Stochastic Coherence Over Attention Trajectory For Continuous Learning In Video Streams

Matteo Tiezzi[1] , **Simone Marullo**[1,2] , **Lapo Faggi**[1,2] , **Enrico Meloni**[1,2] ,
Alessandro Betti[3] , **Stefano Melacci**[1]
[1]DIISM, University of Siena (Italy)
[2]DINFO, University of Florence (Italy)
[3] Inria, Lab I3S, MAASAI, Universitè Côte d'Azur (France),
mtiezzi@diism.unisi.it, {simone.marullo,lapo.faggi}@unifi.it, meloni@diism.unisi.it,
alessandro.betti@inria.fr, mela@diism.unisi.it

Abstract

Devising intelligent agents able to live in an environment and learn by observing the surroundings is a longstanding goal of Artificial Intelligence. From a bare Machine Learning perspective, challenges arise when the agent is prevented from leveraging large fully-annotated dataset, but rather the interactions with supervisory signals are sparsely distributed over space and time. This paper proposes a novel neural-network-based approach to progressively and autonomously develop pixel-wise representations in a video stream. The proposed method is based on a human-like attention mechanism that allows the agent to learn by observing what is moving in the attended locations. Spatio-temporal stochastic coherence along the attention trajectory, paired with a contrastive term, leads to an unsupervised learning criterion that naturally copes with the considered setting. Differently from most existing works, the learned representations are used in open-set class-incremental classification of each frame pixel, relying on few supervisions. Our experiments leverage 3D virtual environments and they show that the proposed agents can learn to distinguish objects just by observing the video stream. Inheriting features from state-of-the art models is not as powerful as one might expect.

1 Introduction

In the context of Artificial Intelligence, the idea of designing *agents that exist in an environment and perceive and act* [Russell and Norvig, 2009] is a longstanding goal that introduces a huge number of challenges. While Machine Learning solutions applied to Computer Vision might help in pursuing such a goal, most of their outstanding results are obtained in well-defined vision tasks, leveraging huge collections of supervised data or exploiting pretrained backbones [Ranftl *et al.*, 2021]. Recent Self-Supervised Learning schemes aim at robust representations without human intervention, still exploiting very large image collections [Jing and Tian, 2020].

Supplementary material: https://arxiv.org/abs/2204.12193.

A lot of issues arise when trying to exploit neural models from other vision tasks in order to design a visual agent that learns while watching a video stream, especially when the agent is expected to parsimoniously interact with humans to get information on what it sees. Pretrained models might not always help in capturing properties of entities that belong to the particular environment in which the agent lives [Kornblith *et al.*, 2019], and they are subject to inductive biases. Moreover, the agent must be able to learn synchronously with the continuous video stream, and the target classes are not known in advance. This setting not only implies redundancy in visual information, but it also introduces constraints in the data order, that cannot be shuffled as commonly done to implement stochastic gradient descent. To this regard, the scientific community is progressively paying more attention to continual learning [Parisi *et al.*, 2019].

An often neglected element of crucial importance is the focus of attention, which guides the agent in wild visual scenes and attributes precise locations to the human-machine interaction. For example, consider an agent that asks for or receives a specific supervision in a crowded scene, or whenever there is a linguistic interface to exchange information with the human. Without contextualizing the dialogue to what is being precisely observed, the interaction is hardly meaningful. In particular, we are referring to the simulation of *human-like visual attention trajectories* [Zanca *et al.*, 2020], which is different from popular neural attention models [Chaudhari *et al.*, 2019], that are learnt to cope with a certain task, still relying on large datasets processed offline. Supervisions can be in the form of a class/instance label about what is being observed, without a precise indication on the boundaries of what is supervised (as depicted in Fig. 1), differently from several Computer Vision tasks [Long *et al.*, 2015].

In this paper, (*i*) we propose a novel approach to online learning from a video stream, that is rooted on the idea of using a scanpath-based focus of attention mechanism [Zanca *et al.*, 2020] to explore the video and to drive the learning dynamics in conjunction with motion information. The attention coordinates offer a precise location for interaction purposes, and its trajectory has been recently proved to efficiently select the most salient information of the video stream when learning with deep architectures [Tiezzi *et al.*, 2020].

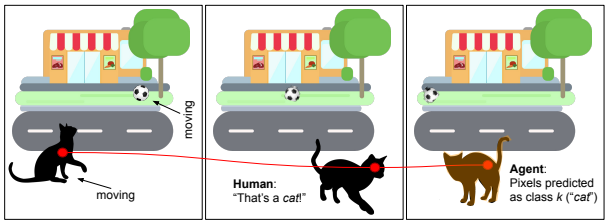

Figure 1: Stream with static and moving elements. The focus of attention (red circle) disambiguates what is observed, making the supervision well contextualized (2nd pic). The agent learns (1) pixel-wise representations that are coherent over the focused moving region, and (2) to make predictions (3rd pic). A stochastic sampling procedure yields an *attention graph* that facilitates learning.

