AdaGlimpse: Active Visual Exploration with Arbitrary Glimpse Position and Scale

Adam Pardyl [1,2] Michał Wronka [2] Maciej Wołczyk [1] Kamil Adamczewski [1] Tomasz Trzciński [1,3,4] Bartosz Zieliński [1,2]

[1]IDEAS NCBR [2]Jagiellonian University [3]Warsaw University of Technology [4]Tooploox

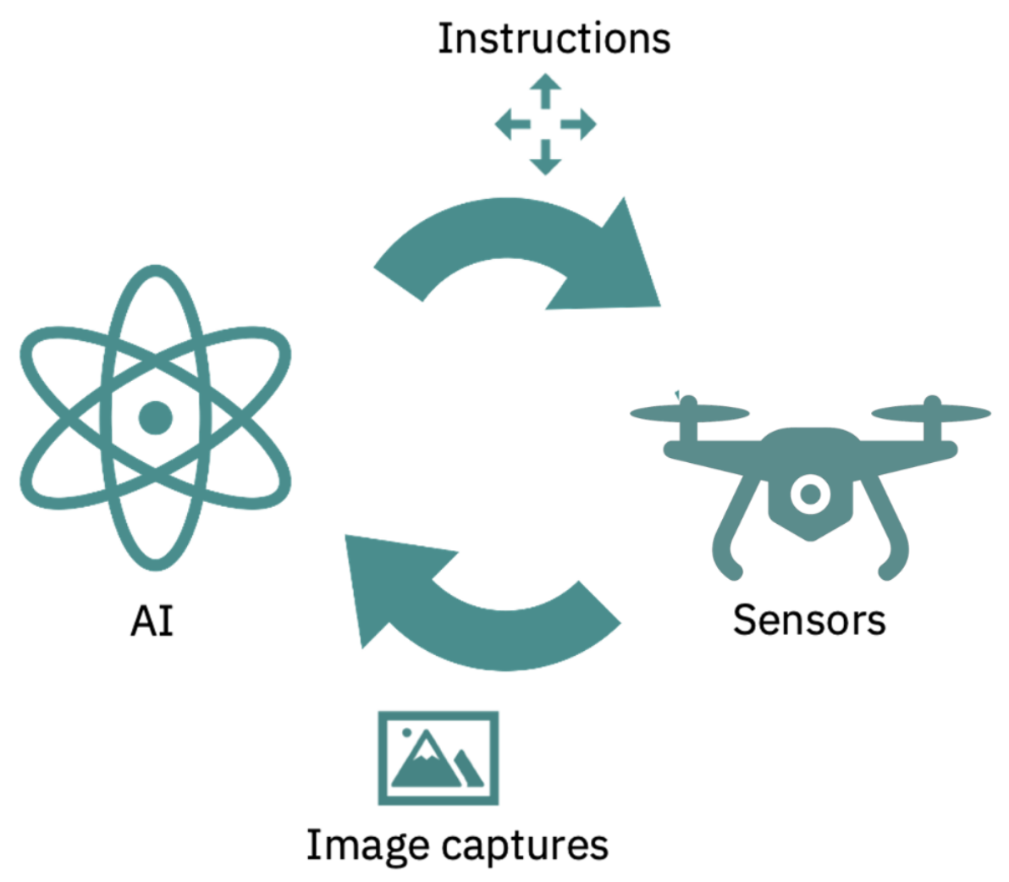
#3128

Motivation: explore agent's surroundings quickly

Open world environments pose a challenge for common computer vision methods:

- Machine learning solutions for computer vision typically assume complete input data.
- Real-world embodied agents like robots and UAVs face limitations due to restricted fields of view and operational time.
- Capturing high-resolution images of entire scenes is inefficient, as not all image areas hold equal amount of information.

