SnapPix: Efficient-Coding–Inspired In-Sensor Compression for Edge Vision

Weikai Lin[1,*] Tianrui Ma[2,3,*] Adith Boloor[3] Yu Feng[4,†] Ruofan Xing[5]
Xuan Zhang[3,5] Yuhao Zhu[1]
[1]University of Rochester [2]Institute of Computing Technology, CAS [3]Washington University in St. Louis
[4]Shanghai Jiao Tong University [5]Northeastern University
wlin33@ur.rochester.edu, matianrui@ict.ac.cn, adith@wustl.edu, y-feng@sjtu.edu.cn,
{xing.ruo, xuan.zhang}@northeastern.edu, yzhu@rochester.edu
*Both authors contributed equally to this research. †Work done while at University of Rochester.

Abstract

Energy-efficient imaging is essential for edge sensing, where energy is dominated by in-sensor data readout and wireless transmission. In-sensor compression can reduce this cost but faces challenges in hardware overhead, information loss, and task specificity. Inspired by the mammalian visual system, we present SnapPix *[16], an in-sensor compression system that uses coded exposure (CE) for lightweight, sensor-compatible compression; learns a task-agnostic CE pattern by maximizing decorrelation among coded pixels based on efficient coding theory; and co-designs tile-repetitive CE patterns with Vision Transformers, augmented by reconstruction-based pre-training. Evaluating on action recognition and video reconstruction,* SnapPix *outperforms state-of-the-art video-based methods while reducing edge energy by up to 15.4×. Code is available at: https://github.com/horizon-research/SnapPix.*

1. Introduction

Energy-efficient imaging is essential for edge sensing scenarios, such as traffic monitoring [13], remote satellite sensing [8], and AR/VR devices [27], where the sensor node must transmit data to a server for processing. In these scenarios, sensing energy is dominated by in-sensor data readout [7, 18] and wireless data transmission [5].

To reduce the readout and transmission energy, recent work has explored in-sensor compression [7, 18, 20, 28, 29], which aggressively samples raw pixels before digitization and transmission. However, in-sensor compression faces three challenges: **C1**: some methods perform complex computation (e.g., convolutions, feature extraction) inside the sensor [15, 17, 18, 20], incurring significant hardware overhead; **C2**: aggressive compression saves energy but risks information/accuracy loss; and **C3**: existing methods tailor the compression pattern to a specific downstream task [7, 14, 18, 21], making it difficult to support diverse downstream tasks using the same compressed data.

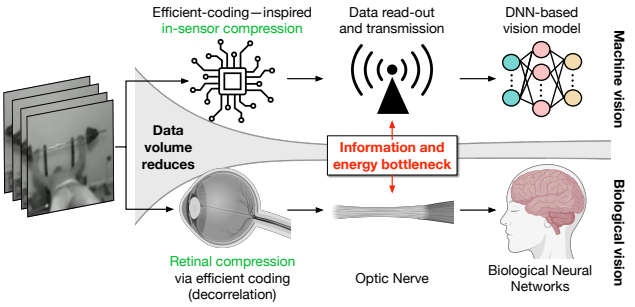

Fig. 1. SnapPix reduces edge sensing energy through in-sensor compression by decorrelating output pixel values. This is inspired by the mammalian visual system, where the retina compresses information by decorrelating the retinal output neurons.

To address the energy-information dilemma, we draw inspiration from the mammalian visual system (Fig. 1), where the retina efficiently compresses information through decorrelating its output neurons [2, 4]. We propose SnapPix [16], a general in-sensor compression system that is lightweight in sensor hardware, maximizes information preservation, and supports diverse downstream tasks. SnapPix has three key contributions:

Lightweight In-Sensor Compression via Coded Exposure (Sec. 2.1). To minimize sensor hardware overhead incurred by in-sensor compression (**C1**), SnapPix builds on coded exposure (CE) [14, 21, 22, 28], a technique that selectively exposes pixels spatially and temporally and is highly compatible with existing sensor architectures. To facilitate real-world deployment, we further propose minimal augmentations to stacked CMOS image sensors to support CE with negligible area overhead. We omit the hardware design details in this paper and refer readers to the full version of SnapPix [16] for a complete discussion.