We propose to learn representations that are coherent over the temporal attention trajectory during slow movements of the simulated gaze, that are likely to cover visual patterns with the same semantics. Attention is paired with information coming from motion, that intrinsically suggests the spatial bounds of the attended area. This leads to a spatio-temporal unsupervised criterion that enforces coherence in the representations learned while observing what is moving (Fig. 1). In order to avoid trivial solutions, we augment the criterion with a contrastive term that favours the development of different representations in what is inside the moving area and what is right outside of it. Thanks to a graph-based formalization of this approach, we define a stochastic procedure that introduces variability in the information provided to the learning algorithm and leads to faster processing, also mitigating the effects of noisy motion information. (*ii*) We consider the case in which the human intervention is rare, and each supervision is about the coordinates of a single pixel with its class/instance label, thus not a signal that can strongly drive the features development. Target classes are not known in advance, and our model includes an open-set approach to avoid making predictions about unknown elements. In order to cope with continual learning, we propose a template-based schema with a dynamic update procedure that is synchronous with the processed stream and efficiently handled by modern hardware. (*iii*) Dealing with this setting introduces the further challenge of how to evaluate the artificial agent. We exploit the growing activity in realistic 3D Virtual Environments [Meloni *et al.*, 2020], designing from scratch (and sharing) three ad-hoc streams with different visual difficulty levels. (*iv*) We compare the learned representations with those from state-of-the art pretrained models that exploited a massive supervision, showing that the latter are not as powerful as one might expect in the considered setting.

2 Related Work

Online, continual, open set learning. An online learning model progressively learns from a stream of data, continuously adapting to every new processed input instance [Hoi *et al.*, 2018]. In the specific case of continual learning (also known as life-long, continuous or incremental learning [Parisi *et al.*, 2019]), the goal of the agent is not fixed a priori but changes over time. In this paper, the goal of the agent is to learn to predict the class labels in every pixel of a video stream. Such goal is not fully defined in advance, since the agent becomes aware of classes in function of what the human supervisor tells him, being closer to task-free cases [Aljundi *et al.*, 2019]. Open-set classifiers [Scheirer *et al.*, 2012] can distinguish between examples belonging to different training classes and they can detect whether data do not belong to any of them, that is the case of what we propose. Our work is class-incremental [Geng *et al.*, 2020], due to the progressive inclusion of new classes after human intervention.[1]

Focus of attention. Several attempts to model *human-like* focus of attention mechanisms were presented [Borji and Itti, 2012], that not only differ in the way they are implemented, but also in the nature of the predicted attention (i.e., a temporal trajectory rather than saliency) [Borji, 2019]. Recently, an unsupervised dynamical model was proposed [Zanca *et al.*, 2020] that can be applied both to static images and videos, also studied in the context of online learning in deep networks [Tiezzi *et al.*, 2020]—without any loss of generality, it is the model we consider here.

Learning invariant features. Typical convolutional neural architectures require high sample complexity to learn representations invariant to factors that are not informative for the task at hand. Some solutions disentangle *what* and *where* features, each of them separately encoding informative and uninformative factors of variation [Burt *et al.*, 2021]. This paper follows the perspective in which the attention trajectory, paired with motion information, offers a compact way to implicitly constrain the agent to learn invariances. This idea is linked to recent studies about learning invariance to motion in unsupervised learning over time [Betti *et al.*, 2020].

Semantic segmentation. Semantic segmentation aims at associating a class label to each pixel of a given image. Deep architectures for this task usually rely on supervised learning from offline data, including fully convolutional networks [Long *et al.*, 2015], models based on transposed convolutions, dilated convolutions, upsampling and/or unpooling (U/V-net architectures [Ronneberger *et al.*, 2015]), transformers [Ranftl *et al.*, 2021]. We will compare with these models.

Virtual environments. The significantly improved quality of the rendered scenes and the intrinsic versatility of 3D Virtual Environments have quickly increased their popularity in the Machine Learning community (e.g., [Gan *et al.*, 2020]). SAILenv [Meloni *et al.*, 2020] is a recently proposed environment based on Unity3D, specifically aimed at easy customization and interface to Machine Learning libraries. SAILenv yields pixel-level motion information from Unity3D, and provides utilities to ease the generation of ad-hoc scenes for continual learning scenarios [Meloni *et al.*, 2021], making it well suited for what we propose.

3 Model

The basic concepts of the proposed unsupervised feature extractor are summarized in the example of Fig. 2. Our method

[1] However, it differs from the protocol of the open-world scenario, that goes beyond what we describe in this paper [Geng *et al.*, 2020]. One/few shot supervised models [Min *et al.*, 2021] also learn new classes from few examples, exploiting prior knowledge.

