

# Reviewing continual learning from the perspective of human-level intelligence

Yifan Chang, Wenbo Li, Jian Peng, Bo Tang, Yu Kang, Yinjie Lei, Yuanmiao Gui, Qing Zhu, Yu Liu, Haifeng Li *Member, IEEE,*

arXiv:2111.11964v1 [cs.LG] 23 Nov 2021

**Abstract**—Humans continual learning (CL) ability is closely related to Stability Versus Plasticity Dilemma that describes how humans achieve ongoing learning capacity and preservation for learned information. The notion of CL has always been present in artificial intelligence (AI) since its births. This paper proposes a comprehensive review on CL. Different from previous reviews that mainly focus on the catastrophic forgetting phenomenon in CL, this paper surveys CL from a more macroscopic perspective based on the Stability Versus Plasticity mechanism. Analogous to biological counterpart, “smart” AI agents are supposed to i) remember previously learned information (information retrospection); ii) infer on new information continuously (information prospection); iii) transfer useful information (information transfer), to achieve high-level CL. According to the taxonomy, evaluation metrics, algorithms, applications as well as some open issues are then introduced. Our main contributions concern i) recheck CL from the level of artificial general intelligence; ii) provide a detailed and extensive overview on CL topics; iii) present some novel ideas on the potential development of CL.

**Index Terms**—Continual learning, artificial general intelligence, information retrospection, information prospection, information transfer.

## I. INTRODUCTION

WHAT is intelligence? Many theories [1] have been developed to define it, for example, intelligence is the resultant of the process of acquiring, storing in memory, retrieving, combining, comparing, and using in new contexts information [2]; intelligence is goal-directed adaptive behavior [3] and so on [4]. While different in terms of perspective, a common and central idea can be noticed: the ability to mold our cognitive system to deal with the always changing

This work was supported by the National Nature Science Foundation of China under Grant 41871302, the Development Program of China under Grant 2018YFB1004600 and Anhui Provincial Natural Science Foundation under Grant 2108085J19. Corresponding author: Haifeng Li, email: lihaifeng@csu.edu.cn.

Y. Chang and Y. Kang are with the State Key Laboratory of Fire Science, Department of Automation and Institute of Advanced Technology, University of Science and Technology of China, Hefei, China.

W. Li and Y. Gui are with Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031.

B. Tang is with the Department of Electrical and Computer Engineering at Mississippi State University, USA.

Y. Lei is with the college of Electronics and Information Engineering at Sichuan University, Chengdu, China.

Q. Zhu is with the Faculty of Geosciences and Environmental Engineering of the Southwest Jiaotong University, Chengdu, China.

Y. Liu is with the Institute of Remote Sensing and Geographic Information System, Peking University, Beijing, China

J. Peng and H. Li are with the School of Geosciences and Info-Physics, Central South University, South Lushan Road, Changsha, 410083, China.

## Stability Versus Plasticity Dilemma

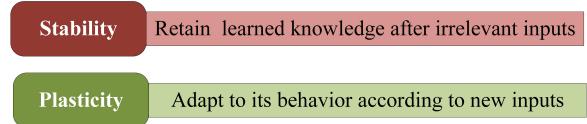


Fig. 1. The interpretation of Stability Versus Plasticity Dilemma [8].

demanding circumstances [5], which is strictly related to the ability of continual learning (CL).

Humans have the extraordinary ability to learn continually from experience. Not only can we remember previously learned knowledge and apply these skills to new situations, we can also abstract useful knowledge and use these as the foundation for later learning. The CL ability is closely owed to Stability Versus Plasticity Dilemma which explores how a learning system remains adaptive in response to input, yet remains stable in response to irrelevant input [6]. Stability is the ability to retain the learned information on the old tasks and plasticity is the ability to adapt to a new task [7], as shown in Figure 1.

Analogous to our biological counterpart, achieving adaption is also a grand goal of artificial intelligence (AI) [8]. The aim is to build an artificial CL agent that adapts to a sophisticated understanding of the world from its own experience through the incremental development of skills and knowledge [8].

Up to present, many progress has been witnessed in CL field and many paper have summarized CL, for example, Van et al. [9], Thrun et al. [10] and Farquhar et al. [11] surveyed on certain aspects of CL, such as applications and evaluation metrics; Parisi et al. [18] provided a comprehensive review on the catastrophic forgetting problem. However, these reviews mainly focus on a specific issue on CL. In fact, to achieve the aim of artificial general intelligence (AGI), an ideal CL agent is supposed to perform well in the whole process of CL, not only on the forgetting problem for the past encountered information. To bridge this gap, this paper reviews CL on a macroscopic level based on the Stability Versus Plasticity mechanism. The advantages of examining CL in the context of human-level intelligence lie in two reasons: on the one hand, AI agents are conceptually derived from the biological neural networks, and thus, such an examination may facilitate the validation. What's more important, it can clearly show how far the current work is from the final destination in CL, providing introspection for existing research while serving as inspiration for future work.

As mentioned above, Stability Versus Plasticity mechanism highlights on both memory retention and future learning. Analogous to the mechanism, an AI agent is supposed to perform well in the three aspects in CL process, as shown in Figure 2 (a): i) Information retrospection: remembering previously learned information; ii) Information prospection: infer on new information continuously; iii) Information transfer: transferring useful information. Specifically:

- *Information retrospection:* It is long-term memory of events, facts, knowledge and skills that have happened in the past [12]. One big challenge in information retrospection is the issue of catastrophic forgetting, the tendency of AI agents to completely and abruptly forget previously learned information upon learning new information [13] [14]. This has motivated many ideas to address the problem such as reducing overlapped regions among different information, repeating past information, expanding network architecture to accommodate more information, and corresponding algorithms has also been proposed, as shown in Figure 2 (b).
- *Information prospection:* It is defined as the ability to remember to carry out intended actions in the future [15]. For AI agents, they are expected to deduce on future learning based on learned experience. The *|±smart|±* AI agents can accelerate learning speed, achieve high generalization, learn from limited data, and reduce convergence speed on future learning. Many learning paradigms are combined with CL such as incremental metalearning, incremental fewshot learning, incremental active learning and so on, to facilitate information prospection, as shown in Figure 2 (c).
- *Information transfer:* Information transfer aims to improve the learning in a task through the transfer of information from a related task [16]. The circulating and cross-utilized information can be exploited to facilitate future learning, and be supplemented to enhance the established knowledge system. Current CL learning problem are still now difficult to encapsulate and isolate into single domains or tasks. This has motivated many transfer learning techniques in CL algorithms, as shown in Figure 2 (d).

According to the taxonomy, this paper is organized as follows. Section 2 introduces the evaluation metrics in terms of information retrospection, information prospection and information transfer. Section 3 categorizes and describes the mainstream CL algorithms according to their contribution to the three aspects. Section 4 introduces the applications of CL. Section 5 presents recommendation regarding CL to promote a more extensive exploration in this field. Section 6 presents the concluding remarks.

## II. EVALUATION METRICS

The metrics to evaluate CL system lacks consensus, and the almost all metrics exclusively focus on the forgetting problem. In this section, a comprehensive summary on the performance of CL systems, in terms of information retrospection, information prospection and information transfer, is presented. The

visualization of the evaluation metrics is presented in Figure 3.

Information retrospection pertains to memory access to events and information that occurred or were learned in the past [17]. The loss of past information results in an abrupt deterioration in the performance pertaining to earlier knowledge after the learning of new information (catastrophic forgetting) in CNNs [13]. Information retrospection can be validated by the following three metrics.

- *Forgetting extend:* Forgetting is defined as the biggest knowledge difference for a specific task between the model has acquired continually till task  $k$  and a previous one acquired before  $k$  in the past [18]. It represents the performance degradation on past tasks given its current state. For a classification problem, the forgetting for the  $j$ -th task after the model has been incrementally trained up to task  $k$  ( $j < k$ ) is computed as Equation 1.

$$f_j^k = \max_{l \in \{1, \dots, k-1\}} a_{l,j} - a_{k,j} \quad (1)$$

- *Memory capacity:* Memory capacity is defined as the number of learned tasks the model can remember at the current state. Recalling long sequence of previous tasks represents the outlasting memory retain ability. In fact, a simple fully connected model will show no forgetting on a two-task [19]. Accordingly, testing CL agent on extremely long sequences of tasks is necessary.
- *(Average) accuracy:* Accuracy pertains to the index of a model trained on training data and evaluated on test data [20]. In CL scenario, it is always computed for each previous task on the test data at the end of training on the current task. Average accuracy (ACC) is computed for all tasks at the end of a model's training for the current task continually, as shown in Equation 2.

$$A_k = \frac{1}{k} \sum_{j=1}^k a_{k,j} \quad (2)$$

Where  $A_k$  is the ACC on all the task after the last task  $k$  is learned;  $a_{k,j}$  is the accuracy evaluated on the task  $j$ , after the model has been trained with the final task  $k$ .

Information prospection refers to the ability of a CL model to incorporate new information. The loss of memory prospection leads to a decreased ability to accumulate new information. The performance of memory prospection can be considered in terms of three aspects.

- *Learning Curve Area (LCA):* The LCA is a graphical representation of the rate of improvement in performing a task as a function of time or rate of change [21]. This index represents the amount of time required by a model to acquire new information. Specifically, the learning curve in CL is regarded as convergence curve which describes an average performance at the  $j$ -th mini-batch after the model has been trained for all the  $k$  tasks, as shown in Equation 3.

$$Z_b = \frac{1}{T} \sum_{k=1}^T a_{k,b,k} \quad (3)$$

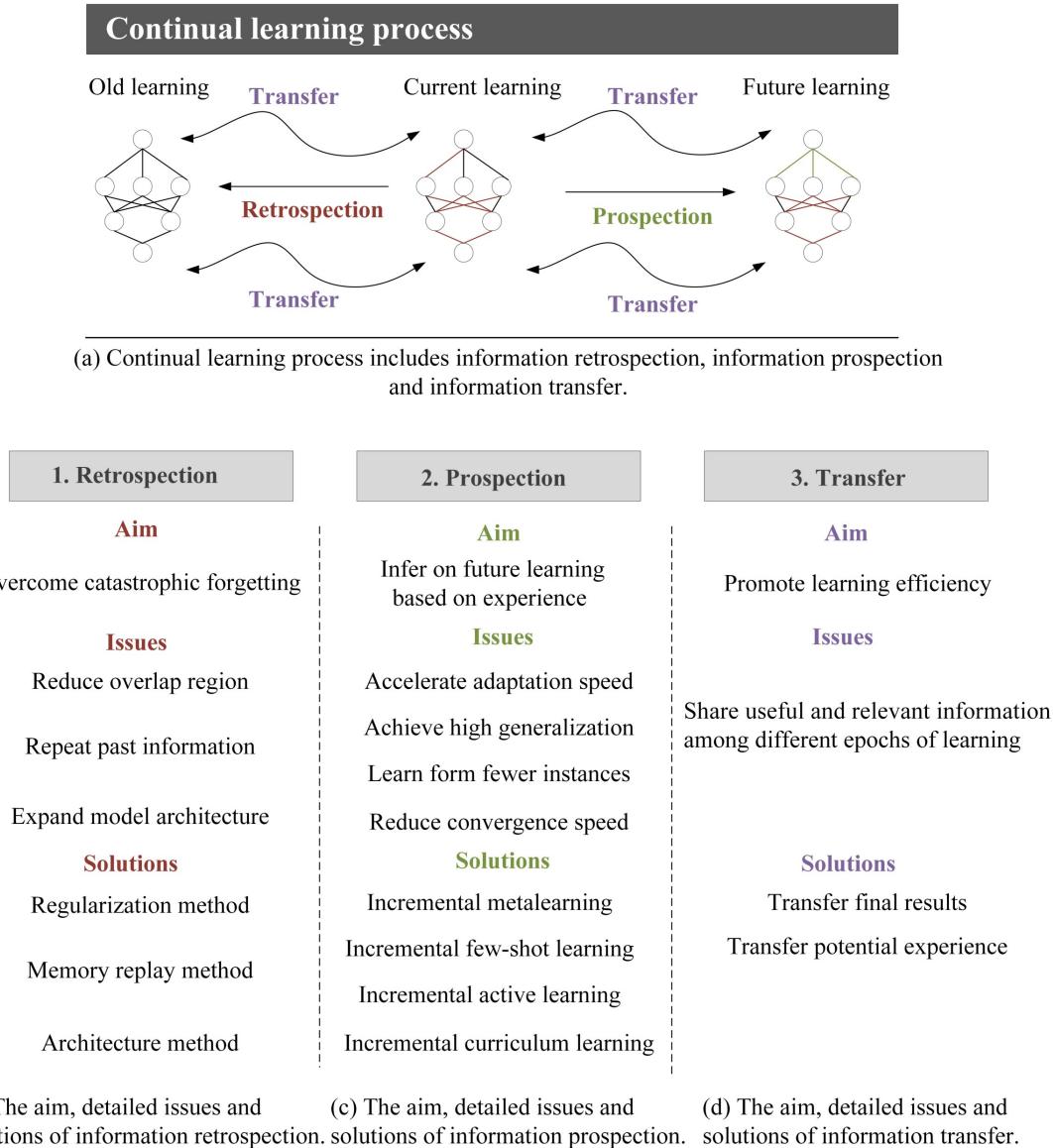


Fig. 2. The learning paradigm of an ideal CL agent.

