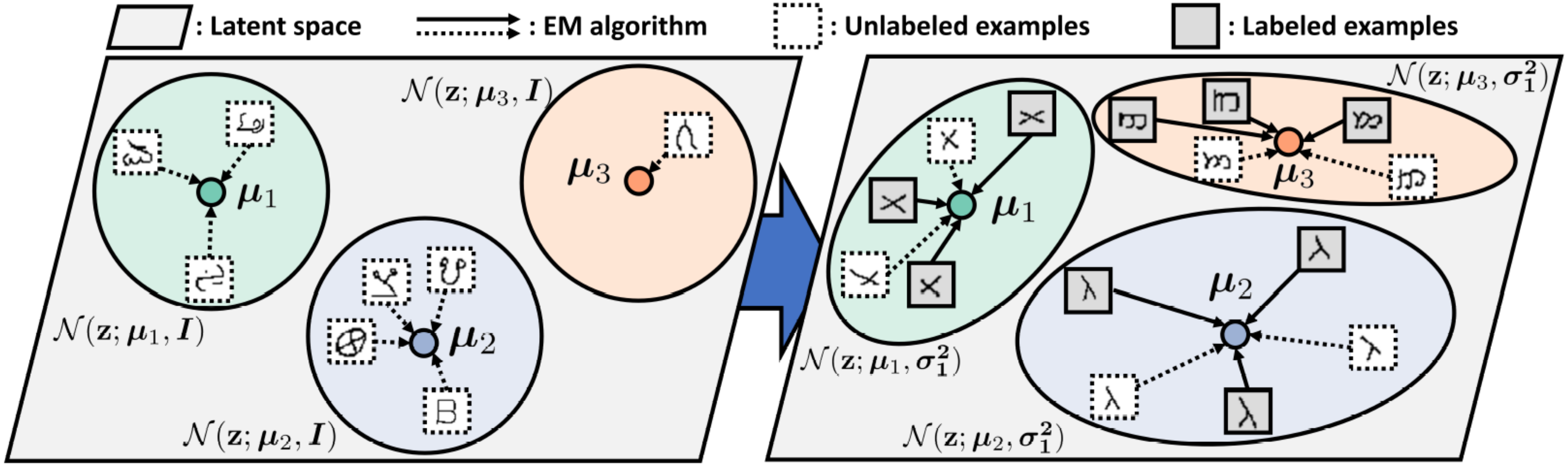


- **Question:** How to **progressively improve a pseudo-labeling** strategy during meta-learning?

[1] Hsu et al., Unsupervised Learning via Meta-learning, ICLR 2019
[2] Khodadadeh et al., Unsupervised Meta-learning for Few-shot Image Classification, NeurIPS 2019

Previous Approaches to Construct Synthetic Tasks

2. Utilizing generative models [1-3]
   - They generate synthetic tasks via generative models like VAE
   - **Limitation:** Rely on the quality of generated samples which are cumbersome to scale into large-scale

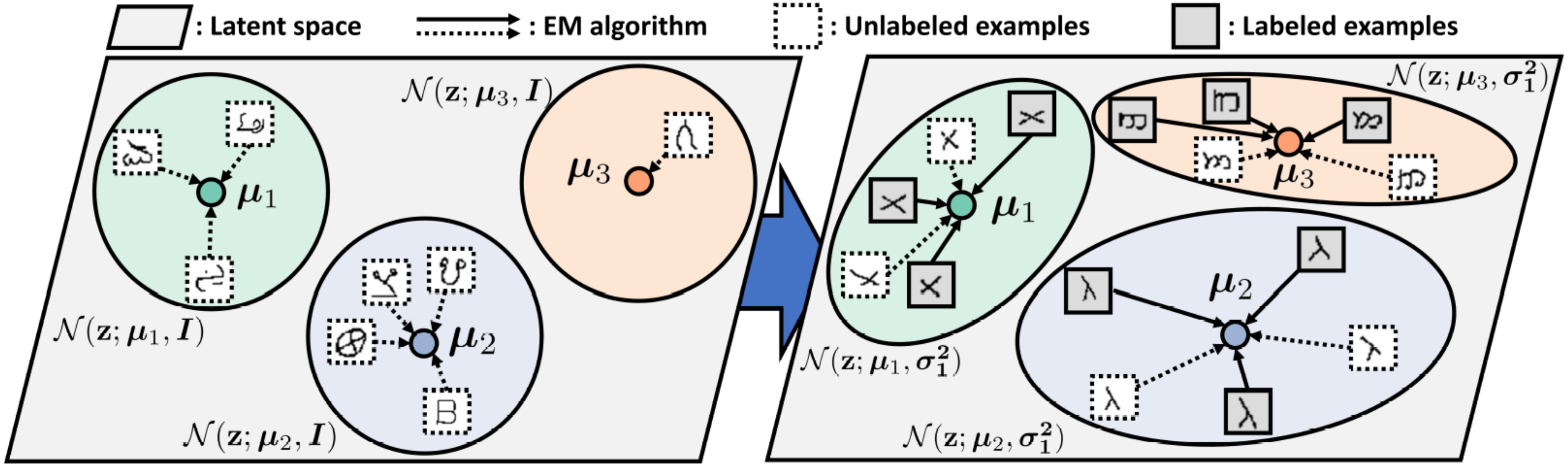


(a) Unsupervised Meta-training (b) Supevised Meta-test

[1] Khodadadeh et al., Unsupervised Meta-learning through Latent-space Interpolation in Generative Models, ICLR 2021
[2] Lee et al., Meta-GMVAE: Mixture of Gaussian VAE for Unsupervised Meta-learning, ICLR 2021
[3] Kong et al., Unsupervised Meta-learning via Latent Space Energy-based Model of Symbol Vector Coupling, NeurIPSW-MetaLearn 2021

Previous Approaches to Construct Synthetic Tasks

2. Utilizing generative models [1-3]
   - They generate synthetic tasks via generative models like VAE
   - **Limitation:** Rely on the quality of generated samples which are cumbersome to scale into large-scale



(a) Unsupervised Meta-training (b) Supevised Meta-test

- **Question:** How to **construct diverse tasks** without generative models?

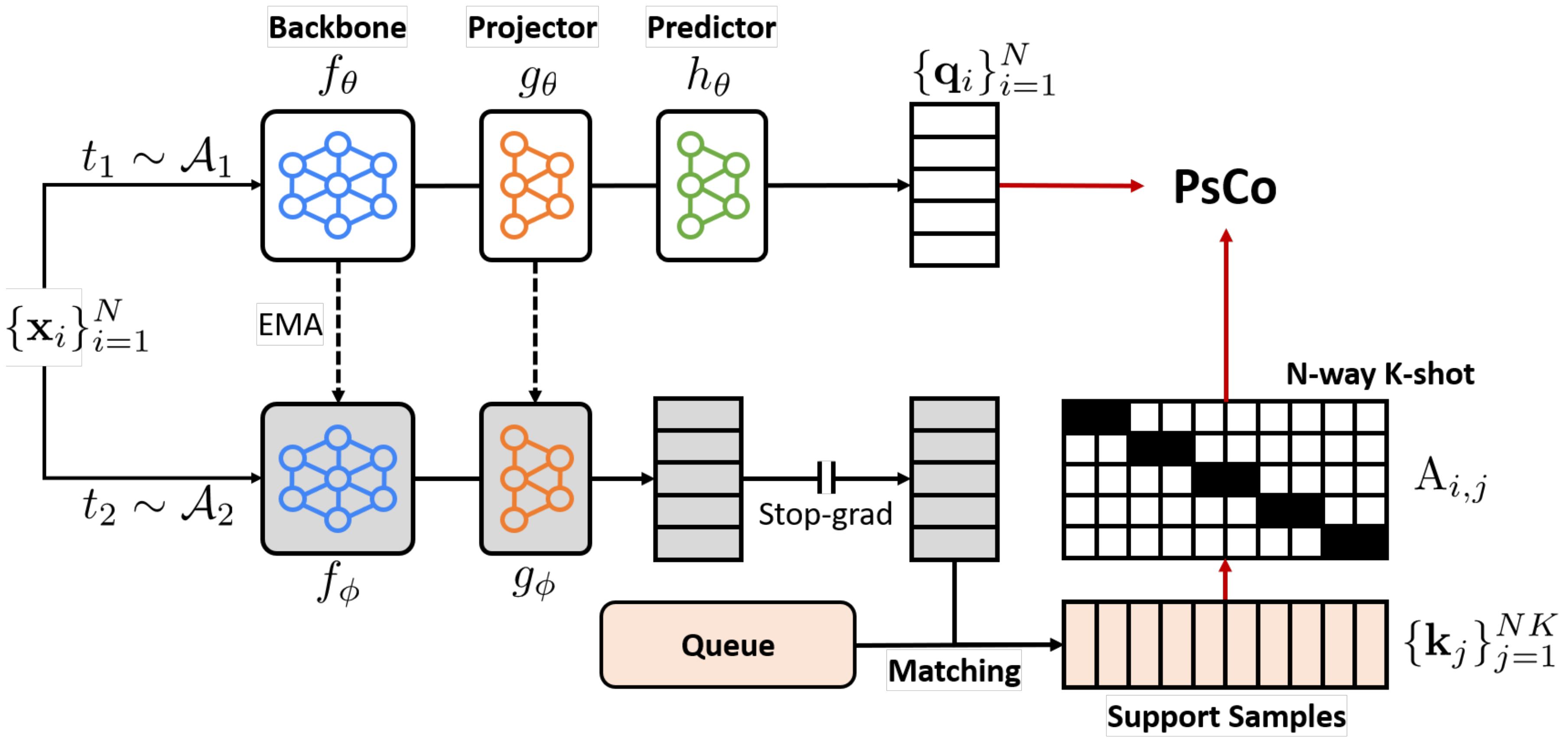
[1] Khodadadeh et al., Unsupervised Meta-learning through Latent-space Interpolation in Generative Models, ICLR 2021
[2] Lee et al., Meta-GMVAE: Mixture of Gaussian VAE for Unsupervised Meta-learning, ICLR 2021
[3] Kong et al., Unsupervised Meta-learning via Latent Space Energy-based Model of Symbol Vector Coupling, NeurIPSW-MetaLearn 2021

Method: Pseudo-supervised Contrast (PsCo)

🤔 How to **progressively improve a pseudo-labeling** strategy during meta-learning?

