Journal Pre-proofs

Article

Engram-Driven Videography

Lu Fang, Mengqi Ji, Xiaoyun Yuan, Jing He, Jianing Zhang, Yinheng Zhu, Tian Zheng, Leyao Liu, Bin Wang, Qionghai Dai

PII: S2095-8099(22)00057-1
DOI: https://doi.org/10.1016/j.eng.2021.12.012
Reference: ENG 937

To appear in: *Engineering*

Received Date: 10 June 2021
Revised Date: 3 December 2021
Accepted Date: 8 December 2021

Please cite this article as: L. Fang, M. Ji, X. Yuan, J. He, J. Zhang, Y. Zhu, T. Zheng, L. Liu, B. Wang, Q. Dai, Engram-Driven Videography, *Engineering* (2022), doi: https://doi.org/10.1016/j.eng.2021.12.012

Research
Artificial Intelligence—Article

Engram-Driven Videography

Lu Fang [a,b,c,d], Mengqi Ji [c,e], Xiaoyun Yuan [a], Jing He [e], Jianing Zhang [a], Yinheng Zhu [a], Tian Zheng [a], Leyao Liu [a], Bin Wang [f], Qionghai Dai [b,c,d,e,*]

[a] *Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*
[b] *Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China*
[c] *Institute for Brain and Cognitive Science, Tsinghua University, Beijing 100084, China*
[d] *Beijing Laboratory of Brain and Cognitive Intelligence, Beijing Municipal Education Commission, Beijing 100010, China*
[e] *Department of Automation, Tsinghua University, Beijing 100084, China*
[f] *Hangzhou Hikvision Digital Technology Co., Ltd., Hangzhou 310012, China*

Corresponding Author.
E-mail address: qhdai@tsinghua.edu.cn (Q. Dai).

ARTICLE INFO
Article history:
Received 10 June 2021
Revised 21 October 2021
Accepted 8 December 2021
Available online

Keywords:
Instance-level videography
Memory engram
Large-scale dynamic scene
Feedforward and feedback
Entropy equilibrium

ABSTRACT
Sensing and understanding large-scale dynamic scenes require a high-performance imaging system. Conventional imaging systems pursue higher capability by simply increasing the pixel resolution via stitching cameras at the expense of a bulky system. Moreover, they strictly follow the feedforward pathway: that is, their pixel-level sensing is independent of semantic understanding. Differently, a human visual system owns superiority with both feedforward and feedback pathways: The feedforward pathway extracts object representation (referred to as memory engram) from visual inputs, while, in the feedback pathway, the associated engram is reactivated to generate hypotheses about an object. Inspired by this, we propose a dual-pathway imaging mechanism, called engram-driven videography. We start by abstracting the holistic representation of the scene, which is associated bidirectionally with local details, driven by an instance-level engram. Technically, the entire system works by alternating between the excitation–inhibition and association states. In the former state, pixel-level details become dynamically consolidated or inhibited to strengthen the instance-level engram. In the association state, the spatially and temporally consistent content becomes synthesized driven by its engram for outstanding videography quality of future scenes. The association state serves as the imaging of future scenes by synthesizing spatially and temporally consistent content driven by its engram. Results of extensive simulations and experiments demonstrate that the proposed system revolutionizes the conventional videography paradigm and shows great potential for videography of large-scale scenes with multi-objects.

1. Introduction

Recent decades have witnessed various developments in videography technology (e.g., from analog to digital, from single lens to multi-camera, and from megapixel-level to gigapixel-level). Even though the imaging resolution has significantly improved, the basic sensing and understanding pipeline remains unchanged; that is, the conventional videography system strictly follows a feedforward pathway by naively accumulating pixel-level information to increase the image resolution and progressively abstract global content. Therefore, the image quality of such videography methods is highly limited by the physical sampling resolution of the photography. Stitching cameras [1–3] and scanning strategies [4–8] can significantly boost the count of physically sampled pixels, especially for gigapixel videography. However, most such videography strategies simply fuse spatio-temporal photographic outputs, rather than taking advantage of the inherent instance-level representation with spatio-temporal consistency. This leads to high hardware complexity and tremendous data redundancy for the spatio-temporally dense sampling.

In contrast, a high-level human visual system uses past observations in service of the present or future, with a dual-pathway mechanism [9]. Specifically, it extracts and consolidates a memory engram, defined as the representation of a stable and semantic memory in the brain [9–12], and subsequently reactivates the associated engram through a feedback pathway for memory retrieval. In particular, upon engram formation [9–12], local detailed information is initially transmitted to the hippocampus. Then, recurrent associations between the hippocampal–cortical networks gradually strengthen and consolidate the cortical engram, and the global semantic information is extracted by the prefrontal cortex (PFC) [9,13]. The PFC then

discriminates whether the incoming visual information corresponds to a previously stored cortical engram, and determines the subsequent procedure of consolidation or inhibition. Taken together, these dynamic and bidirectional information transfers realize efficient, robust, and adaptive visual perception and understanding.

Inspired by this, we propose a dual-pathway imaging mechanism denoted "engram-driven videography." It starts by abstracting a holistic representation of a scene, which is bidirectionally associated with the local details, as driven by an instance-level engram. Analogically, the visual information is first captured by virtual eyes and pre-processed by the PFC module. The PFC module determines which information is already stored in the videographic memory. We call this the excitation–inhibition state, where the pixel-level details are dynamically consolidated or inhibited to strengthen the instance-level engram. Such a dynamic system mimics the human memory mechanism and can be programmed to maximize uncertainty, similar to the principle of entropy towards the equilibrium of the system [14,15]. Intuitively, to effectively maintain a dynamic videographic memory with a limited size, it is generally encouraged to increase the uncertainty of the memorized content by maximizing the entropy. Subsequently, in the association state, a constant-level engram is retrieved to synthesize the spatio-temporally consistent content for high-performance videography of future scenes. The schematic of engram-driven videography is shown in Fig. 1.

Experiments on both computer synthetic and real-world images/videos demonstrate that our engram-driven videography can significantly outperform traditional videographic approaches. It can generate high-quality instance-level results with only 5% high-resolution observations and 95% low-resolution observations. In addition, with the help of the associated engram, we can recover the high-resolution details of small instances. We believe that the engram-driven videography will open new directions for image/video capturing, understanding, and storage, leading to next-generation visual sensation systems.

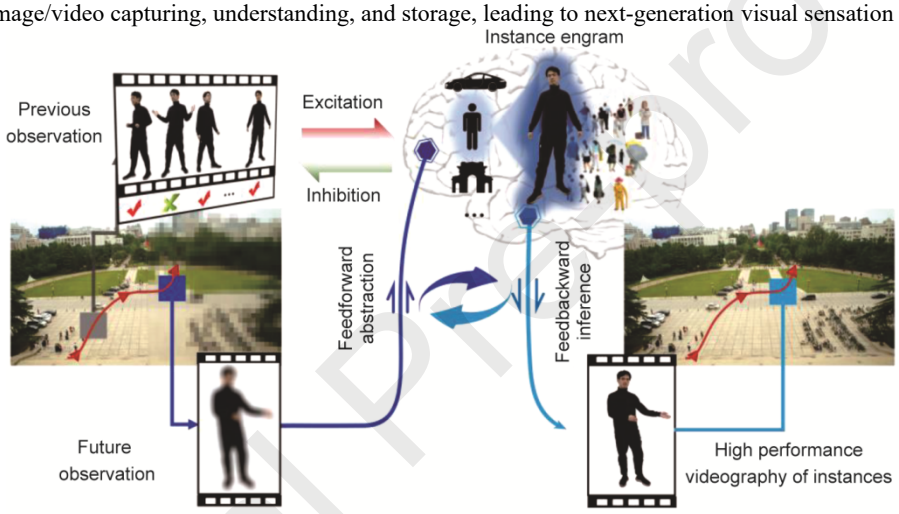


Fig. 1. Schematic of engram-driven videography. It consists of an excitation–inhibition state, where the pixel-level details get dynamically consolidated or inhibited to strengthen the instance-level engram, and an association state, where the spatially and temporally consistent content get synthesized as driven by the engram for high-performance imaging of future scenes.