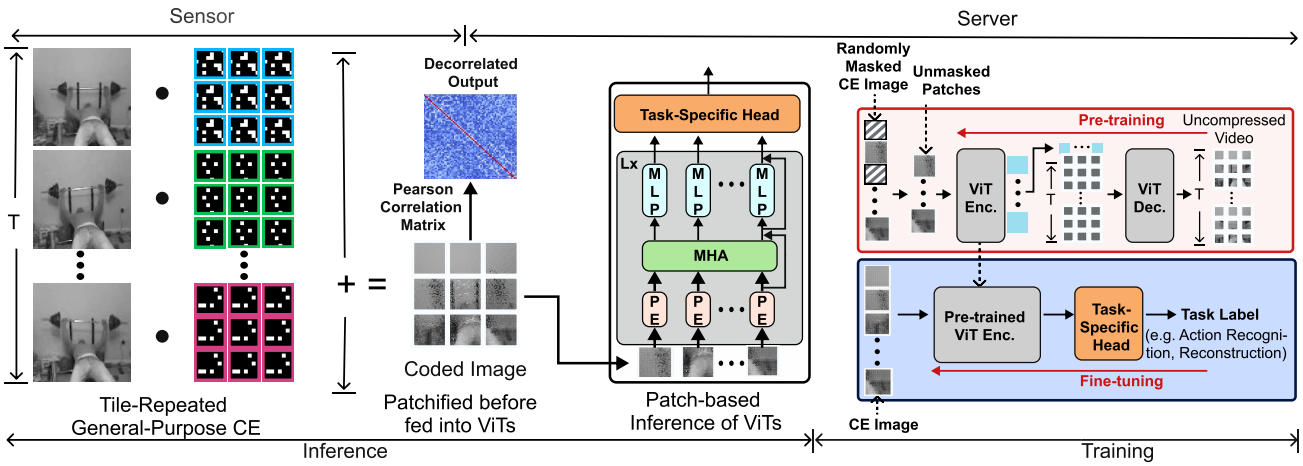

Fig. 2. The end-to-end pipeline of SnapPix with in-sensor CE for compression and a ViT-based vision model for downstream tasks. The CE pattern is trained task-independently using decorrelation, while the downstream model is co-designed with CE patterns and pre-trained for CE-encoded inputs.

General CE Pattern via Decorrelation (Sec. 2.2). To maximize information preservation under CE compression (**C2**) and enable diverse downstream tasks (**C3**), we draw on efficient coding theory in neuroscience [2, 4]: SnapPix learns a task-agnostic CE pattern that maximizes decorrelation among output coded pixels, thereby maximizing information density rather than tailoring to a specific task.

Co-Design CE and Vision Model (Sec. 2.3). After obtaining CE-compressed data, downstream vision models must decode and process it. However, CE introduces shift-variant pixel characteristics, where each pixel carries different amounts of information depending on its exposure, making standard shift-invariant CNNs suboptimal for decoding and processing CE-compressed data [14, 21].

To address this, SnapPix co-designs the CE pattern with Vision Transformers (ViTs) [6]: we constrain the CE pattern to be tile-repetitive, where pixel exposure varies within each tile (i.e., ViT patch) but repeats across tiles.

Since ViTs naturally process each patch independently via patch embedding and MLPs, they can naturally handle the within-tile pixel variations introduced by CE. This co-design preserves CE's flexibility in designing exposure patterns while ensuring alignment between the compressed data and the model's processing structure.

To further enhance the ViT's ability to extract information from CE-compressed data, we design an MAE-like [10, 26] reconstruction-based pre-training tailored for CE inputs, which teaches the model to recover the original video from a single coded image.

2. Method

The overview of SnapPix is shown in Fig. 2. It features a tile-repetitive coded exposure (CE) pattern that compresses the input video into a single coded image (Sec. 2.1). The

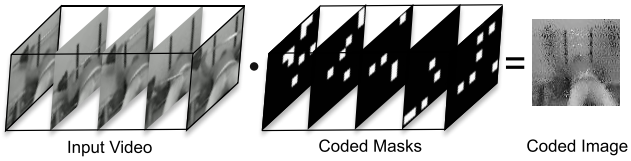

Fig. 3. Coded exposure with 5 exposure slots. In each slot, pixels are selectively exposed, controlled by a coded mask. The values at all exposure slots are integrated pixel-wise to form a coded image.

CE pattern is learned to maximize decorrelation among the coded pixels, ensuring task-agnostic information preservation (Sec. 2.2). A co-designed Vision Transformer then processes the coded image, augmented by a reconstruction-based pre-training that enhances its ability to extract information from CE-compressed inputs (Sec. 2.3). The resulting system supports diverse downstream tasks including video reconstruction and action recognition.