🤔 How to **construct diverse tasks** without generative models?

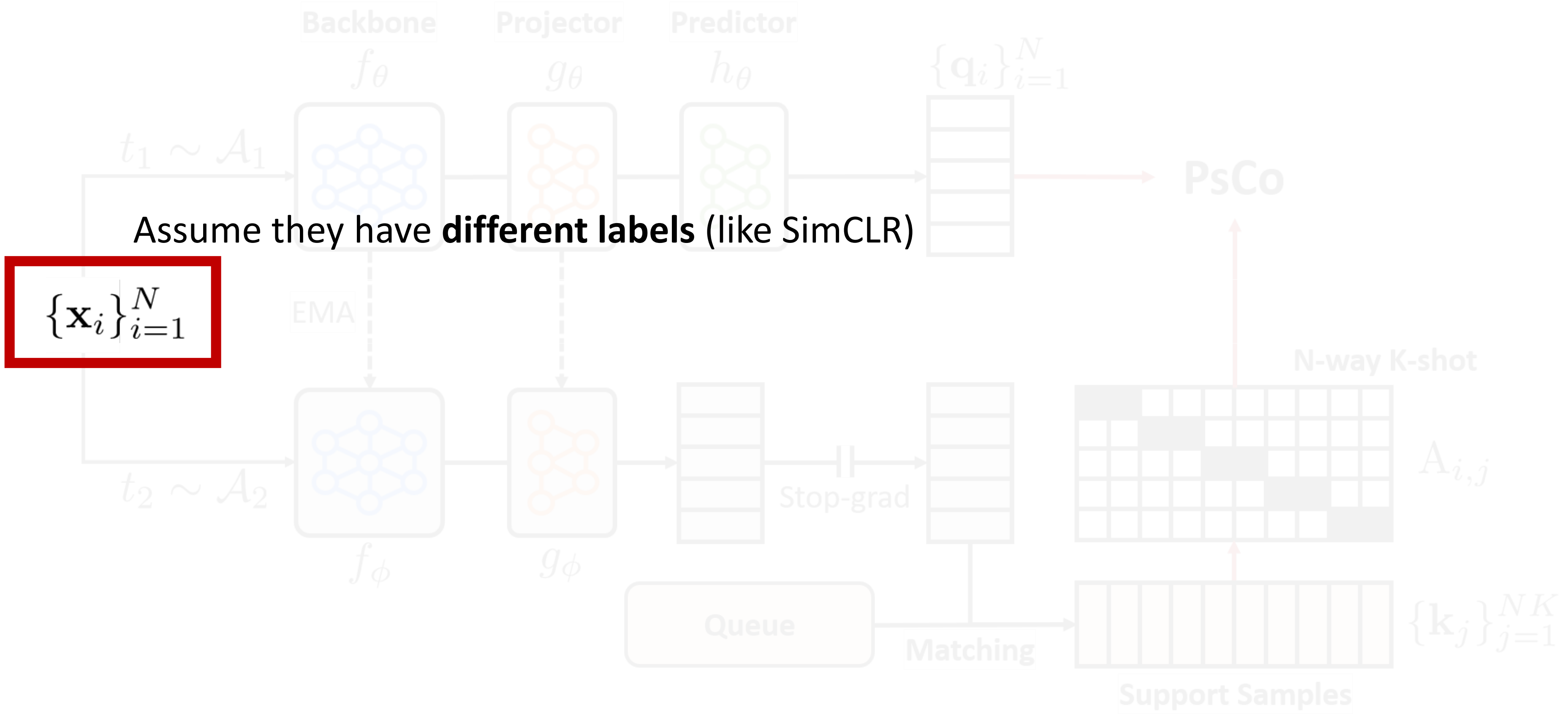
Idea: Construct pseudo-tasks via momentum representations and **apply contrastive learning**



Method: Pseudo-supervised Contrast (PsCo)

Idea: Construct pseudo-tasks via momentum representations and **apply contrastive learning**

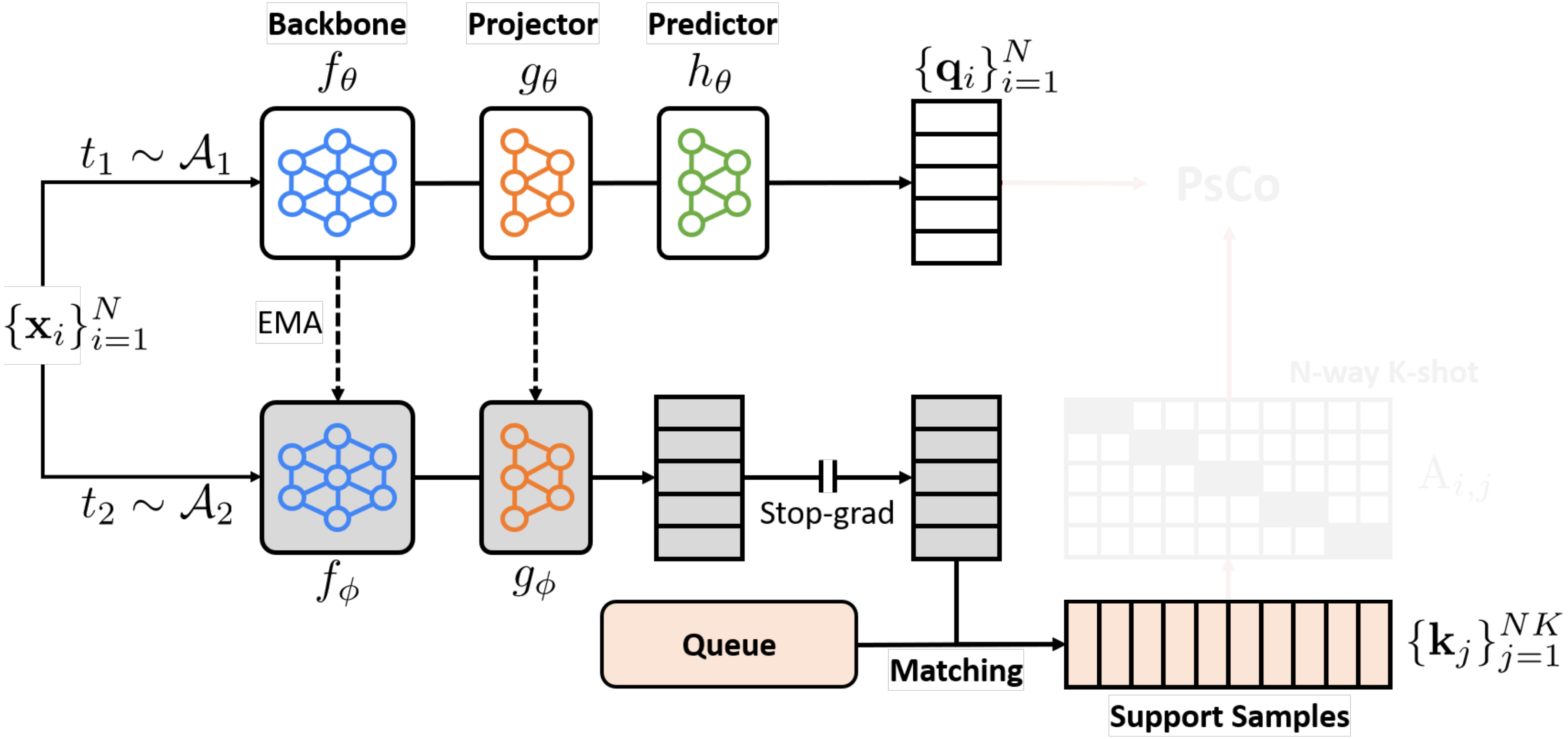
- $\{\mathbf{x}_i\}_{i=1}^N$: **query** samples for **N-way K-shot** task



Method: Pseudo-supervised Contrast (PsCo)

Idea: Construct pseudo-tasks via momentum representations and **apply contrastive learning**