2. Related work

A large number of studies have considered human visual engrams, super-resolution algorithms, and high-resolution imaging systems. These are reviewed in the following subsections.

2.1. Human visual engram

The human brain perceives and processes visual information at two levels [16]. At a low level, the cerebral visual cortex processes local visual features such as color, contrast, direction, and motion through feedforward hierarchical structure information. In the high-level visual processing, the engram is an indispensable component of the conscious experience. Object recognition, in particular, relies on the observer's previous engram [16].

The term engram was coined by Semon more than 100 years ago, and refers to the representation of a more stable memory [9–12]. The prevailing view is that an engram in the brain may change with time [9–12]. Specifically, visual information is initially encoded in parallel in hippocampal–cortical engrams, and the recurrent associations between the hippocampal–cortical networks gradually strengthen the cortical engram for memory consolidation. Finally, the PFC abstracts the global semantic information from the pre-existing cortical engram, and then discriminates whether the incoming visual information corresponds to a previously stored cortical engram for the subsequent decision procedure of consolidation or inhibition [13]. In information theory, Shannon entropy measures the amount of information held in data; the proposed memory entropy is named based on reference to this idea. In particular, in our system, the memory entropy encodes the uncertainty of the instance-level visual information in the engram. Memory entropy can be regarded as an extension of Shannon entropy in the field of imaging and memory, as we do not calculate the entropy directly. Instead, we use the distance between the feature vectors to

represent the relative entropy change.

2.2. Super-resolution algorithms

Image/video super resolution approaches aim to recover high-resolution details from low-resolution inputs. The most common super-resolution method is single image super-resolution (SISR). Early works only used low-level priors such as sparsity [17–20] or exemplar patches [21,22], whereas the deep neural network has shown great performance. Kim et al. [23] first proposed a three-layer neural network for single-image super-resolution, and further improved it by using a 20-layer deep neural network [23], recursive structures [24,25], and a dense structure [26]. However, the mean squared error losses from these methods usually led to over-smooth results with no details. Thus, perceptual losses were introduced into image/video super-resolution. Johnson et al. [27] first proposed perceptual losses for real-time style transfer and super-resolution by combining conventional pixel-wise losses and Visual Geometry Group (VGG)-like feature spaces losses. Ledig et al. [28] presented super resolution generative adversarial network (SRGAN), a generative adversarial network for photo-realistic 4× natural images at super-resolution.

The performance of SISR is limited, especially under a large resolution gap (> 8×), because the high-frequency details are lost during down-sampling and are unrecoverable under general priors. Therefore, reference-based super-resolution (RefSR) has been proposed. Boominathan et al. [29] adopted a digital single-lens reflex (DSLR) camera image as a reference, and presented a patch-matching-based method for super-resolved low-resolution light field images. Wu et al. [30] improved the algorithm by proposing a better patch-matching algorithm combined with dictionary-learning-based reconstruction. Wang et al. [31] iterated the patch-matching step to enrich the patch database and improve the reconstruction quality. Zhang et al. [32] presented the super-resolution neural texture transfer (SRNTT) to conduct multilevel patch matching in the neural space. The methods mentioned above only used the information of local patches. This often led to poor super-resolution results under real data from hybrid camera systems, owing to the easily failed patch matching. Therefore, Zheng et al. [33,34] proposed CrossNet and CrossNet++, an end-to-end neural network containing a novel two-stage cross-scale warping module to build the correspondence between the input and reference. Compared with patch-based approaches, the warping approach can find more reliable correspondences for the entire image when the input and reference images are close.

2.3. High-resolution imaging systems

In addition to super-resolution algorithms, researchers have developed powerful imaging systems to increase the resolution. Kopf et al. [35] designed a motor-controlled camera mount for static gigapixel image capture. Brady et al. [3] built the world's first gigapixel camera "AWARE2" with a spherical objective lens and 98 micro-optics; it produced a 0.96 gigapixel image in a single shot. The objective lenses were carefully designed and manufactured to minimize aberrations [36,37]. In contrast, Cossairt et al. [38] used a simple optical design and post-capture deconvolution stage to increase the resolution. Owing to the large computational complexity caused by the extremely high resolution, AWARE2 can only capture three frames per minute. A similar idea has also been employed in microscope design for gigapixel-level whole-mouse brain imaging [39]. To reduce the computational complexity, Yuan et al. [40] proposed multiscale gigapixel videography, which dramatically reduced the bandwidth requirement and was capable of handling an 8× resolution gap. Zhang et al. [1] extended this idea to 3D videography for virtual reality (VR) applications. However, even with such a strategy, a large number of cameras is still required. The spatial redundancy is reduced, whereas the temporal redundancy still exists.

Our new engram imaging system considers both spatial and temporal redundancies. The history inputs are encoded and buffered in "memory" as references, and are used to super-resolve further inputs.

3. Method

In this paper, we propose engram-driven videography, including a feedforward pathway for consolidating the interesting information of a scene, and a feedback pathway for associating the engram for future content inference and synthesis. Our main contribution lies in a high-level neuromorphic imaging system consisting of the functional modules for an engram neural circle [9,10,12]. As shown in Figs. 2 and 3, our imaging system consists of three main modules: an instance-level observation module, instance-level engram module, and engram processing unit (EPU), corresponding to the hippocampus, cortical memory, and PFC in the human brain, respectively. The former two are responsible for low-level and high-level information representation, whereas the latter is responsible for information abstraction and control.

In the feedforward pathway (Fig. 2), an entropy equilibrium policy is proposed for consolidating the valid information from the low-level observations to the engram and inhibiting meaningless information. In the feedback path (Fig. 3), the pre-generated engram is associated with new observations and used to synthesize high-quality future content.

3.1. Videographic memory entropy

Similar to the Universe, our brains might be programmed to maximize disorder (similar to the principle of entropy), and our memory mechanism could simply be a side effect. Accordingly, we used entropy to quantify the functioning of the photographic memory in the human brain (i.e., to characterize the degree of the underlying uncertainty of the instance-level engram). Entropy is a term used to describe the progression of a system from order to disorder (i.e., random variables with

small entropies have a high level of predictability, and hence a low level of uncertainty) [41].

The visual observations abstracted into the engram are defined as a continuous-valued random vector $\bar{\boldsymbol{x}} \in \mathbb{R}^N$ in an N-dimensional feature space with a probability density function $f(\boldsymbol{x})$. The photographic memory entropy $H(\boldsymbol{X})$ is defined as follows, and is used to measure the degree of uncertainty that the information content $\bar{\boldsymbol{x}}$ comprises. Storing redundant observations in the engram should be avoided, because the visual system is highly robust to predicting one sample from a similar one.

$$H(\boldsymbol{X}) = \mathrm{E}\{-\log f(\bar{\boldsymbol{x}})\} \tag{1}$$

where E is the expected value operator.

The next question is how to measure the prediction confidence for the neighboring regions in the feature space. Intuitively, when two samples $\boldsymbol{x_1}, \boldsymbol{x_2}$ get closer in the feature space, it is more reasonable to represent one sample using the other. Therefore, we use a multivariate Gaussian $g_i(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ with a mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$, (i.e., $\mathcal{N}(\boldsymbol{x}_2; \boldsymbol{x}_1, \boldsymbol{\Sigma}_i)$ to represent the prediction confidence of $\boldsymbol{x}_2$ using $\boldsymbol{x}_1$. For the case where the memory contains multiple visual observations, $f(\boldsymbol{x})$ can be defined as follows:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} c_i \cdot g_i(\boldsymbol{x}) = \sum_{i=1}^{N} c_i \cdot \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{2}$$

where c_i is non-negative weighting coefficients, with $\boldsymbol{\Sigma}_i c_i = 1$.