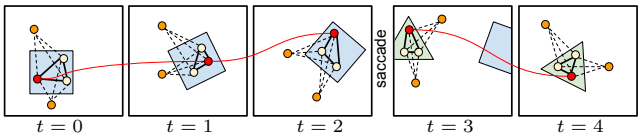

Figure 2: A rectangle rotates toward the right, attracting the attention a_t (red nodes) that explores it. Then, a triangle enters the scene, and the simulated gaze quickly moves on it (saccade). The red path links nodes on which we enforce *temporal coherence* (not in saccades). For each t, some coordinates (yellow) are sampled in the *moving* region that includes a_t, while other points (orange) are sampled outside of it. Solid lines (positive edges) link *spatially coherent* samples; dashed lines (negative edges) are about the *contrastive* term.

is based on the assumption that a *human-like attention* trajectory [Zanca *et al.*, 2020] generally spans the important location of the stream, and the gaze moves more slowly within areas with uniform semantic properties, whose bounds can be further guessed by motion information. Hence, we enforce both *temporal* and *spatial coherence* constraints to force pixel embeddings to be consistent in time and space, with respect to locations virtually connected by the attention trajectory and by motion, respectively. We exploit a *contrastive loss* to avoid trivial solutions, and a *stochastic approach* to lighten the computational burden. Given the pixel embeddings, a template-based classifier learns how to classify each of them in an open-set class-incremental setting (Fig. 3).

Let us indicate with X the set $\mathbb{R}^{w\times h\times c}$, and consider a continuous and potentially life-long video stream $V \in (X)^{\mathbb{N}}$ that, at the discrete time index $t \in \mathbb{N} \mapsto V_t \in X$, yields the video frame V_t at the resolution of $w \times h$ pixels with c channels. We also consider a neural network model that implements the function $f(\cdot,\omega)\colon X \to F$, with $F = \mathbb{R}^{w\times h\times d}$, where ω indicates the weights and biases of the net. The network is designed to process a frame V_t, and $f(V_t,\omega)$ is an encoded representation of the frame where, for each pixel, we have a vector with d components (Fig. 3). The network evolves over time and ω_t are the weights and biases at time t. We denote with f_x the output of f restricted to the pixel at 2D coordinates $x \in Z^\diamond := \{1,\dots,w\} \times \{1,\dots,h\}$. In online learning from a video stream V, the network weights ω_{t+1} are obtained updating the previous ω_t with a law that depends on the gradient of a suitable loss function L with respect to ω.

Human-like attention. The first pillar sustaining our model is a function $t \mapsto a(t)$, yielding an explicit estimate of the 2D coordinates where human would focus the attention at each time t. Among the various attention-prediction models yielding temporal trajectories [Borji and Itti, 2012], the recent unsupervised model [Zanca *et al.*, 2020] achieved state-of-the art results in simulating human-like attention. The authors showed that visual cues of the frame can act as gravitational masses. The equation of the potential of the gravitational field, $\varphi(x,t) := -(2\pi)^{-1}\int_Z \log\|x-z\|\mu(z,t)\,dz$, is at the basis of the attention model, where Z is the continuous set of frame coordinates and $\mu(x,t)$ is the total mass at (x,t). In particular, the magnitude of the brightness and an optical-flow-based measure of motion activity can be modeled as masses, and their impact is controlled by two positive scalars α_b and α_m, respectively, and an inhibitory signal is

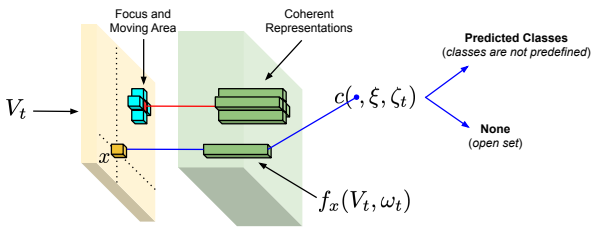

Figure 3: Pixels of frame V_t are encoded into new representations by the neural net f. Inference involves all the coordinates (example for x), and the classifier c can predict one or more classes or nothing (open-set). Classes are not known in advance. Learning is only about the attended moving area, developing coherent representations over time and space.

also included. The focus of attention is modeled as a point-like particle subject to the above potential, and its trajectory $a(t)$ is determined integrating the following equation with initial conditions $a(0) = a^0$ and $\dot{a}(0) = a^1$,

$$\ddot{a}(t) + \rho\dot{a}(t) + \nabla\varphi(a(t),t) = 0 \qquad t > 0, \tag{1}$$

where the dissipation is controlled by $\rho > 0$ and $\nabla\varphi$ is the spatial gradient of the potential. The gaze performs *fixations* in locations of interest, with relatively low speed movements. *Smooth pursuit* consists of slow tracking movements performed to follow a considered stimulus. Differently, *saccades* are fast movements to relocate the attention toward a new target. In the most naive case, the first two categories of patterns can be distinguished by the latter exploiting a thresholding procedure on velocity, based on $\nu > 0$.