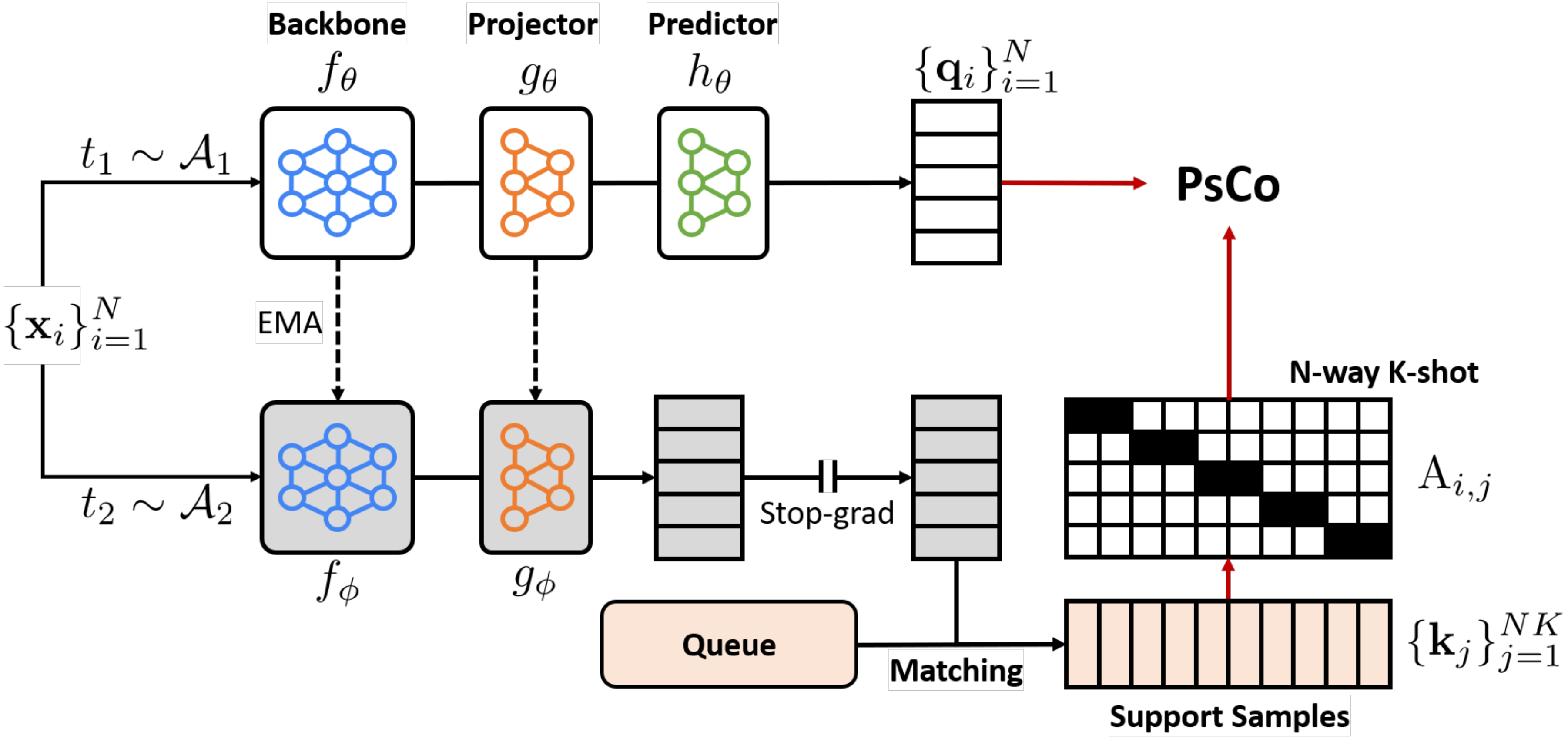
- $\{\mathbf{x}_i\}_{i=1}^N$: **query** samples for **N-way K-shot** task
- Select appropriate **K-shot support** samples from momentum queue



Method: Pseudo-supervised Contrast (PsCo)

Idea: Construct pseudo-tasks via momentum representations and **apply contrastive learning**

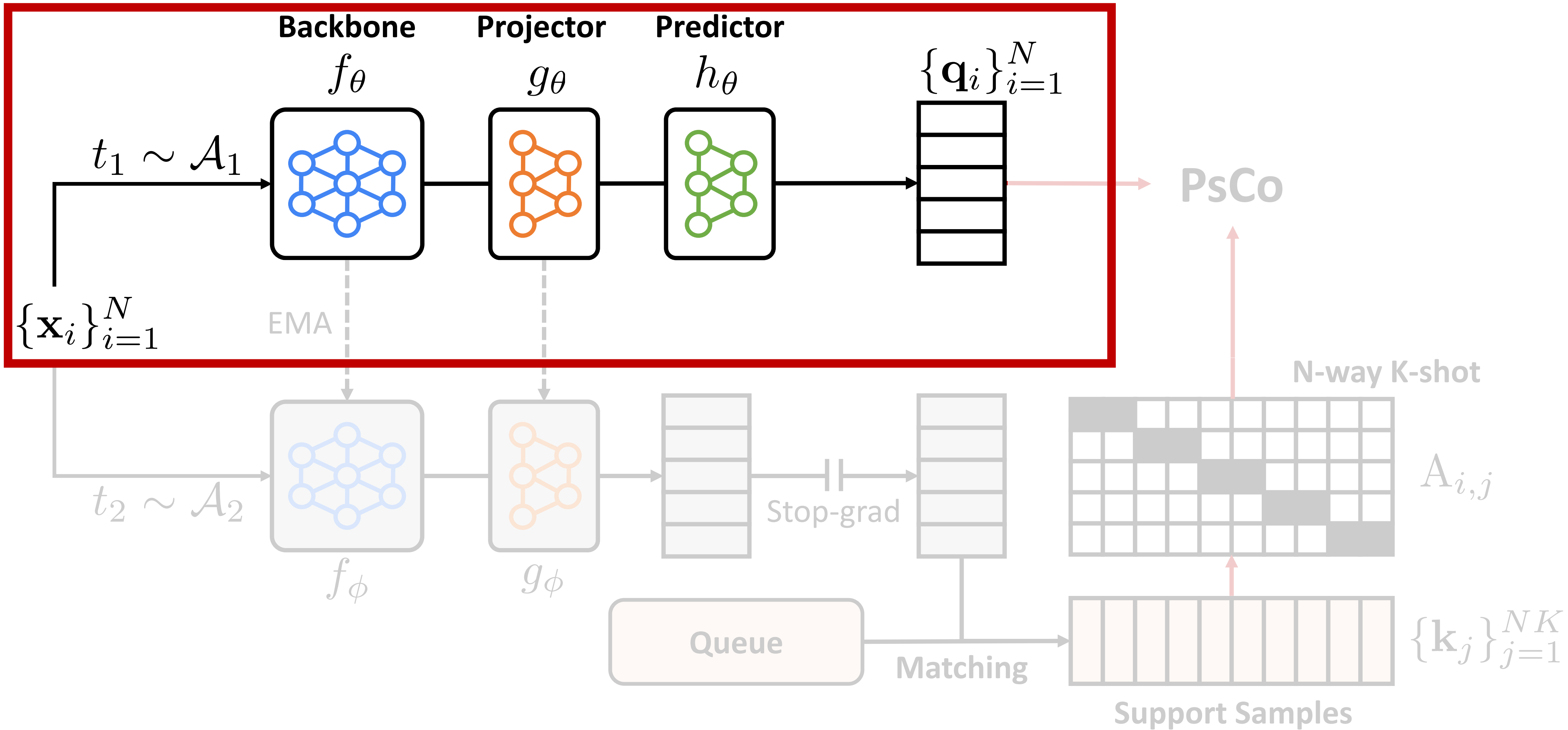
- $\{\mathbf{x}_i\}_{i=1}^{N}$: **query** samples for **N-way K-shot** task
- Select appropriate **K-shot support** samples from momentum queue
- Supervised contrastive learning for pseudo-labeled tasks: **Pseudo-supervised Contrast (PsCo)**



Method: Pseudo-supervised Contrast (PsCo)