Similar to the human working memory, which maintains a limited amount of information, the proposed photographic memory also has a limited capacity, so that it can be quickly accessed to serve the needs of ongoing videography tasks. Accordingly, we propose an engram update mechanism for stimulating the progression of the system from order to disorder. Specifically, it is programmed to maximize the information content; simultaneously, the entropy of our videography system can only increase.

Here we define two states, the memory state and instantaneous state, as follows:

$$f_{\mathrm{mem}}(\boldsymbol{x}) = \sum_{i=1}^{N-1} c_i \cdot g_i(\boldsymbol{x}) + c_k \cdot g_k(\boldsymbol{x}) \tag{3}$$

$$f_{\mathrm{inst}}(\boldsymbol{x}) = \sum_{i=1}^{N-1} c_i \cdot g_i(\boldsymbol{x}) + c_q \cdot g_q(\boldsymbol{x}) \tag{4}$$

where the kth sample is the most redundant in the memory, and the qth sample represents an incoming query sample.

Intuitively, after the query sample replaces the most useless one in the size-limited memory, and if the system entropy increases (i.e., $H_{\mathrm{mem}}(\boldsymbol{X}) < H_{\mathrm{inst}}(\boldsymbol{X})$), the engram update mechanism tends to encourage such an update. In contrast, if the newly observed information decreases the system entropy (i.e., $H_{\mathrm{mem}}(\boldsymbol{X}) \geq H_{\mathrm{inst}}(\boldsymbol{X})$), such a process will be inhibited.

The key idea for the engram update mechanism is to retain the information that increases the entropy of the photographic memory system. For the implementation, the critical part is in evaluating the entropy to guide the engram update strategy. Even though the entropy of an N-dimensional Gaussian has a simple closed-form expression, the entropy generally cannot be calculated in the closed form for Gaussian mixtures, owing to the logarithm of the sum of exponential functions. Following the approximation method [52], the bounds of the mixture entropy are as follows:

$$\hat{H}_{\mathrm{BD}} \leq H(\boldsymbol{X}) \leq \hat{H}_{\mathrm{KL}} \tag{5}$$

Here, $\hat{H}_{\mathrm{BD}}$ and $\hat{H}_{\mathrm{KL}}$ are defined on Chernoff α-divergence and Kullback–Leibler (KL)-divergence, respectively, which decrease together with the distance $\min_{i,j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$.We can conclude that it is highly likely that the entropy $H(\boldsymbol{X})$ is proportional to its bounds (i.e., $\hat{H}_{\mathrm{BD}}$ and $\hat{H}_{\mathrm{KL}}$).

Therefore, to maintain the equilibrium of the system, we evaluate the instantaneous videographic memory entropy $H_{\mathrm{inst}}(\boldsymbol{X})$ for each observation, and compare it with $H_{\mathrm{mem}}(\boldsymbol{X})$. Eventually, the pixel-level details are dynamically consolidated or inhibited, so as to strengthen the instance-level engram in the excitation-inhibition state.

3.2. Feedforward pathway

Fig. 2 illustrates the feedforward pathway for the engram generation. The pixel-level details are first captured by a virtual eye (camera) from a large-scale scene, and are pre-processed and saved as instance-level observations (hippocampus in the human brain), including foreground dynamic objects and interesting backgrounds. Here, faster region-based convolutional neural network (R-CNN) [42] and mask R-CNN [43] are used for bounding box detection and segmentation, respectively. Subsequently, the high-level semantic information is extracted from these low-level observations using the encoder. We compute the feature vector using five convolutional layers, each with 64 filters of size 5×5. The strides are set to one for the first two layers and two for the following three layers, leading to coarse-to-fine multiscale feature maps. The feature vector is then computed by concatenating the feature maps. Subsequently, an entropy equilibrium policy is used to decide whether to consolidate the information to the engram or to inhibit it.

(1) Virtual eye. Although the human eye has approximately 0.576 gigapixels, it only has very high resolution in the center of the field of view (FoV) [44,45], and uses saccadic movements to generate an all-clear wide-FoV image in the brain, leading to highly efficient information capturing and integration. Thus, we designed a hybrid camera system to mimic it: A global camera with wide-FoV captures low-resolution whole scenes, whereas several local cameras with long-focus lens tracks are used to capture interesting details.

(2) Entropy equilibrium. To reach the videographic memory equilibrium, the state with the maximum entropy, we can directly maximize the minimal pairwise distance $\min_{i,j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$, which is approximately proportional to both the videographic memory entropy $H_{\mathrm{mem}}(\boldsymbol{X})$ and the instantaneous one $H_{\mathrm{inst}}(\boldsymbol{X})$. As depicted in Fig. 2, the feature vectors $\boldsymbol{\mu}_{i=1,\ \ldots,\ N}$ are

produced by the encoder and transmitted to the videographic memory to update the instance-level engram. Intuitively, if a query sample $\boldsymbol{\mu}_q$ can increase the videographic memory entropy (i.e., $H_{\text{mem}}(\boldsymbol{X}) < H_{\text{inst}}(\boldsymbol{X})$), the engram update mechanism tends to encourage such an update. In contrast, if the newly observed information decreases the system entropy, i.e., $H_{\text{mem}}(\boldsymbol{X}) \geq H_{\text{inst}}(\boldsymbol{X})$, such a process will be inhibited. Notably, we do not need to calculate the entropy. Instead, as described in the previous section, the minimal pairwise distance $\min_{i,j} \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$ can reflect the relative changes in the entropy. In our experiment, we used the perceptual embeddings as feature vectors (i.e., $\boldsymbol{\mu}_{i=1,\ldots,N}$).

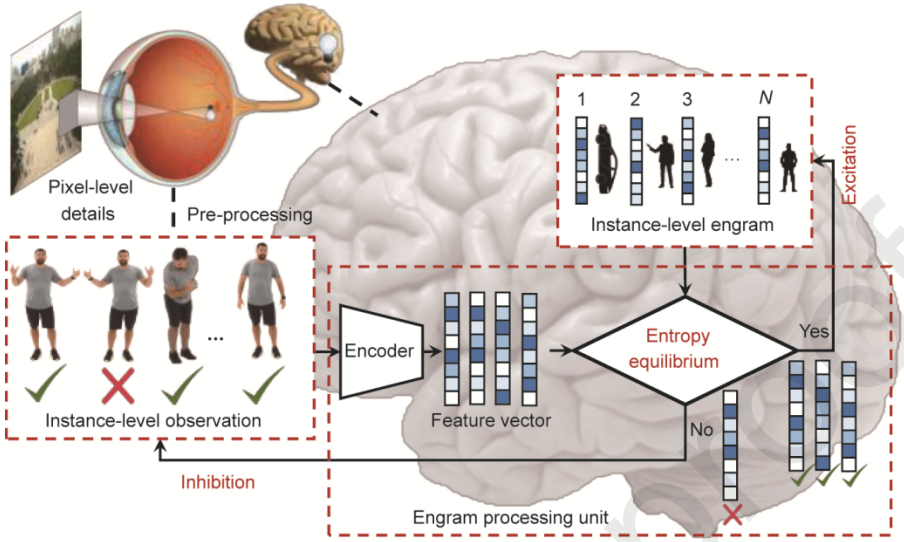


Fig. 2. Feedforward pathway for engram generation. The large-scale scene is captured by virtual eyes, and is pre-processed and saved in instance-level observations first. After that, the EPU abstracts the low-level information and consolidates it to an instance-level engram (excitation state) or inhibits it (inhibition state), based on the entropy equilibrium model.

3.3. Feedback pathway

After generating the engram, we used a designed feedback pathway for the engram-driven videography synthesis (Fig. 3). In this approach, the pixel-level details are captured and pre-processed to instance-level observations. In contrast to the feedforward pathway focused on engram generation, the instance-level engram helps synthesize high-resolution images from low-quality instance-level observations. Technically, an engram association module is trained to retrieve the best-matched feature vector from the engram (denoted the "engram vector"). Finally, we concatenate the two feature vectors and use a synthesis module to generate high-resolution images [33,34].