Temporal coherence. During fixations, the attention spans a certain part of an object, with limited displacements of the gaze. Similarly, during smooth-pursuit the attended moving area has uniform semantic properties. Differently, during saccades, the attention switches the local context, shifting toward something that might or might not belong to the same object. We implement the notion of *temporal coherence* defining the loss function L_T that $(i.)$ restricts learning to the attention trajectory, filtering out the information in the visual scene [Tiezzi *et al.*, 2020] and $(ii.)$ avoids abrupt changes in the feature representation during fixations and smooth pursuit,

$$L_T(\omega,\hat{\omega},t) := \delta_t \left\| f_{a_t}(V_t,\omega) - f_{a_{t-1}}(V_{t-1},\hat{\omega}) \right\|^2, \tag{2}$$

where $\|\cdot\|$ is the Euclidean norm, and δ_t is equal to 0 in case of saccades, otherwise it is 1.[2] When plugged into the online optimization, $\omega \leftarrow \omega_t$ and $\hat{\omega} \leftarrow \omega_{t-1}$. This loss is forced to zero for $t = 0$, and it will be paired with other penalties described in the following, thus avoiding trivial solutions.

Spatial coherence. Temporal coherence is not enough to capture the spatial extension of the encoded data, since it is only limited to the attention trajectory. Motion naturally provides such spatial information, and we follow this intuition designing agents that learn by observing moving attended regions.[3] As a matter of fact, motion plays a twofold role, being crucial in defining the attention masses and as a mean to

[2] We also consider the case of feature vectors with unitary Euclidean norm—see supplementary material.

[3] Filtering out camera motion, e.g., in agents aware of how the camera moves (devices with sensors) or using software techniques.

extend the notion of coherence to a frame region, i.e., *spatial coherence*. Formally, for each fixed $t \in \mathbb{N}$, we indicate with $S_t \subseteq Z^\diamond$ the set of frame coordinates that belong to the region of connected moving pixels that includes a_t.[4] We introduce what we refer to as *spatial attention graph* at t, $\mathcal{G}_t$, with a node for each pixel of the frame ($\forall x \in Z^\diamond$) and with two types of edges, referred to as positive and negative edges. Positive edges link pairs of nodes whose coordinates belong to S_t, while negative edges link nodes of S_t to nodes outside the moving region. The positive edges of the attention graph allow us to introduce a spatial coherence loss L_S,

$$L_S(\omega, t) := \frac{1}{2} \sum_{x,z \in S_t, x \neq z} \|f_x(V_t, \omega) - f_z(V_t, \omega)\|^2 , \quad (3)$$

that encourages the agent to develop similar representations inside the attended moving area S_t, while $\omega \leftarrow \omega_t$. The notion of learning driven by the fulfilment of spatio-temporal coherence over the attention trajectory (L_S and L_T of Eq. 2 and Eq. 3) is the second pillar on which our model is built.

Contrastive loss. In order to prevent the development of trivial constant solutions, which fulfill the spatio-temporal coherence, we add a *contrastive* loss L_C that works the opposite way L_S does. In particular, L_C exploits negative edges of $\mathcal{G}$ to foster different representations between what is inside the moving area and what is outside of it,

$$L_C(\omega, t) := \left(\sum_{x \in S_t, z \in O_t} \|f_x(V_t, \omega) - f_z(V_t, \omega)\|^2 + \varepsilon \right)^{-1} , \quad (4)$$

where $O_t = Z^\diamond \setminus S_t$ is composed of frame coordinates not in S_t and $\varepsilon > 0$ avoids divisions by zero. We notice that this contrastive loss is different from InfoNCE [Oord *et al.*, 2018].

Stochastic coherence. These spatial and contrastive losses are plagued by two major issues. First, the number of pairs in Eq. 3 and Eq. 4 is large, being it $(1/2)|S_t|(|S_t| - 1)$ and $|S_t|(wh - |S_t|)$, respectively, making the computation of the loss terms pretty cumbersome. Secondly, whenever an exact copy of a moving object appears in a not-moving part of the scene, there will be pixels of the first instance that are fostered to develop different representations with respect to pixels of the second one, what we refer to as collision. However, this clashes with the idea of developing a common representation of the object pixels. Hence, we replace $\mathcal{G}_t$ with the subgraph $\tilde{\mathcal{G}}_t$, that is the *stochastic spatial attention graph* at time t, composed by nodes that are the outcome of a stochastic subsampling of those belonging to $\mathcal{G}_t$. In particular, the node set of $\tilde{\mathcal{G}}_t$ is the union of $\tilde{S}_t \subseteq S_t$ and $\tilde{O}_t \subseteq O_t$, where a_t is guaranteed to belong to the former. Edges of $\tilde{\mathcal{G}}_t$ are the positive and negative edges of $\mathcal{G}_t$ connecting the subsampled nodes. The key property of the stochastic graph is that the number of positive and negative edges is a chosen $e > 0$.[5] The set $\tilde{S}_t$ is populated by uniformly sampling nodes in S_t, ensuring that a_t is always present, while $\tilde{O}_t$ is populated by sampling from a Gaussian distribution centered in a_t and with variance σ, discarding samples $\notin O_t$.[6] Large σ's lead to sampling data also far away from the focus of attention, while small σ's will generate samples close to the boundary of the moving region. In the example of Fig. 2 we emphasized how the attention bridges multiple instances of $\tilde{\mathcal{G}}_t$ over time, yielding a *stochastic attention graph*. Such a graph reduces the probability of collisions, both due to the random sampling and to the control on the sampled area by means of σ, and it introduces variability in the loss functions also when computed on consecutive frames. Moreover, the impact of imperfect segmentation of the moving region is reduced, since only some pixels are actually exploited, re-sampled at every frame. Since the number of pairs is bounded, the stochastic graph makes the formulation suitable for real-time processing.