Where  $a_{i,b,k}$  be the accuracy evaluated on the task  $j$ , after the model has been trained with the  $b$ -th mini-batch of task  $k$ .

- **Intransigence:** Intransigence describes the inability of a model to learn new tasks [18]. To quantify the intransigence on the  $k$ -th task (denoted as  $I_k$ ), the performance of the model trained on the  $k$ -th task in the incremental manner (denoted as  $a_{k,k}$ ) is compared with that of standard model which has access to all the datasets at all times (denoted as  $a_k^*$ ), as shown in Equation 4.

$$I_k = a_k^* - a_{k,k} \quad (4)$$

- **Scalability:** Scalability is defined as the ability to ensure dynamic and efficient resource expansion or reuse when the weights are already saturated in the model [22]. To absorb as much new information as possible, it is necessary for a CL model to allocate a new set of weights to learn complimentary representations for a new task

based on the sharing of common representations with old tasks.

Information transfer refers to the flexible knowledge transformation between retrospective and prospective information. In CNNs, the circulation of information can accelerate the learning of new information while reoptimizing the learning of old information. Memory transfer can be validated from two metrics, with the one to measure the impact that the new ones on the old ones, and vice versa.

- **Backward transfer (BWT)** [20]: The BWT indicates the influence that learning a new task has on the performance of former tasks. This index describes the extent to which a new task weakens the previous tasks. Learning new tasks may have two contrasting effects on the previous tasks: positive backward transfer (PBT) or negative backward transfer (NBT), which correspond to the promotion or degradation of the previous performance, respectively.

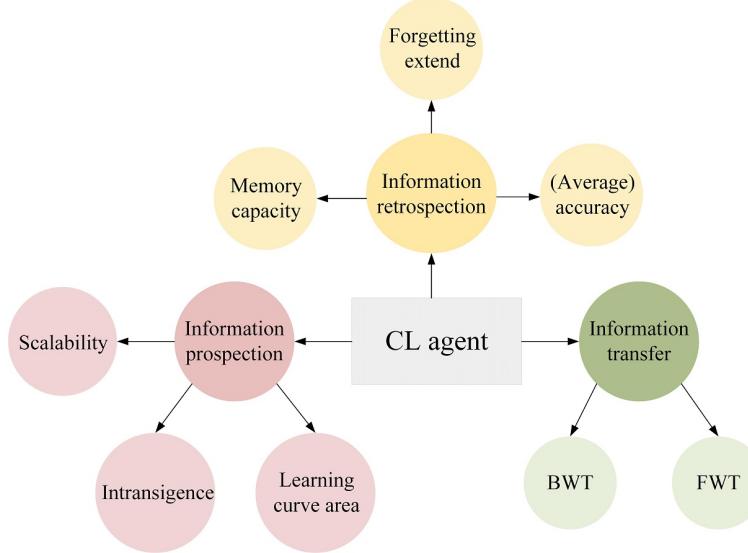


Fig. 3. Visualization of the evaluation metrics for a CL agent in terms of information retrospection, information prospection and information transfer.

The equation of BWT is in Equation 5.

$$BWT = \frac{1}{k-1} \sum_{i=1}^{k-1} a_{k,j} - a_{i,i} \quad (5)$$

- *Forward transfer (FWT)* [20]: FWT describes the influence that learning a current task has on the performance on a future task. In particular, a positive FWT occurs when the model can perform zero-shot learning, likely by exploiting the structure specific to the task. The equation of BWT is in Equation 6.

$$FWT = \frac{1}{k-1} \sum_{i=2}^k a_{i-1,i} - \bar{b}_i \quad (6)$$

Where  $\bar{b}_i$  is the vector of test accuracies for each task at random initialization.

### III. CL ALGORITHMS

This section summarizes the mainstream CL algorithms referring to memory retrospection, memory prospection and memory transfer.

#### A. Information retrospection

Information retrospection is the most basic function of CL agents. The learned knowledge would likely be promptly forgotten, and new information would be difficult to acquire without it. Currently, many methods are available to preserve past information, and these methods can be categorized as regularization, memory and architecture methods.

1) *Regularization methods*: Regularization methods limit the catastrophic forgetting phenomenon by imposing constraints as a regular term on the update of the weights in CNNs to help the network identify a set of weights that can provide proper mappings from each learned input to the output. The weights optimized using regularization methods may not be optimal for each pattern but include acceptable

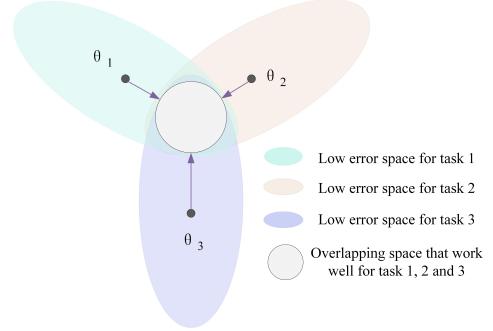


Fig. 4. Working mechanism of regularization method.

response values for all patterns, as shown in Figure 4. Regularization methods are based on the concepts of Hebbian learning mechanism, according to which, synapses are dynamically organized and updated according to different stimuli [23]. High-frequency stimuli initiate long-term potentiation (LTP) in synapses, causing memory enhancement, whereas low-frequency stimuli result in long-term depression (LTD), causing memory degradation [23], as shown in Figure 5 (a).

Elastic weight consolidation (EWC) [24] is a pioneering and the most cited regularization method. EWC quantifies the importance of the learned parameters by estimating the Fisher information relative to the objective likelihood and preserves parameters with high importance values by restricting their drastic changes against new tasks. However, EWC assumes that the Fisher information matrix (FIM) [25] is diagonal, which is unrealistic in the original parameter space. R-EWC [26] enhances this diagonal assumption of EWC through a reparameterization strategy. This strategy rotates the parameter space by singular value decomposition (SVD) [27] such that the output of the forward pass is unchanged, but the FIM computed from the gradients during the backward pass is approximately diagonal. In this rotated parameter space, EWC can effectively optimize the new task. Although

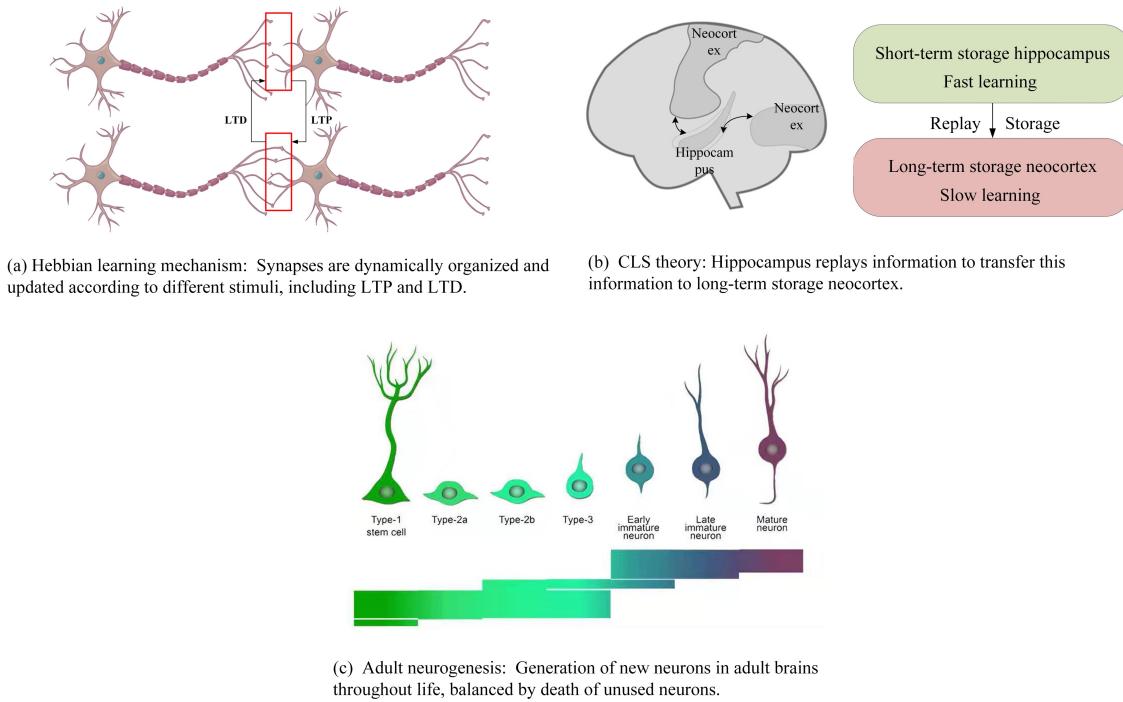


Fig. 5. The biological mechanisms for regularization method, memory replay method and architecture method.

several methods similar to EWC have been proposed, these approaches calculate the parameter importance using different techniques. For example, according to the memory-aware synapsis (MAS) [28] technique, the change in an important weight can influence the output function of the model more significantly than changes in unimportant weights, and thus, this approach computes the importance of the weight by measuring the magnitude of the gradient of a parameter when it is perturbed. Compared with EWC, an apparent advantage of MAS is that it can update the model in an unsupervised and online manner by avoiding the need for labeled data. Synaptic intelligence (SI) [29] computes the path integral of the gradient vector field as the weight importance along the entire learning trajectory. The most notable difference between SI and EWC is that SI computes the importance online and along the entire learning trajectory, whereas EWC computes the importance in a separate phase at the end of each task. Orthogonal weight modification (OWM) [30] optimizes the update direction of parameters in the direction orthogonal to all the previous input spaces. This strategy can avoid mutual interference among different tasks. To protect the most important weights, ABLL [31] leverages an autoencoder to capture the submanifold that contains the most informative features pertaining to a past task. When training for a new task, the features projected onto this submanifold are controlled to not be drastically updated. Rather than focusing on the weights in all layers of CNNs, LFL [32] restricts the drastic changes in the learned parameters in the final hidden activations to preserve the previously learned input-output mappings and maintain the decision boundaries. IMM [33] progressively matches the Gaussian posterior distribution of the CNNs trained on the old and new tasks and uses various transfer learning techniques to

render the Gaussian distribution smooth and reasonable.

Overall, the advantage of regularization methods is that it is computationally effective even with a small training period, low storage occupancy and low computational complexity. However, this approach is not scalable, resulting in a performance degradation when the number of classes increases gradually. In addition, the learning result is highly dependent on the relevance between tasks because the parameters for future tasks are bounded to the parameter regions generated by the past tasks.

2) *Memory replay methods:* The memory method consolidates old information by replaying them when models are updated for new tasks. The replayed old experience can be used to constrain the parameter updates such that the loss pertaining to previous tasks is not aggravated to retain connections for previously learned memories. The memory method originates from the hippocampal replay mechanism in complementary learning systems (CLS) [34], as shown in Figure 5 (b). Biological research has indicated that the hippocampus in rodents replays information that they experience in the daytime during sleep or rest [35] to transfer this information to long-term storage.

Memory replay methods can be broadly categorized into two types of methods, namely, rehearsal and generative replay methods, according to whether the replayed past data are real data.