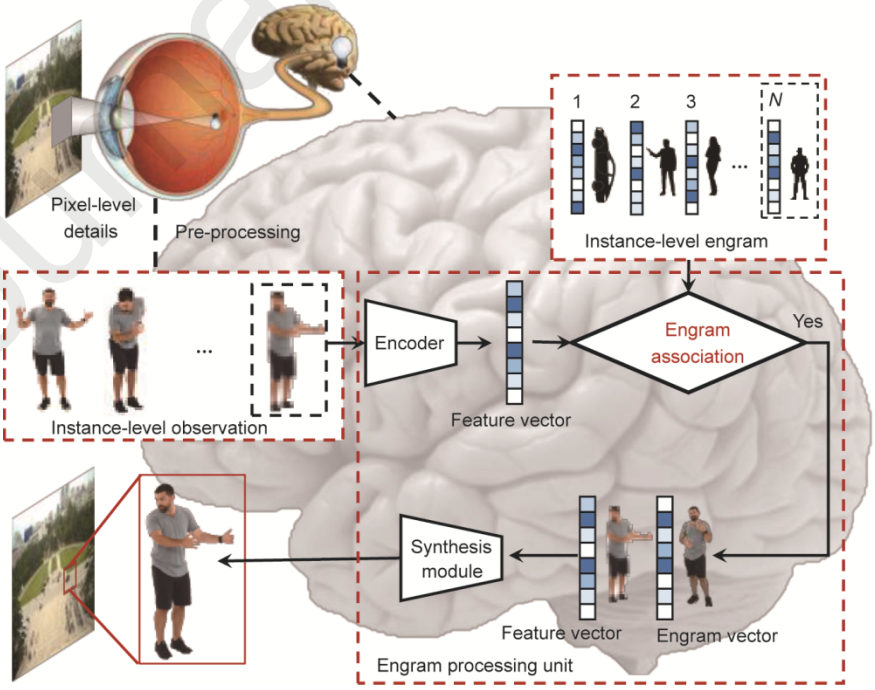


Fig. 3. Feedback pathway for engram-driven videography synthesis. An engram association module is used to match the feature vector between the observation and engram. These two vectors are then concatenated, and are input to the synthesis module.

(1) Engram association. The purpose of engram association is to find the engram vector for generating the best high-resolution image. This means that the selected engram vector should be close to the feature vector of the observation. Although there are already a large number of loss functions both in the low-level pixel domain and high-level semantic domain [27], none of them can handle a large resolution/quality gap. Thus, we implement the engram association module using our new pairwise loss function, as shown in Fig. 4(a). The feature vector of the observation is generated by the encoder, whereas the engram vector is extracted from the high-level engram directly. A perceptual loss is then used to estimate the similarity score. This step is applied to all of the engram–input pairs to select the best engram vector.

As it is very difficult to obtain the ground truth of a similarity score, we adopt the pairwise ranking loss to train the network [46]. During the generation of the training data, we examine all of the observation–engram pairs and use the peak signal-to-noise ratio (PSNR) of the synthesized image to rank all of the engram vectors. With this ranking, we train the network by comparing two observations–engram pairs using the pairwise ranking loss.

(2) Synthesis module. As illustrated in Fig. 4(b), we recover the high-quality image using a synthesis module. A warping module is used to build a dense correspondence between the feature vector and engram vector. We modify the FlowNet structure to support multichannel feature maps [47,48]. Finally, we concatenate the two vectors and use a decoder to generate the final high-resolution image. The decoder consists of one deconvolutional layer with 64 filters (size 4 × 4, stride 2). The loss function L is defined between the decoder output and ground truth, as follows:

$$L = \rho(I_H - I_{\text{gt}}) \tag{6}$$

where $\rho(x) = \sqrt{x^2 + 0.001^2}$ is the Charbonnier penalty function [49]; I_{gt} denotes the ground truth; and I_H represents the decoder output.

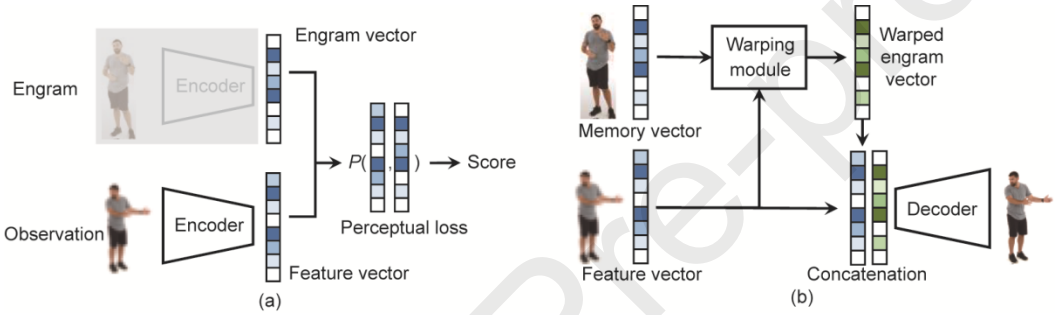


Fig. 4. Network structure of two sub-modules in the feedback pathway. (a) Association module for engram extraction; (b) synthesis module. *P*: perceptual loss.

4. Experiment

We verified our engram imaging system on both computer synthetic data and real-world data, and showed evident improvements for both compared to conventional imaging systems. We also conducted ablation studies to demonstrate the effectiveness of each module in our neural network.

4.1. Data preparation

To better verify the effectiveness of our engram imaging system, we used the Unreal Engine to render synthetic data for the simulation verification. We also rendered virtual human models on the scene using predefined paths; these were used as the simulated interesting instances of the synthetic video. The main camera was set to a 90-degree FoV with 8192 × 8192 30-fps resolution. The high-resolution videos were first rendered as a ground truth, and then were down-sampled as a low-resolution wide-FoV video. We also cropped a high-resolution small block from the ground truth as a high-resolution small-FoV video (mimicking the center of the human eye).

In addition to computer synthetic data, we tested our system on real captured videos. The real-world data were generated using the gigapixel video dataset PANDA [50]. As PANDA covers a wide-FoV with an extremely high resolution, we cropped a small moving block located on interesting regions or objects as a high-resolution small-FoV video, and down-sampled the entire video as a low-resolution wide-FoV video.

4.2. Comparison with conventional imaging system

Fig. 5 demonstrates the results from our engram imaging system and the conventional imaging system on the computer synthetic scene with quantitative evaluation in Table 1. Five persons are covered by the whole wide-FoV video, and the high-resolution video block is designed to scan all the persons circularly. Each scanning cycle for one person contains one high-resolution frame and 19 low-resolution frames. The results show that our system can dramatically improve the visual quality. In the conventional imaging system, the person becomes blurry, especially the faces, once the local camera moves to other persons. In our system, the persons can be kept at high resolution for a long time with the help of the engram. We demonstrate

faces with different view angles for two representative persons. Our image system can recover the details of the faces and hairs, such as the eyes, nose, and mouth.

