Associative Memory with Modern Hopfield Networks for Incremental and Open-Set Visual Learning

Jelena Epifanic, Kristína Malinovská

Faculty of Mathematics, Physics and Informatics
Comenius University Bratislava
epifanic1@uniba.sk, kristina.malinovska@fmph.uniba.sk

Abstract

In dynamic environments, robust representations and memory mechanisms are important for continuous learning and adaptation. In this study, we explore Modern Hopfield Networks (MHNs) as associative memory for incremental visual recognition. We propose a memory-based framework built on a pretrained self-supervised DINOv3 encoder with prototype-based or exemplar-based memory. Memory is updated incrementally and MHN-based refinement can optionally be applied to query embeddings before retrieval. For classification, we compare cosine similarity with MHN-based attention scoring. The method is evaluated under the OWOD split protocol using ground-truth object regions. The observed trade-off between known and unknown object recognition shows that the memory dynamics play a critical role in balancing stability and plasticity, where stronger associative retrieval improves familiarity-based classification but reduces sensitivity to novel or out-of-distribution inputs.

1 Introduction

Continuous learning in artificial intelligence is essential for real-world applications, where systems must autonomously acquire new knowledge, adapt to dynamic environments, and preserve previously learned information (Parisi et al., 2019). Biological systems serve as an inspiration, since humans are capable of learning incrementally and retaining stable representations over time despite changing inputs. In practice, ANNs still struggle with catastrophic forgetting, where new information interferes with earlier representations (Wang et al., 2023). In vision systems, some biological ideas such as stable representations and memory consolidation have been used as motivation for incremental learning (Liu et al., 2019), although the complete mechanisms are still not fully understood.

To address this challenge, we explore Modern Hopfield Networks (MHNs) (Ramsauer et al., 2021) as an associative memory mechanism for open-world perception, combined with large-scale self-supervised visual representations. Our framework builds on a pretrained DINOv3 (Siméoni et al., 2025) encoder for feature extraction and supports both prototype-based and exemplar-based memory representations. This allows different strategies for storing and updating knowledge over time, depending on how much detail is preserved.

We investigate incremental visual recognition under the OWOD (Joseph et al., 2021) protocol. The model must continuously learn new object categories while still recognizing both known and previously unseen ones. We assume ground-truth object regions are available, so the focus is more on classification and open-set recognition rather than detection.

Inspired by biological learning principles, we try to balance plasticity and stability through explicit memory updates and retrieval. The results indicate that memory structure and update strategy may significantly affect retrieval quality, classification accuracy, and unknown class detection, though the effect varies across settings.

Our contributions are as follows:

- We propose a unified memory-based framework that combines MHN retrieval with self-supervised visual representations for open-world recognition under the OWOD split protocol using ground-truth object regions.
- We provide a systematic comparison of prototype-based and exemplar-based memory under incremental update settings.
- We analyze the impact of MHN-based query refinement and classification on open-world visual recognition.

2 Related work

Ramsauer et al. (2021) connect energy-based memory models with modern deep learning architectures. They revisit classical Hopfield networks by introducing continuous states. In this formulation, memory is represented as attractor states in an energy landscape.

The authors also show that Modern Hopfield Networks (MHNs) are closely related to the softmax attention mechanism used in transformer architectures. From

this perspective, attention can be interpreted as a form of associative memory, where a query interacts with stored key–value patterns. This makes MHNs a natural bridge between energy-based models and large-scale representation learning.

In continual learning, these properties are particularly relevant, as the main challenge is balancing stability of previously acquired knowledge with plasticity for new information. Memory-based approaches are commonly used to reduce catastrophic forgetting by explicitly retaining or aggregating past experiences (Kudithipudi et al., 2022). From this perspective, MHNs provide a natural mechanism for associative memory retrieval, which can be interpreted as a computational abstraction of similarity-based recall processes in cognitive systems.

Similar ideas appear in prototype- and exemplar-based representations. Prototype models compress experience into a central representation per class, while exemplar-based models preserve individual instances and rely on similarity-based retrieval over stored memories. Different granularity representations originate from cognitive science as competing models of human categorization. Prototypes capture central tendencies, while exemplars preserve instance-level variability (Nosofsky, 1992). Rather than being fundamentally different mechanisms, more recent work suggests they can be viewed as different regimes of a shared similarity-based retrieval process (Smith, 2014).