The rehearsal strategy stores data from previously learned tasks with a memory buffer and interleaves them with the training data of the current task to jointly train the model, as shown in Figure 6 (a). Incremental classifier and representation learning (iCaRL) [36] is a representative approach. iCaRL employs an exemplar set including the most representative

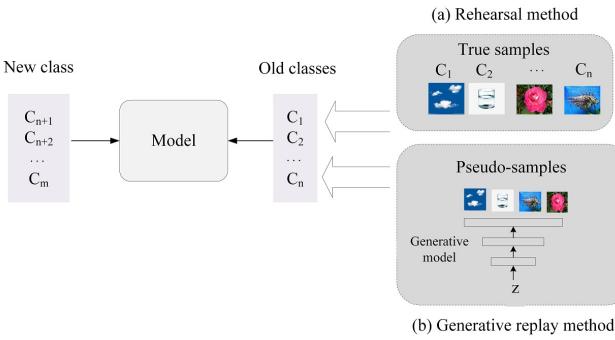


Fig. 6. Working mechanism of the memory method [39], including the rehearsal and generative replay methods.

samples of each previous class. These samples can most accurately approximate the average feature vector over all training examples. Next, the approach applies the nearest mean classifier [37] strategy to predict a label for a new image based on the exemplar set. iCaRL can incrementally learn many classes over a long period; however, it updates the exemplar set and classifier independently. End-to-end incremental learning (EEIL) [38] overcomes the limitation of iCaRL by jointly learning the classifier and features. The approach achieves end-to-end learning by formulating an integrated cross-distilled loss based on the distillation loss to extract representative samples from old data and using the cross-entropy loss to simultaneously learn new classes. Subsequently, the imbalance problem between previous and new data is considered. The unified classifier (UC) [39] adopts a cosine normalization-based classifier to eliminate the significant difference in the bias and weights between old and new tasks. Subsequently, the approach employs a less-forget constraint strategy to ensure that the features of the old samples in the new and old models do not significantly differ. According to large-scale incremental learning (LSIL) [40], the class imbalance problem causes the classifier to classify an image into the category with a larger amount of data; therefore, the approach introduces a bias correction (BiC) layer to correct the bias regarding the output logits for the new classes. In the incremental learning with dual memory (IL2M) [41] approach, a dual memory is introduced to alleviate the negative effect of the imbalance problem. The first memory stores the exemplar images of past classes. The second memory stores the initial class statistics in a highly compact format as it is assumed that the initially learned classes are best modeled. Experimental results show that the initial class statistics stored in the second memory can help the model overcome the problem of imbalanced datasets in past classes and rectify the associated prediction scores. In addition to the sample imbalance problem, certain approaches have been developed to optimize the memory storage utilization. For example, MECIL [42] aims to optimize the exemplar set management. This approach retains low-fidelity exemplar samples rather than the original high-fidelity samples in the memory buffer to ensure that the limited memory can store more exemplars. Certain other approaches attempt to enhance the stored items of previous tasks. For example, instead of directly rehearsing the stored examples, the gradient episodic

memory model (GEM) [20] stores gradients of the previous task to define inequality constraints regarding the loss to ensure that the loss does not increase with respect to that in previous tasks.

However, the rehearsal strategy relies on stored data, which is undesirable for several reasons. First, data storage is not always possible in practice due to safety or privacy concerns. Second, the approach is difficult to scale up to address problems involving many tasks. Finally, the rehearsal method is questionable in terms of the neuroscience perspective because the brain does not directly store data, such as all pixels of an image. As an alternative, generative replay (GR) has been proposed to rehearse past data without having access to them. In contrast to restoring the exact samples in old tasks, this approach uses a separate generative model to generate a pseudosample of old tasks, as shown in Figure 6 (b). Deep generative replay (DGR) [43] introduces generative adversarial networks (GANs) [44] to mimic past data. The generated pseudodata and their responses pertaining to the past model are paired to represent old tasks, and the data are interleaved with new data to update the model. DGR achieves promising continual learning results; however, the generator must be repeatedly trained using a mix of samples synthesized for previous categories and real samples of new classes. Therefore, certain researchers attempted to reduce the computational overhead for the replayed data based on DGR. For example, the deep generative memory (DGM) [45] approach eliminates the reuse of previous knowledge by introducing learnable connection plasticity for the generator. The approach designs task-specific binary sparse masks for the learnable units of the generator weights. The gradients of weights in each layer of the generator are multiplied by the reverse of the cumulated mask to prevent the overwriting of previous knowledge. Rather than replaying the real samples, BIR [46] replays the internal or hidden representations of past tasks. The replayed representation is generated by the context modulated feedback connection of the network. GRFC [47] reduces the computational cost by integrating the generative model with the main model through generative feedback connections. In addition, because combining DGR with a distillation strategy can enhance the performance, the approach labels the input data of the current task by using the model trained for the previous tasks as soft targets and uses the resulting input-target pairs as the pseudodata. DGM, BIR and GRFC are more scalable than DGR in the case of complicated problems involving many tasks or complex inputs. It has been highlighted that reliable data help retain the corresponding past information, and thus, certain researchers focused on enhancing the quality of the replayed data. For example, MeRGAN [48] uses a conditional GAN, in which the category is used as an input to guide the replay process, thereby avoiding less biased sampling for past categories and generating more reliable past data. CloGAN [49] uses an auxiliary classifier to filter a portion of distorted replays to block inferior images from entering the training loop.

In general, the memory method achieves satisfactory results in continual learning, and many inherent issues, such as the efficacy of memory buffers, have been considered with

TABLE I  
ADVANTAGES AND DISADVANTAGES OF THE REGULARIZATION METHODS, MEMORY REPLAY METHODS AND ARCHITECTURE METHODS

CL methods	Advantage	Disadvantage
Regularization method	Small training period Low storage occupancy Low computational complexity	Low scalability Medioocre performance
Memory method	Promising performance Small training period (Generative replay)	Large training period (Rehearsal) High storage occupancy Sample imbalance
Architecture method	Low storage occupancy	Low scalability Medioocre performance High computational complexity

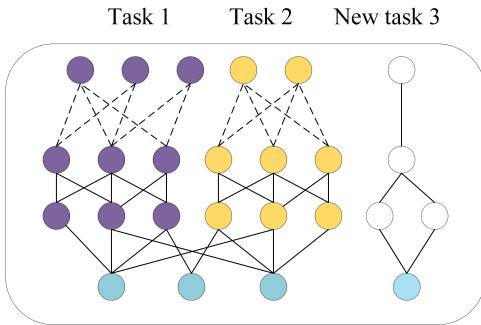


Fig. 7. Working mechanism of the architecture method.

the research progress. The memory replay method is robust against catastrophic forgetting when the relevant experience is elaborately selected. However, an inherent problem is the long training time when training models with mixed data. In addition, the problem of imbalanced samples when rehearing or replaying past data has not been fully solved.

3) *Architecture methods*: The architectural strategy realizes memory retrospection by flexibly expanding model architectures to accommodate more information while preserving part of the architectures trained for previous tasks. Different subsets of model parameters are assigned to different tasks in architecture method, as shown in Figure 7. Research on the architecture method is motivated by the adult neurogenesis theory [50], as shown in Figure 5 (c). Adult neurogenesis is the formation of functional, mature neurons from neural stem cells in specific brain regions in adults. In these regions, new neurons are generated throughout life and integrated into established neuronal circuits [50].

There are two main categories of architectural strategies, namely, fixed networks and dynamic networks, based on whether the model structure is expanded. The fixed network strategy only allows inner network adjustments, such as changes in the weights and activations. PackNet [51] is a typical fixed network technique. This approach employs a weight-based pruning technique [52] to release redundant parameters across all layers after the training for a task. The remaining parameters retain the information for old tasks, and the released parameters are updated for new tasks. PathNet [53] employs evolutionary strategies [54] to select pathways that determine the parameters in the network to be retained

or updated. The approach fixes the parameters along a path learned on the old task and re-evolves a new population of paths for the new task.

In contrast to enforcing a predefined architecture, a dynamic network allows the weights to be added or removed according to the future tasks. PNN [55] expands the architecture by allocating novel subnetworks with a fixed capacity to be trained using the new information. OIFL [56] is based on adaptive feature learning that adds features for samples with a high loss and subsequently merges the similar features to prevent based on a denoising autoencoder. Instead of enforcing a predefined architecture, Adanet [57] automatically evolves to learn network structures. Starting from a simple linear model, Adanet adds more units and additional layers, as necessary. The added units are carefully selected and penalized according to rigorous estimates from several theories of statistical learning [58] [59] to seek high-quality models with minimal expert intervention. EDIL [60] employs a training algorithm that grows the network capacity in a hierarchical manner. Similar classes are grouped and self-organized into levels, with models cloned from previous classes and trained in parallel to accelerate the training. The neurogenesis deep learning (NDL) model [61] adds new neurons to the autoencoder and employs intrinsic replay to reconstruct old samples and preserves their learned weights. Following this idea, a dynamically expandable network (DEN) selectively [62] retrains the old network and expands its capacity when necessary, thereby dynamically deciding the optimal capacity of the network being trained online. RPSnet [63] adopts a random path selection algorithm that progressively chooses optimal paths for the new tasks while encouraging parameter sharing and reuse.

Generally, architecture methods exhibit a low storage occupancy since they do not store any extra samples or parameters. However, such methods are not scalable because the network architecture may not be sufficient or be randomly expanded as the number of tasks increases. In addition, the computational complexity is slightly high because the approaches recompute the pathways in the network for each new task.

In conclusion, the three types of methods to achieve memory retrospection are generally effective and exhibit characteristic merits and unavoidable defects, as summarized in Table I. These algorithms are summarized in Table II.

TABLE II  
SUMMARY ON REGULARIZATION METHODS, MEMORY REPLAY METHODS AND ARCHITECTURE METHODS.

Method	Algorithm	Keypoint	Reference
Regularization method	EWC	Measure weight importance	[24]
	R-EWC	Rotates parameter space	[26]
	MAS	Label free	[28]
	SI	Online training	[29]
	OWM	Orthogonal weight modification	[30]
	ABLL	Capture informative feature	[31]
	LFL	Focus on hidden layer	[32]
	IMM	Bayesian moment matching	[33]
Memory replay method - Rehearsal	iCaRL	Representation Learning, Nearest-Mean-of-Exemplars Classification	[36]
	EEIL	Cross-distilled loss function	[38]
	UC	Cosine Normalization	[39]
	LSIL	Store initial class statistics	[40]
	IL2M	Bias correction method	[41]
	MECIL	Low-fidelity exemplar	[42]
	GEM	Store previous gradient	[20]
Memory replay method - Generative replay	DGR	Introduces GAN	[43]
	DGM	Task-specific binary mask	[45]
	BIR	Replay hidden representation	[46]
	GRFC	Distillation and Replay-through-Feedback	[47]
	MeRGAN	Use conditional GAN	[48]
	CloGAN	Filter out distorted replay	[49]
Architecture method	PackNet	Iterative pruning and re-training technique	[51]
	PathNet	Computationally cost-effective	[53]
	PNN	Expand network architecture	[55]
	OIFL	Optimize feature set	[56]
	Adanet	Auto structure search	[57]
	EDIL	Expand network hierarchically	[60]
	NDL	Applied in AE	[61]
	DEN	Selective retraining	[62]
	RPSnet	Random search	[63]

### B. Information prospection

Preserving learned information is important but not sufficient in the CL scenario. The abovementioned methods can address forgetting aspects; however, they may not be able to facilitate future learning. Constant incorporation and efficient acquisition of new information are necessary in a long-term learning process. Only a few CL approaches have been specifically designed to leverage memory prospection, although certain interdisciplinary fields have combined CL with other learning paradigms, such as active learning and few-shot learning have witnessed remarkable enhancements in the promotion of memory prospection.

1) *Incremental metalearning*: Metalearning [64] is most commonly understood as learning to learn and refers to the process of enhancing a learning algorithm over multiple learning episodes. This approach uses the metadata regarding past experiences, such as hyperparameters, to promptly learn new experiences. The aim is to progressively increase learning efficiency while learning an increasing number of tasks, which is also the objective of CL in memory prospection. An

increasing number of researchers have designed or adopted metatraining strategies such as OML [65], MAML [66] and Reptile [67] to enhance the memory prospection in CL. These metatraining algorithms can finetune the model optimization to enable prompt adaptation to new tasks.

An example of applying OML in continual metalearning pertains to the work of Javed and White [65]. To address the continual learning prediction problem, the model is divided into two parts, specifically, the RLN that learns the features of inputs by representation learning, and the PLN that predicts a class for inputs based on the learned representations. In the PLN, the parameters are updated in a metalearning manner guided by online aware metalearning (OML). OML can learn a highly sparse and well-distributed representation by exploiting the large capacity of the representation. This aspect can help reduce forgetting because each update changes only a small number of weights, which in turn only affect a small number of inputs. Experimental results show that the OML is robust to interference under online updates and promotes future learning.

In addition to OML, MAML [66] and its variant FOMAML

[66] have been utilized. For example, Gupta et al. [68] optimize the OML objective in an online way through a multistep MAML procedure. The authors indicate that the gradient alignment among old tasks does not degrade while a new task is learned in OML; therefore, it is necessary to avoid repeated optimization of the inter-task alignment between old tasks to enable acceleration. Considering this aspect, the authors propose C-MAML to focus only on aligning the gradients of the current task and average gradient of the previous tasks. To prevent forgetting, the authors adopt La-MAML, in which the learning rates of the inner loop are clipped to positive values to avoid gradient ascension and interfering parameter updates, thereby mitigating catastrophic forgetting. SeqFOMAML [69] uses FOMAML to promptly learn tasks in the sequence and discourages interference between tasks. This approach first learns a prior by MAML to initialize the model parameters and relies on the plain stochastic gradient descent in the continual learning process over a sequence of tasks. Experimental results show that SeqFOMAML exhibits a superior performance for metageneralization to longer sequences. MAML is extended to Continual-MAML in the OSAKA framework [70]. Continual-MAML has two stages, namely, the pretraining phase, which initializes the model with parameters learned by the MAML, and the CL phase, which adapts the learned parameter initialization to solve new tasks. When a change in the distribution is detected, Continual-MAML adds the new knowledge into the learned initialization. With the introduction of Continual-MAML, OSAKA can rapidly solve new tasks and remember old tasks.