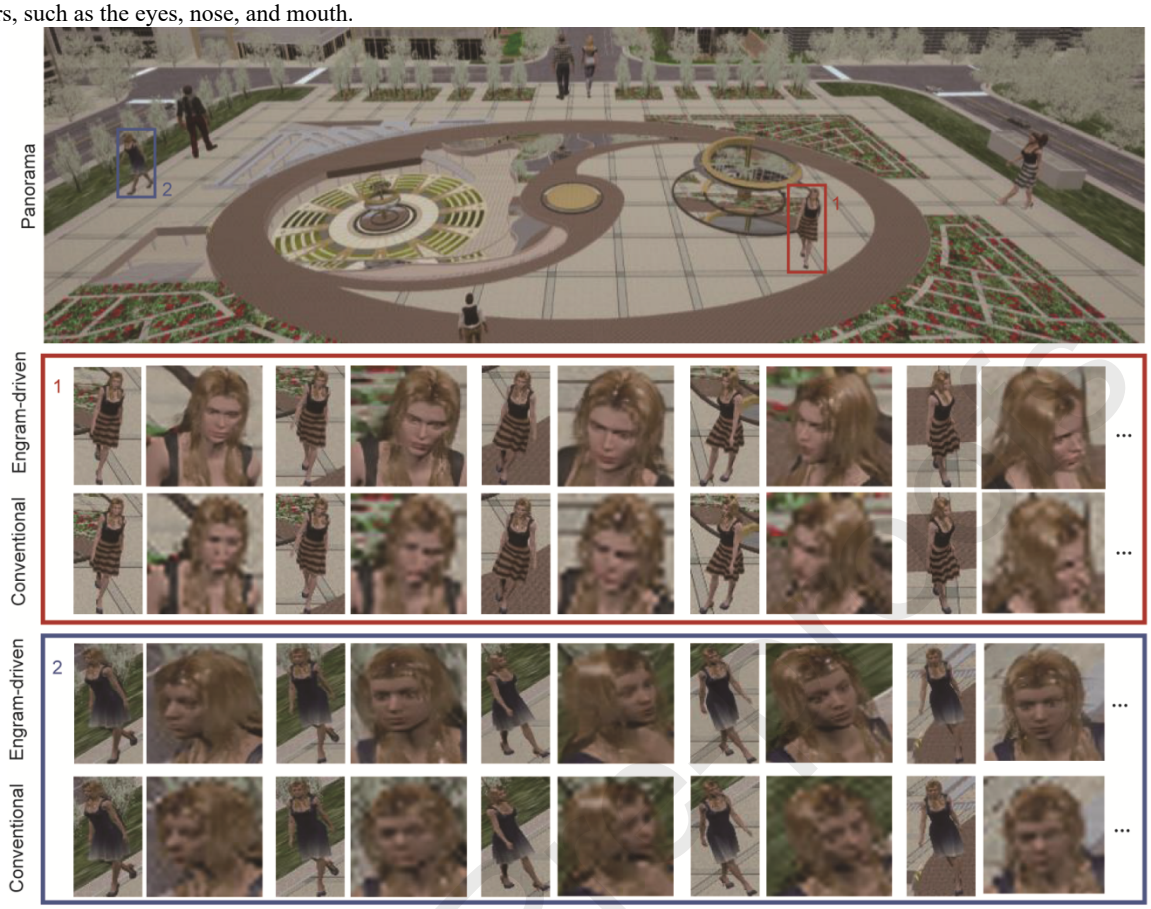


Fig. 5. Results on synthetic data. Two representative persons are highlighted.

Table 1
Quantitative comparison with state-of-the-art methods for different sequences.

Sequence	Method [7–9]	PSNR (dB)
Synthesis	LIIF	31.84
	SRNTT	31.87
	MDSR	32.19
	Ours	32.32
Real-world	LIIF	31.94
	SRNTT	32.12
	MDSR	32.17
	Ours	32.53
PANDA	LIIF	32.24
	SRNTT	32.35
	MDSR	33.36
	Ours	34.29

LIIF: local implicit image function; MDSR: multi-scale deep super-resolution system.

We also tested our results using real-world data (Fig. 6). This video was captured on the campus of the Harbin Institute of Technology, Shenzhen. The same scanning strategy was used to generate the data. Several typical frames of two representative persons are cropped and shown below the panoramic image. The results show that our system can successfully recover small details, such as fingers. Our system can recover different expressions from very low-resolution observations with the help of the engram.

In addition to dynamic persons, some static objects also benefit from our imaging system, as shown in Fig. 7. Compared with the state-of-the-art SISR, multi-scale deep super-resolution system [51], and SRNTT algorithms [32], our method successfully restores the recognizable Chinese characters (red block); this is almost impossible without the engram. The quantitative results are presented in Table 1. The quantitative results show that our method achieves a higher PSNR than the

local implicit image function and SRNTT approaches, especially on the PANDA dataset.

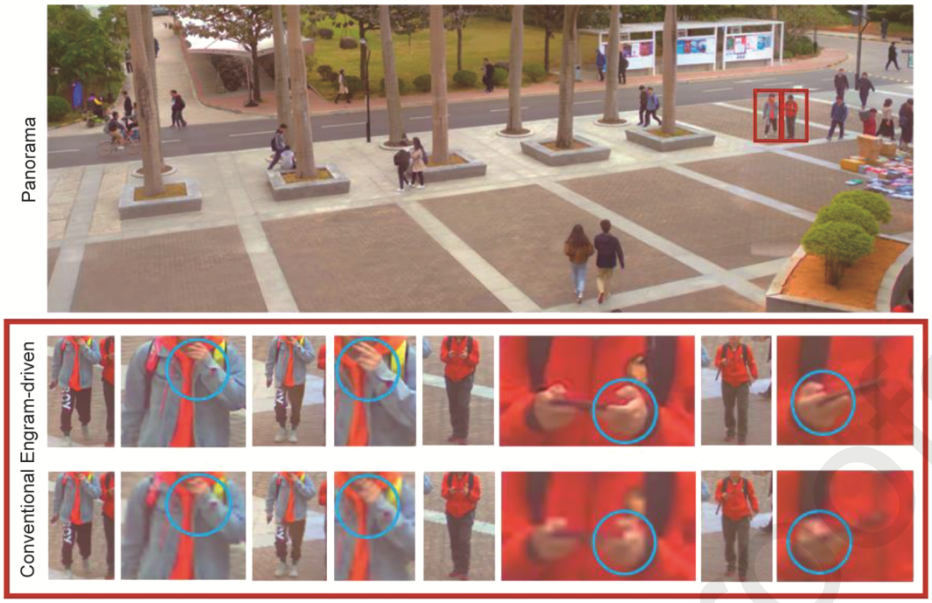


Fig. 6. Results on real-world data (dynamic). Two representative persons are highlighted. The red and blue trajectories denote their moving path.

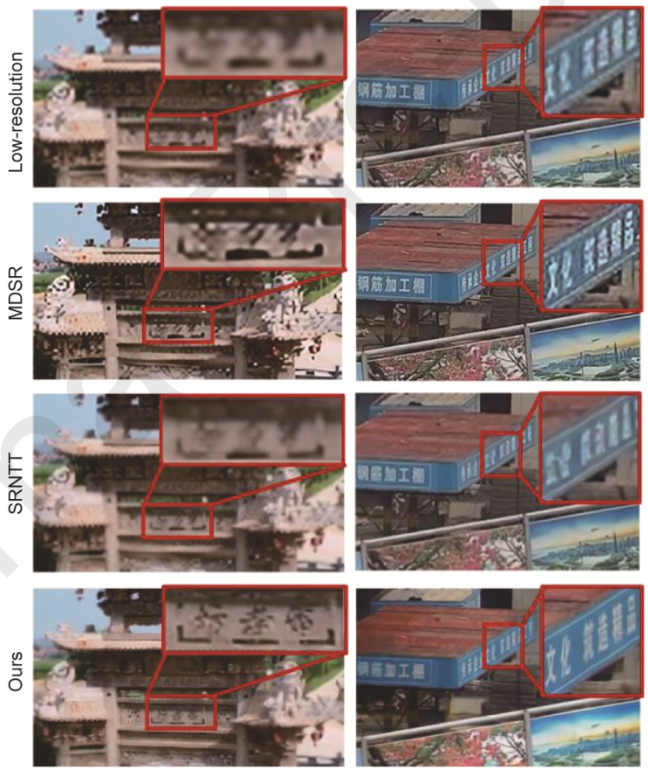


Fig. 7. Results on real-world data (static objects). From top to bottom, low resolution observation, results from state-of-the-art SISR, multi-scale deep super-resolution system (MDSR) [51], and RefSR SRNTT algorithms [32], and those from our engram imaging system.

