CLIFF: Continual Learning for Incremental Flake Features in 2D Material Identification

Sankalp Pandey[1], **Xuan Bac Nguyen**[1],
Nicholas Borys[2], **Hugh Churchill**[3], **Khoa Luu**[1]
[1]Department of Electrical Engineering and Computer Science, University of Arkansas, AR
[2]Department of Physics and Astronomy, University of Utah, UT
[3]Department of Physics, University of Arkansas, AR
{sankalpp, xnguyen, hchurch, khoaluu}@uark.edu
{nicholas.borys}@utah.edu

Abstract

Identifying quantum flakes is crucial for scalable quantum hardware; however, automated layer classification from optical microscopy remains challenging due to substantial appearance shifts across different materials. In this paper, we propose a new Continual-Learning Framework for Flake Layer Classification (CLIFF) [1]. To our knowledge, this is the first systematic study of continual learning in the domain of two-dimensional (2D) materials. Our method enables the model to differentiate between materials and their physical and optical properties by freezing a backbone and base head trained on a reference material. For each new material, it learns a material-specific prompt, embedding, and a delta head. A prompt pool and a cosine-similarity gate modulate features and compute material-specific corrections. Additionally, we incorporate memory replay with knowledge distillation. CLIFF achieves competitive accuracy with significantly lower forgetting than naive fine-tuning and a prompt-based baseline.

1 Introduction

Characterizing the layer counts of two-dimensional (2D) material flakes is crucial for the fabrication of van der Waals heterostructures, which facilitate a range of studies and applications, especially those in quantum mechanics [11, 14, 16]. Typically, researchers find flakes through repetitive, manual optical microscopy searches, and the samples must be transferred to an Atomic Force Microscope (AFM) for thickness measurement, which doubles the manual effort involved and significantly limits the complexity and scalability of heterostructure construction. Deep learning approaches aim to automate flake layer classification of exfoliated 2D materials but present poor versatility. In this work, we present, to our knowledge, the first systematic study of continual learning for 2D flake thickness classification. We establish our evaluation as a material-incremental, continual learning problem setting, where new materials arrive sequentially.

The Challenges of Automated Flake Identification. The difficulty in automating the identification of 2D materials is estimating the layer count of a flake from optical microscopy images. Importantly, the subtle visual characteristics for layer count classification are highly dependent on factors that vary in real-world laboratory settings. This variability makes it challenging to train a typical deep learning model, motivating the use of a continual learning approach to preserve learned information.

Limitations of Prior Work. Prior work in automated flake characterization has several limitations. Traditional machine learning methods [10, 7] often fail to capture the subtle, non-linear visual

[1]The code is available at https://github.com/uark-cviu/quantumflake.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

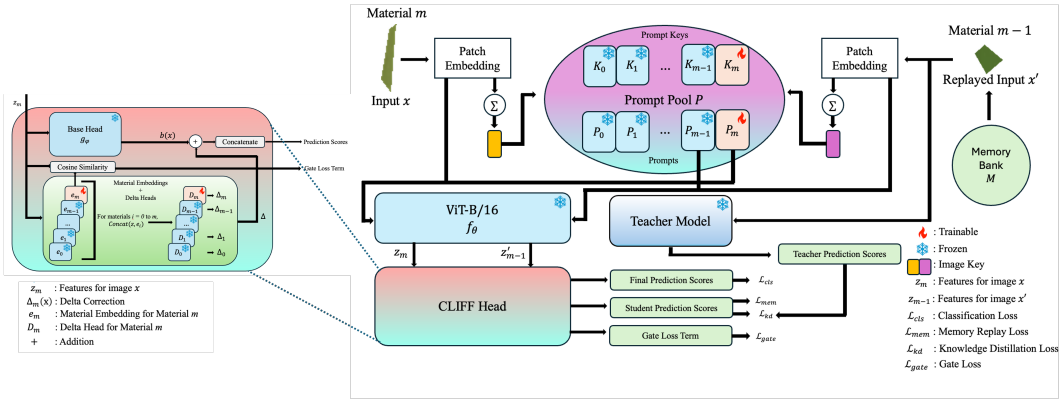

Figure 1: The proposed CLIFF approach.

differences between flake thicknesses. While deep learning approaches have shown promise [24, 9, 19, 13, 22, 18], they often overfit to specific materials or imaging conditions. When typical static models are fine-tuned on new materials, they suffer from catastrophic forgetting, rendering them impractical for real-world laboratory environments. In CL, forgetting is mitigated by methods such as rehearsal-based approaches [8, 4, 5], knowledge distillation [3, 1, 23, 28], self-supervision [21, 2, 20], and architectural methods that add new parameters [29, 15, 27, 12]. Our work takes inspiration from prompt-based CL [26, 25].