Furthermore, Reptile [67] has gained attention. For example, Riemer et al. propose meta-experience replay (MER), which can facilitate continual learning by maximizing useful transfer and minimizing interference. The objective is aimed at encouraging the network to share parameters when the gradient directions align and keep parameters separate when interference is caused by the gradients in opposite directions. To this end, this approach interleaves new inputs with the examples sampled from the replay buffer and later updates the parameters in the model based on Reptile by using these mixed data. Reptile can maximize the inner product between gradients of different minibatches from the same task, corresponding to enhanced generalization. By combining experience replay with optimization-based metalearning Reptile, the updated parameters are more likely to be transferred and less likely to incur interference with respect to past examples in the MER.

Instead of applying a specific metalearning algorithm, He et al. [71] propose a framework in which a wide range of meta-algorithms can be embedded, such as MAML, LEO [72], CAVIA [73] and CNP [71]. The aim of the framework is to exploit metalearning to accelerate the recovery of lost performance rather than focusing on remembering previous tasks, aided by the explicit inference of the current observed task. The authors design a metaframework known as What & How functions, which consists of an encoder or task inference network that predicts the current task representation based on the context data, and a decoder that maps the task representation to a task specific model. By using the continual learning

strategy BGD [74], the metaparameters in the framework can be updated continually to learn a sequence of tasks.

In general, metalearning enables prompt adaptation to new tasks in CL due to its powerful generalization ability. An increasing number of classic metalearning algorithms have been utilized in CL, and novel metalearning strategies specifically designed for CL are being gradually examined.

*2) Incremental few-shot learning:* Few-shot learning [75] refers to the ability to learn to recognize new concepts based on only a few samples. The research on few-shot learning mainly focuses on achieving a high generalization over new tasks with a limited number of training data. Historically, most methods pertaining to CL are based on the assumption that relatively large batches of training data are available. However, this assumption substantially limits the scalability of these approaches to the open-ended accommodation of novel classes with limited training data. Incremental few-shot learning is an promising approach to address this dilemma in the CL scenario. The key aspect in incremental few-shot learning is to prevent overfitting for new classes with only a few training samples on the basis of preserving old information.

To address the overfitting problem, few-shot learning adopts episodic training [76]. Specifically, tasks are sampled from support examples on which the learner is guided to learn. Subsequently, the learner is evaluated on query examples with an objective function to determine how well the learner can guide the learner to generalize to new tasks. Finally, the learner is optimized using the objective function. Through repetitive episodic training, the learner can gradually generalize over few-shot tasks. This training method has been widely incorporated in incremental few-shot learning.

Open-ended CentreNet (ONCE) [77] is designed to incrementally detect novel class objects with few examples. First, a feature extractor is trained with abundant base class training data and frozen to preserve the learned information. Subsequently, a generator that can synthesize class-specific codes for novel classes is trained, and finally, an object locator is trained to detect objects. The class-specific codes are specifically designed for each new task, and the information does not change the original weights in the model; thus, the forgetting problem can be addressed. When training the generator, ONCE adopts an episodic metalearning strategy in which a support set is constructed to train the model to learn and a query set is used to train the model to generalize. Cheraghian et al. [78] propose a semantic-specific information method for incremental few-shot classification tasks. Specifically, given an image as input, the semantic word vectors are estimated, and the similarity of the predicted word vectors with the word vectors from the set of possible class labels is measured to obtain the final class. The semantic information for novel classes is generated by the combination of multiple embedding features and global features trained via episodic training. In this way, the model can attain a high generalization when classifying both base and novel classes. Xiang et al. [79] adopt incremental few-shot learning for pedestrian attribute recognition. Specifically, the authors design APGM that extracts the multiple-attributed information from feature embedding and produces classification weights for the novel attributes. Subsequently, the

TABLE III  
SUMMARY ON INCREMENTAL METALEARNING, INCREMENTAL FEW-SHOT LEARNING, INCREMENTAL ACTIVE LEARNING.

Method	Algorithm	Keypoint	Reference
Incremental metalearning	Javed and White	RLN and PLN	[65]
	Gupta et al	Avoid repeated optimization	[68]
	SeqFOMAML	learn a prior initialization by MAML	[69]
	Continual-MAML	OSAKA framework	[70]
	MER	Update parameters by Reptile	[67]
Incremental few-shot learning	He et al	Universial metaframework	[71]
	ONCE	Meta-learned generator network	[77]
	Cheraghian et al	Meta-learned semantic information detector	[78]
	Xiang et al	Meta-learned feature abstracter	[79]
Incremental active learning	CBCL	Generate the centroids	[80]
	Ahmed et al	Two-level sample selection strategy	[87]
	MEN	Modified entropy Learning strategy	[88]
	Lin et al	Independent metrics	[90]
	Zhu et al	Bayesian network	[92]

approach samples  $N$ -way  $K$ -shot tasks from the base class as the support set and treats the sampled attributes from the support set as fake novel attributes. Next, the classification weights for fake novel attributes are generated under the guidance of APGM. Finally, the classification performances of the classification weights on the query set are evaluated, and the gradient is backpropagated to update the APGM. Since the APGM module does not interfere with the base model, the performances over the base data are not hampered. CBCL [80] involves the generation of a set of concepts in the form of centroids for each class using a clustering approach known as Agg-Var clustering. After generating the centroids, to predict the label of a test image, the distance of the feature vector of the test image to the  $n$  closest centroids is used. Since CBCL stores the centroids for each class independent of the other classes, the decrease in the overall classification accuracy is not catastrophic when new classes are learned. CBCL has been implemented to incremental robotic object detection tasks in which a robot learns different categories of objects from only a few examples. Experimental results show that even after learning 22 different categories of objects, the robot can correctly recognize previously learned objects at an accuracy of approximately 90%.

Incremental few-shot represents an emerging research direction, and it has been widely examined since the 2020 year [81]. Current incremental few-shot algorithms have considerably enhanced memory prospection by exploring and exploiting the structural information among different tasks.

3) *Incremental active learning*: Active learning algorithms are used in situations in which abundant unlabeled data are available but manual labeling is expensive. Active learning can interactively query a user (or another information source) to label new data points [82]. The key concept is that an algorithm can achieve a higher accuracy with fewer training instances if it is allowed to choose the data from which it learns.

Incremental active learning draws on this idea to enhance

memory prospection in CL. When active learning is performed, CL can be realized without learning each instance in future learning. The most useful data for current learning can be selected based on query strategies in active learning, for instance, by using the uncertainty sampling strategy [83], expected model change strategy [84] and variance reduction and Fisher information ratio strategy [85]. In this way, informative and critical data are preserved, whereas redundant and useless data are removed, thereby increasing the efficiency of future learning. To this end, designing appropriate query selection strategies for each new specific scene while preserving old scenes is important in incremental active learning, as shown in Figure 10. The subsequent section describes the ways in which different query selection strategies are applied to improve CL.

Ahmed et al. [86] present a two-level sample selection strategy that can enable motion planning networks (MPNets) [87] to learn from streaming data. To address the forgetting problem, the GEM method and replayed memory are employed to ensure lifelong learning. When updating the episodic memory in GEM and replayed memory, the approach first collects the training data by actively asking for demonstrations on problems in which MPNet fails to find a path. Second, the approach prunes the expert demonstrations to fill the episodic memory to enable the approximation of the global distribution from the streaming data. The online MPNet is successfully applied to robotic motion planning and navigation and exhibits comparable planning performance in tasks with approximately 80% less data than traditional approaches.

Shi et al. [88] introduce a novel active learning strategy known as modified entropy (MEN) to achieve incremental atrial fibrillation detection. Transfer learning is used to address forgetting. While continuously updating the model, the MEN selects the most useful information from massive medical data as the learning set. Specifically, the features and uncertainty of the predicted results are integrated to select the informative samples. Traditionally, samples with a high entropy are believed to be abundant in information. MEN assumes that

samples with a high RR intervals [89] series but low entropy are also informative. Next, the model is continuously finetuned using the most informative samples selected by MEN.

Lin et al [90] propose a sampling strategy specifically designed for the semantic segmentation of large-scale ALS point clouds [91] in urban areas. To identify the most informative parts of a point cloud for model training, this approach quantitatively assesses both data-dependent and model-dependent uncertainties by using three independent metrics. The data-dependent uncertainty is estimated through the point entropy and segment entropy. Segment entropy considers the interactions among neighboring points. The model-dependent uncertainty is estimated by mutual information, and the disagreements produced by different model parameters are evaluated. Experimental results demonstrate that all three metrics can help select informative point clouds that can be generalized to point clouds through terrestrial mobile laser scanners or indoor scenes.

Furthermore, Zhu et al. [92] introduce an active and dynamic selection method for crop disease diagnosis algorithms. The authors selected a subset of symptoms that are most relevant to the diagnosis of crop diseases based on a Bayesian network. In this approach, the symptoms are represented by the nodes of Bayesian networks, and another node pertaining to  $\{\text{diseases}\}$  is introduced, with each value of this node representing one crop disease. A Markov blanket [93] is used to obtain the symptom nodes that most significantly influence the disease. Experiments show that the proposed method is suitable for diagnosing not only crop diseases but also other diseases by extending the values of the disease variable and symptom variables to other disease situations.

Overall, incremental active learning algorithms can increase the efficiency of grasping new information in CL. Future work can be aimed at developing additional real-time incremental active learning algorithms because the current methods are slightly time-consuming in the context of the query strategy.

*4) Incremental curriculum learning:* In curriculum learning, the models first consider easy examples of a task and the task difficulty gradually increase over a sequence of tasks [94]. The objective of curriculum learning is to solve the last task, whereas the objective of CL is to be able to solve all tasks. In fact, the aim of finding a globally optimum CL can be achieved by gradual optimization in curriculum learning. Specifically, curriculum learning first optimizes a smoothed objective pertaining to easy examples, gradually adjusts the objective while maintaining the loss at a local minimum, and finally obtains the parameters that are globally optimal for all data.

The core of adopting curriculum learning in CL is to rearrange the sequence of examples according to their complexity. Intuitively, selecting a simpler example and building on this example to learn increasingly sophisticated representations are less time consuming than directly attempting to manage an unlearnable example when encountering a batch of new data. In addition, starting with easy samples can help the model avoid  $\{\text{noisy}\}$  data, which reduces the convergence speed. Although difficult examples are more informative than easy examples, they are likely not useful because they confuse the

learner rather than helping it establish the correct location of the decision surface [94]. In addition, experiments demonstrate that the generalization error pertaining to training on simple data is lower than that pertaining to training on randomly selected data. Therefore, the curriculum strategy can increase the learning efficiency, enhance the generalization ability and increase the convergence speed for acquiring new knowledge in CL.

According to this analysis, curriculum learning can considerably enhance the memory prospection of CL models if the learning sequence for the new data is ordered. Regretfully, only a few incremental curriculum learning algorithms have been developed, and the potential advantages of curriculum learning in the context of CL should be explored in the future.

Generally, efficiently grasping new knowledge is important for a CL model to manage situations involving unknown tasks. Multiple learning paradigms are demonstrated to be effective in leveraging memory prospection when they are incorporated into CL. The LCA, intransigence and scalability are superior to those of plain CL algorithms. These new learning methods are evaluated according to their performance and potential reported in the original papers, as shown in Table III.

### C. Information transfer

The actual world is structured, and most components are correlated; new situations look confusingly similar to old ones. In such a scenario, past information can be drawn on to facilitate future learning, and new information can be supplemented to enhance the established knowledge system. Therefore, whether appropriate knowledge is transferred to leverage related learning is an important evaluation metric related to CL.

According to the level of transferred information, the memory transfer techniques can be mainly divided into two categories. The first category involves directly transferring the outputs of old tasks, and the second category involves transferring hidden information such as feature maps, similarities and attributes across tasks.

The simplest way is to transfer the final results of previous tasks. For example, learning without forgetting (LwF) [95] records the output of old tasks for new data and applies a knowledge distillation loss  $0$  as a regularizer to minimize the bias of stored outputs for old tasks while adapting to the new task. iCaRL [36] forces the model to reproduce the results of old classes when the model is updated for the new task, and it uses an exemplar set to store representative exemplars of old tasks. Similarly, EEIL [38], LSIL [40], and IL2M [41] transfer the previous information in classification layers for the old classes.

