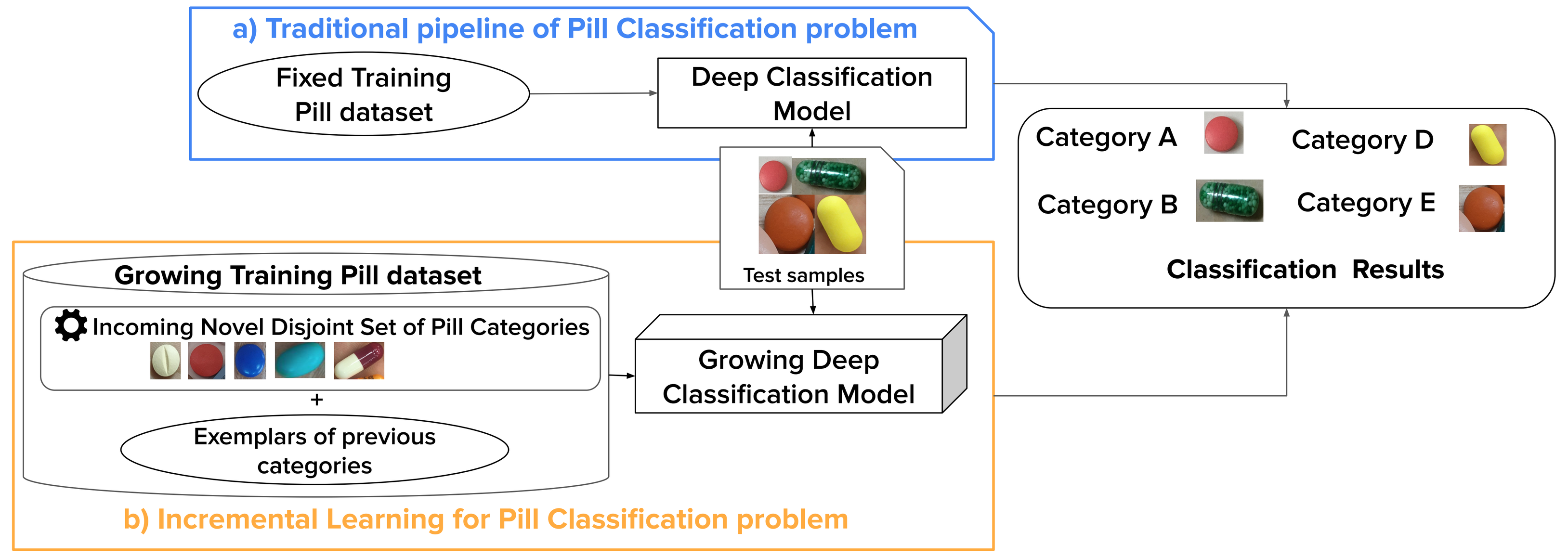
ACCV 2022 Macau
4-8 DECEMBER 2022
Macau SAR, China

Multi-stream Fusion for Class Incremental Learning in Pill Image Classification

[1, 2] Trong-Tung Nguyen, [2] Huy-Hieu Pham, [3] Phi-Le Nguyen, [3] Thanh-Hung Nguyen, [2, 4] Minh N. Do

[1] John von Neumann Institute, University of Science, VNU-HCM, Vietnam, [2] VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam,
[3] Hanoi University of Science and Technology, Vietnam [4] University of Illinois at Urbana-Champaign, United States