Cumulative loss. We define L as the cumulative loss at a certain time instant t, where the contributes of Eq. 2, Eq. 3, Eq. 4 are weighted by the positive scalars λ_T, λ_S and λ_C,

$$L(\omega, \hat{\omega}, t) := \lambda_T L_T(\omega, \hat{\omega}, t) + \lambda_S \tilde{L}_S(\omega, t) + \lambda_C \tilde{L}_C(\omega, t). \quad (5)$$

The losses $\tilde{L}_S$ and $\tilde{L}_C$ are the stochastic counterparts of L_S and L_C, respectively, in which the sets $\tilde{S}_t$ and $\tilde{O}_t$ are used in place of S_t and O_t. We define ∇L to be the gradient of the loss with respect to its first argument, that drives the online learning process, $\omega_0 = \varpi$, $\omega_{t+1} = \omega_t - \alpha \nabla L(\omega_t, \omega_{t-1}, t)$, $t \in \mathbb{N}$, with $\alpha > 0$ (learning rate) and ϖ some random initialization of the parameters of the network. We remark that other recent learning schemes for temporal domains could be exploited as well [Tiezzi *et al.*, 2020].

Pixel-wise classification. A human supervisor occasionally provides a supervision at coordinates a_t about a certain class y_t.[7] Let us define an open-set classifier $c(\cdot, \xi, \zeta) \colon F \mapsto Q$ that predicts the class-membership scores (belonging to a generic set indicated with Q), over a certain number of classes that can be attached to each pixel-level representation. The main parameters of the classifier are collected in ζ, and when all the membership scores are below threshold ξ the classifier assumes to be in front on an unknown visual element and it does not provide a decision (*open-set*)—see Fig. 3. Whenever a supervision on a never-seen-before class is received, the classifier becomes capable of computing the membership score of such class for all the following time steps (*class-incremental*). We consider the case in which supervisions are extremely rare, not offering a suitable basis for gradient-based learning of c or further refinements of f. The previously described unsupervised learning process is what is crucial to learn compact representations that can be easily classified, since it favours pixels of the same object to be represented in similar ways over time and space.

The most straightforward way to implement the open-set c is with a distance-based model, storing the feature vectors

[4] Moving pixels are selected as detailed in the suppl. material.

[5] Once we select e, if $s := |\tilde{S}_t|$ and $o := |\tilde{O}_t|$, we have $e = s(s-1)/2$ positive edges and $e = so$ negative ones. We approximated the solution with $s = \lfloor (1 + \sqrt{1 + 8e})/2 \rfloor$ and $o = \lceil e/s \rceil$.

[6] We selected σ to be $\propto \sqrt{|S_t|}$ by means of an integer spread factor $\beta \geq 1$. We repeat the Gaussian sampling until we collect the target number of points in $\tilde{O}_t$, up to a max number of iterations.

[7] Notice that only one pixel gets a supervision a at time t, thus it is different from few-shot semantic segmentation [Min *et al.*, 2021].

associated to the supervised pixels as templates.[8] This allows the model to not make predictions when the minimum distance from all the templates is greater than ξ. We indicate with (k_t, y_t) a supervised pair where $k_t = f_{a_t}(V_t, \omega_t)$ is the template at coordinates a_t of frame V_t. The intrinsic dependence of k_t on the time index could make templates become outdated during the agent life, for example due to the evolution of the system, so that for some $t' > t$ we might have $\|k_t - f_{a_t}(V_t, \omega_{t'})\| \gg 0$, leading to potentially wrong predictions. In order to solve this issue, we propose to dynamically update the templates. Modern hardware and software libraries for Machine Learning are designed to efficiently exploit batched computations In online learning, this feature is commonly not exploited, since only one data sample becomes available for each t. We indicate with B_t a special type of mini-batch of frames, $B_t = \{V_t\} \cup \{V_r,\ r \in H_t\}$, composed of the current V_t and the frames associated to supervised time instants whose indices are in H_t. Due to the tiny number of supervisions, storing supervised frames does not introduce any critical issues, and a maximum size $b > 1$ for each mini-batch can also be defined, populating B_t with up to $b-1$ previously supervised frames, chosen with or without priority. This way, batched computations can be efficiently exploited to keep templates up-to-date.