However, retaining the results of past tasks in an invariant manner involves two problems: the model is prone to overfitting against the transferred examples, and the low relatedness among the tasks significantly increases the amount of forgetting. To exploit the useful information, certain methods propose transferring the relevant experience among tasks. Several algorithms transfer the previous feature information; for example, LFL [32] regularizes the  $l_2$  distance between the

TABLE IV  
SUMMARY ON INFORMATION TRANSFER METHOD.

Method	Algorithm	Keypoint	Reference
Direct result transfer	LWF	Knowledge distillation loss	[95]
	iCaRL	Reproduce old results	[36]
Relevant experience transfer	LFL	Transfer hidden activations	[32]
	P&C	Knowledge base	[96]
	PNN	Reuse low-level visual features	[55]
	Expert Gate	Gating autoencoders	[97]
	A-GEM	Task descriptors	[98]

final hidden activations and preserves the previously learned input-output mappings by computing additional activations with the parameters of the old tasks. In this way, the target network learns to extract features that are similar to the features extracted by the source network. Progress & compress (P&C) [96] introduces a knowledge base to transfer the learned features. The learned knowledge is constantly distilled into the knowledge base in the compression phase, and the learned features are selected and reused for new learning in the progress phase. PNN [55] supports the transfer of features across sequences of tasks. This approach enables the reuse of all aspects from low-level visual features to high-level policies to facilitate transfer to new tasks by utilizing lateral connections to previously learned models. In this manner, PNN achieves a richer composition, in which prior knowledge is no longer transient and can be integrated in each layer of the feature hierarchy. Expert Gate [97] selects the most appropriate specialist model when learning new tasks. This approach introduces gating autoencoders that inherently decide which of the most relevant prior models is to be used for training a new expert and captures the relatedness between tasks. A dubbed averaged GEM (A-GEM) [98] introduces task descriptors that describe the attributes of a task. The task descriptors are shared across tasks, and thus, a model can promptly recognize a new class provided a descriptor specifies certain attributes that it has learned.

Information transfer is a useful technique in CL, and it can be flexibly embedded into many CL algorithms to enhance the efficiency of the learning process. Current memory transfer techniques in the memory retrospection method migrate the previous information pertaining to the updated model to consolidate past memory, facilitating BWT and FWT. The summary on these transfer techniques is presented in Table IV.

#### IV. APPLICATION

A reliable and robust application is expected to pass tests in real-world scenarios in which data are dynamic and variable. Traditional training paradigms for models are confined in a fixed and static setting, resulting in their poor reaction to new surroundings. Therefore, to realize model applications, it is necessary to address the conflict between the inadequate response and demand for flexible adjustment according to different environments. The enhancement of the reaction is based on the preservation and utilization of learned experiences and

efficient acquisition of new knowledge. In this context, CL is a promising and appropriate learning paradigm for models in applications. This chapter provides an overview of the major applications in which CL plays an important role.

##### A. Object detection

With the increasing deployment of object detection globally, the variety and novelty of objects to be detected are expected to increase beyond the designated range of classes, requiring the existing models to be updated to detect additional classes. Incremental object detection can identify and localize additional instances of semantic objects of a certain class. This approach broadly consists of selecting region proposals, extracting features, and predicting the object class for new objects while preserving the old objects.

The early work in this domain focused on the continuous location of region proposals. For example, Shmelkov et al. [99] used knowledge distillation to measure the discrepancy of distillation proposals between the old and new networks. This discrepancy is added as an additional term in the standard cross-entropy loss function for new classes. By minimizing the final loss function, the model can balance the interplay between the detection of old and new classes. CIFRCN [100] involves an end-to-end class incremental detection method based on distilling of the RPN [101]. Continuous bounding box identification through RPN accelerates incremental object detection. Chen et al. [102] note that the confidence in the coarse labels of the initial model contains abundant knowledge of the learned class and can help the model retain the old class information in the ROI. Therefore, the authors innovatively use the confidence loss to extract the confidence information of the initial model to suppress forgetting.

Certain approaches extensively consider continual feature extraction. For example, RILOD [103] introduces an extra feature distillation loss in addition to the classification and bounding box loss. Feature distillation prevents dramatic changes in the features extracted from a middle neural network layer, which further reduces catastrophic forgetting and increases the model accuracy. RODEO [104] involves migrating the replay strategy to overcome forgetting in incremental object detection. This approach uses a memory buffer to store compressed representations of feature maps of old examples. Next, the new input is combined with a reconstructed subset of samples from the replay buffer when training the model over the new examples.

Certain approaches emphasize the continual classification loss in object detection because it is considered that the classifier misclassifies objects of old classes that lack class annotations as background when the model is adapted to new class images. IncDet [105] designs the pseudobounding box annotation of previously learned classes to replace the class annotations. Specifically, for each image in new datasets, IncDet predicts old class objects with the old model and obtains pseudobounding boxes for old classes. Next, the old class pseudobounding boxes and new class annotated bounding boxes are jointly used to incrementally finetune the model.

Overall, object detection fully exploits the existing CL methods in combination with knowledge distillation techniques. These incremental object detection algorithms have achieved promising results, in the context of single-stage networks such as RetinaNet [106] and YOLO [107] [108] and two-stage networks such as Fast RCNN [109] and Faster RCNN [101]. Most of the existing studies focus on maintaining satisfactory information retrospection and information transfer ability; however, with the deepening of research, the memory prospection problem has been gradually recognized. For example, IncDet [105] tests the mAP for new classes by adding them simultaneously to test the intransigence and LCA ability of the model.

### B. Image segmentation

Incremental image segmentation is aimed at realizing pixel-level labeling with a set of continuous object categories [110]. In contrast to object detection, in continuous semantic segmentation, each image contains pixels belonging to multiple classes, and thus, the labeling is dense. In addition, the pixels associated with the background during a learning step may be assigned to a specific object class in subsequent steps or vice versa, making the problem highly complicated. Therefore, an effective image segmentation model must classify objects in a fine-grained manner and handle background classes sequentially in addition to the standard requirement for incremental object detection. Most incremental objection detection approaches realize memory retrospection by exploiting the transfer learning technique.

Michieli et al. [110] use a masked cross-entropy loss between the logits produced by the output layer in the previous model and current model to transfer the final results and apply  $l_2$  loss to transfer the intermediate level of the feature space. This work is similar to the mainstream methods associated with incremental object detection.

Certain approaches divide traditional transfer learning into more detailed subprocesses. Mazen et al. [111] propose a coarse-to-fine semantic segmentation method. This approach transfers the previously gained knowledge, acquired on a simple semantic segmentation task with coarse classes, to a new model involving more fine-grained and detailed semantic classes. This approach exploits domain sharing at the feature extraction level and can thus provide insight for seeking commonalities and differences across tasks in incremental segmentation tasks.

Notably, the existing studies have not considered the incremental transfer of background information. Cermelli et al.

[112] report that the semantics associated with the background class change over time and may be assigned to a specific object class in subsequent learning, causing the inaccuracy of segmentation results. To address this problem, the researchers enhance the traditional cross-entropy and distillation loss by considering the background information and initialized the classifier parameters to prevent biased predictions toward the background class.

Another problem in transfer learning is the imbalance problem since classes in the training data are highly imbalanced in most cases when determining the training patches to be transferred. Tasar et al. [113] alleviate the sample imbalance problem by selecting previous training data with high importance values. This method can achieve excellent segmentation results for remote sensing data.

Klingner et al. [114] note that the existing approaches are either restricted to settings in which the additional classes have no overlap with the old ones or rely on labels for both old and new classes. The authors introduce a generally applicable technique that learns new data solely from labels for the new classes and outputs of a pretrained teacher model. The segmentation model is trained in stage 1 on dataset 1 for a set 1 of classes. Later, the model is extended in stage 2 by additional classes 2 and dataset 2, such that the model outputs both old and new classes.

Generally, the process of incremental image segmentation is similar to that of incremental objection detection, with a higher emphasis on updating background classes. Currently, transfer learning is the most widely applied method in the specific field, and it can achieve promising results in information retrospection.

### C. Face recognition

Face recognition [115] is widely implemented in real-world situations, such as e-passports, ID cards and entrance guarding frameworks. Owing to the increasing number of people involved and facial changes due to aging, sunlight, angle, occlusion, expressions, and make-up, the original or static face database is either insufficient or inappropriate for future events. A solution to this problem is to ensure that face recognition systems learn continuously to accommodate unknown faces and known faces with spatial and temporal variation. A substantial number of algorithms for continuous face detection were introduced twenty years ago, and many of them achieved promising results in memory retrospection, memory prospection and memory transfer.

To facilitate memory retrospection, Toh et al. [116] present an adaptive face recognition system based on memory storage. The initial idea is to accumulate and retrain all previous data when training new face images; however, this aspect is infeasible for large datasets. Therefore, the approach focuses only on misclassified samples. The approach uses RAN-LTM [117] to store the weights in hidden layers and create a memory item for misclassified samples. These pairs are trained with newly given training data to suppress forgetting. Kim et al. [118] propose an incremental method for classic short-term memory (STM) and long-term memory (LTM) algorithms for face

recognition. In this approach, the STM uses a recall algorithm to recall existing data from the LTM before incrementing new incoming data. La Torre et al. [119] report an incremental adaptive ensemble of classifiers in which new samples are trained and combined with the previously trained classifiers by using an iterative boolean combination (IBC).

Many researchers have focused on memory prospection for the latter task. For example, Chen et al. [120] adopt a constraint block NMF (CBNMF) to rapidly learn additional samples as well as additional class labels by increasing the between-class distances while reducing within-class distances based on a divergence criterion. Jia et al. [121] propose an incremental version of a manifold learning technique known as Laplacian eigenmaps, in which the adjacency matrix is updated, followed by the projection of the new sample onto the lower dimensional space. This fast and efficient feature extraction technique accelerates the subsequent learning. Ye and Yang [122] propose an incremental sparse representation classification (SRC) strategy. When new data arrive, the groups or classes to which these batch data belong are updated, while the remaining classes are retained. The high efficiency of this approach in updating the local dictionaries of the related class during training accelerates the learning process. Zhao et al. [123] state that the existing incremental principal component analysis (IPCA) [124] methods cannot manage the approximation error, which deteriorates the results of the subsequent recognition. To address this limitation, the authors proposed SVDU-IPCA based on the idea of an SVD [125] updating algorithm. SVDU-IPCA achieves excellent face recognition results with an extremely small degradation in the accuracy.

To flexibly transfer the knowledge learned in the training process, Hisada et al. [126] extend the incremental linear discriminant analysis (LDA) [127] for the multitask recognition problem. Several classifications are performed with the same input; the knowledge from each task is extended to the remaining tasks and incrementally updated over the incoming inputs. The proposed algorithm can perform several levels of recognition, such as name, age, and gender. Sakai [128] propose an incremental subspace generation approach known as the Monte Carlo subspace method that uses a row-incremental singular value decomposition (RiSVD) to select useful a priori knowledge to support current learning. Liwicki et al. [129] propose an incremental PCA over the Krein space [130], which can provide valuable insights for the classifier geometry. In the incremental update, the existing Krein subspace is used to create a complementary subspace based on the mapping functions of the new data. The correlation between the old and new inputs significantly enhances the recognition performance. Kang and Choi [131] suggest an incremental version of the support vector domain description (SVDD) [132] to find a minimum volume sphere that could enclose all or most of the data. In the incremental version, the authors proposed utilizing new data that contribute to the existing class descriptions to update the support vectors and discarding the remaining data.

In conclusion, incremental face recognition has been fully exploited. First, several techniques have been proposed to enhance memory retrospection, memory prospection and memory transfer. Second, in addition to traditional CL methods,

novel strategies from other domains have been integrated. Third, many detailed and specific problems have been solved, for example, targeted to illuminance invariance. Furthermore, the reduction in the computational complexity of incrementation has been considered [133] [128], which is of significance for large face datasets.

#### D. Image generation

Image generation refers to the task of generating new images from an existing dataset by learning high-dimensional data distributions. Image generation is widely implemented in fields with low data availability. With the requirement of diversified and big data, a model that can consecutively generate data according to different scene requirements is desirable. Continual image generation aims to continually generate new categories of images by remembering the old generation process [134]. A robust continuous image generation model can capture the difference between multiple datasets in terms of the data distribution. Many strategies to protect memory retrospection, such as regularization and generative replay methods, have been applied.

In terms of regularization methods, CLGAN [135] is a pioneering and representative approach. This technique employs EWC in the generator of a GAN to assess the weight importance and elastically updates the weights in the generator when learning new information. Weights are confined to a region benefiting the performance for all data classes to prevent forgetting. However, it is difficult for CLGAN to seek a shared feature space when the data distribution varies significantly or the number of tasks is large. The performance of CLGAN is tested and shown in Figure 8 (a).