Step 1: **Compute query representations**

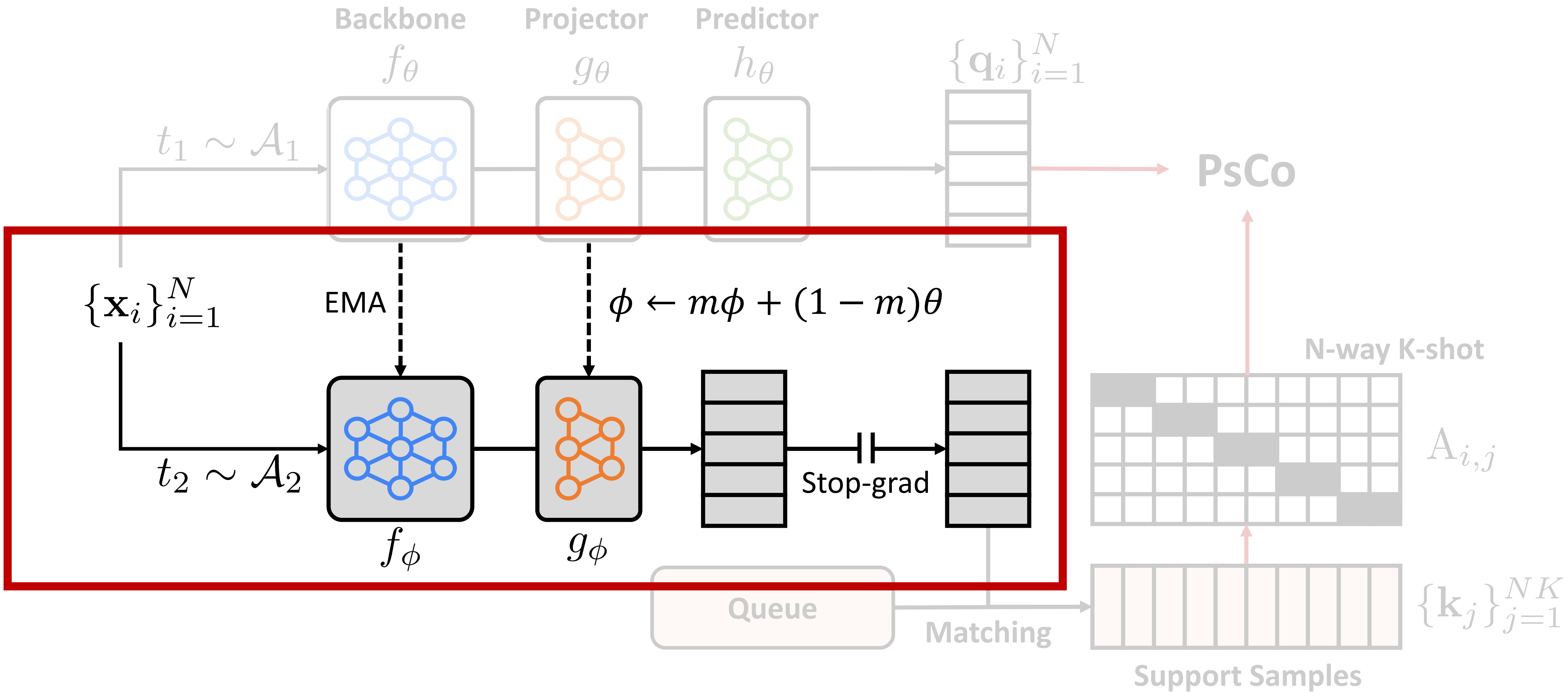
- Use **strong** augmentations and **online** encoder



Method: Pseudo-supervised Contrast (PsCo)

Step 2: **Compute momentum representations of queries**

- Use **weak** augmentations (to find an accurate pseudo-label) and **momentum** encoder



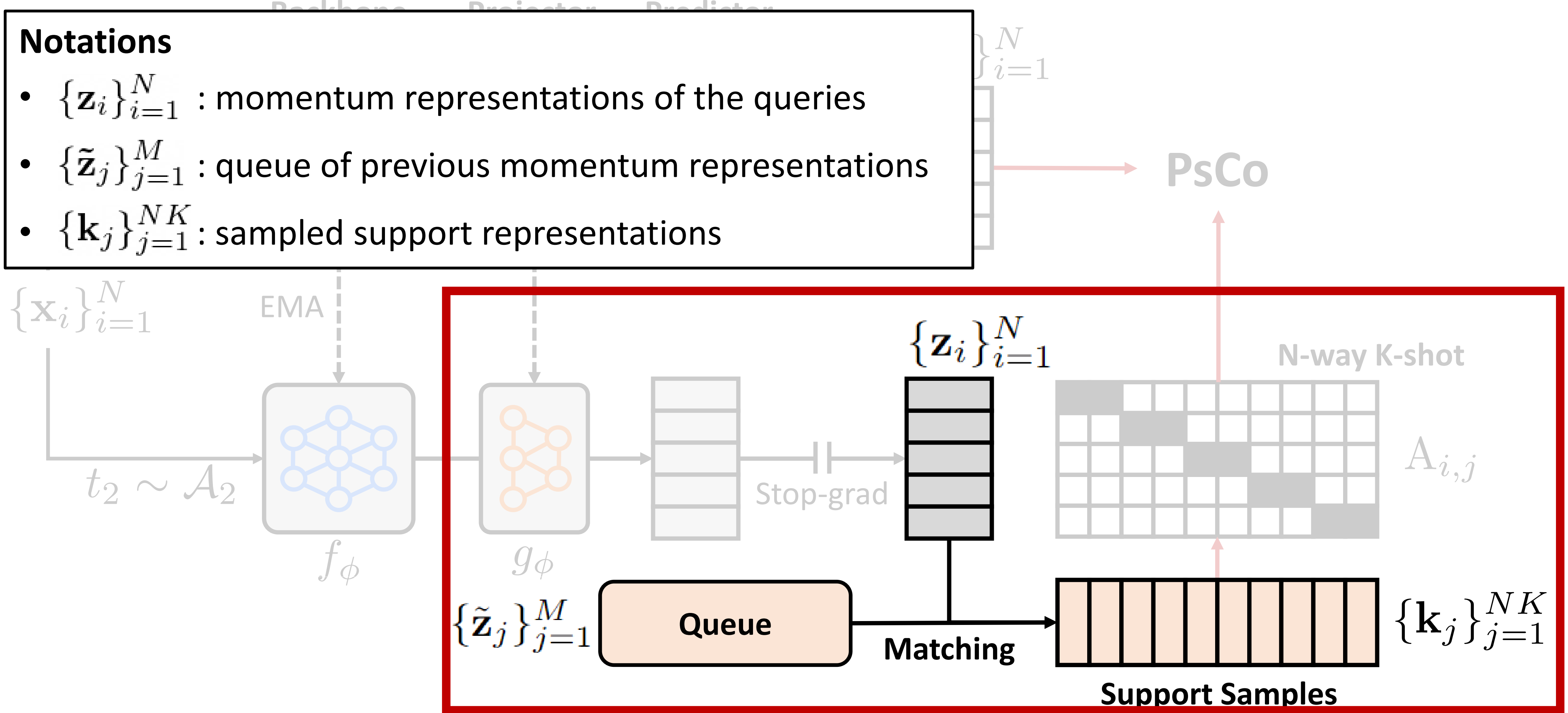
Method: Pseudo-supervised Contrast (PsCo)

Step 3: **Sample support representations from Queue via a matching algorithm**

- Use **momentum queue** with **matching** algorithm (Sinkhorn-Knopp + Top-k sampling)

Notations

- $\{\mathbf{z}_i\}_{i=1}^{N}$: momentum representations of the queries
- $\{\tilde{\mathbf{z}}_j\}_{j=1}^{M}$: queue of previous momentum representations
- $\{\mathbf{k}_j\}_{j=1}^{NK}$: sampled support representations

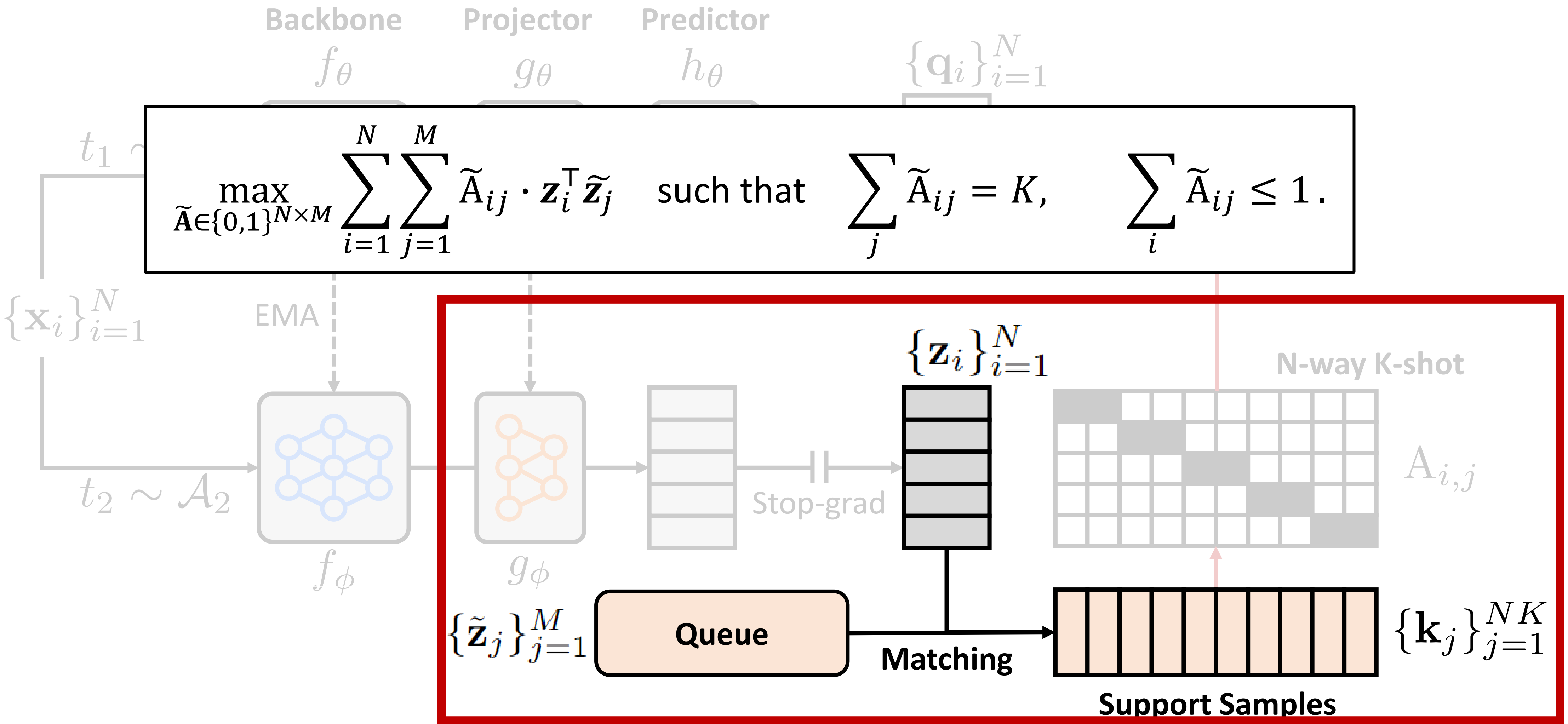


Method: Pseudo-supervised Contrast (PsCo)