4 Experiments

In order to create the right setting to evaluate what we propose, we need continuous streams able to provide pixel labels at spatio-temporal coordinates that are not defined in advance.

Virtual environments. We consider photo-realistic 3D Virtual Environments within the SAILenv platform [Meloni *et al.*, 2020], that include pixel-wise semantic labeling and motion vectors of potentially endless streams, creating (from scratch) three 3D scenes to emphasize different and challenging aspects on which we measure the skills of the agents.[9]. The agent observes the scene from a fixed location, and some objects of interest move, one at the time, along pre-designed smooth trajectories while rotating and getting closer to/farther from the camera. We denote with the term *lap* a complete route traveled by each object to its starting location.

We designed three different scenes. (*i*) EMPTYSPACE: four photo-realistic textured object from the SAILenv library (chair, laptop, pillow, ewer) move over a uniform background. The goal is to distinguish them in a non-ambiguous setting. (*ii*) SOLID: a gray-scale environment with three white solids (cube, cylinder, sphere) is considered. Due to the lack of color-related features, the agent must necessarily develop the capability of encoding information from larger contexts around each pixel. (*iii*) LIVINGROOM: the objects from EMPTYSPACE are placed in a photo-realistic living room composed by other non-target objects (i.e., an heterogeneous background with a couch, tables, staircase, door, floor), and multiple static instances of the objects of interest. Samples of are shown in Fig. 4 (top) and, in what follows, n is the number of target objects in each stream, while $m = n + 1$ is the total number of categories (including the "unknown" class).

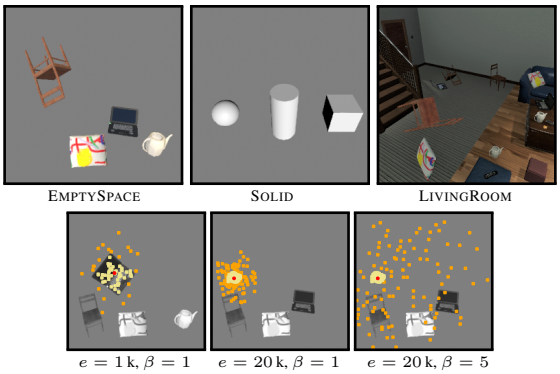

Figure 4: Top: sample frames from the three 3D scenes/streams (objects rotate and scale while moving). Bottom: samples of stochastic spatial graphs in EMPTYSPACE (same color patterns of Fig. 2).

Setup. We created three pre-rendered 2D visual streams by observing the moving scenes (256×256 pixels, with ≈ 51 k, 12 k and 20 k frames, respectively, corresponding to 31 completed laps for each object), both in grayscale (BW) and color (RGB)—SOLID is BW only. The agent learns by watching the first 25 laps per object and, only in the subsequent laps, receives a total of 3 supervisions (k_t, y_t) per object, spaced out by at least 100 frames. Learning stops when all the objects complete 30 laps and, finally, performances are measured in the last lap, considering the F1 score (averaged over the m categories), either along the attention trajectory or in the whole frame area. We evaluated the proposed approach considering two different families of deep convolutional architectures yielding d output features, referred to as HOURGLASS (UNet-like [Ronneberger *et al.*, 2015]) and FCN-ND (6-layer Fully-Convolutional without any downsamplings [Sherrah, 2016]), respectively (see the supplementary material for all the details). We compared the obtained features against those produced by massively pretrained state-of-the-art models in Semantic Segmentation. We considered the Dense Prediction Transformer (DPT) [Ranftl *et al.*, 2021] and DEEPLABV3 [Chen *et al.*, 2017] with ResNet101 backbone, exploiting both the features produced by the penultimate layer in the classification heads (-C suffix) and the ones obtained by the backbones (i.e., upsampling the representations if needed, -B suffix). In this way, we investigate both lower-level features based on backbones pretrained on millions of images (ImageNet), and task-specialized higher-level features for semantic segmentation (COCO [Lin *et al.*, 2014] and ADE20k [Zhou *et al.*, 2019] datasets—the latter explicitly includes the categories of the considered textured objects). As BASELINE model we considered the case in which the pixel representations are left untouched (i.e., pixel color/brightness).

Parameters. Parameters of the attention model were either fixed ($\alpha_m = 1$), or adapted according to a preliminary run of the model (α_b, ρ). For each video stream, we searched for the model hyper-parameters that maximize the F1 along the attention trajectory, measured during the 30-th laps (hyper-params grids and the best selected are reported in the suppl. material). Fig. 4 (bottom) shows the effects of the parameters e and β in modeling the stochastic coherence graph.

[8] We tested the squared Euclidean distance and cosine similarity.