2.1. Coded Exposure for In-Sensor Compression

SnapPix compresses data in-sensor via coded exposure (CE) [22, 28], which selectively exposes pixels across spatial dimensions and multiple frames, integrating the exposures into a single coded image before readout (Fig. 3). Given a video sequence Y of dimensions $T \times H \times W$, CE applies a binary mask M to produce a coded image:

$$X(i,j) = \sum_{t=1}^{T} M(i,j,t) \cdot Y(i,j,t) \qquad (1)$$

achieving a data reduction factor of T (e.g., $T = 16$ in our experiments).

2.2. General-Purpose Sampling via Decorrelation

Rather than learning task-specific CE patterns [14, 21], SnapPix learns a task-agnostic pattern by maximizing

decorrelation among coded pixels. The CE mask is optimized by minimizing:

$$\mathcal{L}_{\text{Cor}} \triangleq \frac{1}{P(P-1)} \sum_{i=1}^{P} \sum_{j \neq i} C_{ij}^2 \tag{2}$$

where C_{ij} is the Pearson correlation coefficient between coded pixels i and j within a tile of P pixels. We decorrelate within local tiles because proximal pixels exhibit the strongest correlation. Before computing correlations, each tile is preprocessed to have zero mean (contrast encoding), which prevents training collapse where all exposure slots remain closed. The mask is binarized and optimized using STE [3]. More details can be found in the full version [16].

2.3. Co-design CE and Vision Model

Co-designed CE Pattern and Vision Transformer. Pixels in coded images carry varying amounts of information due to different exposures. Standard convolutions treat all pixels uniformly, leading to suboptimal performance. Prior work [21] proposed Shift-Variant Convolution (SVC) to address this, but it incurs a 4× slowdown.

Instead, we co-design the CE pattern and vision backbone. SNAPPIX adopts a *tile-repetitive* CE pattern: pixel exposure varies within each tile but repeats across tiles. We use Vision Transformers (ViTs) [6] as the backbone, which naturally divide inputs into patches and process pixels within each patch differently via patch embedding and MLPs (Fig. 2). The CE tile size is set to match the ViT patch size (8×8), and each pixel value is normalized by its number of exposure slots to reduce variation across pixels.

CE-Optimized ViT Pre-training. Inspired by masked pre-training [10, 26], we design a "coded image-to-video" pre-training task. Starting from a CE-coded image, we randomly mask 85% of tiles and train the model to reconstruct the original video sequence (Fig. 2, right). This forces the model to learn both spatial structure (predicting masked tiles) and temporal dynamics (upsampling from CE-coded information).

3. Evaluation

Setup. We evaluate on three action recognition (AR) datasets: Something-Something v2 (SSV2) [9], Kinetics-400 (K400) [11], and UCF-101 [23]. Videos are downsampled to 112 pixels on the shorter dimension, converted to grayscale, and CE-compressed with $T = 16$ exposure slots into a single 112×112 coded image. We provide two variants: SNAPPIX-B (ViT-B, 87M params) and SNAPPIX-S (ViT-S, 22M params). All baselines are reproduced using the same data preprocessing and training recipe for fair comparison. Full experimental details are in [16].

Baselines. We compare our decorrelated CE pattern with the following task-agnostic CE patterns ($T = 16$ in all cases):

- LONG EXPOSURE: All pixels exposed in all slots.
- SHORT EXPOSURE: All pixels exposed every 8th frame.
- RANDOM: Each pixel exposed randomly with a 50% probability per exposure slot.
- SPARSE RANDOM: Each pixel exposed once randomly across T exposure slots.

For AR, we also compare against three prior systems:

- SVC2D [21]: A CE-based AR model with SVC and an end-to-end learned CE pattern.
- C3D [25]: A strong video-based AR model, treated as an upper bound in prior CE works [14, 21].
- VIDEOMAEV2-ST [26]: A state-of-the-art ViT-based video AR model, adjusted to match SNAPPIX-B's speed.

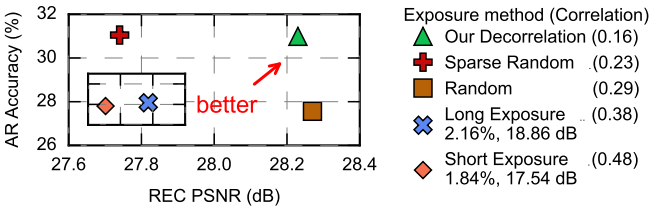

Fig. 4. Comparison of task-agnostic CE patterns using AR and REC results on SSV2. The legend shows the Pearson correlation coefficients.

Table 1. Comparison with previous systems on action recognition. Top 3 highlighted (red, orange, yellow).