Step 3: **Sample support representations from Queue via a matching algorithm**

- **Matching:** How to sample supports that are semantically similar to queries while all samples are different?

$$\max_{\widetilde{\mathrm{A}} \in \{0,1\}^{N \times M}} \sum_{i=1}^{N} \sum_{j=1}^{M} \widetilde{\mathrm{A}}_{ij} \cdot \boldsymbol{z}_i^{\top} \widetilde{\boldsymbol{z}}_j \quad \text{such that} \quad \sum_{j} \widetilde{\mathrm{A}}_{ij} = K, \qquad \sum_{i} \widetilde{\mathrm{A}}_{ij} \leq 1\,.$$

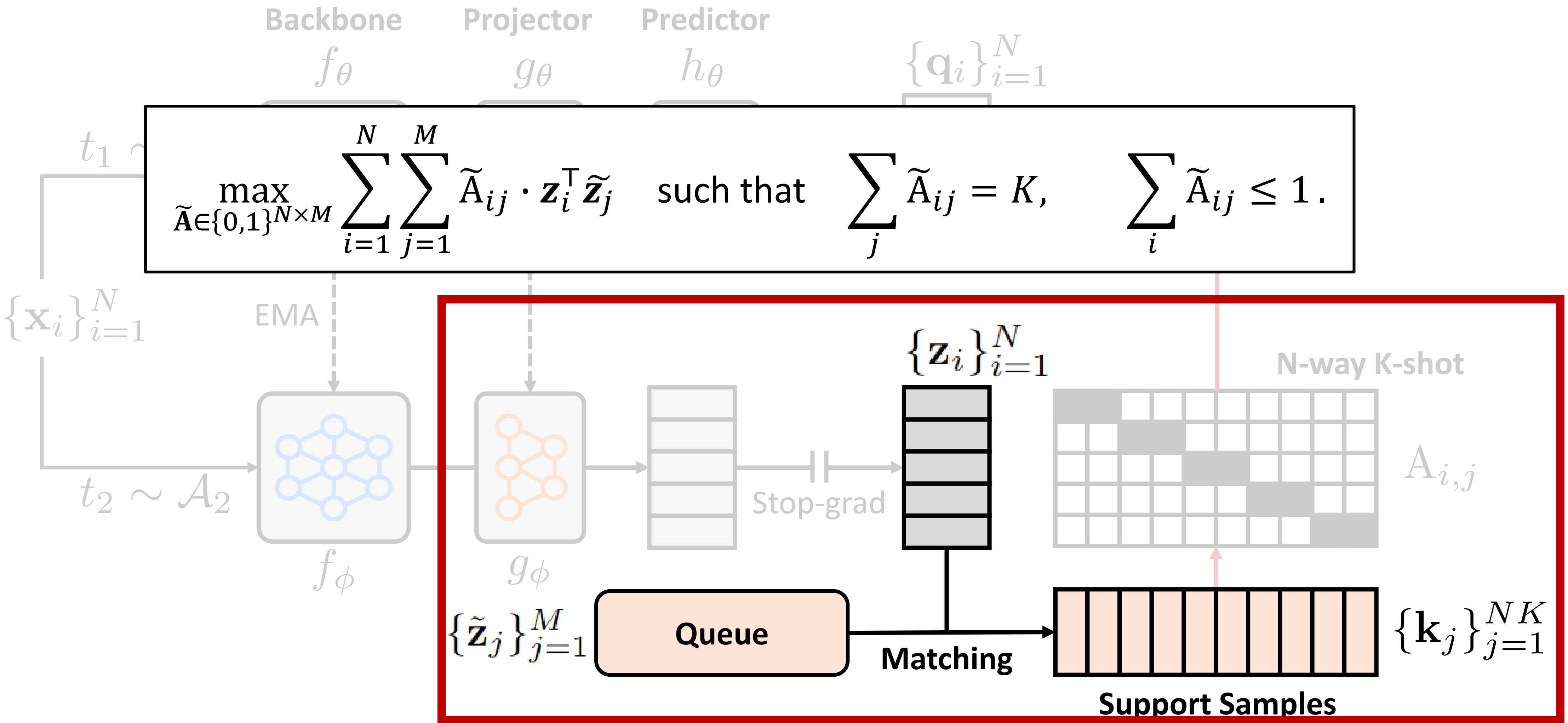


Method: Pseudo-supervised Contrast (PsCo)

Step 3: **Sample support representations from Queue via a matching algorithm**

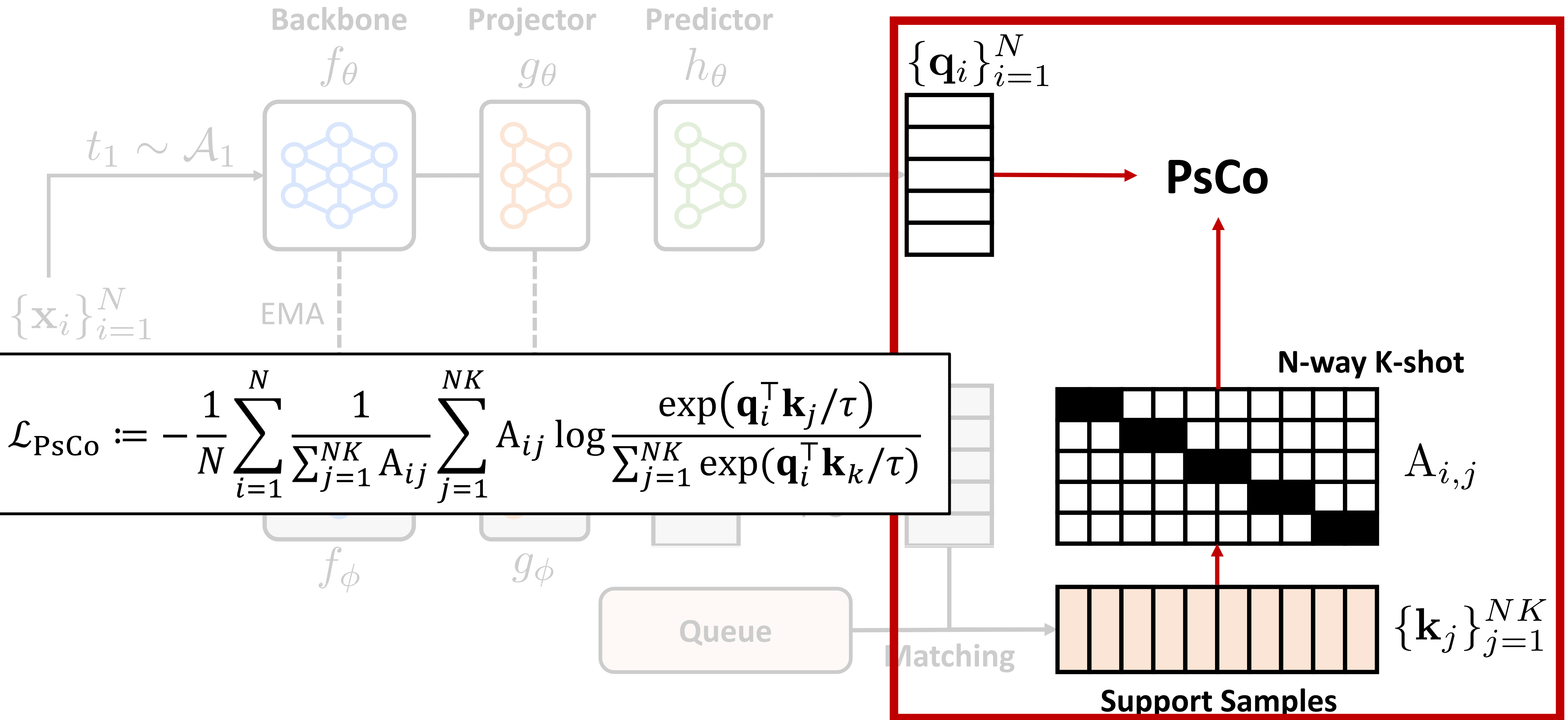
- <u>Sinkhorn-Knopp + Top-k sampling</u>

$$\max_{\widetilde{\mathrm{A}}\in\{0,1\}^{N\times M}} \sum_{i=1}^{N}\sum_{j=1}^{M} \widetilde{\mathrm{A}}_{ij}\cdot \boldsymbol{z}_i^{\top}\widetilde{\boldsymbol{z}}_j \quad \text{such that} \quad \sum_{j}\widetilde{\mathrm{A}}_{ij}=K, \quad \sum_{i}\widetilde{\mathrm{A}}_{ij}\le 1\,.$$