[9] Code, data and selected hyper-parameters can be downloaded at https://github.com/sailab-code/cl_stochastic_coherence

	EMPTYSPACE		SOLID	LIVINGROOM	
	BW	RGB	BW	BW	RGB
DPT-C	0.70	0.76	0.54	0.38	0.39
DPT-B	**0.73**	**0.77**	0.53	0.29	0.32
DEEPLAB-C	0.46	0.57	0.56	0.19	0.26
DEEPLAB-B	0.61	0.72	0.66	0.28	0.35
BASELINE	0.39	0.61	0.32	0.52	**0.65**
HOURGLASS	**0.73**±0.03	**0.77**±0.05	**0.68**±0.02	**0.59**±0.02	0.45±0.12
FCN-ND	0.69±0.03	0.57±0.08	0.58±0.07	0.39±0.02	0.38±0.02
DPT-C	0.66	0.67	0.64	0.35	0.39
DPT-B	**0.71**	0.69	**0.68**	**0.39**	0.39
DEEPLAB-C	0.49	0.61	0.57	0.31	0.34
DEEPLAB-B	0.70	0.65	0.66	0.34	**0.44**
BASELINE	0.50	0.45	0.18	0.10	0.23
HOURGLASS	0.55±0.03	**0.71**±0.03	0.50±0.01	0.31±0.04	0.25±0.07
FCN-ND	0.60±0.05	0.51±0.07	0.48±0.03	0.24±0.01	0.28±0.03

Table 1: Top: F1 score (mean ± std) measured along the attention trajectory. Competitors are not affected by the model initialization (no std). Bottom: F1 considering all the pixels.

Results. Table 1 (top) shows the F1 along the attention trajectory measured during the latest object lap, averaged over 3 runs with different initialization. The proposed learning mechanism is competitive and able to overcome models pretrained on large supervised data. This is mostly evident in the case of HOURGLASS, while FCN-ND is less accurate, mostly due to the lack of implicit spatial aggregation. Table 1 (bottom) is about the F1 computed considering predictions on all the pixels of all the frames. This measure is not directly affected by the attention model, and while our model performs worse, it is still on par with some of the competitors, with the exception of LIVINGROOM (RGB). We investigated our results, showing in Fig. 5 some sample predictions comparing HOURGLASS with transformers (DPT-B). Indeed, state-of-the-art models have troubles in recognizing closer objects and also in discriminating from the background, due to the fact that their pixel representations are typically strongly affected by a large context. When presenting objects in unusual orientations, likely different from what observed during the fully supervised training, they tend to perform badly. Differently, our model adapts to the video stream, learning more coherent representations when the object transforms. Overall, the attention model allows to focus on what is more important, and, although just a tiny number of supervisions are provided, our model can online-learn to make predictions that competes with the massively offline trained competitors.

In-depth studies and ablations. In order to evaluate the sensitivity of the proposed approach to the key elements of the considered setup, we selected the HOURGLASS model of Table 1. Fig. 6 reports results of experiments in which we changed the number of supervisions per object, the number of edges e (per type) in the stochastic graph, we disabled the temporal coherence, and we changed the length of the streams (discarding SOLID in which differences were less appreciable). Even with a single supervision, the model is able

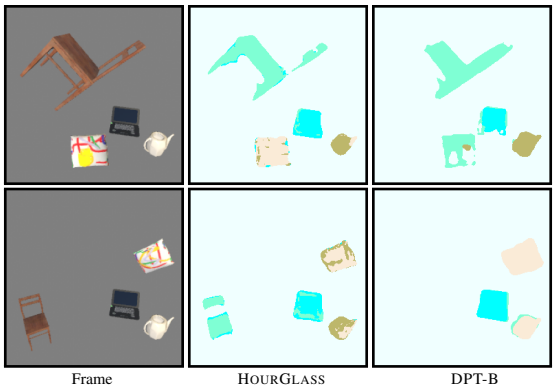

Figure 5: Predictions in two frames, comparing HOURGLASS (our) with DPT-B (transformers) in EMPTYSPACE stream. Different colors indicate different predictions.

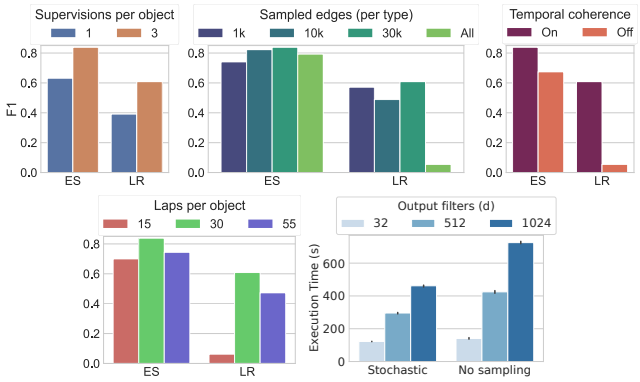

Figure 6: First 4 plots (left-to-right, top-to-bottom): in-depth experiments on EMPTYSPACE (ES) and LIVINGROOM (LR) RGB streams (F1 on focus trajectory). Last plot: timings (3 runs) for stochastic sampling ($e = 10$ k) and no subsampling at all, varying d.