Model	Input	Accuracy (↑)			Inference /sec (↑)
		UCF-101	SSV2	K400	
SNAPPIX-S (ours)	CE	74.65%	42.38%	47.58%	2282
SNAPPIX-B (ours)	CE	79.14%	45.21%	54.11%	760
SVC2D [21]	CE	41.16%	23.05%	26.09%	2135
C3D [25]	Video	62.70%	33.48%	41.66%	541
VideoMAEv2-ST [26]	Video	72.54%	39.84%	41.99%	750

Decorrelation Outperforms Other Task-Agnostic CE Patterns. We compare our decorrelation-based CE pattern against other task-agnostic patterns on SSV2 for both AR and video reconstruction (REC), as shown in Fig. 4. Our pattern achieves the best overall performance across both tasks, while patterns with higher pixel correlation (e.g., LONG EXPOSURE, SHORT EXPOSURE) perform significantly worse. The relative performance across strategies corresponds to their correlation coefficients, confirming that decorrelation is an effective objective for learning general CE patterns.

Action Recognition Results. Tbl. 1 summarizes the comparison against CE-based (SVC2D [21]) and video-based methods (C3D [25], VIDEOMAEV2-ST [26]). SNAPPIX-B and SNAPPIX-S achieve the highest AR accuracy across all datasets, nearly doubling the accuracy of prior CE-based methods.

They also surpass video-based models, previously considered upper bounds for CE methods [14, 21], while executing faster. It might be surprising that we surpass video-based baselines in accuracy. The reason stems from our use of a single coded image as input, whereas video-based methods must process 16-frame sequences. This reduced input burden allows us to deploy a significantly larger network within the same inference time budget, ultimately enabling higher accuracy despite operating on less input data.

Energy Savings. Using CamJ [19] for energy modeling, we evaluate the edge power of SNAPPIX under two scenarios. In the *edge-server* scenario, the edge sensor is responsible solely for data capture, with all inference offloaded to a remote server; the dominant energy cost is therefore data transmission. In the *pure-edge* scenario, the edge node performs both sensing and ViT inference, and we measure the total on-device energy consumption.

SNAPPIX with $T = 16$ reduces ADC/MIPI and wireless transmission energy by 16×. In edge-server scenarios, SNAPPIX achieves 7.6× energy saving with short-range passive WiFi [12] and 15.4× with long-range backscatter LoRa [24]. When the edge node has a mobile GPU (Jetson Xavier [1]), SNAPPIX-S achieves 1.4× and 4.5× energy savings over VIDEOMAEV2-ST and C3D, respectively, as it processes a single coded image rather than a full video.

Ablation Study. We perform an ablation study by removing various components from Sec. 2.2 and Sec. 2.3, using SNAPPIX-S as a baseline and the SSV2 dataset and AR task for evaluation:

- Removing pre-training reduces accuracy by 11.39%.
- Replacing the decorrelated pattern with a random pattern further decreases accuracy by 3.43%.
- Replacing the tile-repetitive CE pattern with a global pattern reduces accuracy by 23.74%.

4. Conclusion

We presented SNAPPIX, an in-sensor compression system that addresses three key challenges in edge vision: sensor hardware overhead (**C1**), information loss under compression (**C2**), and task specificity (**C3**). SNAPPIX leverages coded exposure for lightweight, sensor-compatible compression, learns a task-agnostic CE pattern through decorrelation inspired by efficient coding theory, and co-designs tile-repetitive CE patterns with Vision Transformers augmented by reconstruction-based pre-training. SNAPPIX achieves action recognition accuracy surpassing video-based methods while reducing edge energy by up to 15.4×.

Acknowledgment

The work is partially supported by NSF Award #2416375.

References

[1] Nvidia reveals xavier soc details. https://www.forbes.com/sites/moorinsights/2018/08/24/nvidia-reveals-xavier-soc-details/amp/. 4

[2] Horace B Barlow et al. Possible principles underlying the transformation of sensory messages. *Sensory communication*, 1(01):217–233, 1961. 1, 2

[3] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013. 3

[4] Yang Dan, Joseph J Atick, and R Clay Reid. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *Journal of neuroscience*, 16(10):3351–3362, 1996. 1, 2

[5] Harsh Desai, Matteo Nardello, Davide Brunelli, and Brandon Lucia. Camaroptera: A long-range image sensor with local inference for remote sensing applications. *TECS*, 21 (3):1–25, 2022. 1

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 3

[7] Yu Feng, Tianrui Ma, Yuhao Zhu, and Xuan Zhang. Blisscam: Boosting eye tracking efficiency with learned in-sensor sparse sampling. In *ISCA*, pages 1262–1277, 2024. 1