Introduction



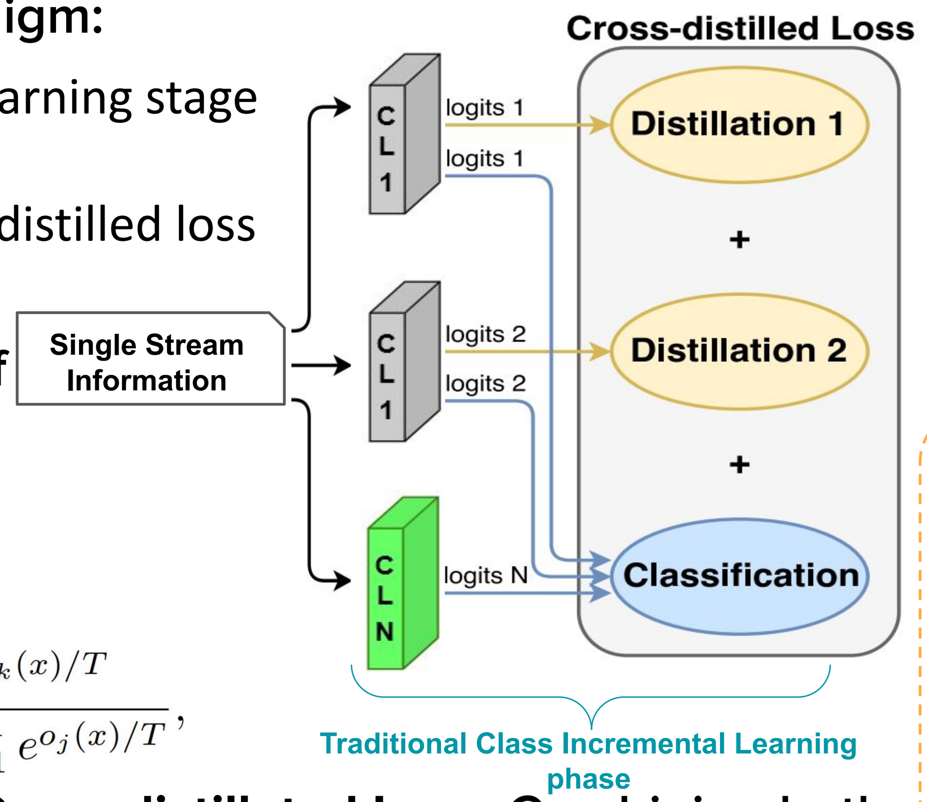
Towards a flexible learning paradigm for pill classification problem, we contribute to:

1. Introduce a novel IL framework for incorporating additional stream of information via fusion techniques.
2. Extend our framework to perform in-depth studies of stream utilization and fusion mechanisms.
3. Achieve state-of-art result in various task settings for IL-based pill classification problem.

3/ Single stream exemplar-based IL learning paradigm (baseline):

A simple baseline approach for Pill IL learning is to take only the original RGB image stream as input to the common IL learning paradigm:

- Each CL_i is responsible for a set of categories at learning stage i^{th}.
- Model weights are optimized based on the cross-distilled loss function computed as follows:



Traditional Class Incremental Learning phase

IL-based distillation Loss: Distilling knowledge of old model M_{i-1} to updated model M_i.

$$L_d = \sum_{x \in V^t_{train}} \sum_{k=1}^{t-1} -\hat{\pi}_k(x) \log\left[\pi_k(x)\right],$$

$$\hat{\pi}_k(x) = \frac{e^{\hat{o}_k(x)/T}}{\sum_{j=1}^{t-1} e^{\hat{o}_j(x)/T}}, \quad \pi_k(x) = \frac{e^{o_k(x)/T}}{\sum_{j=1}^{t-1} e^{o_j(x)/T}},$$

IL-based cross-entropy Loss: Learning current category set of categories for task CL_i.

$$L_c = \sum_{x \in V^t_{train}} \sum_{k=1}^{t} -y(x) \log\left[p_k(x)\right],$$

Cross-distillated Loss: Combining both losses with weighted parameter α controlling the balance.