Problem Motivation. Addressing these limitations is critical for developing robust automated systems. Although training on all materials is feasible on a small, fixed set of materials and can provide strong performance, it is not a practical strategy for a real-world scientific workflow. In laboratory settings, new materials may be introduced over time, and retraining a large model from scratch with each new dataset can be computationally heavy, especially for larger datasets. Similarly, domain adaptation methods are relevant for handling distribution shifts, but they are typically designed for a single source and target domain. This can be insufficient for a growing sequence of new material domains encountered in a laboratory setting. Therefore, we formulate flake layer count classification as a material-incremental continual learning problem where the model must learn new materials while retaining performance on previously seen materials. To address this, we propose a continual learning (CL) framework that preserves prior knowledge by freezing a backbone and base head trained on a reference material and learning a per-material addition for each new material.

Contributions of this Work. We present, to the best of our knowledge, the first study of continual learning for 2D material flake thickness classification and formulate a material-incremental benchmark. Our proposed CLIFF approach is a novel continual learning framework for 2D material flake layer count classification across multiple materials. We evaluate our approach and compare it against joint training, naive fine-tuning, and another Learning-to-Prompt (L2P) method [26].

2 Methodology

We propose CLIFF, a continual learning framework for 2D flake layer classification. For each subsequent material, the framework learns a small set of new components: a dedicated prompt pool, a material embedding, and a delta head that models a material-specific correction. A prompt pool adapts the frozen backbone's features during training on a new material by prepending learned tokens to the input patch sequence. We also replay a small number of stored samples and use knowledge distillation. The final classification head computes predictions for all seen materials in parallel, guided by an auxiliary loss on material identity. This allows our approach to perform task-agnostic classification without needing material labels at test time.

2.1 Base Training on a Reference Material

Let $D_0 = \{(x_i, y_i)\}_{i=1}^{N_0}$ be the reference material dataset, where x_i represents the i-th input image and $y_i \in \{\text{Few}, \text{Mono}, \text{Thick}\}$ is the corresponding thickness label. A Vision Transformer (ViT) [6]

backbone f_θ with parameters θ and a linear classification head g_ϕ with parameters ϕ are trained by optimizing the base loss function shown in Eqn. (1) as follow,

$$\mathcal{L}_{base} = \frac{1}{N_0} \sum_{i=1}^{N_0} \mathrm{CE}\big(b(x_i),\, y_i\big) = \frac{1}{N_0} \sum_{i=1}^{N_0} \mathrm{CE}\big(g_\phi(f_\theta(x_i)),\, y_i\big), \tag{1}$$

where $\mathrm{CE}(\cdot, \cdot)$ denotes the cross-entropy loss function and $b(x_i) = g_\phi(f_\theta(x_i))$ represents the base classification logits for image x_i.

2.2 Incremental Learning for New Materials

When a new material m arrives, we introduce three new learnable components: a prompt P_m, a material embedding e_m, and a delta head D_m.

Prompting. For each material m, we learn a separate prompt pool P_m. Each is initialized as a new set of learnable parameters, consisting of prompt tokens and their corresponding keys, with values drawn from a random uniform distribution. During training, the model selects a set of these prompt tokens by choosing the top-k tokens according to cosine similarity between the input's CLS embedding and the prompt keys and prepends them to the sequence of image patch embeddings The prompt tokens and their corresponding keys are optimized via backpropagation based on the final task loss. We denote the feature output from the prompted backbone as Eqn. (2):

$$z_m = f_\theta(x; P_m). \tag{2}$$

CLIFF Head. The CLIFF head processes the prompted features $z_m \in \mathbb{R}^d$, where d is the backbone's feature dimension. It maintains an embedding table $E \in \mathbb{R}^{M \times d_e}$ containing a unique embedding vector e_i for each of the M seen materials, and d_e is the embedding dimension. At inference, CLIFF evaluates all material-specific delta heads in parallel. As described in Eqn. (3), for each material i, a multilayer perceptron (MLP), the delta head $D_i : \mathbb{R}^{d+d_e} \to \mathbb{R}^C$, computes a residual correction $\Delta_i(x) \in \mathbb{R}^C$:

$$\Delta_i(x) = D_i\big(\mathrm{Concat}[z_m,\, e_i]\big). \tag{3}$$

Here, $C = 3$ is the number of thickness classes. The final output is a single, large logit vector $L(x)$ created by concatenating the corrected logits for all M materials, as calculated using Eqn (4):

$$L(x) = \mathrm{Concat}_{i=1}^{M}\big(b(x) + \Delta_i(x)\big) \in \mathbb{R}^{CM}. \tag{4}$$