4.3. Effectiveness of the engram association module

We also conducted experiments to verify the effectiveness of the engram association module. Two factors were considered: the engram buffer size and the association strategy. For the former, we chose 2%, 5%, 10%, and 20% of the video size, and tested their performance. For the association strategy, we compared our trained strategy model with a random selection method. Fig. 8 illustrates the results of the study. As expected, increasing the engram size benefits the imaging quality, but the improvement becomes minor when over 10%. Our trained engram association module shows superior performance to the

random method, especially when the engram size is small.

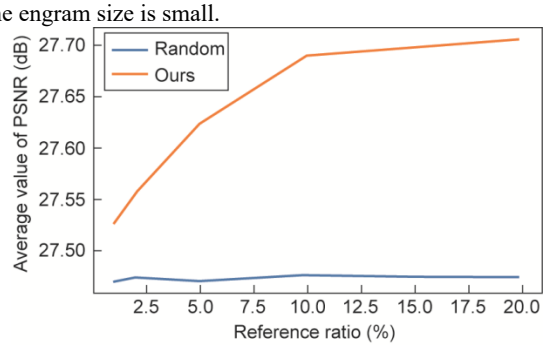


Fig. 8. Plots show the effectiveness of our engram association module.

4.4. Ablation study

An ablation study was conducted to verify the effectiveness of the three modules in our engram imaging system. There are three important modules in our system, EPU, instance-level observation module, and instance-level engram module, corresponding to three regions in the human brain.

(1) Without EPU. The EPU corresponds to the PFC of the human brain. As described in Section 4.2, we tested a random method and untrained perceptual distance, but the references were still selected from the same person. Without the EPU, the engram association module can hardly find the correct engram vector belonging to the same person or object. Thus, we randomly selected the engram vectors from the entire engram. Engram vectors could also be selected from other persons or objects.

(2) Without instance-level engram. The instance-level engram module corresponds to the cortical memory in the human brain, and is responsible for saving the abstracted and consolidated engram. Without it, we can no longer consolidate the useful feature vectors and inhibit meaningless ones. In addition, the instance-level observations can only save short-term content (similar to the hippocampus of the human brain). Hence, we used random patches from the entire image frame of nearby frames as the engram vectors in this subsection.

(3) Without instance-level observation. Without instance-level observation, the system degrades to a SISR system, so the feature vector of the observation itself is used as the engram vector.

Fig. 9 shows the results of the ablation study. As expected, the full system has the highest PSNR and best visual quality (distinguishable face). Without any of the three modules, the reconstruction quality drops dramatically, and the faces become very blurry.

Fig. 9. Results of ablation studies. w/o: with/without.

5. Conclusions

To achieve high-performance imaging, we proposed an engram-driven videography system with dual feedforward and feedback pathways. The feedforward pathway extracts instance-level representations to form the engram used by the feedback pathway to synthesize future high-resolution images, similar to that in the human visual system. In the feedforward pathway, a videographic entropy equilibrium concept is proposed for deciding whether to consolidate the information to the instance-level engram or inhibit it, leading to a compact and highly efficient representation. In the feedback pathway, a ranking-based engram association module is combined with a feature domain warping and synthesis module to generate spatially and temporally consistent high-resolution content. Experiments on both computer synthetic and real-world images/videos

demonstrate that our engram-driven videography can generate high-quality results using only 5% high-resolution observations and 95% low-resolution observations. In addition, with the help of the associated engram, we can recover the high-resolution details of small instances. Such techniques can also be used for image/video super-resolution, low-cost gigapixel imaging, VR/augmented reality (AR) content representation and compression (foveated rendering), and so on. We believe that the engram-driven videography will open new avenues for image/video capturing, understanding, and storage, leading to a next-generation visual sensation system.

Acknowledgments

The authors would like to thank Yijie Deng, Xuechao Chen, Xuecheng Chen, and Puhua Jiang for their great help, as well as the Shuimu Tsinghua Scholar Program. Project funded by Natural Science Foundation of China (62125106, 61860206003, and 62088102), in part by Shenzhen Science and Technology Research and Development Funds (JCYJ20180507183706645), in part by Beijing National Research Center for Information Science and Technology (BNRist) under Grant (BNR2020RC01002), China Postdoctoral Science Foundation (2020TQ0172, 2020M670338, and YJ20200109), and Postdoctoral International Exchange Program (YJ20210124).