$$L = \alpha L_d + (1-\alpha) L_c$$

Approach

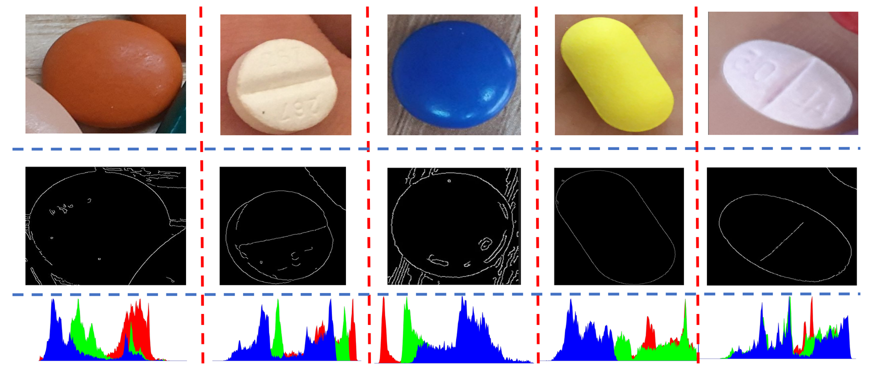
1/ Multi-stream Class Incremental Learning Framework:

M = Base method **X** + Feature stream **Y** + Fusion mechanism **Z**

- Base method X: Represents for any CIL exemplar-based methods.
- Feature Stream Y: Additionally discriminative pill features.
- Fusion mechanism Z: Techniques of fusing multiple streams.

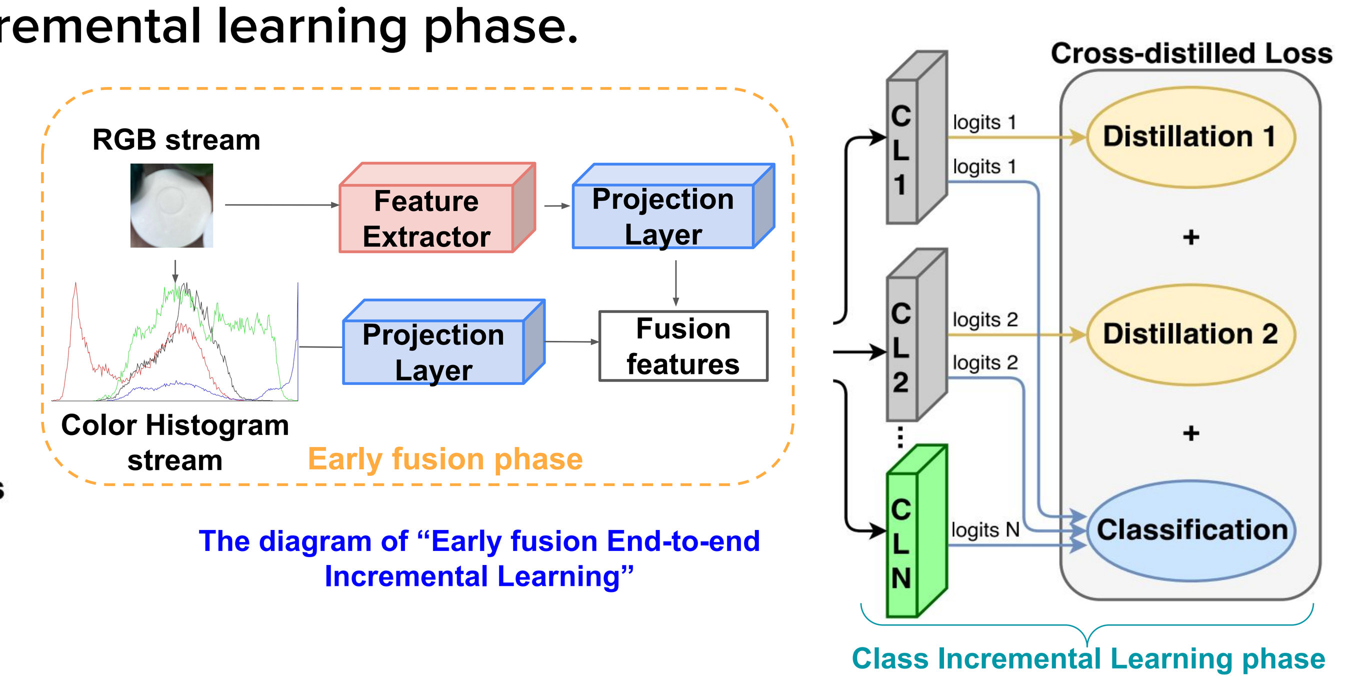
2/ Additional Stream Usage:

We observed that there could be multiple pill features which are potential to help the incremental learning model more effectively discriminate hard pill instances. Those are color histogram and edge signals of pill instances which are shown below:

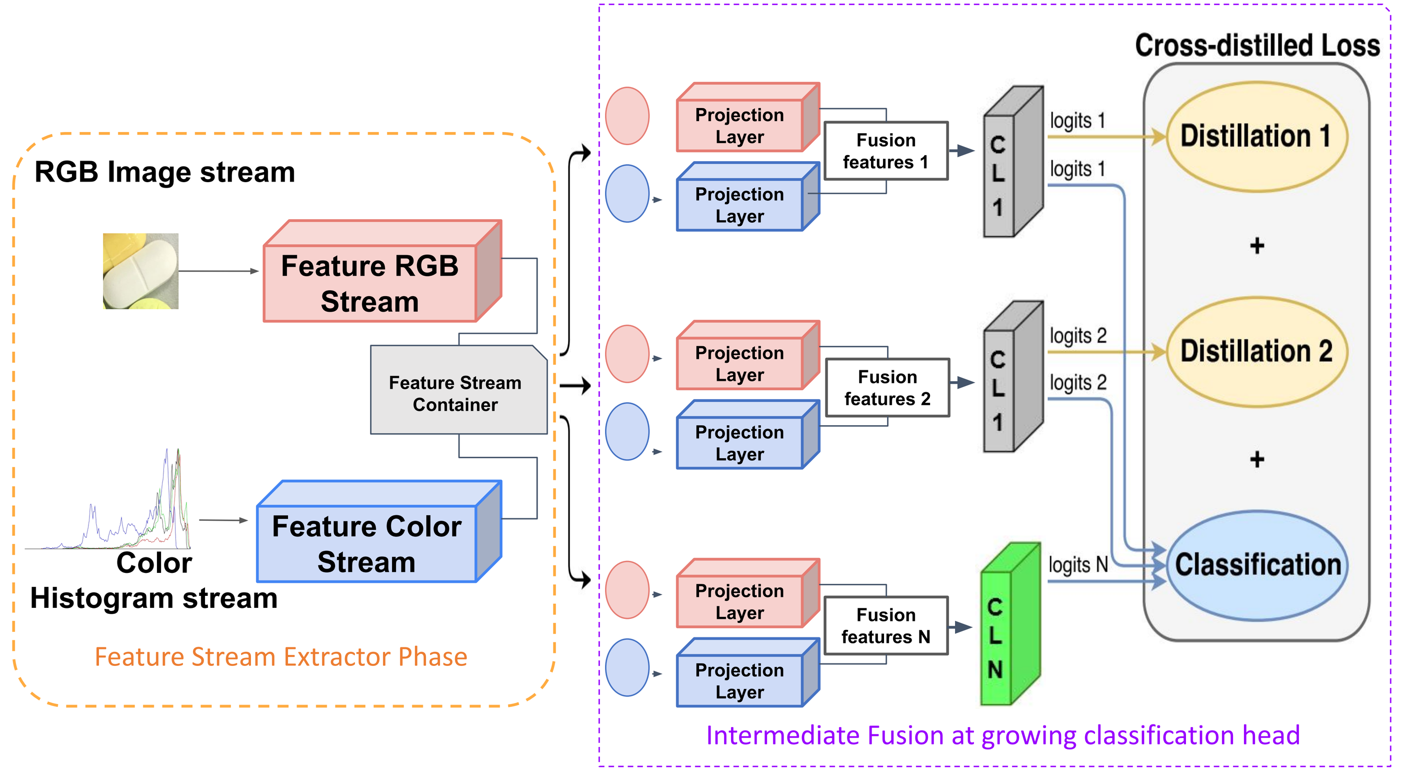


4/ Incorporation of fusion techniques with stream information:

Early fusion mechanism:
Stream information are aligned prior to the incremental learning phase.