This interpretation naturally aligns with Modern Hopfield Networks. MHNs operate over an energy landscape where stored patterns act as attractors. Depending on retrieval sharpness, the same mechanism can behave more like a compressed prototype system or a distributed exemplar system. MHN-based retrieval naturally spans both regimes by operating directly on similarity in representational space.

Despite these connections, the role of Modern Hopfield Networks as an associative memory mechanism in incremental open-world visual recognition remains underexplored. In particular, it is not well understood how memory granularity affects stability–plasticity trade-offs in continual learning settings. To investigate this behavior, we evaluate our method under the OWOD protocol (Joseph et al., 2021). We follow the standard split setup on the COCO dataset (Lin et al., 2014), where object-level annotations are incrementally introduced as new classes. This provides a test-bed for continual learning with expanding label spaces and varying memory representations.

3 Our method

Motivated by biological continual learning systems that maintain both stable and adaptable representations through memory consolidation and incremental updates, we propose a memory-based framework for incremental visual recognition. The proposed approach combines large-scale self-supervised visual representations with an associative memory mechanism. We use a pretrained DINOv3 encoder for feature extraction and a memory module parameterized as either class prototypes or class exemplars. The pipeline consists of: (i) feature extraction from object regions, (ii) supervised memory construction from labeled support data, (iii) incremental memory updates that allow new classes to be integrated without retraining from scratch, and (iv) open-world inference with optional MHN-based query refinement and unknown rejection.

3.1 Memory Construction

Given an input image and corresponding object regions (from annotations), each region is encoded using a pretrained self-supervised DINOv3 backbone (Siméoni et al., 2025). The encoder outputs a feature embedding $z \in \mathbf{R}^d$, which is L2-normalized before being stored in memory:

$$\hat{z} = \frac{z}{\|z\|_2}. \tag{1}$$

Let $M = \{(p_j, c_j)\}_{j=1}^{N_p}$ denote memory support items, where p_j is normalized prototype or exemplar embedding and c_j its associated class label. We consider two memory granularities:

Prototype-based memory. For each class, a fixed number of memory items is initialized (e.g., k-means centroids or selected examples).

Exemplar-based memory. Support embeddings are stored directly without aggregation preserving instance-level detail, though at higher memory cost.

This design reflects a trade-off between compact semantic representation and fine-grained instance preservation.

Memory is built from labeled support samples grouped by class. For each class, support embeddings are added to memory according to the selected representation strategy. Class-wise confidence statistics is maintained for open-set thresholding:

$$\tau_c = \mu_c - k\sigma_c, \tag{2}$$

where μ_c and σ_c are class c confidence mean and standard deviation, and k is a tunable margin parameter.

These thresholds are later used for unknown rejection during inference.

3.2 Incremental Memory Update

To support continual learning, memory is updated incrementally as new labeled samples arrive. In prototype-based memory, updates follow a stability-preserving exponential moving average (EMA):

$$p \leftarrow \text{norm}\left((1-\alpha)p + \alpha\hat{z}\right), \tag{3}$$

applied when similarity exceeds a predefined threshold (update vs. append decision). In exemplar-based memory, new embeddings are directly stored without modification. This improves continuous adaptation while preserving past class representations.

3.3 Inference

At inference time, query embeddings can optionally be refined using a Modern Hopfield retrieval step over memory items (Ramsauer et al., 2021).

Given a query embedding $\hat{z}$ and memory matrix P, retrieval is computed as:

$$w = \text{softmax}(\beta\hat{z}P^\top), \quad r = wP, \tag{4}$$

where β controls the retrieval sharpness.
The query is then refined via interpolation:

$$\tilde{z} = \text{norm}\big((1-\lambda)\hat{z} + \lambda r\big), \tag{5}$$

where λ controls the influence of retrieved memory. This step is optional and introduces an associative recall mechanism before classification, where two different strategies are employed:

Cosine-based classification. The predicted class is assigned using nearest-neighbor search in cosine similarity space. This serves as a baseline.

MHN-based classification. Given a query embedding $\hat{z}$, attention weights are computed as

$$w_j = \text{softmax}(\beta s_j), \quad s_j = z^\top p_j \tag{6}$$

Class scores are obtained by summing attention weights over memory items belonging to each class. The predicted class corresponds to the class with the highest aggregated score. The confidence is defined as the normalized attention mass of the predicted class, i.e., the fraction of total attention assigned to that class. In this setting, the parameter m acts as a global minimum-confidence threshold.