Optimization with Rehearsal and Distillation. With f_θ and g_ϕ frozen, we learn the new components (P_m, e_m, D_m) for a new material. We maintain a small memory buffer $\mathcal{M}$ of samples from past tasks. The total loss for a sample (x, y) from the current task m is calculated as shown in Eqn. (5).

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{gate}}\mathcal{L}_{\text{gate}} + \lambda_{\text{mem}}\mathcal{L}_{\text{mem}} + \lambda_{\text{kd}}\mathcal{L}_{\text{kd}}. \tag{5}$$

Here, $\mathcal{L}_{\text{cls}}$ is the standard cross-entropy loss on the current task's data. An auxiliary gate loss, $\mathcal{L}_{\text{gate}}$, is computed using cosine similarity between features and material embeddings to improve material identification. The cosine-similarity gate supervises material awareness/prompt selection but does not mask or disable any heads at inference. $\mathcal{L}_{\text{mem}}$ is the cross-entropy loss for samples replayed from the memory buffer $\mathcal{M}$. Finally, $\mathcal{L}_{\text{kd}}$ is a knowledge distillation loss on replayed samples that aligns the current model's outputs with those of a frozen "teacher" model from the previous task.

2.3 Complexity

The total number of parameters increases linearly with the number of materials M.The approximate per-material parameter count is

$$\underbrace{KLd}_{\text{prompt tokens}} + \underbrace{Kd}_{\text{prompt keys}} + \underbrace{d_e}_{\text{embedding}} + \underbrace{(d+d_e)h + hC}_{\text{delta head}}, \tag{6}$$

where K is the prompt pool size, L is the prompt length, d is the backbone's feature dimension, d_e is the embedding dimension, h is the delta head's hidden dimension, and C is the number of classes. The total parameter count scales as $O(M)$. At inference, evaluating all heads incurs an $O(M)$ computational cost per image. For rehearsal, a memory buffer that stores n RGB images per class at a resolution of 224×224 requires approximately $M \cdot C \cdot n \cdot 224^2 \cdot 3$ bytes of storage.

Table 1: Sequential task performance on four materials: BN (T1), Graphene (T2), MoS_2 (T3), and WTe_2 (T4). We report per-task accuracy, final average accuracy, and forgetting.

	Trained on	Tested on (Accuracy %)			
		T1	**T2**	**T3**	**T4**
Joint Training	**Ensemble**	92.68	92.04	90.91	92.82
	Summary	Avg. Accuracy: 92.11%			
Naive Fine-tuning	**T1**	91.46	-	-	-
	T2	10.98	86.09	-	-
	T3	8.54	22.23	83.77	-
	T4	4.88	3.20	0.65	62.68
	Summary	Avg. Accuracy: 17.85%		Forgetting: 84.20%	
L2P [26]	**T1**	85.37	-	-	-
	T2	60.98	82.34	-	-
	T3	57.32	52.88	87.66	-
	T4	53.66	21.23	1.30	71.77
	Summary	Avg. Accuracy: 36.99%		Forgetting: 59.73%	
Ours	**T1**	90.24	-	-	-
	T2	86.59	79.33	-	-
	T3	64.63	77.70	79.87	-
	T4	56.10	44.79	44.16	82.78
	Summary	**Avg. Accuracy: 56.96%**		**Forgetting: 34.80%**	

3 Experimental Setups and Implementation Details

Datasets. For our study, we use the dataset from Masubuchi *et al.*[17]. We address material-incremental layer classification over four materials: BN (base), graphene, MoS_2, and WTe_2. Each image x is labeled as Few, Mono, or Thick with its material type (e.g., `Mono_BN`). For training, we use standard augmentations such as random horizontal and vertical flips, rotations, and color jittering.