References

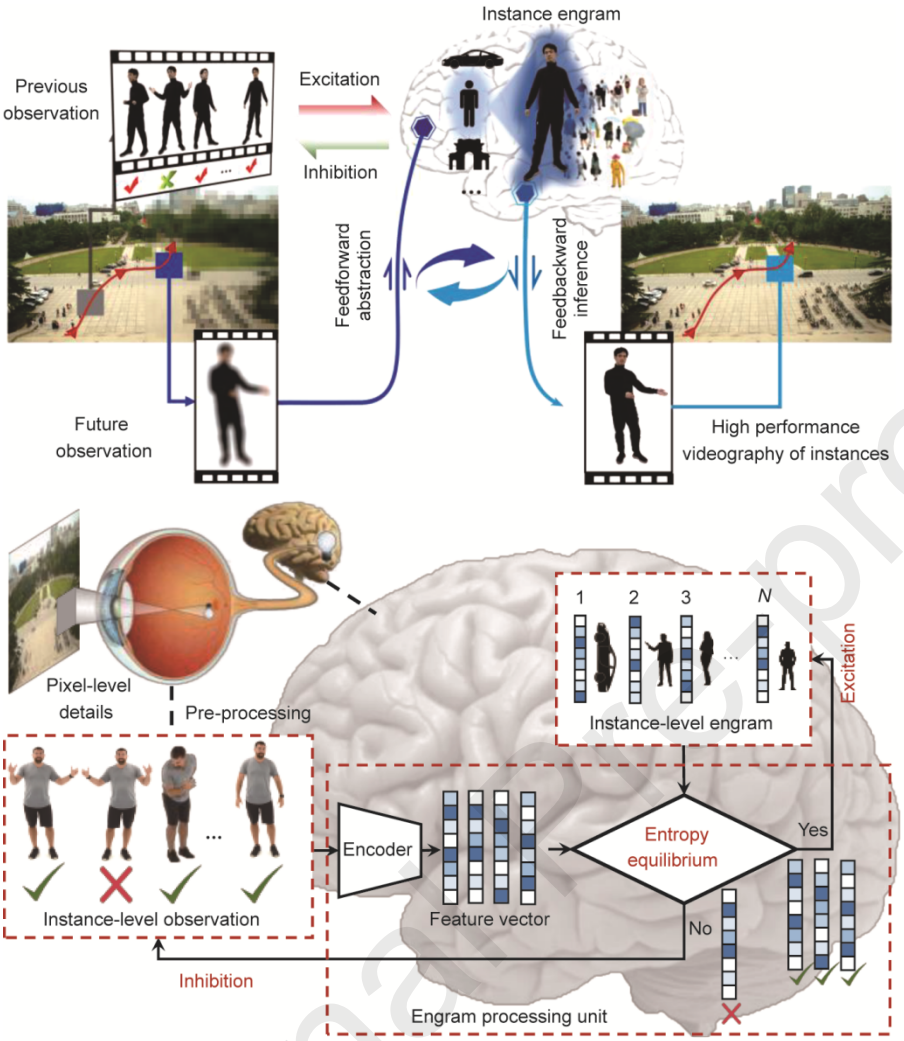
[1] Zhang J, Zhu T, Zhang A, Yuan X, Wang Z, Beetschen S, et al. Multiscale-VR: multiscale gigapixel 3D panoramic videography for virtual reality. In: Proceedings of 2020 IEEE International Conference on Computational Photography (ICCP); 2020 Apr 24–26; St. Louis, MO, USA. New York: IEEE; 2020. p. 1–12.
[2] Li F, Yu J, Chai J. A hybrid camera for motion deblurring and depth map super-resolution. In: Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition; 2008 Jun 23–28; Anchorage, AK, USA. New York: IEEE; 2008. p. 1–8.
[3] Brady DJ, Gehm ME, Stack RA, Marks DL, Kittle DS, Golish DR, et al. Multiscale gigapixel photography. Nature 2012;486(7403):386–9.
[4] Li G, Zhao Y, Ji M, Yuan X, Fang L. Zoom in to the details of human-centric videos. 2020. arXiv:2005.13222.
[5] Xu Y, Deng Z, Wang M, Xu W, So AMC, Cui S. Voting-based multiagent reinforcement learning for intelligent IoT. IEEE Internet Things J 2021;8(4):2681–93.
[6] Zhang J, Koppel A, Bedi AS, Szepesvari C, Wang M. Variational policy gradient method for reinforcement learning with general utilities. 2020. arXiv:2007.02151.
[7] Ilie A, Welch G. Online control of active camera networks for computer vision tasks. ACM Trans Sens Netw 2014;10(2):1–40.
[8] Gu J, Hitomi Y, Mitsunaga T, Nayar S. Coded rolling shutter photography: flexible space-time sampling. In: Proceedings of 2010 IEEE International Conference on Computational Photography (ICCP); 2010 Mar 29–30; Cambridge, MA, USA. New York: IEEE; 2010. p. 1–8.
[9] Josselyn SA, Tonegawa S. Memory engrams: recalling the past and imagining the future. Science 2020;367(6473):eaaw4325.
[10] Tonegawa S, Morrissey MD, Kitamura T. The role of engram cells in the systems consolidation of memory. Nat Rev Neurosci 2018;19(8):485–98.
[11] Tonegawa S, Liu X, Ramirez S, Redondo R. Memory engram cells have come of age. Neuron 2015;87(5):918–31.
[12] Josselyn SA, Köhler S, Frankland PW. Finding the engram. Nat Rev Neurosci 2015;16(9):521–34.
[13] Frankland PW, Bontempi B. The organization of recent and remote memories. Nat Rev Neurosci 2005;6(2):119–30.
[14] Dudai Y. The neurobiology of consolidations, or, how stable is the engram? Annu Rev Psychol 2004;55:51–86.
[15] Marr D. A theory for cerebral neocortex. Proc R Soc Lond B 1970;176(1043):161–234.
[16] Kandel ER, Schwartz JH, Jessell TM. Principles of neural science. 4th ed. New York: McGraw-hill; 2000.
[17] Kim KI, Kwon Y. Single-image super-resolution using sparse regression and natural image prior. IEEE Trans Pattern Anal Mach Intell 2010;32(6):1127–33.
[18] Yang J, Wang Z, Lin Z, Cohen S, Huang T. Coupled dictionary training for image super-resolution. IEEE Trans Image Process 2012;21(8):3467–78.
[19] Cao F, Cai M, Tan Y, Zhao J. Image super-resolution via adaptive l_p ($0 < p < 1$) regularization and sparse representation. IEEE Trans Neural Networks Learn Syst 2016;27(7):1550–61.
[20] Yu J, Gao X, Tao D, Li X, Zhang K. A unified learning framework for single image super-resolution. IEEE Trans Neural Networks Learn Syst 2013;25(4):780–92.
[21] Yang J, Lin Z, Cohen S. Fast image super-resolution based on in-place example regression. In: Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition; 2013 Jun 23–28; Portland, OR, USA. New York: IEEE; 2013. p. 1059–66.
[22] Freeman WT, Jones TR, Pasztor EC. Example-based super-resolution. IEEE Comput Graphics Appl 2002;22(2):56–65.
[23] Kim J, Lee JK, Lee KM. Accurate image super-resolution using very deep convolutional networks. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. New York: IEEE; 2016. p. 1646–54.
[24] Tai Y, Yang J, Liu X. Image super-resolution via deep recursive residual network. In: Proceedings of 2017 IEEE conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. New York: IEEE; 2017. p. 3147–55.
[25] Kim J, Lee JK, Lee MK. Deeply-recursive convolutional network for image super-resolution. In: Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. New York: IEEE; 2016. p. 1637–45.
[26] Tong T, Li G, Liu X, Gao Q. Image super-resolution using dense skip connections. In: Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. New York: IEEE; 2017. p. 4799–807.
[27] Johnson J, Alahi A, Li FF. Perceptual losses for real-time style transfer and super-resolution. In: Proceedings of European Conference on Computer Vision (ECCV); 2016 Oct 11–14; Amsterdam, The Netherlands. Springer; 2016. p. 694–711.
[28] Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, et al. Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA. New York: IEEE; 2017. p. 4681–90.