[8] Akshay Gadre, Zachary Machester, and Swarun Kumar. Adapting lora ground stations for low-latency imaging and inference from lora-enabled cubesats. *TOSN*, 20(5):1–30, 2024. 1

[9] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, pages 5842–5850, 2017. 3

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 2, 3

[11] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[12] Bryce Kellogg, Vamsi Talla, Shyamnath Gollakota, and Joshua R Smith. Passive wi-fi: Bringing low power to wi-fi transmissions. In *NSDI*, pages 151–164, 2016. 4

[13] Jeremy Klotz and Shree K Nayar. Minimalist vision with freeform pixels. In *ECCV*, pages 329–346, 2024. 1

[14] Sudhakar Kumawat, Tadashi Okawara, Michitaka Yoshida, Hajime Nagahara, and Yasushi Yagi. Action recognition from a single coded image. *TPAMI*, 45(4):4109–4121, 2022. 1, 2, 3, 4

[15] Robert LiKamWa, Yunhui Hou, Yuan Gao, Mia Polansky, and Lin Zhong. Redeye: Analog convnet image sensor architecture for continuous mobile vision. In *ISCA*, pages 255–266, 2016. 1

[16] Weikai Lin, Tianrui Ma, Adith Boloor, Yu Feng, Ruofan Xing, Xuan Zhang, and Yuhao Zhu. Snappix: efficient-coding–inspired in-sensor compression for edge vision. In *2025 62nd ACM/IEEE Design Automation Conference (DAC)*, pages 1–7. IEEE, 2025. 1, 3

[17] Tianrui Ma, Weidong Cao, Fei Qiao, Ayan Chakrabarti, and Xuan Zhang. Hogeye: neural approximation of hog feature extraction in rram-based 3d-stacked image sensors. In *ISLPED*, pages 1–6, 2022. 1

[18] Tianrui Ma, Adith Jagadish Boloor, Xiangxing Yang, Weidong Cao, Patrick Williams, Nan Sun, Ayan Chakrabarti, and Xuan Zhang. Leca: In-sensor learned compressive acquisition for efficient machine vision on the edge. In *ISCA*, pages 1–14, 2023. 1

[19] Tianrui Ma, Yu Feng, Xuan Zhang, and Yuhao Zhu. Camj: Enabling system-level energy modeling and architectural exploration for in-sensor visual computing. In *ISCA*, pages 1–14, 2023. 4

[20] Gopikrishnan R Nair, Pragnya S Nalla, Gokul Krishnan, Jonghyun Oh, Ahmed Hassan, Injune Yeo, Kishore Kasichainula, Mingoo Seok, Jae-sun Seo, Yu Cao, et al. 3d in-sensor computing for real-time dvs data compression: 65nm hardware-algorithm co-design. *SSC-L*, 2024. 1

[21] Tadashi Okawara, Michitaka Yoshida, Hajime Nagahara, and Yasushi Yagi. Action recognition from a single coded image. In *ICCP*, pages 1–11. IEEE, 2020. 1, 2, 3, 4

[22] Dikpal Reddy, Ashok Veeraraghavan, and Rama Chellappa. P2c2: Programmable pixel compressive camera for high speed imaging. In *CVPR*, pages 329–336. IEEE, 2011. 1, 2

[23] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3

[24] Vamsi Talla, Mehrdad Hessar, Bryce Kellogg, Ali Najafi, Joshua R Smith, and Shyamnath Gollakota. Lora backscatter: Enabling the vision of ubiquitous connectivity. *IMWUT*, 1(3):1–24, 2017. 4

[25] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, pages 4489–4497, 2015. 3

[26] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *CVPR*, pages 14549–14560, 2023. 2, 3

[27] XREAL. Xreal air 2. https://us.shop.xreal.com/products/xreal-air-2. 1

[28] Michitaka Yoshida, Toshiki Sonoda, Hajime Nagahara, Kenta Endo, Yukinobu Sugiyama, and Rin-ichiro Taniguchi. High-speed imaging using cmos image sensor with quasi pixel-wise exposure. *TCI*, 6:463–476, 2019. 1, 2

[29] Tianyi Zhang, Kishore Kasichainula, Dong-Woo Jee, Injune Yeo, Yaoxin Zhuo, Baoxin Li, Jae-sun Seo, and Yu Cao. Improving the efficiency of cmos image sensors through in-sensor selective attention. In *ISCAS*, pages 1–4. IEEE, 2023. 1