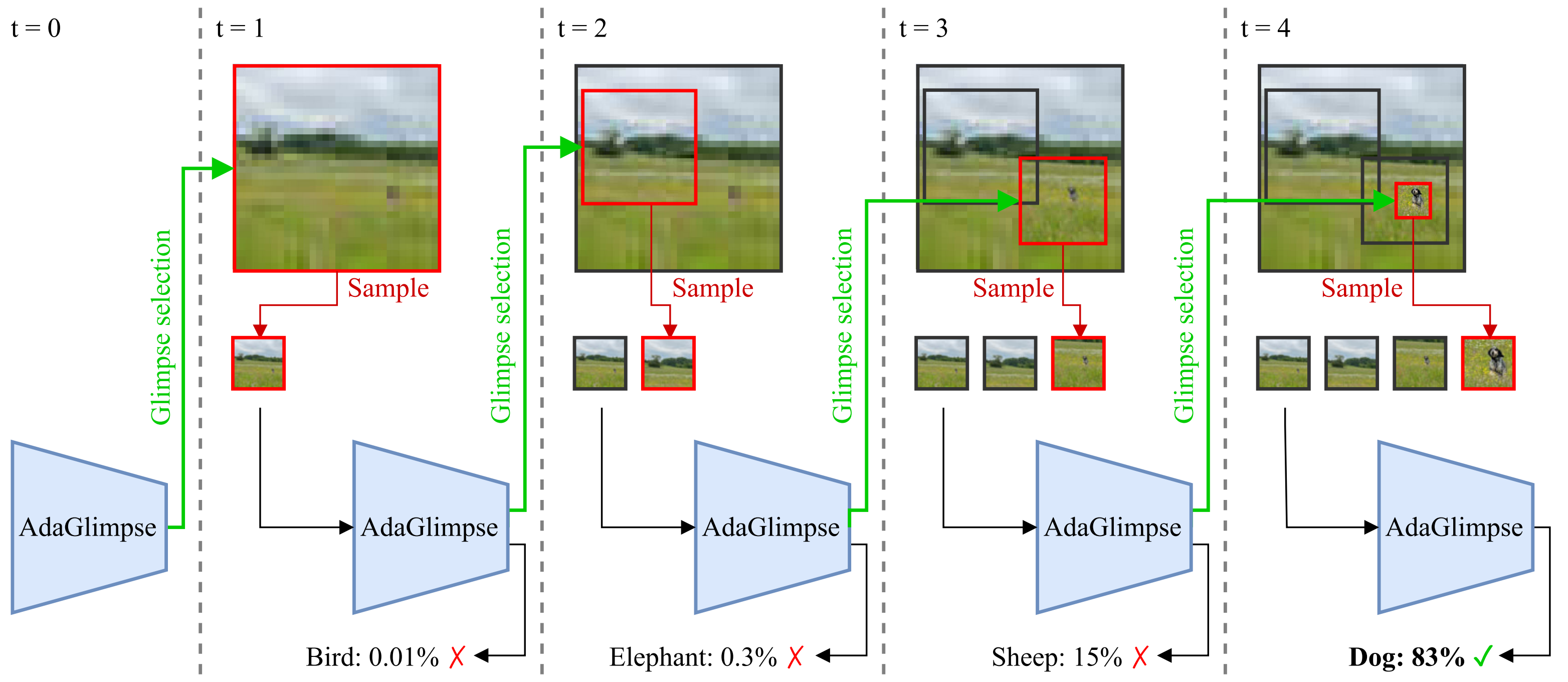
Idea: let the vision model directly guide navigation



Active Visual Exploration addresses how an agent should select visual information from its environment:

- Instead of sampling and analyzing the entire environment at the highest resolution, the agent dynamically chooses sampling locations based on insights from previous exploration steps.
- This process is inspired by how humans instinctively move their heads and eyes to explore.
- Current methods for AVE utilise fixed-sized observations selected from a grid of possible actions, and therefore fail to fully exploit the capabilities of modern hardware.
- Existing hardware can provide a glimpse of any position and scale, e.g., using a simple pan-tilt-zoom camera for robotic platforms. Similarly, UAVs can alter their position and altitude freely.