Evaluation Protocol. We train and evaluate using the following materials in order: BN (T1), graphene (T2), MoS_2 (T3), and WTe_2 (T4). We report per-material accuracy at each step, final macro-average accuracy, and forgetting, which is the average drop from each material's peak accuracy to its final accuracy.

Implementation Details. All experiments are performed using a Vision Transformer backbone (ViT-B/16). We train for 15 epochs per task with a batch size of 32, using an Adam optimizer with a learning rate of 5×10^{-5} for the delta heads and 1×10^{-4} for the prompts. We use a memory bank of 100 samples per class, $\lambda_{kd} = 1.0$, 128-dimensional material embeddings, and 30 prompts of length 8 per material. We evaluate our approach against three baselines: (1) Joint Training, an upper-bound model trained with data from all four materials simultaneously rather than sequentially; (2) Naive Fine-tuning, a sequential strategy that updates the full model for each new task; and (3) L2P (Learning to Prompt), a prompt-based continual learning method that keeps the backbone frozen while learning a shared pool of prompts.

Method Comparisons. Table 1 summarizes quantitative performance. Joint training serves as an upper bound. Naive fine-tuning suffers from severe catastrophic forgetting, while L2P shows substantial forgetting despite improving on the naive baseline. In contrast, CLIFF yields a significantly higher final average accuracy and the lowest forgetting.

3.1 Ablation Studies

Our ablation experiments study the contribution of each key component and its configurations. We explore the removal of primary components in Table 2 to investigate their contributions. The removal of memory replay and knowledge distillation results in a massive drop in performance, showing that rehearsal is the core mechanism for retaining knowledge. Removing only prompts leads to

Table 2: Effectiveness of CLIFF on the material-incremental benchmark with different configurations.

Memory	KD	Prompts	Avg. Acc. (%)	Forgetting (%)
		✓	18.80	79.64
✓	✓		57.44	32.78
✓	✓	✓	56.95	34.80

a slightly lower forgetting and a higher average accuracy. This result reveals a plasticity-stability trade-off, where prompts allow higher peak accuracy on new tasks (e.g., 79.87% on Task 3 vs. 75.97% without prompts), while the CLIFF Head provides strong underlying stability, resulting in slightly lower overall forgetting when it operates alone. Additionally, the impact of the memory buffer size is quantitatively shown by Table 3. Performance degrades as the buffer shrinks, but even a small memory of 20 samples per class provides a substantial benefit over having no memory at all.

Table 3: Impact of memory buffer size on final average accuracy and forgetting.

Mem. Size	Avg. Acc. (%)	Forgetting (%)
0	18.80	79.64
20	31.47	56.12
40	35.95	53.12
60	42.56	46.38
80	49.85	37.64
100	56.95	34.80

4 Conclusion

This paper has introduced CLIFF, a continual learning framework for 2D material layer classification that adapts to new materials by learning material-specific information while retaining prior knowledge through memory rehearsal with knowledge distillation. The strong performance of CLIFF makes it a valuable tool for practical laboratory use. It is well-suited to accelerate the identification of promising flakes across different materials, significantly reducing the manual effort required for expert verification. Our experiments have demonstrated that CLIFF substantially improves average accuracy and reduces forgetting compared to naive fine-tuning and a strong prompt-based baseline. This work has presented the first systematic study of continual learning for this problem, bridging the gap between current deep learning models and the practical needs of real-world laboratories.

Limitations: Although CLIFF significantly reduces catastrophic forgetting and improves the applicability of continual learning for real-world material science, a concern with this architecture is scalability. CLIFF adds new parameters for each new material and computes corrections for all seen materials at inference time. While this proves to be feasible at the current scope of this work, the linear growth in parameters could lead to heavy computation across hundreds of materials. Additionally, performance is dependent on the memory buffer, which introduces a storage overhead and assumes replayed samples are sufficient to account for feature space shifts. Future work could investigate parameter-sharing techniques to address scalability, explicit feature alignment to handle feature space shifts, and knowledge transfer between optically similar materials to reduce data dependency.

Broader Impacts: CLIFF improves the practicality and usability of automated quantum flake characterization. By enabling deep learning models to adapt to new materials efficiently without forgetting old knowledge, this approach accelerates the pace of discovery in 2D material analysis and promotes the broader application of robust AI systems in scientific research.