The diagram of "Early fusion End-to-end Incremental Learning"



Intermediate fusion mechanism:

- The bottleneck of early fusion lies in the pre-incremental fusion layer. This harness the flexibility of learning features for independent CIL task.
- We propose to relocate the fusion layer to the incremental layer.
- In this way, the learning features are separated for independent task which is beneficial for learning optimal features for each task.
- At the incremental learning phase, each projection layer is designed to help features from different stream explore appropriate embedding space which could be aligned later.

Results and Conclusion

1/ Experimental Protocol:

Task-agnostic Setting

Incremental accuracy: measure how well model performs at each task setting.

Incremental Forgetting Rate: measure how much model forgets at each task setting.

2/ Dataset

VAIPE-Pill: Our collected dataset with 7,294 images of 262 categories from different scenarios in real-world setting.

Characteristic	Training set	Testing set	Total
Number of images	6,461	833	7,294
Number of pill categories	262	262	262
Instances per category	179.75	23.56	203.2
Image size (pixel×pixel, mean)	3,311 × 3,276	3,276 × 3,469	3,300 × 3,40
Instances per image	7.28	7.4	7.3
Number of bounding box annotations	47,097	6,174	53,271
Number of categories per image	5.18	5.76	5.32

3/ Experimental Results:

Average accuracy and forgetting rate

Metric	Method	Task Settings		
		N=5	N=10	N=15
Average acc. (%) ↑ $\bar{\mathcal{A}} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{A}_i$	EEIL [4]	63.83	62.40	57.41
	EEIL-CG-IMIF	**70.80**	**64.85**	**60.93**
	BiC [22]	53.83	55.75	53.77
	BiC-CG-IMIF	**65.53**	**63.59**	**54.83**
	LUCIR [13]	69.63	62.90	55.49
	LUCIR-CG-IMIF	**76.85**	**69.94**	**64.97**
Forgetting rate. (%) ↓ $\bar{\mathcal{F}} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{F}_i$	EEIL [4]	49.82	45.46	48.27
	EEIL-CG-IMIF	**46.68**	**44.64**	**46.23**
	BiC [22]	20.05	30.50	26.93
	BiC-CG-IMIF	**7.75**	**22.01**	**27.35**
	LUCIR [13]	44.13	44.32	47.11
	LUCIR-CG-IMIF	**33.15**	**37.88**	**39.79**