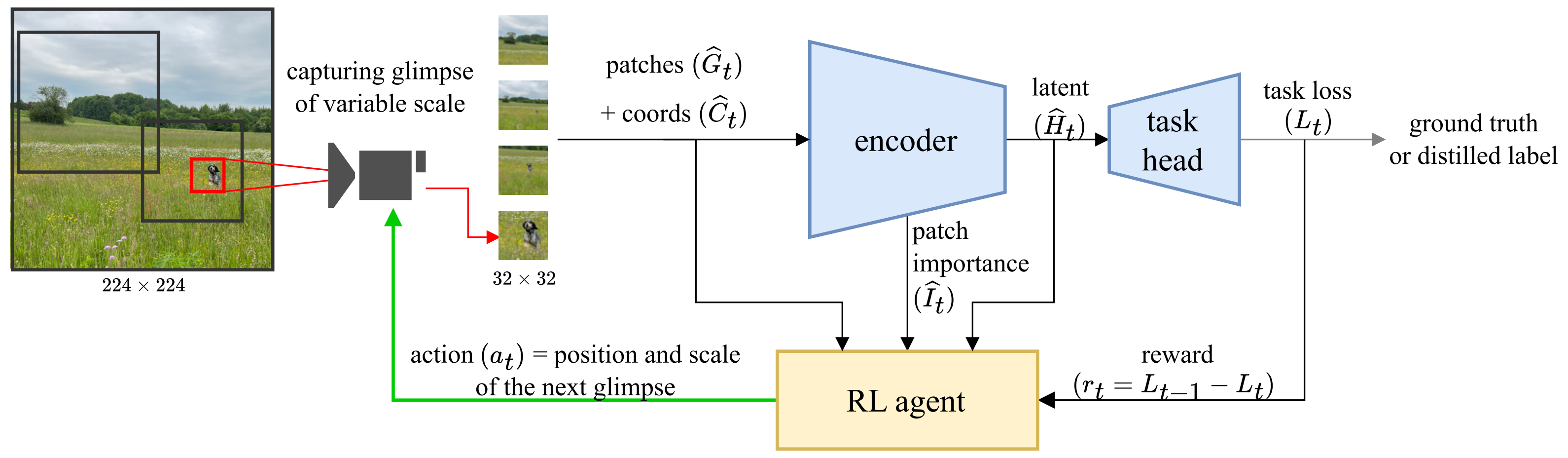
Our solution: Adaptive active visual exploration



AdaGlimpse processes glimpses of arbitrary position and scale, fully exploiting the capabilities of modern hardware:

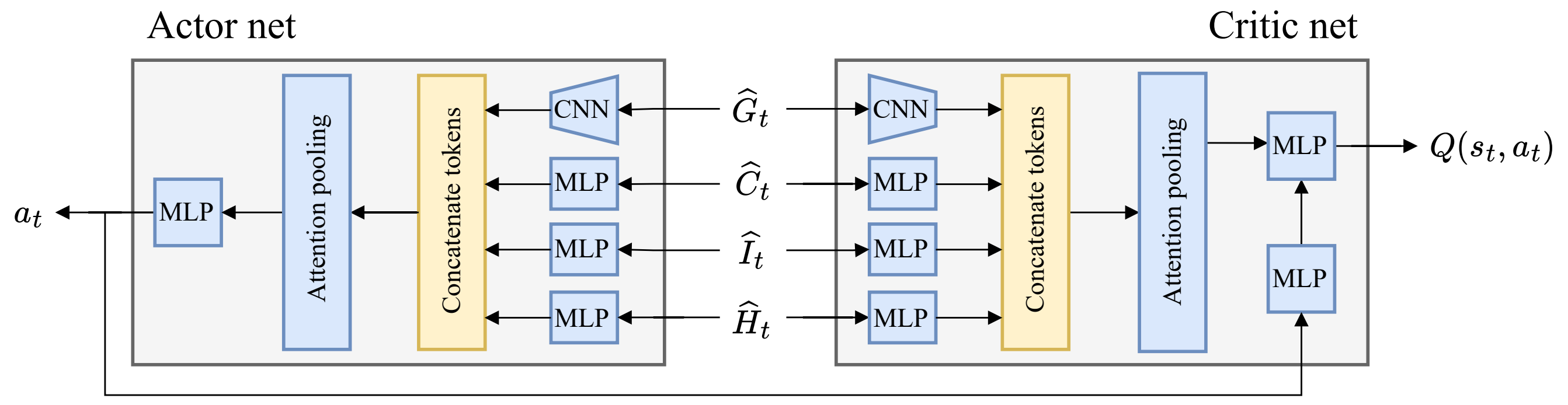
- At each step, the model uses previously acquired observations to predict both the target (in the above example the class of the image) and the best position and scale for the next observation, selected from a continuous space.
- In this above example, AdaGlimpse selects a low-resolution glimpse of the whole environment. Based on this glimpse, it predicts a bird with probability 0.01, too low to make the final decision. Instead, it selects the second glimpse by zooming in to the upper left corner. The process repeats four times until the probability of the predicted class is higher than a specified threshold.