Acknowledgments. This work is partly supported by MonArk NSF Quantum Foundry, supported by the National Science Foundation Q-AMASE-i program under NSF award No. DMR-1906383. It acknowledges the Arkansas High-Performance Computing Center for providing GPUs.

References

[1] P. Buzzega, M. Boschini, A. Porrello, D. Abati, and S. Calderara. Dark experience for general continual learning: a strong, simple baseline, 2020.

[2] H. Cha, J. Lee, and J. Shin. Co^{2}l: Contrastive continual learning, 2021.

[3] A. Chaudhry, A. Gordo, P. K. Dokania, P. Torr, and D. Lopez-Paz. Using hindsight to anchor past knowledge in continual learning, 2021.

[4] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny. Efficient lifelong learning with a-gem, 2019.

[5] A. Chaudhry, M. Rohrbach, M. Elhoseiny, T. Ajanthan, P. K. Dokania, P. H. S. Torr, and M. Ranzato. On tiny episodic memories in continual learning, 2019.

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[7] E. Fix and J. Hodges. *Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties*. USAF School of Aviation Medicine, 1951.

[8] T. L. Hayes, N. D. Cahill, and C. Kanan. Memory efficient experience replay for streaming learning, 2019.

[9] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[10] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[11] K. James Singh, T. Ahmed, P. Gautam, A. S. Sadhu, D.-H. Lien, S.-C. Chen, Y.-L. Chueh, and H.-C. Kuo. Recent advances in two-dimensional quantum dots and their applications. *Nanomaterials*, 11(6):1549, 2021.

[12] Z. Ke, B. Liu, and X. Huang. Continual learning of a mixed sequence of similar and dissimilar tasks, 2021.

[13] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[14] M. C. Lemme, D. Akinwande, C. Huyghebaert, and C. Stampfer. 2d materials for future heterogeneous electronics. *Nature communications*, 13(1):1392, 2022.

[15] X. Li, Y. Zhou, T. Wu, R. Socher, and C. Xiong. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting, 2019.

[16] J. Liu, Y. Ding, M. Zeng, and L. Fu. Chemical insights into two-dimensional quantum materials. *Matter*, 5(7):2168–2189, 2022.

[17] S. Masubuchi, E. Watanabe, Y. Seo, S. Okazaki, T. Sasagawa, K. Watanabe, T. Taniguchi, and T. Machida. Deep-learning-based image segmentation integrated with optical microscopy for automatically searching for two-dimensional materials. *npj 2D Materials and Applications*, 4(1):3, 2020.

[18] H.-Q. Nguyen, X. B. Nguyen, S. Pandey, T. Faltermeier, N. Borys, H. Churchill, and K. Luu. φ-adapt: A physics-informed adaptation learning approach to 2d quantum material discovery, 2025.

[19] X. B. Nguyen, A. Bisht, B. Thompson, H. Churchill, K. Luu, and S. U. Khan. Two-dimensional quantum material identification via self-attention and soft-labeling in deep learning. *IEEE Access*, 2024.

[20] X.-B. Nguyen, H.-Q. Nguyen, S. Y.-C. Chen, S. U. Khan, H. Churchill, and K. Luu. Qclusformer: A quantum transformer-based framework for unsupervised visual clustering, 2024.

[21] Q. Pham, C. Liu, and S. Hoi. Dualnet: Continual learning, fast and slow, 2021.

[22] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[23] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. icarl: Incremental classifier and representation learning, 2017.

[24] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

[25] Z. Wang, Z. Zhang, S. Ebrahimi, R. Sun, H. Zhang, C.-Y. Lee, X. Ren, G. Su, V. Perot, J. Dy, and T. Pfister. Dualprompt: Complementary prompting for rehearsal-free continual learning, 2022.

[26] Z. Wang, Z. Zhang, C.-Y. Lee, H. Zhang, R. Sun, X. Ren, G. Su, V. Perot, J. Dy, and T. Pfister. Learning to prompt for continual learning, 2022.

[27] M. Wortsman, V. Ramanujan, R. Liu, A. Kembhavi, M. Rastegari, J. Yosinski, and A. Farhadi. Supermasks in superposition, 2020.

[28] Y. Wu, Y. Chen, L. Wang, Y. Ye, Z. Liu, Y. Guo, and Y. Fu. Large scale incremental learning, 2019.

[29] T. Zhao, Z. Wang, A. Masoomi, and J. Dy. Deep bayesian unsupervised lifelong learning, 2021.