Following the open-world setting described earlier, a prediction is accepted as known only if its confidence score exceeds a class-dependent threshold (and, in MHN mode, optionally a global minimum-confidence term). Otherwise, the sample is assigned to an unknown category.

4 Experiments

4.1 Dataset and Protocol

We follow the standard task-wise split setting (Gupta et al., 2022) on the COCO dataset (Lin et al., 2014), with four incremental stages ($t \in \{1, 2, 3, 4\}$), and report results on the corresponding test partitions. We use ground-truth object bounding boxes as input regions and do not evaluate object proposal or detection performance. All methods are evaluated at the region level, where each ground-truth object crop is treated as an independent recognition instance.

Unknown recognition uses the class-aware confidence threshold as defined in Eq. 2. In MHN mode, an additional minimum-confidence constraint m is applied. We further evaluate the effect of memory representation (prototype vs exemplar), classifier type (cosine vs MHN), and MHN query refinement.

Performance is evaluated on ground-truth instances under the OWOD incremental protocol. We report known-class accuracy, unknown recall (U-Recall), and open-set error rate (OSE).

4.2 Experimental Setup

We evaluate configurations varying along three factors: (i) memory representation (prototype vs. exemplar), (ii) classification strategy (cosine similarity as a baseline vs. MHN-based classification), and (iii) optional MHN-based query refinement.

This results in a total of eight configurations, summarized in Table 1. All setups are evaluated under the OWOD protocol across tasks T1–T4.

Row	Memory	Classifier	Refinement
1	Prototype	Cosine (baseline)	–
2	Prototype	MHN	–
3	Prototype	Cosine	MHN
4	Prototype	MHN	MHN
5	Exemplar	Cosine (baseline)	–
6	Exemplar	MHN	–
7	Exemplar	Cosine	MHN
8	Exemplar	MHN	MHN

Tab. 1: Evaluated configurations across memory representation, classifier type, and optional MHN-based query refinement.

All experiments use frozen DINOv3 (Siméoni et al., 2025) features extracted from object regions. Memory is built from labeled support samples (n_{support}) using either: (i) prototype-based aggregation, or (ii) direct exemplar storage.

During training, optional refinement updates memory using additional labeled samples. This allows gradual adaptation without full retraining.

Unless stated otherwise, we keep the encoder, data splits, and extracted features fixed across all experiments, and vary only memory and inference configurations. Key hyperparameters include the number of support samples (n_{support}), number of prototypes, update strategy (`ema` or `append only`), and the unknown-threshold parameter k. To ensure stability of the exemplar memory, we limit storage to at most 100 samples per class.

5 Results

For the first experiment, we keep the following parameters fixed: the MHN sharpness parameter $\beta = 30.0$, the unknown threshold parameter $k = 0.5$, and the minimum confidence threshold $m = 0.35$.

The results show a consistent trade-off between known-class recognition and open-set performance across both memory representations. For exemplar-based memory (Table 2), MHN-based classification improves known-class accuracy compared to the cosine baseline (from 7.84% to 19.65%). A further improvement is obtained when MHN-based query refinement is included (up to 25.33%). However, this gain is accompanied by a strong decrease in unknown-class performance, where U-Recall drops from 81.53% to 33.74% and OSE increases accordingly. A similar behavior is observed for prototype-based memory. MHN classification improves known-class accuracy (61.98% to 66.74%) and slightly increases unknown recall. However, MHN-based refinement has a strong negative effect on open-set behavior, reducing U-Recall to 3.19% in the cosine+refine setting and significantly increasing OSE. The results are consistent with the idea that prototype-based regime enforce stronger abstraction, while exemplar-based regime preserve variability necessary for open-set discrimination.

Overall, these results suggest that MHN-based retrieval improves recognition performance on known classes, but also introduces a bias toward predicting known categories (Fig. 1). This effect becomes more pronounced when MHN query refinement is applied. Interestingly, despite this bias, the person class still shows a tendency to be misclassified as unknown, indicating class-specific ambiguity.