AdaGlimpse architecture

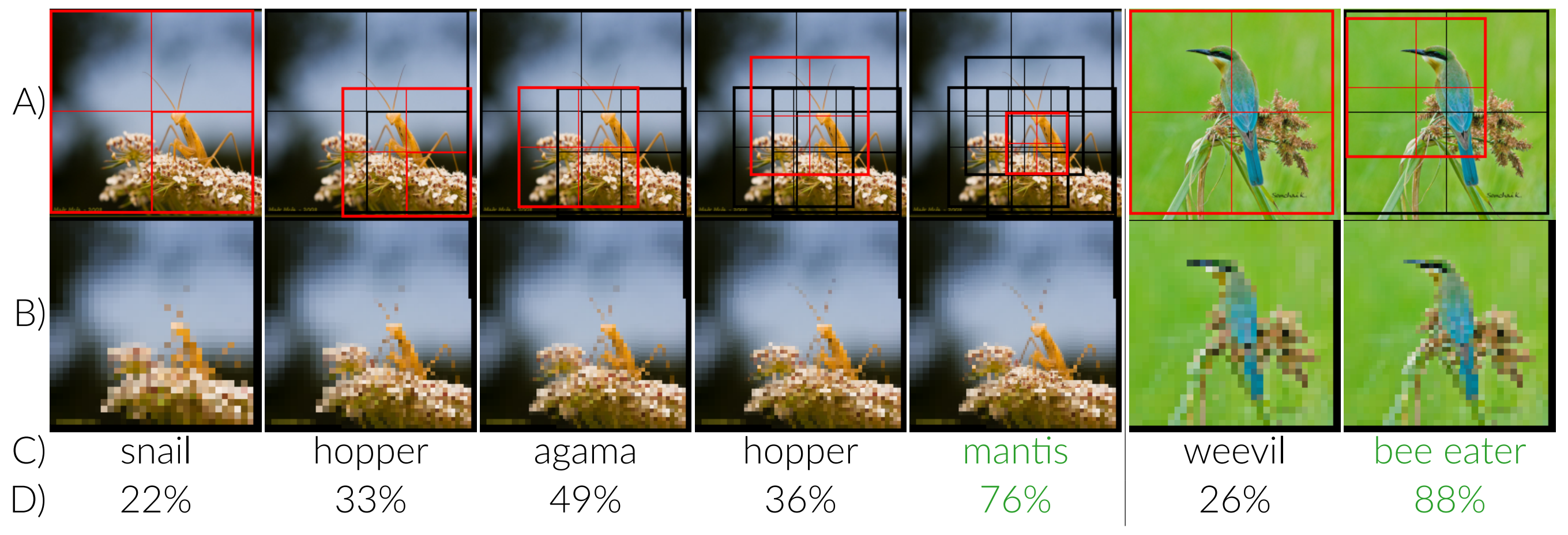

224 × 224
capturing glimpse of variable scale
32 × 32
patches $(\widehat{G}_t)$
+ coords $(\widehat{C}_t)$
encoder
latent $(\widehat{H}_t)$
task head
task loss (L_t)
ground truth or distilled label
patch importance $(\widehat{I}_t)$
action (a_t) = position and scale of the next glimpse
RL agent
reward $(r_t = L_{t-1} - L_t)$

AdaGlimpse consists of two parts: a vision transformer-based encoder with a task-specific head and a Soft Actor-Critic RL agent:

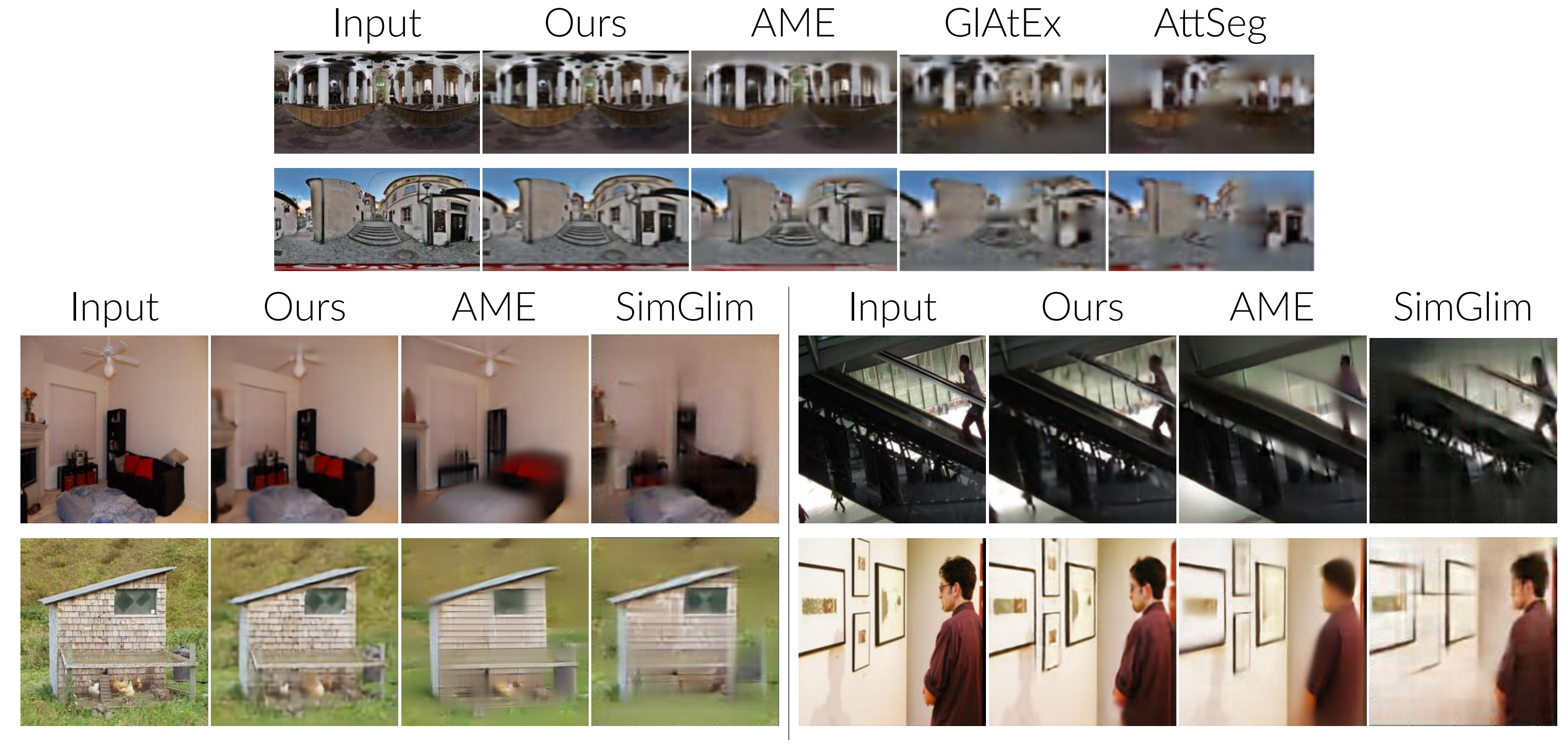
- The encoder combines information from all previous observations, creating a single representation of the seen environment.
- The task head is a linear layer for classification task and a transformer for reconstruction and segmentation.
- The RL agent actor consists of an attention pooling layer and an MLP, which predicts the next action. For training, a second network is used to act as a critic in the SAC framework.



Exploration examples



Glimpse selection step-by-step: AdaGlimpse explores 224×224 images from ImageNet with 32×32 glimpses, zooming in on objects of interest and stopping the process after reaching 75% predicted probability.



Reconstruction quality for SUN360 (top) and ADE20K (bottom): Sample reconstructions of our method compared with baselines.