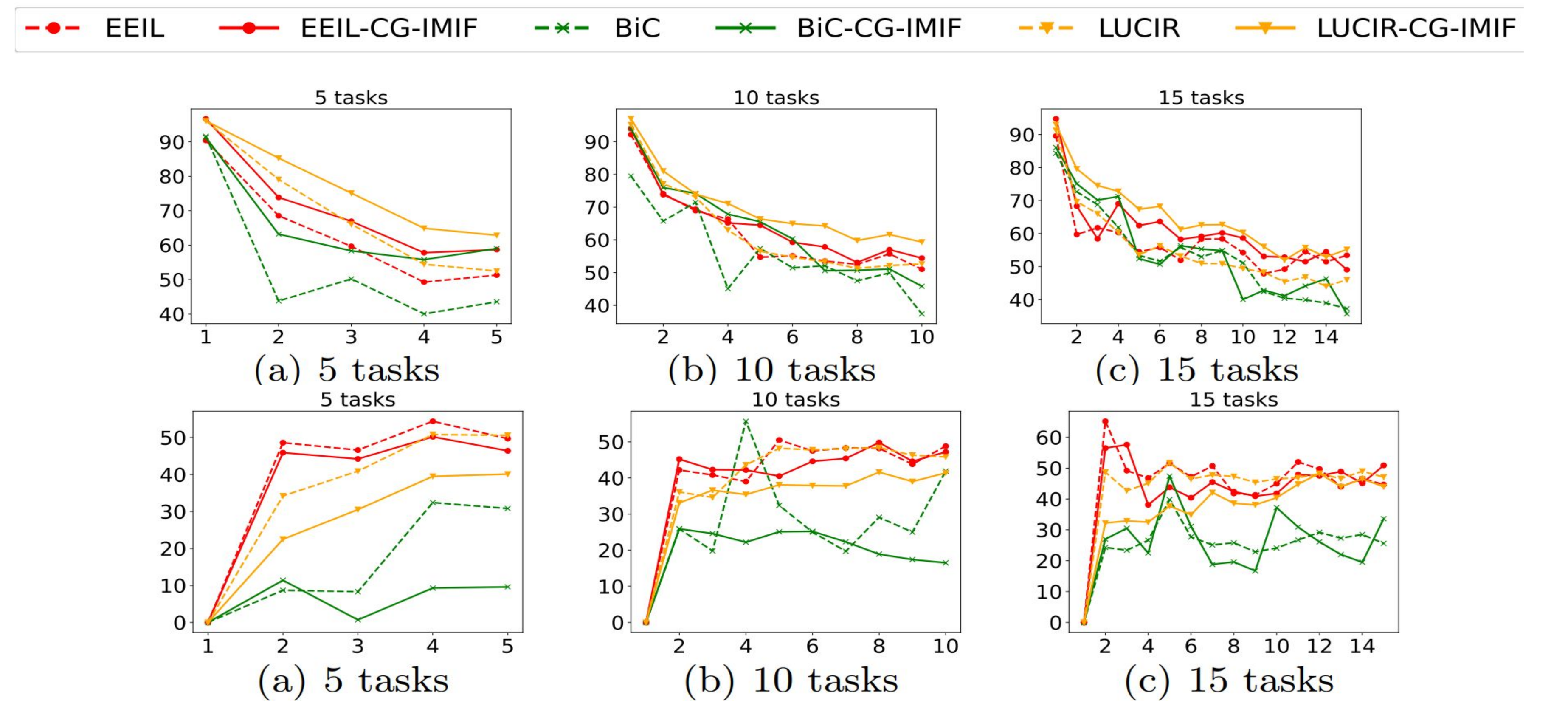
◇ Using the similar exemplar settings and selection for fair comparison.

4/ Ablation Study

Variant of Combination	Average acc. (%) ↑			Forgetting rate. (%) ↓		
	N=5	N=10	N=15	N=5	N=10	N=15
RGB only	69.93	62.90	55.49	44.13	44.32	47.1
RGB-Edge + Early	70.94	63.90	55.28	42.4	42.13	45.80
RGB-Edge + Intermediate	**72.58**	**68.38**	**62.90**	**38.78**	**38.19**	**41.02**
RGB-Color + Early	73.58	64.57	53.56	37.825	42.86	46.15
RGB-Color+ Intermediate	76.85	69.94	64.97	33.15	37.88	39.79
RGB-Edge-Color + Early	69.99	63.33	56.34	42.35	44.17	46.24
RGB-Edge-Color+ Intermediate	**73.65**	**68.32**	**62.15**	**36.30**	**38.48**	**40.58**

3/ Experimental Results (cnt):

Incremental accuracy and forgetting rate



Conclusion:

- We introduced an incremental learning multi-stream framework for pill classification task.
- We explored and proposed an efficient intermediate fusion mechanism which achieve state-of-the-art result across different task settings.
- Our framework is flexible and could be extended to include additional information stream as well as base IL methods.