Method: Pseudo-supervised Contrast (PsCo)

Step 4: **Meta-training a pseudo few-shot task via supervised contrastive learning**



$$\mathcal{L}_{\text{PsCo}} \coloneqq -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\sum_{j=1}^{NK}\mathrm{A}_{ij}}\sum_{j=1}^{NK}\mathrm{A}_{ij}\log\frac{\exp\left(\mathbf{q}_i^\top\mathbf{k}_j/\tau\right)}{\sum_{j=1}^{NK}\exp(\mathbf{q}_i^\top\mathbf{k}_k/\tau)}$$

Method: Pseudo-supervised Contrast (PsCo)

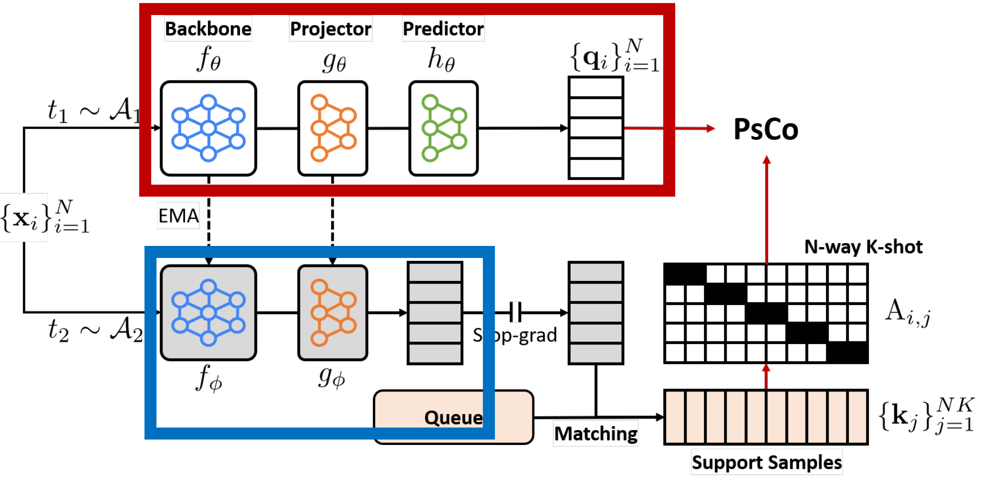
Step 5: **Meta-testing with prototypes (i.e., average of support representations)**

$$\mathcal{L}_{\text{PsCo}} = -\frac{1}{N}\sum_i \frac{1}{\tau_{\text{PsCo}}}\mathbf{q}_i^\top \underbrace{\boxed{\left(\frac{1}{K}\sum_j \mathrm{A}_{i,j}\mathbf{z}_j\right)}}_{\text{Prototype of supports}} + \text{term not depending on } \mathbf{A}.$$

Method: Pseudo-supervised Contrast (PsCo)

Step 5: **Meta-testing with prototypes (i.e., average of support representations)**

$$\mathcal{L}_{\text{PsCo}} = -\frac{1}{N}\sum_i \frac{1}{\tau_{\text{PsCo}}}\mathbf{q}_i^\top \underbrace{\boxed{\left(\frac{1}{K}\sum_j \mathrm{A}_{i,j}\mathbf{z}_j\right)}}_{\text{Prototype of supports}} + \text{term not depending on } \mathbf{A}.$$



Prediction scheme:

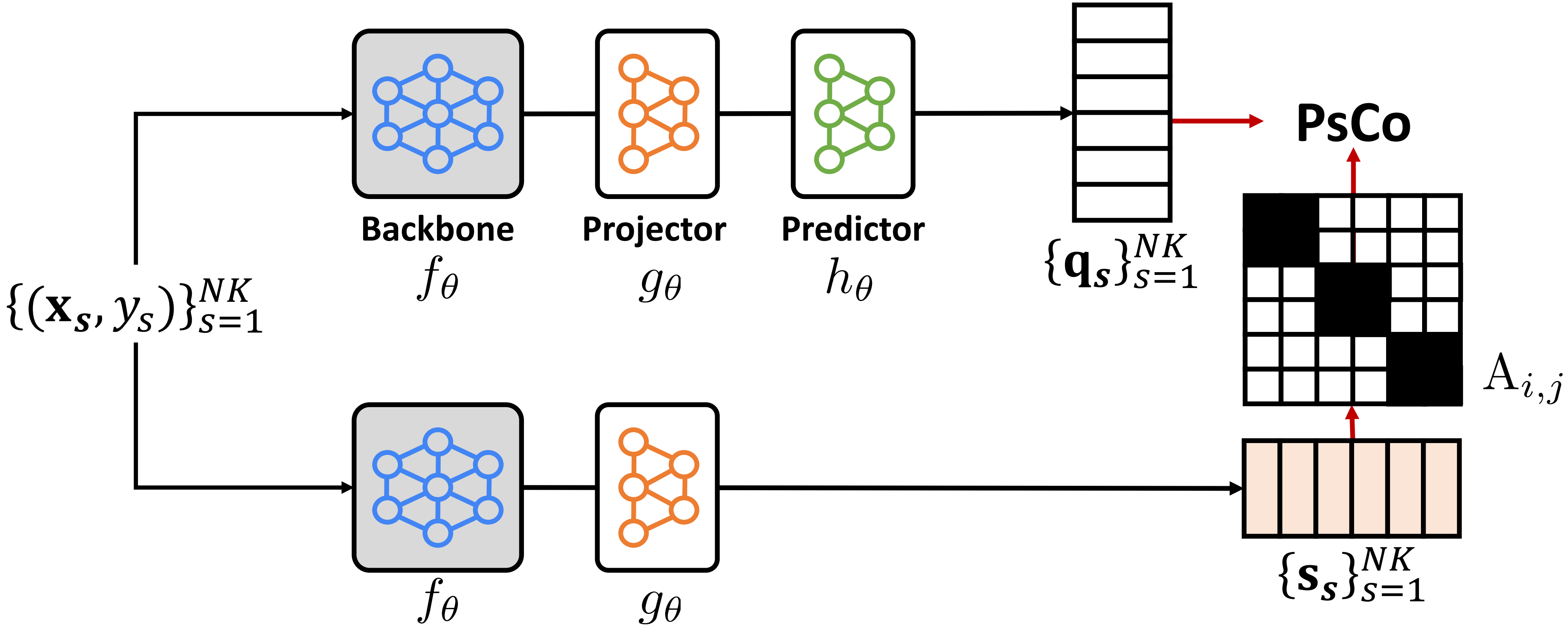
- **Support representation**: $\mathbf{z}_s := \text{Normalize}\left(g_\theta \circ f_\theta(\mathbf{x}_s)\right)$
- **Query representation**: $\mathbf{q}_q := \text{Normalize}\left(h_\theta \circ g_\theta \circ f_\theta(\mathbf{x}_q)\right)$
- $\hat{y} := \arg\max_y \mathbf{q}_q^\top \mathbf{c}_y$ where $\mathbf{c}_y := \text{Normalize}(\sum_s \mathbf{1}_{y_s=y} \cdot \mathbf{z}_s)$
- No momentum network here

Method: Pseudo-supervised Contrast (PsCo)

Step 5: **Meta-testing with prototypes (i.e., average of support representations)**

Adaptation scheme for cross-domain problems:

- Treat each support sample as a query
- Freeze the backbone f_θ and **optimize** only the **projector** g_θ and the **predictor** h_θ
- **E.g., 3-way 2-shot task**



Experiment: Standard Few-shot Classification

- PsCo achieves state-of-the-art performance on **standard** few-shot benchmarks
  - **Omniglot** and **mini-ImageNet**

	Omniglot (way, shot)				miniImageNet (way, shot)			
Method	**(5,1)**	**(5,5)**	**(20,1)**	**(20,5)**	**(5,1)**	**(5,5)**	**(5,20)**	**(5,50)**
Training from Scratch	52.50	74.78	24.91	47.62	27.59	38.48	51.53	59.63
Unsupervised meta-learning								
CACTUs-MAML	68.84	87.78	48.09	73.36	39.90	53.97	63.84	69.64
CACTUs-ProtoNets	68.12	83.58	47.75	66.27	39.18	53.36	61.54	63.55
UMTRA	83.80	95.43	74.25	92.12	39.93	50.73	61.11	67.15
LASIUM-MAML	83.26	95.29	-	-	40.19	54.56	65.17	69.13
LASIUM-ProtoNets	80.15	91.10	-	-	40.05	52.53	61.09	64.89
Meta-GMVAE	94.92	97.09	82.21	90.61	42.82	55.73	63.14	68.26
Meta-SVEBM	91.85	97.21	79.66	92.21	43.38	58.03	67.07	72.28
PsCo (Ours)	**96.37**	**99.13**	**89.64**	**97.07**	**46.70**	**63.26**	**72.22**	**73.50**
Self-supervised learning								
SimCLR	92.13	97.06	80.95	91.60	43.35	52.50	61.83	64.85
MoCo v2	92.66	97.38	82.13	92.35	41.92	50.94	60.23	63.45
SwAV	93.13	97.32	82.63	92.12	43.24	52.41	61.36	64.52
Supervised meta-learning								
MAML	94.46	98.83	84.60	96.29	46.81	62.13	71.03	75.54
ProtoNets	98.35	99.58	95.31	98.81	46.56	62.29	70.05	72.04

Experiment: Cross-domain Few-shot classification

- PsCo achieves state-of-the-art performance on **cross-domain** few-shot benchmarks
  - Small-scale experiments (**Conv5** pretrained on **mini-ImageNet**)

(a) Cross-domain few-shot benchmarks similar to miniImageNet.