Quantitative results

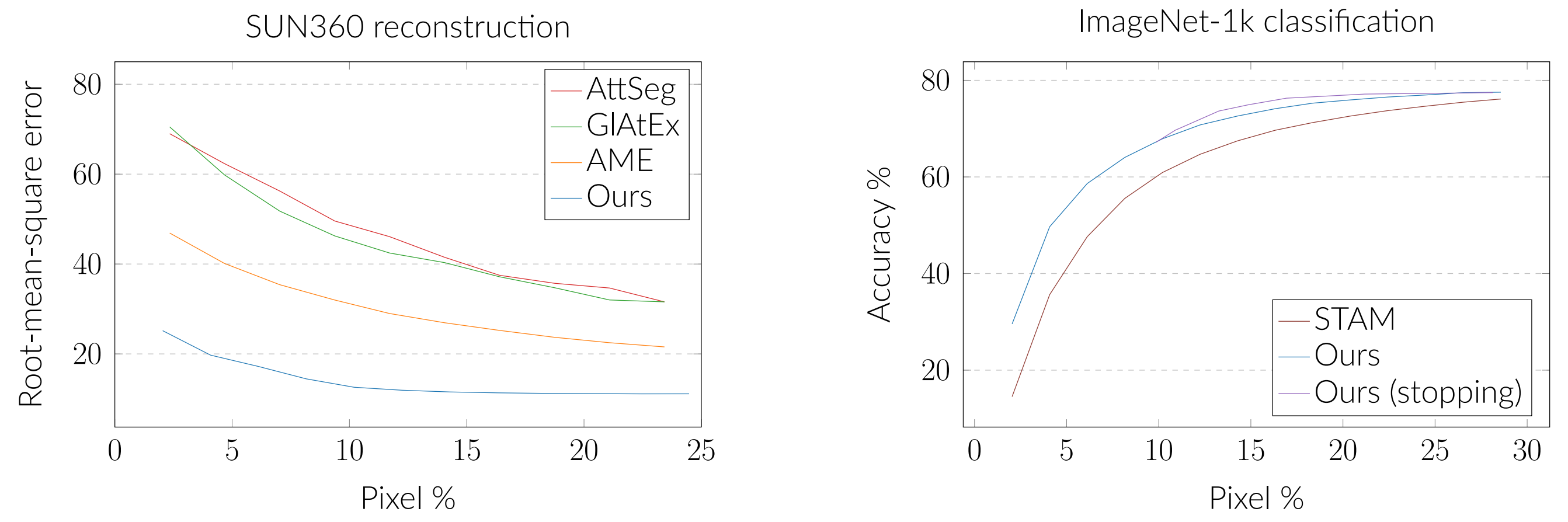
Method	Accuracy %	Glimpses	Regime	Pixel %
DRAM	67.50	8×77^2	full+simple	94.53
GFNet	75.93	5×96^2	full+simple	91.84
Saccader	70.31	6×77^2	full+simple	70.90
TNet	74.62	6×77^2	full+simple	70.90
STN	71.40	9×56^2	full+simple	56.25
PatchDrop	76.00	$\sim 8.9 \times 56^2$	full+simple+stopping	~55.63
STAM	76.13	14×32^2	simple	28.57
Ours	**77.54**	14×32^2	adaptive	28.57
Ours	**76.30**	$\sim 8.3 \times 32^2$	adaptive+stopping	~**16.94**

Classification results: Accuracy obtained by our model for classification task against baselines on ImageNet-1k dataset. Note that Pixel % denotes the percentage of image pixels known to the model.

Method	SUN	ADE	COCO	Image res.	Glimpses	Regime	Pixel %
AME	29.8	30.8	32.5	128×256	8×32^2	simple	25.00
Ours	**11.1***	**14.0***	**14.5***	224×224	12×32^2	adaptive	24.49
AttSeg	37.6	36.6	41.8	128×256	8×48^2	retinal	18.75
GlAtEx	33.8	41.9	40.3	128×256	8×48^2	retinal	18.75
AME	23.6	23.8	25.2	128×256	8×48^2	retinal	18.75
SimGlim	26.2	27.2	29.8	224×224	37×16^2	simple	18.75
AME	23.4	26.2	28.6	224×224	37×16^2	simple	18.75
Ours	**11.1***	**14.2***	**14.7***	224×224	9×32^2	adaptive	18.36
AME	37.9	40.7	43.2	128×256	8×16^2	simple	6.25
Ours	**17.6***	**20.5***	**21.5***	224×224	12×16^2	adaptive	6.12
Ours	**17.2***	**20.7***	**21.4***	224×224	3×32^2	adaptive	6.12

Reconstruction results: RMSE obtained by our model for reconstruction task against baselines on ImageNet-1k, SUN360, ADE20K and MS COCO.

Data efficiency



Percentage of image pixels observed: Figures present the relationship between the amount of pixels observed by the model relative to the full scene resolution (pixel %), and its performance. AdaGlimpse outperforms competitive solutions, requiring significantly less information to achieve the same performance.

Read more at

This paper has been supported by the Horizon Europe Programme (HORIZON-CL4-2022-HUMAN-02) under the project "ELIAS: European Lighthouse of AI for Sustainability", GA no. 101120237. This research was funded by National Science Centre, Poland (grant no. 2023/49/N/ST6/02465, 2022/47/B/ST6/03397, 2022/45/B/ST6/02817, and 2023/50/E/ST6/00469). The research was supported by a grant from the Faculty of Mathematics and Computer Science under the Strategic Programme Excellence Initiative at Jagiellonian University. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2023/016558.