to distinguish the target objects in EMPTYSPACE, while in the more cluttered LIVINGROOM it benefits from multiple supervisions, as expected. Our proposal is better than using a non-stochastic criterion (LIVINGROOM), and works well even with a limited value for e. Moreover, the agent benefits from relatively longer streams (going from 15 to 30 laps), that allow it to develop more coherent representations (we tuned the model in the 30-lap case, that is the reason for the slight performance drop in 55-lap), and temporal coherence has an important role in the overall results. For completeness, we highlighted in Fig. 6 (last) the computational benefits, in terms of time (one lap per object), brought by the stochastic subsampling, showing that they are more evident for large d.

5 Conclusions

We presented a novel approach to the design of agents that continuously learn from a visual environment dealing with an open-set class-incremental setting, leveraging a focus of attention mechanism and spatio-temporal coherence. We devised an innovative way of benchmarking this class of algorithms using 3D Virtual Environments, in which our proposal leads to results that, on average, are comparable to those obtained using state-of-the art strongly-supervised models.

Acknowledgements

This work was partly supported by the PRIN 2017 project RexLearn, funded by the Italian Ministry of Education, University and Research (grant no. 2017TWNMH2).

References

[Aljundi *et al.*, 2019] Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. In *Proc. of the IEEE/CVF CVPR*, pages 11254–11263, 2019.

[Betti *et al.*, 2020] Alessandro Betti, Marco Gori, and Stefano Melacci. Learning visual features under motion invariance. *Neural Networks*, 126:275–299, 2020.

[Borji and Itti, 2012] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE TPAMI*, 35(1):185–207, 2012.

[Borji, 2019] Ali Borji. Saliency prediction in the deep learning era: Successes and limitations. *IEEE TPAMI*, 2019.

[Burt *et al.*, 2021] Ryan Burt, Nina N. Thigpen, Andreas Keil, and Jose C. Principe. Unsupervised foveal vision neural architecture with top-down attention. *Neural Networks*, 141:145–159, 2021.

[Chaudhari *et al.*, 2019] Sneha Chaudhari, Varun Mithal, Gungor Polatkan, and Rohan Ramanath. An attentive survey of attention models. *arXiv preprint arXiv:1904.02874*, 2019.

[Chen *et al.*, 2017] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2017.

[Gan *et al.*, 2020] Chuang Gan, Jeremy Schwartz, Seth Alter, Martin Schrimpf, et al. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv:2007.04954*, 2020.

[Geng *et al.*, 2020] Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE TPAMI*, 2020.

[Hoi *et al.*, 2018] Steven CH Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey. *arXiv preprint arXiv:1802.02871*, 2018.

[Jing and Tian, 2020] Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE TPAMI*, 2020.

[Kornblith *et al.*, 2019] Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? In *Proc. of the IEEE/CVF CVPR*, June 2019.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proc. of the IEEE/CVF CVPR*, pages 3431–3440, 2015.

[Meloni *et al.*, 2020] Enrico Meloni, Luca Pasqualini, Matteo Tiezzi, Marco Gori, and Stefano Melacci. Sailenv: Learning in virtual visual environments made simple. In *ICPR*, pages 8906–8913, 2020.

[Meloni *et al.*, 2021] Enrico Meloni, Alessandro Betti, Lapo Faggi, Simone Marullo, Matteo Tiezzi, and Stefano Melacci. Evaluating continual learning algorithms by generating 3d virtual environments. *arXiv preprint arXiv:2109.07855*, 2021.

[Min *et al.*, 2021] Juhong Min, Dahyun Kang, and Minsu Cho. Hypercorrelation squeeze for few-shot segmentation. *arXiv preprint arXiv:2104.01538*, 2021.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Parisi *et al.*, 2019] German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

[Ranftl *et al.*, 2021] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *CoRR*, abs/2103.13413, 2021.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[Russell and Norvig, 2009] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, USA, 3rd edition, 2009.

[Scheirer *et al.*, 2012] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE TPAMI*, 35(7):1757–1772, 2012.

[Sherrah, 2016] Jamie Sherrah. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv preprint arXiv:1606.02585*, 2016.

[Tiezzi *et al.*, 2020] Matteo Tiezzi, Stefano Melacci, Alessandro Betti, Marco Maggini, and Marco Gori. Focus of attention improves information transfer in visual features. In *NeurIPS*, volume 33, pages 22194–22204, 2020.

[Zanca *et al.*, 2020] Dario Zanca, Stefano Melacci, and Marco Gori. Gravitational laws of focus of attention. *IEEE TPAMI*, 42(12):2983–2995, 2020.

[Zhou *et al.*, 2019] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *IJCV*, 127(3):302–321, 2019.