	CUB		Cars		Places		Plantae	
Method	**(5, 5)**	**(5, 20)**	**(5, 5)**	**(5, 20)**	**(5, 5)**	**(5, 20)**	**(5, 5)**	**(5, 20)**
Unsupervised meta-learning								
Meta-GMVAE	47.48	54.08	31.39	38.36	57.70	65.08	38.27	45.02
Meta-SVEBM	45.50	54.61	34.27	46.23	51.27	61.09	38.12	46.22
PsCo (Ours)	**57.38**	**68.58**	**44.01**	**57.50**	**63.60**	**73.95**	**52.72**	**64.53**
Self-supervised learning								
SimCLR	52.11	61.89	37.40	50.05	60.10	69.93	43.42	54.92
MoCo v2	53.23	62.81	38.65	51.77	59.09	69.08	43.97	55.45
SwAV	51.58	61.38	36.85	50.03	59.57	69.70	42.68	54.03
Supervised meta-learning								
MAML	56.57	64.17	41.17	48.82	60.05	67.54	47.33	54.86
ProtoNets	56.74	65.03	38.98	47.98	59.39	67.77	45.89	54.29

(b) Cross-domain few-shot benchmarks dissimilar to miniImageNet.

	CropDiseases		EuroSAT		ISIC		ChestX	
Method	**(5, 5)**	**(5, 20)**	**(5, 5)**	**(5, 20)**	**(5, 5)**	**(5, 20)**	**(5, 5)**	**(5, 20)**
Unsupervised meta-learning								
Meta-GMVAE	73.56	81.22	73.83	80.11	33.48	39.48	23.23	26.26
Meta-SVEBM	71.82	83.13	70.83	80.21	38.85	48.43	**26.26**	28.91
PsCo (Ours)	**88.24**	**94.95**	**81.08**	**87.65**	**44.00**	**54.59**	24.78	27.69
Self-supervised learning								
SimCLR	79.90	88.73	79.14	85.05	42.83	51.35	25.14	**29.21**
MoCo v2	80.96	89.85	79.94	86.16	43.43	52.14	25.24	29.19
SwAV	80.15	89.24	79.31	85.62	43.21	51.99	24.99	28.57
Supervised meta-learning								
MAML	77.76	83.24	71.48	76.70	47.34	55.09	22.61	24.25
ProtoNets	76.01	83.64	64.91	70.88	40.62	48.38	23.15	25.72

Experiment: Cross-domain Few-shot classification

- PsCo achieves state-of-the-art performance on **cross-domain** few-shot benchmarks
  - Large-scale experiments (**ResNet-50** pretrained on **ImageNet**)

Method	CUB	Cars	Places	Plantae	CropDiseases	EuroSAT	ISIC	ChestX
MoCo v2	64.16	47.67	81.39	61.36	82.89	76.96	38.26	**24.28**
+PsCo (Ours)	**76.63**	**53.45**	**83.87**	**69.17**	**89.85**	**83.99**	**41.64**	23.60
BYOL	67.45	45.74	75.43	56.86	80.82	77.70	37.27	24.15
+PsCo (Ours)	**82.13**	**56.19**	**83.80**	**71.14**	**92.92**	**85.33**	**42.90**	**26.05**
Supervised	89.13	75.15	84.41	72.91	90.96	85.64	43.34	25.35

(a) Cross-domain few-shot benchmarks similar to miniImageNet.

Method								
Meta-GMVAE								
Meta-SVEBM								
PsCo (Ours)								
SimCLR								
MoCo v2								
SwAV								
Supervised meta-learning								
MAML	56.57	64.17	41.17	48.82	60.05	67.54	47.33	54.86
ProtoNets	56.74	65.03	38.98	47.98	59.39	67.77	45.89	54.29

(b) Cross-domain few-shot benchmarks dissimilar to miniImageNet.

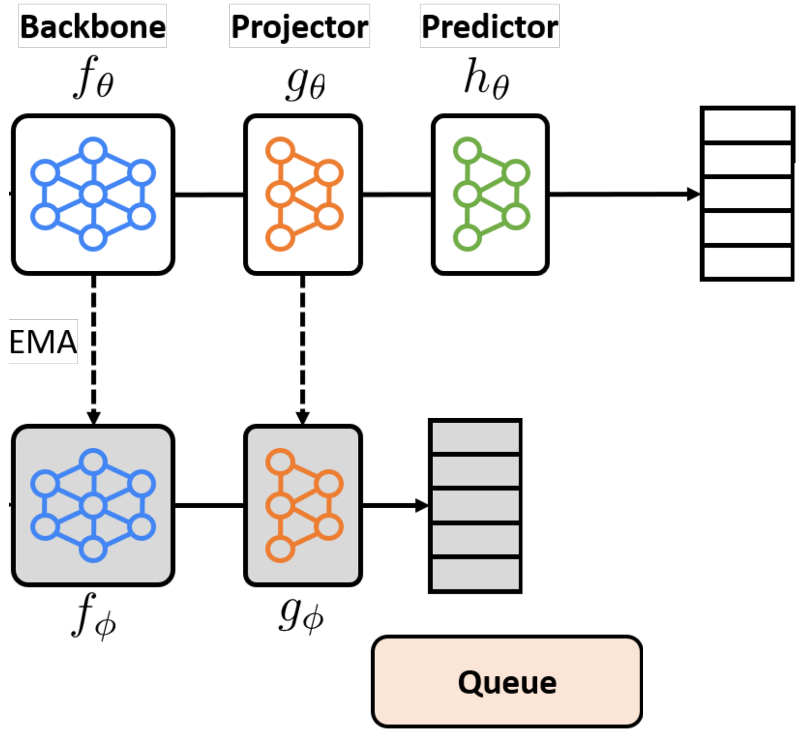
Method							ChestX (5, 5)	ChestX (5, 20)
							23.23	26.26
							26.26	28.91
							24.78	27.69
							25.14	**29.21**
							25.24	29.19
							24.99	28.57
Supervised meta-learning								
MAML	77.76	83.24	71.48	76.70	47.34	55.09	22.61	24.25
ProtoNets	76.01	83.64	64.91	70.88	40.62	48.38	23.15	25.72

Ablation studies

- All components are meaningful
  - **Architecture choices:** Momentum network & predictor enhances pseudo-labeling quality online

Table 4: Component ablation studies on Omniglot.

Momentum	Predictor	Sinkhorn	Top-K sampling	$\mathcal{L}_{\texttt{MoCo}}$	(5, 1)	(5, 5)	(20, 1)	(20, 5)
✓	✓	✓	✓	✓	**96.37**	**99.13**	**89.64**	**97.07**
✗	✓	✓	✓	✓	90.32	96.78	76.17	90.41
✓	✗	✓	✓	✓	90.21	96.86	76.15	90.53

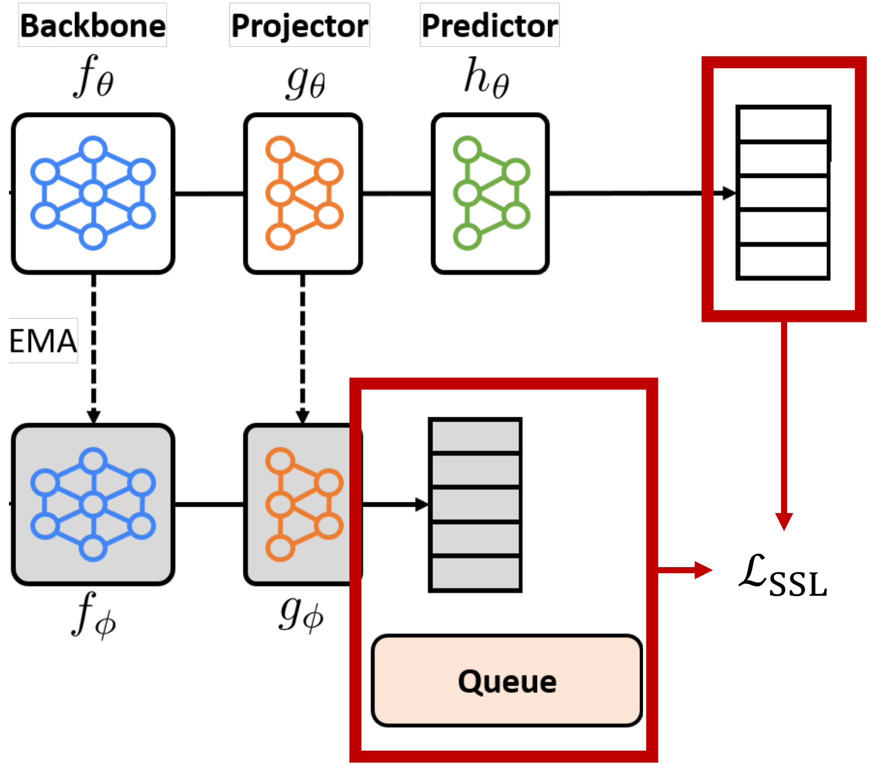


Ablation studies

- All components are meaningful
  - **Architecture choices:** Momentum network & predictor enhances pseudo-labeling quality online
  - Incorporating **loss of self-supervised learning** without additional cost helps to get better representation

Table 4: Component ablation studies on Omniglot.