Method	**Known Acc.**	**U-Recall**	**OSE**
Exemplar-based memory			
Cosine (baseline)	7.84	81.53	18.47
MHN classification	19.65	72.97	27.03
Cosine + MHN refine	19.97	51.29	48.71
MHN + refine	25.33	33.74	66.26
Prototype-based memory			
Cosine (baseline)	61.98	25.30	74.70
MHN classification	66.74	38.54	61.46
Cosine + MHN refine	66.27	3.19	96.81
MHN + refine	65.75	29.45	70.55

Tab. 2: Comparison of exemplar- and prototype-based memory under cosine and MHN-based inference variants.

When we compare the incremental tasks, the same pattern remains stable, as shown in Table 3. Earlier tasks achieve higher known-class accuracy, while later tasks show a gradual shift toward a more balanced but still unstable known–unknown trade-off.

Task	Known Acc. (%)	U-Recall (%)	OSE (%)
T1	82.52	23.80	76.20
T2	70.03	35.64	64.36
T3	59.08	56.17	43.83
T4	55.34	–	–

Tab. 3: Per-task performance for prototype-based memory with MHN classification across OWOD incremental stages T1–T4.

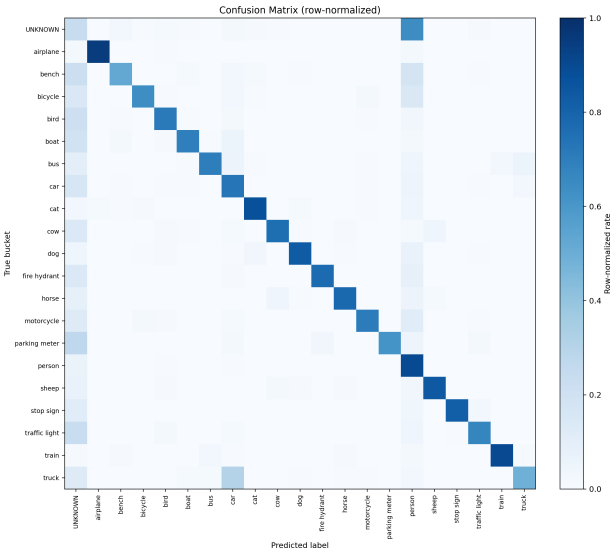

Fig. 1: Confusion matrix of MHN-based classification. The model shows strong diagonal structure for known classes, but a tendency to assign unknown samples to known categories.

We also designed the experiments to analyze the effect of individual parameters on model behavior. In this setting, we increase the MHN sharpness parameter to $\beta = 60.0$, the unknown threshold to $k = 0.7$, and the minimum confidence threshold to $m = 0.6$.

The results, summarized in Table 4, show how these changes affect performance across tasks T1–T4. Increasing β makes the model more confident but significantly reduces unknown recall, leading to strong over-assignment to known classes. Increasing k makes the decision boundary slightly more permissive for unknown detection. Increasing the minimum confidence threshold m improves open-set behavior by rejecting uncertain predictions, but at the cost of lower known-class accuracy. This pattern is consistent with the attractor-based memory systems, where increased attractor strength can lead to overgeneralization toward familiar categories at the expense of sensitivity to novel inputs (Hopfield, 1982).

The results in Table 5, further highlight the role of memory granularity in the stability–plasticity trade-off. Increasing the number of stored exemplars per class leads to a substantial improvement in unknown recall

Task	Known Acc. (%)	U-Recall (%)	OSE (%)
Base	66.74	38.54	61.46
$k = 0.7$	67.03	36.95	63.05
$\beta = 60$	68.67	8.83	91.17
$m = 0.6$	52.13	65.30	34.70

Tab. 4: Prototype-based memory with MHN classification under different parameter settings. The baseline (Base) refers to the default configuration used in the main experiments, with $\beta = 30$, $k = 0.5$, and $m = 0.35$.

(from 33.74% to 52.51%), while only slightly reducing known-class accuracy. This indicates that a denser exemplar memory improves the model's ability to reject unseen categories, likely by providing a more complete coverage of the known-class feature space. Since exemplar memory construction heavily depends on the available support examples, this aspect requires further investigation.

Task	Known Acc. (%)	U-Recall (%)	OSE (%)
Baseline	25.33	33.74	66.26
50 examples/class	22.76	39.40	60.60
100 examples/class	22.38	52.51	47.49

Tab. 5: Exemplar-based memory with MHN classification. Mean performance across tasks for different exemplar settings, with 20 examples per class as the baseline.