The generative replay method is commonly used in image generation because the generator can reproduce past pseudosamples. DGR [43] regenerates data and their corresponding responses on the past model as a pair to represent old tasks, and the pairs are interleaved with new data to update the model. The performance of CLGAN is tested and shown in Figure 8 (b). Certain modern approaches focus on leveraging the quality of regenerated images. For example, MeRGAN [48] uses CGAN, in which the category is used as an input to guide the replay process, thereby avoiding less biased sampling on past categories and generating more reliable past data. CloGAN [49] uses an auxiliary classifier to filter a portion of distorted replay to block inferior images from entering the training loop. BIR [46] replays the internal or hidden representations of past tasks rather than the data. The replayed representation is generated by the context modulated feedback connection of the network. Similar to the replay idea, certain approaches reuse compressed information to represent old tasks rather than repetitive old data. DGM [45] introduces task-specific binary masks to the generator weights when training the model on a class of images. These approaches preserve the mask instead of the data as the old clues. Two variants of DGM, known as DGMw and DGMa [45], have been proposed to ensure sufficient model capacity to accommodate incoming tasks by an adaptive network expansion mechanism.

In addition to memory retrospection, memory transfer has been considered. Memory GAN [136] creates mFiLM and

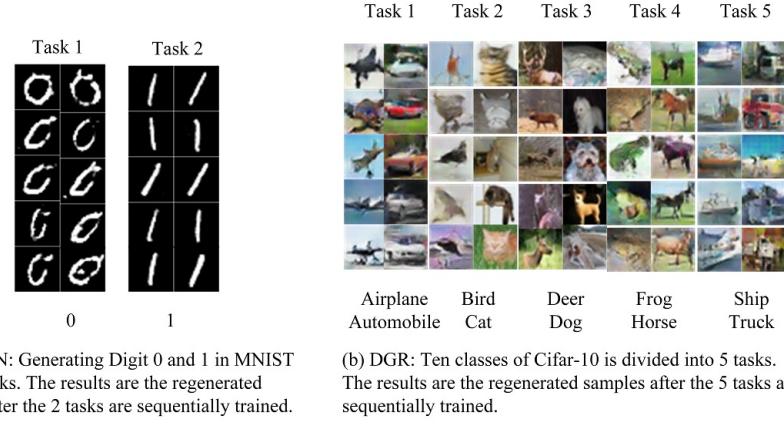


Fig. 8. Incremental image generation results of CLGAN with MNIST and DGR with Cifar-10.

mAdaFM modules to transfer the source information (old task) in the fully connected and convolutional layers, respectively, to the target domains (new task). The mFiLM and mAdaFM modules are designed for each individual task, and their parameters can be compressed to save memory, thereby allowing a long sequence of incremental tasks to be realized. Similarly, Piggyback GAN [137] adopts piggyback filters in which a set of convolutional and deconvolutional filters of previous tasks are factorized. Piggyback GAN can be scale up with the tasks since the filters for previous tasks are not altered and can be restored in a piggyback bank. In addition, the approach leverages weight reuse and adaptation to increase the parameter efficiency when extending GAN to new tasks. GRFC [47] also utilizes a distillation strategy to store past information and reduces the computational cost by integrating the generative model in the main model through generative feedback connections.

More complex tasks such as image-to-image translation have been explored. LiSS [138] achieves continual unpaired image-to-image translation on CycleGAN [139]. The approach distills the knowledge of a reference encoder, which is an exponential moving average of previous encoders in the parameter space, to assist CycleGAN in more effectively disentangling the instances of the objects to be translated. Since the reference encoder can maintain a weighted memory of all the past encoders at the cost of a single additional encoder, LiSS requires low memory and computational cost.

In general, many classic CL algorithms have been successfully applied to continual image generation, and the idea of a replay strategy has been widely studied. However, most of the existing research emphasizes the retention of the retrograde memory and neglects the anterograde memory. In addition, the existing techniques lack backward and forward knowledge transfer to boost bidirectional learning.

### E. Image fusion

Image fusion encloses all data analysis strategies aiming at combining the information of several images obtained with the same platform or by different spectroscopic platforms [140]. Since the source of images is always multiple, fusing

different types of images is required, for example, magnetic resonance imaging (MRI), computerized tomography (CT), positron emission tomography (PET) and single-photon emission computed tomography (SPECT) modalities are always combined together to detect disease in medical operation [141]; multi-spectral images and panchromatic images are always synthetized to detect abundant ground information in the remote sensing community [142].

Continual image fusion aims to solve different fusion problems, such as multi-modal, multi-exposure, multi-focus cases. Many DL methods have been putting forward to solve image fusion problem, but the continual image fusion work is far from satisfactory.

The latest and the only work is from U2Fusion [143], a unified model that is applicable to multiple fusion tasks. In U2Fusion, a feature extractor is first adopted to extract abundant and comprehensive features from source images. Then, the richness of information in features is measured to define the relative importance of these features by EWC. The relative importance represents the similarity relationship between the source images and the fusion result. A higher similarity entails that more information in this source image is preserved in the result, thus leading to a higher information preservation degree. We conducted the U2Fusion by its original EWC strategy and also tried the classical MAS method. Experiment results present that U2Fusion can achieve continual image fusion on multi-modal, multi-exposure, multi-focus cases. The performance of U2Fusion is tested as shown in Figure 9.

In conclusion, several applications were examined long ago, whereas others have been gradually considered in recent years. Therefore, the development level of these applications differs, and their performance on the information retrospection, information prospection and information transfer is shown in Figure 10. Generally, with the combination of CL, the models are exhibiting a higher generalization ability in the open world and are thus more reliable and efficient in the relevant applications.

## V. DISCUSSION

Overall, CL has been fully developed and facilitated developments in other aspects of artificial intelligence. Considering

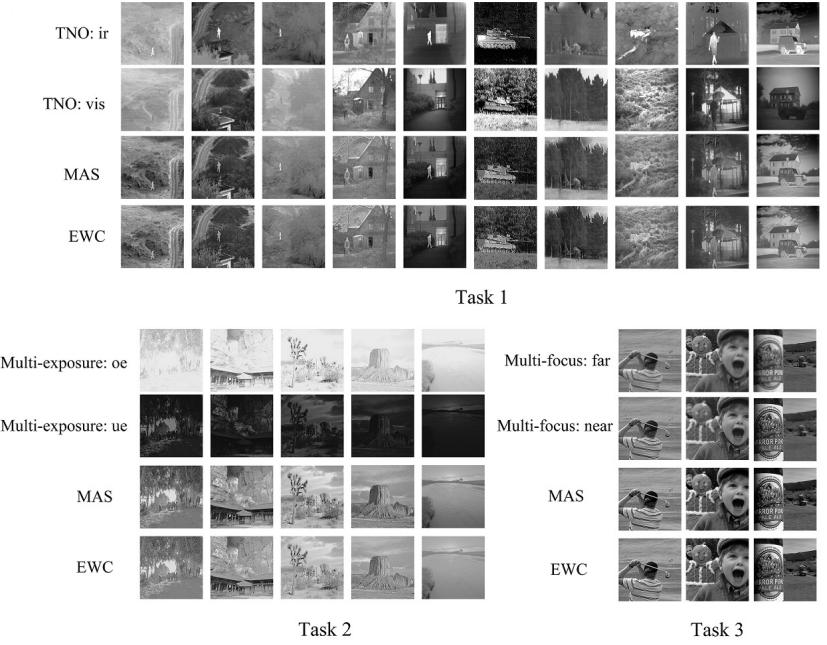


Fig. 9. Incremental image fusion of U2Fusion with three continuous tasks.

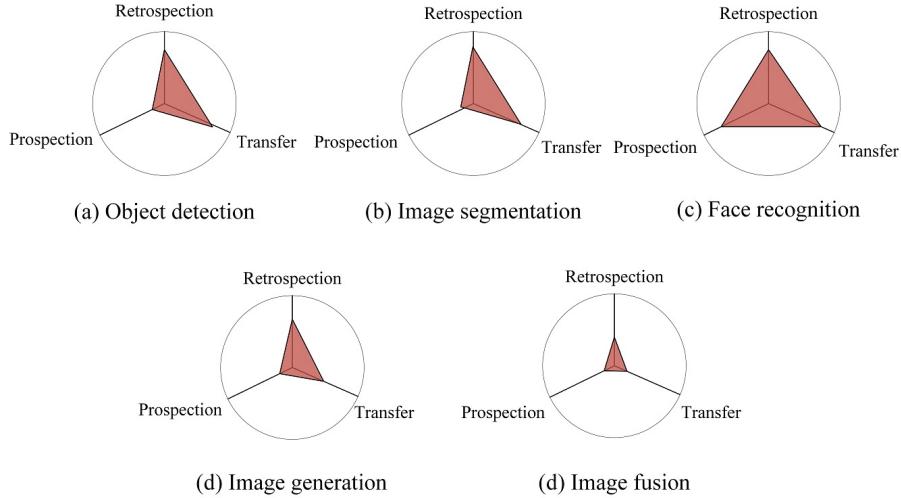


Fig. 10. Valuations of the major CL applications on the performance of information retrospection, information prospection and information transfer..

the continual research advancements, we present several new concepts based on neurobiology knowledge.

To enhance memory retrospection, the following ideas may be considered.

- *Considering the chain reaction associated with the change in weights.* Enhancing the values of certain weights and dampening the action of neighboring weights in a model may help form a  $\pm$ receptive field for inputs, which can avoid a specific task from being interfered with. This insight is inspired by a neuroscience phenomenon termed "lateral inhibition" [144] that describes how an excited neuron reduces

the activity of its neighbors. Lateral inhibition disables the spreading of action potentials from excited neurons to neighboring neurons in the lateral direction, which can create a contrast in the stimulation that allows an enhanced sensory perception. The increased stimulation can help the human brain form the initial memory. The existing CL methods individually and separately regard and adjust the activities of each weight. It must be examined whether the chain reaction resulting from the change in weights can improve memory retention.

- *The modality of past information should be innovated.*

Most of the methods to protect retrograde memory use true or regenerated samples when referring to old data. However, the information that is processed in the brain is abstracted and compressed in the hippocampus [145] [146], which indicates that the modality of learned information is not as simple as its original input. In addition, condensed information can help models focus on essential information and save computational overhead. Therefore, enhancing the form of old information is a promising research direction.

- *The relevance of different past information should be considered.* When constructing a memory buffer for past memory, the existing methods mix the data. However, different categories of data may share similar low-level information and are mutually correlated in other aspects. In addition, biological research indicates that learned information is reorganized between the hippocampus and neocortex, with several neurons being weakened, strengthened and newly formed [147]. Therefore, the old information in the memory buffer should not be a simple integration of the original data and can be structured as a knowledge graph or in other modalities.
- *Preserving past information in an online way.* The existing approaches select or generate old samples after learning the task and before learning a new task, which is an offline technique. Recent neuroscience research shows that hippocampal replay not only exists in rest or sleep but also occurs during the learning of the current task [148]. Therefore, future work can examine techniques to leverage the process of retrograde memory preservation in an online and dynamic way, which can not only accelerate the training of the model but mimic human intelligence.

To facilitate memory prospecting, the following innovative idea can be considered.

- *Considering additional "forgetting" to provide space for information prospecting.* LTD, which represents forgetting, is as important as LTP, which represents the memorization of the synaptic plasticity, because our brain may be exhausted with the influx of information if no forgetting occurs. Moreover, the idea that forgetting might be beneficial for memory maintenance has been frequently expressed [149]. Existing CL models always focus on memorizing important old information, and it must be examined whether active forgetting plays an important role in anterograde memory. By forgetting certain redundant information, the CL models may have additional room to learn new information and more flexibility to optimize this information.

To enable memory transfer, the following changes can be introduced in the model architecture:

- *Designing submodels based on the main CL models to regulate information transfer.* The CLS theory states that the hippocampal system exhibits short-term adaptation and enables the rapid learning of novel information that is played back over time to the neocortical system for long-

term retention [150]. The interplay of hippocampal and neocortical functionalities is crucial to transfer different memories. The cooperation of different areas in the brain indicates that the expansion of the structure of the model is not necessarily restricted to the original one. In fact, more concurrent submodels can be designed to regulate memory circulation in combination with the main models.

## VI. CONCLUSION

CL has attracted considerable attention in the deep learning community, and many effective algorithms have been proposed to achieve CL in ANNs. We reconsider the learning of CL in three aspects: memory retrospection, memory prospecting and memory transfer, homologous to the human CL system. Consolidating information retrospection is a prerequisite in CL; information prospecting represents the fast acquisition and adaptation of novel knowledge; flexibly transferring information promotes the efficiency and intelligence of a CL agent. Although significant progress and achievements have been witnessed in CL, additional illuminating ideas deserve to be introduced. Rethinking CL from the perspective of neuroscience knowledge can help design CL agents that can mimic human intelligence.