Momentum	Predictor	Sinkhorn	Top-K sampling	$\mathcal{L}_{\texttt{MoCo}}$	(5, 1)	(5, 5)	(20, 1)	(20, 5)
✓	✓	✓	✓	✓	**96.37**	**99.13**	**89.64**	**97.07**
✗	✓	✓	✓	✓	90.32	96.78	76.17	90.41
✓	✗	✓	✓	✓	90.21	96.86	76.15	90.53
✓	✓	✓	✓	✗	93.16	97.40	81.03	91.33

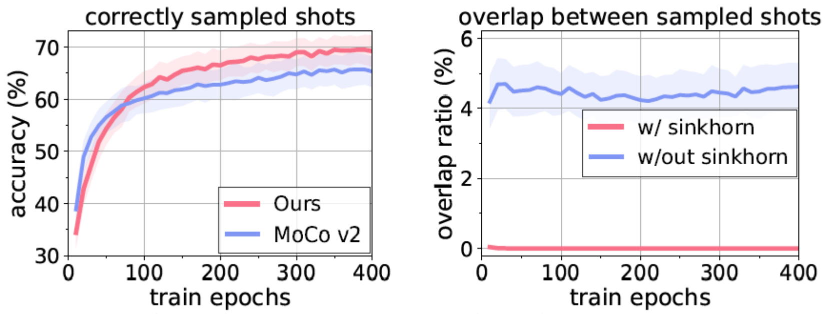


Ablation studies

- All components are meaningful
  - **Architecture choices:** Momentum network & predictor enhances pseudo-labeling quality online
  - Incorporating **loss of self-supervised learning** without additional cost helps to get better representation
  - **Sampling strategy:** Sinkhorn-Knopp & Top-K sampling helps to sample proper few-shot tasks

Table 4: Component ablation studies on Omniglot.

Momentum	Predictor	Sinkhorn	Top-K sampling	$\mathcal{L}_{\texttt{MoCo}}$	(5, 1)	(5, 5)	(20, 1)	(20, 5)
✓	✓	✓	✓	✓	**96.37**	**99.13**	**89.64**	**97.07**
✗	✓	✓	✓	✓	90.32	96.78	76.17	90.41
✓	✗	✓	✓	✓	90.21	96.86	76.15	90.53
✓	✓	✗	✓	✓	95.81	98.94	88.25	96.57
✓	✓	✓	✗	✓	94.95	98.81	86.32	96.05
✓	✓	✓	✓	✗	93.16	97.40	81.03	91.33



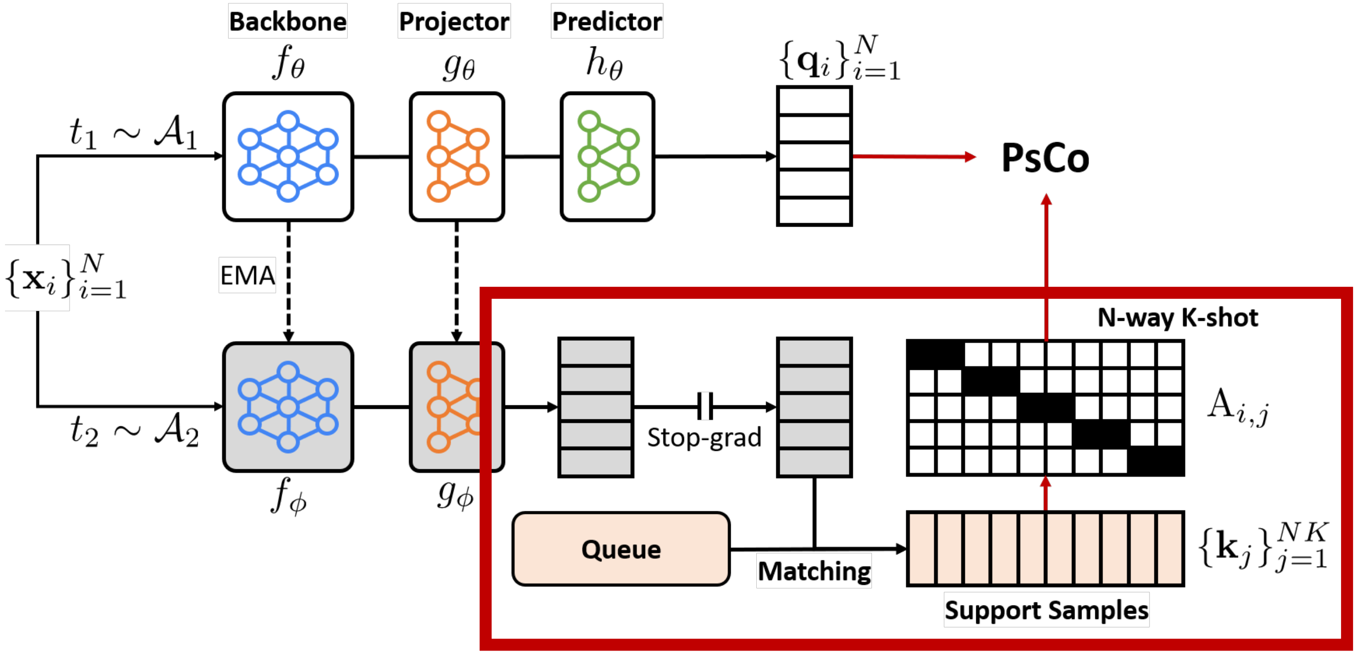
(a) Pseudo-label quality
(b) Shot overlap ratio

Ablation studies

- All components are meaningful
  - **Weak augmentation for $\mathcal{A}_2$** helps to find an accurate pseudo-label assignment matrix $\mathbf{A}$

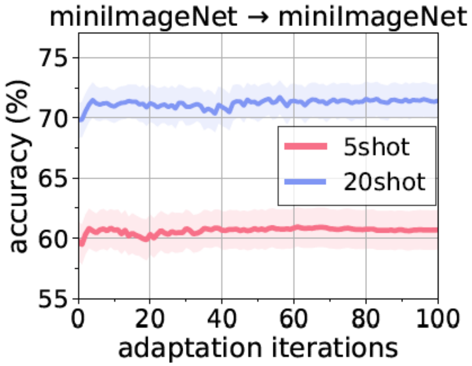
Table 5: The ablation study with varying augmentation choices for $\mathcal{A}_1$ and $\mathcal{A}_2$ on miniImageNet.

$\mathcal{A}_1$	$\mathcal{A}_2$	(5, 1)	(5, 5)	(5, 20)	(5, 50)
Strong	Strong	44.54	60.04	68.61	71.20
Strong	Weak	**46.70**	**63.26**	**72.22**	**73.50**
Weak	Strong	43.56	58.77	67.21	69.46
Weak	Weak	40.68	55.05	63.32	65.82

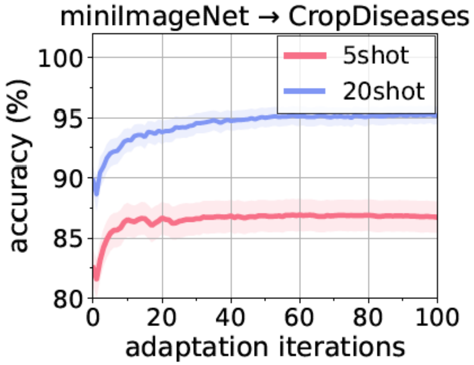


Ablation studies

- New **adaptation scheme** is more **useful in cross-domain**
  - It does not cause over-fitting by optimizing only the projector g_θ and the predictor h_θ



(c) In-domain adaptation



(d) Cross-domain adaptation

Table 11: Before and after adaptation of PsCo in few-shot classification.

Adaptation	miniImageNet	CUB	Cars	Places	Plantae	CropDiseases	EuroSAT	ISIC	ChestX
				5-way 5-shot					
✗	63.26	55.15	42.27	62.98	48.31	79.75	74.73	41.18	24.54
✓	**63.30**	**57.38**	**44.01**	**63.60**	**52.72**	**88.24**	**81.08**	**44.00**	**24.78**
				5-way 20-shot					
✗	72.22	62.35	51.02	70.85	55.91	84.72	78.96	48.53	27.60
✓	**73.00**	**68.58**	**57.50**	**73.95**	**64.53**	**94.95**	**87.65**	**54.59**	**27.69**

Conclusion

We propose **PsCo:** an effective unsupervised meta-learning method for few-shot classification

- PsCo constructs diverse few-shot pseudo-tasks without labels
  utilizing the momentum network and the queue of previous batches in a progressive manner

- We demonstrate the effectiveness of PsCo under various few-shot classification benchmarks
  - PsCo achieves state-of-the-art performance on standard few-shot classification benchmarks
  - PsCo shows superiority on cross-domain few-shot classification benchmarks
  - PsCo is applicable to a large-scale dataset

Thank you for your attention!