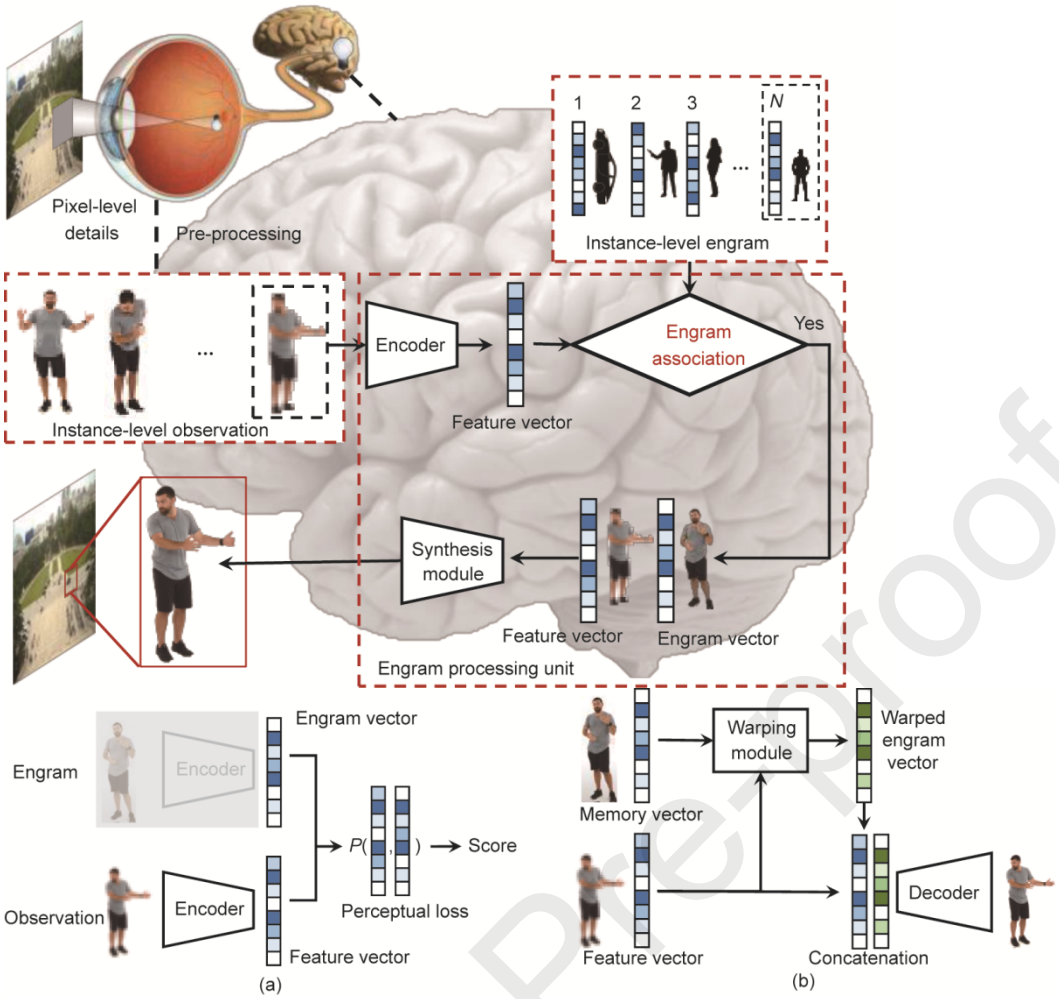
[29] Boominathan V, Mitra K, Veeraraghavan A. Improving resolution and depth-of-field of light field cameras using a hybrid imaging system. In: Proceedings of 2014 IEEE International Conference on Computational Photography (ICCP); 2014 May 2–4; Santa Clara, CA, USA. New York: IEEE; 2014. p. 1–10.
[30] Wu J, Wang H, Wang X, Zhang Y. A novel light field super-resolution framework based on hybrid imaging system. In: Proceedings of 2015 Visual Communications and Image Processing (VCIP); 2015 Dec 13–16; Singapore. New York: IEEE; 2015. p. 1–4.
[31] Wang Y, Liu Y, Heidrich W, Dai Q. The light field attachment: Turning a DSLR into a light field camera using a low budget camera ring. IEEE Trans Visualization Comput Graphics 2016;23(10):2357–64.
[32] Zhang Z, Wang Z, Lin Z, Qi H. Image super-resolution by neural texture transfer. In: Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. New York: IEEE; 2019. p. 7982–91.
[33] Tan Y, Zheng H, Zhu Y, Yuan X, Lin X, Brady D, Fang L. CrossNet++: Cross-scale large-parallax warping for reference-based super-resolution. IEEE Trans Pattern Anal Mach Intell 2021;43(12):4291–305.
[34] Zheng H, Ji M, Wang H, Liu Y, Fang L. Crossnet: An end-to-end reference-based super resolution network using cross-scale warping. In: Proceedings of European Conference on Computer Vision (ECCV); 2018 Sep 8–14; Munich, Germany. New York: IEEE; 2018. p. 88–104.
[35] Kopf J, Uyttendaele M, Deussen O, Cohen MF. Capturing and viewing gigapixel images. In: Proceedings of Special Interest Group on Computer Graphics and Interactive Techniques Conference; 2007 Aug 5–9; San Diego, CA, USA. New York: ACM; 2007. p. 93–es.
[36] Brady DJ, Hagen N. Multiscale lens design. Opt Express 2009;17(13):10659–74.
[37] Marks DL, Brady DJ. Gigagon: a monocentric lens design imaging 40 gigapixels. In: Proceedings of Imaging Systems 2010; 2010 Jun 7–8; Tucson, AZ, USA. OSA; 2010. p. ITuC2.
[38] Cossairt OS, Miau D, Nayar SK. Gigapixel computational imaging. In: Proceedings of 2011 IEEE International Conference on Computational Photography (ICCP); 2011 Apr 8–10; Pittsburgh, PA, USA. New York: IEEE; 2011. p. 1–8.
[39] Fan J, Suo J, Wu J, Xie H, Shen Y, Chen F, et al. Video-rate imaging of biological dynamics at centimetre scale and micrometre resolution. Nat Photonics 2019;13(11):809–16.
[40] Yuan X, Fang L, Dai Q, Brady DJ, Liu Y. Multiscale gigapixel video: A cross resolution image matching and warping approach. In: Proceedings of 2017 IEEE International Conference on Computational Photography (ICCP); 2017 May 12–14; Stanford, CA, USA. New York: IEEE; 2017. p. 1–9.
[41] Vaseghi SV. Advanced digital signal processing and noise reduction. 3rd ed. West Sussex: John Wiley & Sons; 2006.
[42] Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 2017;39(6):1137–49.
[43] He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV); 2017 Oct 22–29; Venice, Italy. New York: IEEE; 2017. p. 2961–9.
[44] Clark RN. Visual astronomy of the deep sky. Cambridge: Cambridge University Press; 1990.
[45] Curcio CA, Sloan KR, Kalina RE, Hendrickson AE. Human photoreceptor topography. J Comp Neurol 1990;292(4):497–523.
[46] Wauthier FL, Jordan MI, Jojic N. Efficient ranking from pairwise comparisons. In: Proceedings of 30th International Conference on Machine Learning; 2013 Jun 16–21; Atlanta, GA, USA. ACM; 2013. p. 109–17.
[47] Dosovitskiy A, Fischer P, Ilg E, Häusser P, Hazirbas C, Golkov V,et al. FlowNet: Learning optical flow with convolutional networks. In: Proceedings of 2015 IEEE International Conference on Computer Vision; 2015 Dec 7–13; Santiago, Chile. New York: IEEE; 2015. p. 2758–66.
[48] Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A, Brox T. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA. New York: IEEE; 2017. p. 2462–70.
[49] Bruhn A, Weickert J, Schnörr C. Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods. Int J Comput Vision 2005;61(3):211–31.
[50] Wang X, Zhang X, Zhu Y, Guo Y, Yuan X, Xiang L, et al. PANDA: a gigapixel-level human-centric video dataset. In: Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. New York: IEEE; 2020. p. 3268–78.
[51] Lim B, Son S, Kim H, Nah S, Lee KM. Enhanced deep residual networks for single image super-resolution. In: Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); 2017 Jul 21–26; Honolulu, HI, USA. New York: IEEE; 2017. p. 136–44.
[52] Kolchinsky A, Tracey BD. Estimating mixture entropy with pairwise distances. Entropy 2017;19(7):361.
[53] Chen Y, Liu S, Wang X. Learning continuous image representation with local implicit image function. In: Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021 Jun 20–25; Nashville, TN, USA. New York: IEEE; 2021. p. 8628–38.

Declaration of interests

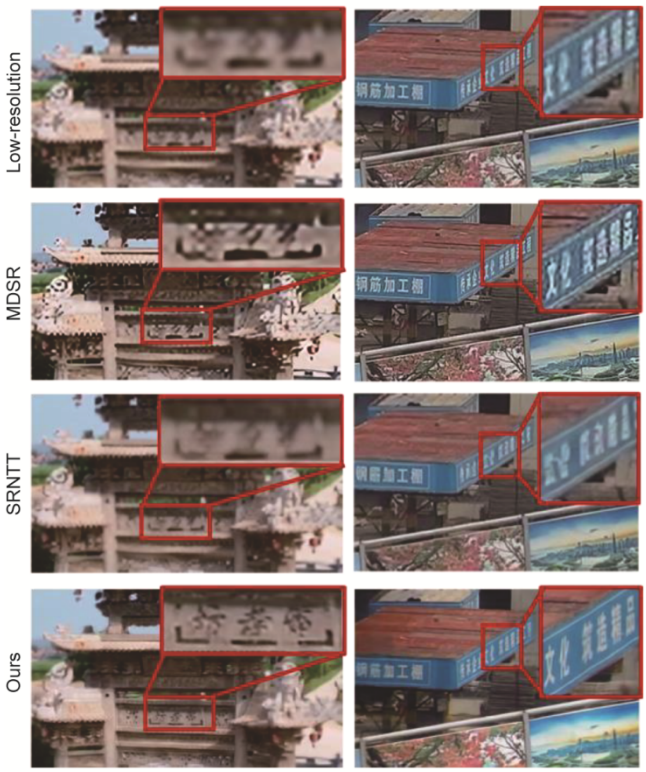
☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

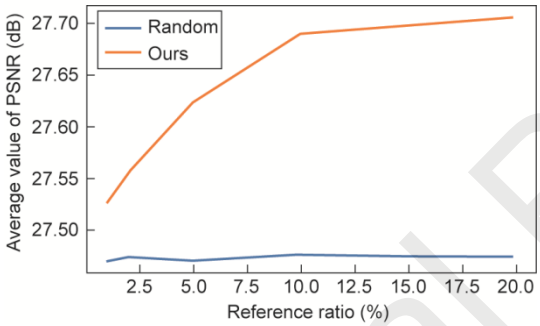
☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:






Panorama
2
1
Engram-driven
Conventional
1
...
...
Engram-driven
Conventional
2
...
...
Panorama
Conventional Engram-driven





Low-resolution observation PSNR: 25.74

Ours PSNR: 27.63

w/o EPU PSNR: 26.07

w/o instance-level engram PSNR: 26.10

w/o instance-level observation PSNR: 25.76