To better understand the MHN mechanism, a qualitative analysis was conducted. The DINO mask visualization shows a clean part-level structure on both the dog and the person, distinguishing it from the background. This confirms that DINOv3 features encode meaningful semantics, which aligns with levels of human organization in figure–ground segmentation. Interestingly, the cosine and MHN attention maps highlight different regions (head/torso/limbs). MHN attention focuses more on the head and central parts of the body, showing stronger weighting on salient features.

6 Conclusion

In this work, we studied how MHN-based associative memory supports incremental open-world visual recognition. Different memory representations originating from cognitive theories, classification strategies, and optional MHN query refinement are studied under the OWOD protocol.

The results show a clear trade-off between known-class recognition and open-set performance. While

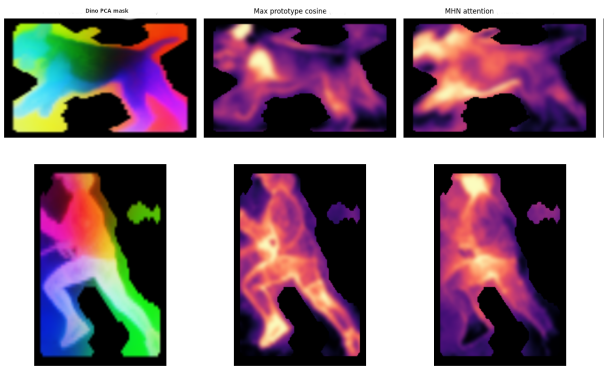

Fig. 2: Dense DINO Feature and Memory Retrieval Visualization

similarity-based retrieval can improve known-class accuracy, it also tends to increase the bias toward known categories, especially when query refinement is used. This leads to lower unknown-class recall and higher open-set error. Stronger associative retrieval increases attraction toward familiar memory states, while reducing sensitivity to novel inputs.

We also find that memory structure has an important effect. Exemplar-based memory improves unknown detection when more samples per class are stored, suggesting that better coverage of the feature space helps rejecting unseen categories. In contrast, more compact memory representations are more stable for known classes but less effective for open-set recognition.

Overall, the results show that retrieval-based memory is a simple and effective approach for open-world recognition, but its behavior depends strongly on memory design and retrieval settings.

Acknowledgement

This paper is based on the research conducted for a master's thesis entitled *Associative Memory with Modern Hopfield Networks for Incremental and Open-Set Visual Learning* under the expert supervision of RNDr. Kristína Malinovská, PhD, Comenius University Bratislava and Prof. Dipl.-Ing. Dr.techn. Markus Vincze, Automation and Control Institute (ACIN), Faculty of Electrical Engineering and Information Technology, Technical University Wien (TUW).

References

Gupta, A., Narayan, S., Joseph, K. J., Khan, S., Khan, F. S. and Shah, M. (2022). Ow-detr: Open-world detection transformer. *arXiv*.

Hopfield, J. J. (1982). Neural networks and physical

systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558.

Joseph, K. J., Khan, S., Khan, F. S. and Balasubramanian, V. N. (2021). Towards open world object detection. *arXiv*.

Kudithipudi, D., Aguilar-Simon, M., Babb, J., Bazhenov, M., Blackiston, D., Bongard, J., Brna, A. P., Raja, S. C., Cheney, N., Clune, J., Daram, A., Fusi, S., Helfer, P., Kay, L., Ketz, N., Kira, Z., Kolouri, S., Krichmar, J. L., Kriegman, S. and Levin, M. (2022). Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, 4(3):196–210.

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L. and Dollár, P. (2014). Microsoft coco: Common objects in context. *arXiv*.

Liu, J., Zhang, H., Yu, T. et al. (2019). Stable maintenance of multiple representational formats in human visual short-term memory. *Nature Communications*, 10:5800.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In *Essays in Honor of William K. Estes*, vol. 1, pp. 149–167.

Parisi, G. I., Kemker, R., Part, J. L., Kanan, C. and Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G. K., Greiff, V., Kreil, D., Kopp, M. and Hochreiter, S. (2021). Hopfield networks is all you need. *arXiv*.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A. and Tolan, J. (2025). Dinov3. *arXiv*.

Smith, J. D. (2014). Prototypes, exemplars, and the natural history of categorization. *Psychonomic Bulletin Review*, 21:312–331.

Wang, L., Zhang, X., Su, H. and Zhu, J. (2023). A comprehensive survey of continual learning: Theory, method and application. *arXiv*.