## REFERENCES

- [1] S. Russell and P. Norvig, “Artificial intelligence: a modern approach,” 2002.
- [2] L. G. Humphreys, “The construct of general intelligence,” 1979.
- [3] A. Eisen, *Handbook of human intelligence*. CUP Archive, 1982.
- [4] R. Feuerstein, R. S. Feuerstein, L. H. Falik, and Y. Rand, *The dynamic assessment of cognitive modifiability: The Learning Propensity Assessment Device: Theory, instruments and techniques, Rev. and exp. ed. of The dynamic assessment of retarded performers*. ICELP Publications, 2002.
- [5] L. Schulz, “The origins of inquiry: Inductive inference and exploration in early childhood,” *Trends in cognitive sciences*, vol. 16, no. 7, pp. 382–389, 2012.
- [6] M. G. Bedia, J. M. Corchado, and L. F. Castillo, “Bio-inspired dynamical tools for analyzing cognition,” in *Encyclopedia of Artificial Intelligence*. IGI Global, 2009, pp. 256–261.
- [7] W. C. Abraham and A. Robins, “Memory retention—the synaptic stability versus plasticity dilemma,” *Trends in neurosciences*, vol. 28, no. 2, pp. 73–78, 2005.
- [8] M. B. Ring *et al.*, “Continual learning in reinforcement environments,” 1994.
- [9] G. M. Van de Ven and A. S. Tolias, “Three scenarios for continual learning,” *arXiv preprint arXiv:1904.07734*, 2019.
- [10] S. Thrun and T. M. Mitchell, “Lifelong robot learning,” *Robotics and autonomous systems*, vol. 15, no. 1-2, pp. 25–46, 1995.
- [11] S. Farquhar and Y. Gal, “Towards robust evaluations of continual learning,” *arXiv preprint arXiv:1805.09733*, 2018.
- [12] J. Ferbinteanu and M. L. Shapiro, “Prospective and retrospective memory coding in the hippocampus,” *Neuron*, vol. 40, no. 6, pp. 1227–1239, 2003.
- [13] M. McCloskey and N. J. Cohen, “Catastrophic interference in connectionist networks: The sequential learning problem,” in *Psychology of learning and motivation*. Elsevier, 1989, vol. 24, pp. 109–165.
- [14] R. Ratcliff, “Connectionist models of recognition memory: constraints imposed by learning and forgetting functions.” *Psychological review*, vol. 97, no. 2, p. 285, 1990.
- [15] J. Ellis and L. Kvavilashvili, “Prospective memory in 2000: Past, present, and future directions,” *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, vol. 14, no. 7, pp. S1–S9, 2000.
- [16] L. Torrey and J. Shavlik, “Transfer learning. handbook of research on machine learning applications,” *IGI Global*, vol. 3, pp. 17–35, 2009.

- [17] N. M. Hunkin, A. J. Parkin, V. A. Bradley, E. H. Burrows, F. K. Aldrich, A. Jansari, and C. Burdon-Cooper, "Focal retrograde amnesia following closed head injury: A case study and theoretical account," *Neuropsychologia*, vol. 33, no. 4, pp. 509–523, 1995.
- [18] A. Chaudhry, P. K. Dokania, T. Ajanthan, and P. H. Torr, "Riemannian walk for incremental learning: Understanding forgetting and intransigence," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 532–547.
- [19] B. Pfülb and A. Gepperth, "A comprehensive, application-oriented study of catastrophic forgetting in dnns," *arXiv preprint arXiv:1905.08101*, 2019.
- [20] D. Lopez-Paz and M. Ranzato, "Gradient episodic memory for continual learning," *Advances in neural information processing systems*, vol. 30, pp. 6467–6476, 2017.
- [21] N. Khan, H. Abboudi, M. S. Khan, P. Dasgupta, and K. Ahmed, "Measuring the surgical j@learning curve: methods, variables and competency," *BjU Int*, vol. 113, no. 3, pp. 504–508, 2014.
- [22] J. Rajasegaran, M. Hayat, S. Khan, F. S. Khan, L. Shao, and M.-H. Yang, "An adaptive random path selection approach for incremental learning," *arXiv preprint arXiv:1906.01120*, 2019.
- [23] J. R. Hughes, "Post-tetanic potentiation," *Physiological reviews*, vol. 38, no. 1, pp. 91–113, 1958.
- [24] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.
- [25] J. Liu, H. Yuan, X.-M. Lu, and X. Wang, "Quantum fisher information matrix and multiparameter estimation," *Journal of Physics A: Mathematical and Theoretical*, vol. 53, no. 2, p. 023001, 2019.
- [26] X. Liu, M. Masana, L. Herranz, J. Van de Weijer, A. M. Lopez, and A. D. Bagdanov, "Rotate your networks: Better weight consolidation and less catastrophic forgetting," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2262–2268.
- [27] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
- [28] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars, "Memory aware synapses: Learning what (not) to forget," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 139–154.
- [29] F. Zenke, B. Poole, and S. Ganguli, "Continual learning through synaptic intelligence," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3987–3995.
- [30] W. Hu, Z. Lin, B. Liu, C. Tao, Z. Tao, J. Ma, D. Zhao, and R. Yan, "Overcoming catastrophic forgetting for continual learning via model adaptation," in *International Conference on Learning Representations*, 2018.
- [31] A. Rannen, R. Aljundi, M. B. Blaschko, and T. Tuytelaars, "Encoder based lifelong learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1320–1328.
- [32] H. Jung, J. Ju, M. Jung, and J. Kim, "Less-forgetting learning in deep neural networks," *arXiv preprint arXiv:1607.00122*, 2016.
- [33] S.-W. Lee, J.-H. Kim, J. Jun, J.-W. Ha, and B.-T. Zhang, "Overcoming catastrophic forgetting by incremental moment matching," *arXiv preprint arXiv:1703.08475*, 2017.
- [34] J. Born and I. Wilhelm, "System consolidation of memory during sleep," *Psychological research*, vol. 76, no. 2, pp. 192–203, 2012.
- [35] J. O'keefe and L. Nadel, *The hippocampus as a cognitive map*. Oxford university press, 1978.
- [36] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.
- [37] C. J. Veenman and M. J. Reinders, "The nearest subclass classifier: A compromise between the nearest mean and nearest neighbor classifier," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 9, pp. 1417–1429, 2005.
- [38] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari, "End-to-end incremental learning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 233–248.
- [39] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin, "Learning a unified classifier incrementally via rebalancing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 831–839.
- [40] Y. Wu, Y. Chen, L. Wang, Y. Ye, Z. Liu, Y. Guo, and Y. Fu, "Large scale incremental learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 374–382.
- [41] E. Belouadah and A. Popescu, "Il2m: Class incremental learning with dual memory," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 583–592.
- [42] H. Zhao, H. Wang, Y. Fu, F. Wu, and X. Li, "Memory efficient class-incremental learning for image classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.
- [43] H. Shin, J. K. Lee, J. Kim, and J. Kim, "Continual learning with deep generative replay," *arXiv preprint arXiv:1705.08690*, 2017.
- [44] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.
- [45] O. Ostapenko, M. Puscas, T. Klein, P. Jahnichen, and M. Nabi, "Learning to remember: A synaptic plasticity driven framework for continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11321–11329.
- [46] G. M. van de Ven, H. T. Siegelmann, and A. S. Tolias, "Brain-inspired replay for continual learning with artificial neural networks," *Nature communications*, vol. 11, no. 1, pp. 1–14, 2020.
- [47] G. M. Van de Ven and A. S. Tolias, "Generative replay with feedback connections as a general strategy for continual learning," *arXiv preprint arXiv:1809.10635*, 2018.
- [48] C. Wu, L. Herranz, X. Liu, J. van de Weijer, B. Raducanu *et al.*, "Memory replay gans: Learning to generate new categories without forgetting," *Advances in Neural Information Processing Systems*, vol. 31, pp. 5962–5972, 2018.
- [49] A. Rios and L. Itti, "Closed-loop memory gan for continual learning," *arXiv preprint arXiv:1811.01146*, 2018.
- [50] C. Zhao, W. Deng, and F. H. Gage, "Mechanisms and functional implications of adult neurogenesis," *Cell*, vol. 132, no. 4, pp. 645–660, 2008.
- [51] A. Mallya and S. Lazebnik, "Packnet: Adding multiple tasks to a single network by iterative pruning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7765–7773.
- [52] S. Han, J. Pool, S. Narang, H. Mao, E. Gong, S. Tang, E. Elsen, P. Vajda, M. Paluri, J. Tran *et al.*, "Dsd: Dense-sparse-dense training for deep neural networks," *arXiv preprint arXiv:1607.04381*, 2016.
- [53] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra, "Pathnet: Evolution channels gradient descent in super neural networks," *arXiv preprint arXiv:1701.08734*, 2017.
- [54] A. Shukla, H. M. Pandey, and D. Mehrotra, "Comparative review of selection techniques in genetic algorithm," in *2015 international conference on futuristic trends on computational analysis and knowledge management (ABLAZE)*. IEEE, 2015, pp. 515–519.
- [55] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," *arXiv preprint arXiv:1606.04671*, 2016.
- [56] G. Zhou, K. Sohn, and H. Lee, "Online incremental feature learning with denoising autoencoders," in *Artificial intelligence and statistics*. PMLR, 2012, pp. 1453–1461.
- [57] C. Cortes, X. Gonzalo, V. Kuznetsov, M. Mohri, and S. Yang, "Adanet: Adaptive structural learning of artificial neural networks," in *International conference on machine learning*. PMLR, 2017, pp. 874–883.
- [58] V. Koltchinskii and D. Panchenko, "Empirical margin distributions and bounding the generalization error of combined classifiers," *The Annals of Statistics*, vol. 30, no. 1, pp. 1–50, 2002.
- [59] M. Kotani, A. Kajiki, and K. Akazawa, "A structural learning algorithm for multi-layered neural networks," in *Proceedings of International Conference on Neural Networks (ICNN'97)*, vol. 2. IEEE, 1997, pp. 1105–1110.
- [60] T. Xiao, J. Zhang, K. Yang, Y. Peng, and Z. Zhang, "Error-driven incremental learning in deep convolutional neural network for large-scale image classification," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 177–186.
- [61] T. J. Draelos, N. E. Miner, C. C. Lamb, J. A. Cox, C. M. Vineyard, K. D. Carlson, W. M. Severa, C. D. James, and J. B. Aimone, "Neurogenesis deep learning: Extending deep networks to accommodate new classes," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 526–533.
- [62] J. Yoon, E. Yang, J. Lee, and S. J. Hwang, "Lifelong learning with dynamically expandable networks," *arXiv preprint arXiv:1708.01547*, 2017.
- [63] J. Rajasegaran, M. Hayat, S. Khan, F. S. Khan, and L. Shao, "Random path selection for incremental learning," *Advances in Neural Information Processing Systems*, 2019.

- [64] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, “Meta-learning in neural networks: A survey,” *arXiv preprint arXiv:2004.05439*, 2020.
- [65] K. Javed and M. White, “Meta-learning representations for continual learning,” *arXiv preprint arXiv:1905.12588*, 2019.
- [66] C. Finn, P. Abbeel, and S. Levine, “Model-agnostic meta-learning for fast adaptation of deep networks,” in *International Conference on Machine Learning*. PMLR, 2017, pp. 1126–1135.
- [67] A. Nichol and J. Schulman, “Reptile: a scalable metalearning algorithm,” *arXiv preprint arXiv:1803.02999*, vol. 2, no. 3, p. 4, 2018.
- [68] G. Gupta, K. Yadav, and L. Paull, “La-maml: Look-ahead meta learning for continual learning,” *arXiv preprint arXiv:2007.13904*, 2020.
- [69] G. Spigler, “Meta-learnt priors slow down catastrophic forgetting in neural networks,” *arXiv preprint arXiv:1909.04170*, 2019.
- [70] M. Caccia, P. Rodriguez, O. Ostapenko, F. Normandin, M. Lin, L. Page-Caccia, I. H. Laradji, I. Rish, A. Lacoste, D. Vázquez *et al.*, “Online fast adaptation and knowledge accumulation (osaka): a new approach to continual learning,” *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [71] X. He, J. Sygnowski, A. Galashov, A. A. Rusu, Y. W. Teh, and R. Pascanu, “Task agnostic continual learning via meta learning,” *arXiv preprint arXiv:1906.05201*, 2019.
- [72] A. A. Rusu, D. Rao, J. Sygnowski, O. Vinyals, R. Pascanu, S. Osindero, and R. Hadsell, “Meta-learning with latent embedding optimization,” *arXiv preprint arXiv:1807.05960*, 2018.
- [73] L. Zintgraf, K. Shiarli, V. Kurin, K. Hofmann, and S. Whiteson, “Fast context adaptation via meta-learning,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 7693–7702.
- [74] C. Zeno, I. Golani, E. Hoffer, and D. Soudry, “Bayesian gradient descent: Online variational bayes learning with increased robustness to catastrophic forgetting and weight pruning,” 2018.
- [75] L. Fei-Fei, R. Fergus, and P. Perona, “One-shot learning of object categories,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 4, pp. 594–611, 2006.
- [76] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra *et al.*, “Matching networks for one shot learning,” *Advances in neural information processing systems*, vol. 29, pp. 3630–3638, 2016.
- [77] J.-M. Perez-Rua, X. Zhu, T. M. Hospedales, and T. Xiang, “Incremental few-shot object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 846–13 855.
- [78] A. Cheraghian, S. Rahman, P. Fang, S. K. Roy, L. Petersson, and M. Harandi, “Semantic-aware knowledge distillation for few-shot class-incremental learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2534–2543.
- [79] L. Xiang, X. Jin, G. Ding, J. Han, and L. Li, “Incremental few-shot learning for pedestrian attribute recognition,” *arXiv preprint arXiv:1906.00330*, 2019.
- [80] A. Ayub and A. Wagner, “Cbcl: Brain inspired model for rgb-d indoor scene classification,” *arXiv preprint arXiv:1911.00155*, vol. 2, no. 3, 2019.
- [81] A. Ayub and A. R. Wagner, “Tell me what this is: few-shot incremental object learning by a robot,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 8344–8350.
- [82] B. Settles, “Active learning literature survey,” 2009.
- [83] D. D. Lewis and W. A. Gale, “A sequential algorithm for training text classifiers,” in *SIGIR’94*. Springer, 1994, pp. 3–12.
- [84] B. Settles, M. Craven, and S. Ray, “Multiple-instance active learning,” *Advances in neural information processing systems*, vol. 20, pp. 1289–1296, 2007.
- [85] D. A. Cohn, Z. Ghahramani, and M. I. Jordan, “Active learning with statistical models,” *Journal of artificial intelligence research*, vol. 4, pp. 129–145, 1996.
- [86] M. C. Y. Ahmed H. Qureshi, Yinglong Miao, “Active continual learning for planning and navigation,” in *2020 Workshop on Real World Experiment Design and Active Learning*.
- [87] A. H. Qureshi, A. Simeonov, M. J. Bency, and M. C. Yip, “Motion planning networks,” in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 2118–2124.
- [88] H. Shi, H. Wang, C. Qin, L. Zhao, and C. Liu, “An incremental learning system for atrial fibrillation detection based on transfer learning and active learning,” *Computer methods and programs in biomedicine*, vol. 187, p. 105219, 2020.
- [89] M. Malik, P. Färbom, V. Batchvarov, K. Hnatkova, and A. Camm, “Relation between qt and rr intervals is highly individual among healthy subjects: implications for heart rate correction of the qt interval,” *Heart*, vol. 87, no. 3, pp. 220–228, 2002.
- [90] Y. Lin, G. Vosselman, Y. Cao, and M. Y. Yang, “Active and incremental learning for semantic als point cloud segmentation,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 169, pp. 73–92, 2020.
- [91] T. Melzer and C. Briese, “Extraction and modeling of power lines from als point clouds,” 2004.
- [92] Y. Zhu, D. Liu, G. Chen, H. Jia, and H. Yu, “Mathematical modeling for active and dynamic diagnosis of crop diseases based on bayesian networks and incremental learning,” *Mathematical and Computer Modelling*, vol. 58, no. 3-4, pp. 514–523, 2013.
- [93] I. Tsamardinos, C. F. Aliferis, A. R. Statnikov, and E. Statnikov, “Algorithms for large scale markov blanket discovery.” in *FLAIRS conference*, vol. 2, 2003, pp. 376–380.
- [94] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 41–48.
- [95] Z. Li and D. Hoiem, “Learning without forgetting,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.
- [96] J. Schwarz, W. Czarnecki, J. Luketina, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, and R. Hadsell, “Progress & compress: A scalable framework for continual learning,” in *International Conference on Machine Learning*. PMLR, 2018, pp. 4528–4537.
- [97] R. Aljundi, P. Chakravarthy, and T. Tuytelaars, “Expert gate: Lifelong learning with a network of experts,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3366–3375.
- [98] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, “Efficient lifelong learning with a-gem,” *arXiv preprint arXiv:1812.00420*, 2018.
- [99] K. Shmelkov, C. Schmid, and K. Alahari, “Incremental learning of object detectors without catastrophic forgetting,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3400–3409.
- [100] Y. Hao, Y. Fu, Y.-G. Jiang, and Q. Tian, “An end-to-end architecture for class-incremental object detection with knowledge distillation,” in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 1–6.
- [101] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.
- [102] L. Chen, C. Yu, and L. Chen, “A new knowledge distillation for incremental object detection,” in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–7.
- [103] D. Li, S. Tasci, S. Ghosh, J. Zhu, J. Zhang, and L. Heck, “Rilon: Near real-time incremental learning for object detection at the edge,” in *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, 2019, pp. 113–126.
- [104] M. Acharya, T. L. Hayes, and C. Kanan, “Rodeo: Replay for online object detection,” *arXiv preprint arXiv:2008.06439*, 2020.
- [105] L. Liu, Z. Kuang, Y. Chen, J.-H. Xue, W. Yang, and W. Zhang, “Incdet: In defense of elastic weight consolidation for incremental object detection,” *IEEE transactions on neural networks and learning systems*, 2020.
- [106] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
- [107] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
- [108] J. Redmon and A. Farhadi, “Yolo9000: better, faster, stronger,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.
- [109] X. Wang, A. Shrivastava, and A. Gupta, “A-fast-rnnc: Hard positive generation via adversary for object detection,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2606–2615.
- [110] U. Michieli and P. Zanuttigh, “Incremental learning techniques for semantic segmentation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.
- [111] M. Mel, U. Michieli, and P. Zanuttigh, “Incremental and multi-task learning strategies for coarse-to-fine semantic segmentation,” *Technologies*, vol. 8, no. 1, p. 1, 2020.
- [112] F. Cermelli, M. Mancini, S. R. Bulo, E. Ricci, and B. Caputo, “Modeling the background for incremental learning in semantic segmentation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9233–9242.

- [113] O. Tasar, Y. Tarabalka, and P. Alliez, "Incremental learning for semantic segmentation of large-scale remote sensing data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 9, pp. 3524–3537, 2019.
- [114] M. Klingner, A. Bär, P. Donn, and T. Fingscheidt, "Class-incremental learning for semantic segmentation re-using neither old data nor old labels," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–8.
- [115] A. K. Jain and S. Z. Li, *Handbook of face recognition*. Springer, 2011, vol. 1.
- [116] S. L. Toh and S. Ozawa, "A face recognition system using neural networks with incremental learning ability," in *Proc. 8th Australian and New Zealand Conf. on Intelligent Information Systems*, 2003, pp. 389–394.
- [117] J. Platt, "A resource-allocating network for function interpolation," *Neural computation*, vol. 3, no. 2, pp. 213–225, 1991.
- [118] S. Kim, R. Mallipeddi, and M. Lee, "Incremental face recognition: hybrid approach using short-term memory and long-term memory," in *International Conference on Neural Information Processing*. Springer, 2012, pp. 194–201.
- [119] M. De-la Torre, E. Granger, P. V. Radtke, R. Sabourin, and D. O. Gorodnichy, "Incremental update of biometric models in face-based video surveillance," in *The 2012 international joint conference on neural networks (IJCNN)*. IEEE, 2012, pp. 1–8.
- [120] W.-S. Chen, B.-B. Pan, B. Fang, and J. Zou, "A novel constraint non-negative matrix factorization criterion based incremental learning in face recognition," in *2008 International conference on wavelet analysis and pattern recognition*, vol. 1. IEEE, 2008, pp. 292–297.
- [121] P. Jia, J. Yin, X. Huang, and D. Hu, "Incremental laplacian eigenmaps by preserving adjacent information between data points," *Pattern Recognition Letters*, vol. 30, no. 16, pp. 1457–1463, 2009.
- [122] J. Ye and R. Yang, "An incremental src method for face recognition," in *Pacific Rim Conference on Multimedia*. Springer, 2015, pp. 170–180.
- [123] H. Zhao, P. C. Yuen, and J. T. Kwok, "A novel incremental principal component analysis and its application for face recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 36, no. 4, pp. 873–886, 2006.
- [124] D. H. Jeong, C. Ziemkiewicz, B. Fisher, W. Ribarsky, and R. Chang, "ipca: An interactive system for pca-based visual analytics," in *Computer Graphics Forum*, vol. 28, no. 3. Wiley Online Library, 2009, pp. 767–774.
- [125] H. Zha and H. D. Simon, "On updating problems in latent semantic indexing," *SIAM Journal on Scientific Computing*, vol. 21, no. 2, pp. 782–791, 1999.
- [126] M. Hisada, S. Ozawa, K. Zhang, and N. Kasabov, "Incremental linear discriminant analysis for evolving feature spaces in multitask pattern recognition problems," *Evolving Systems*, vol. 1, no. 1, pp. 17–27, 2010.
- [127] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," *Institute for Signal and information Processing*, vol. 18, no. 1998, pp. 1–8, 1998.
- [128] T. Sakai, "Monte carlo subspace method: an incremental approach to high-dimensional data classification," in *2008 19th International conference on pattern recognition*. IEEE, 2008, pp. 1–4.
- [129] S. Liwicki, S. Zafeiriou, G. Tzimiropoulos, and M. Pantic, "Efficient online subspace learning with an indefinite kernel for visual tracking and recognition," *IEEE transactions on neural networks and learning systems*, vol. 23, no. 10, pp. 1624–1636, 2012.
- [130] H. Langer and C. Tretter, "A krein space approach to pt-symmetry," *Czechoslovak Journal of Physics*, vol. 54, no. 10, pp. 1113–1120, 2004.
- [131] W.-S. Kang and J. Y. Choi, "Kernel machine for fast and incremental learning of face," in *2006 SICE-ICASE International Joint Conference*. IEEE, 2006, pp. 1015–1019.
- [132] D. M. Tax and R. P. Duin, "Support vector domain description," *Pattern recognition letters*, vol. 20, no. 11-13, pp. 1191–1199, 1999.
- [133] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Online nonnegative matrix factorization with robust stochastic approximation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 7, pp. 1087–1099, 2012.
- [134] K. Gregor, I. Danihelka, A. Graves, D. Rezende, and D. Wierstra, "Draw: A recurrent neural network for image generation," in *International Conference on Machine Learning*. PMLR, 2015, pp. 1462–1471.
- [135] A. Seff, A. Beatson, D. Suo, and H. Liu, "Continual learning in generative adversarial nets," *arXiv preprint arXiv:1705.08395*, 2017.
- [136] Y. Cong, M. Zhao, J. Li, S. Wang, and L. Carin, "Gan memory with no forgetting," *arXiv preprint arXiv:2006.07543*, 2020.
- [137] M. Zhai, L. Chen, J. He, M. Nawhal, F. Tung, and G. Mori, "Piggyback gan: Efficient lifelong learning for image conditioned generation," in *European Conference on Computer Vision*. Springer, 2020, pp. 397–413.
- [138] V. Schmidt, M. N. Sreedhar, M. ElAraby, and I. Rish, "Towards lifelong self-supervision for unpaired image-to-image translation," *arXiv preprint arXiv:2004.00161*, 2020.
- [139] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.
- [140] M. Cocchi, *Data fusion methodology and applications*. Elsevier, 2019.
- [141] J. Du, W. Li, K. Lu, and B. Xiao, "An overview of multi-modal medical image fusion," *Neurocomputing*, vol. 215, pp. 3–20, 2016.
- [142] Z. Wang, D. Ziou, C. Armenakis, D. Li, and Q. Li, "A comparative analysis of image fusion methods," *IEEE transactions on geoscience and remote sensing*, vol. 43, no. 6, pp. 1391–1402, 2005.
- [143] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [144] S.-i. Amari, "Dynamics of pattern formation in lateral-inhibition type neural fields," *Biological cybernetics*, vol. 27, no. 2, pp. 77–87, 1977.
- [145] N. Burgess, J. G. Donnett, and J. O'Keefe, "The representation of space and the hippocampus in rats, robots and humans," *Zeitschrift für Naturforschung C*, vol. 53, no. 7-8, pp. 504–509, 1998.
- [146] C. Pavlides and J. Winson, "Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes," *Journal of neuroscience*, vol. 9, no. 8, pp. 2907–2918, 1989.
- [147] K. Diba and G. Buzsáki, "Forward and reverse hippocampal place-cell sequences during ripples," *Nature neuroscience*, vol. 10, no. 10, pp. 1241–1242, 2007.
- [148] B. Rasch and J. Born, "About sleep's role in memory," *Physiological reviews*, 2013.
- [149] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
- [150] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly, "Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory," *Psychological review*, vol. 102, no. 3, p. 419, 1995.