Research Statement

HE Shengfeng
School of Computing and Information Systems, Singapore Management University
Tel: (65) 6826-4973; Email: shengfenghe@smu.edu.sg
11 (Day) 12 (Month) 2024 (Year)

1. Background

Humans are avid consumers of visual content, engaging daily with videos, digital games, and shared photos on social media. However, there is a clear asymmetry—while nearly everyone can consume visual data, only a select few possess the talent to express themselves effectively through visual mediums. For the rest, attempts at creating or manipulating realistic visual content often fall short, deviating from the natural manifold of images. My research seeks to bridge this gap by exploring human-centric visual properties and interpreting generative models to create and manipulate images while preserving visual realism. Specifically, my work focuses on three key areas:

(1) Developing new methods for machine understanding of multimedia content, structure, semantics, and associated values.
(2) Designing generative models that enable easier creation of visual content and synthetic training data, producing photorealistic outputs (e.g., images, videos, 3D data, multimodal data) for downstream applications.

In the following sections, I outline my contributions to these research areas and conclude with future research directions.

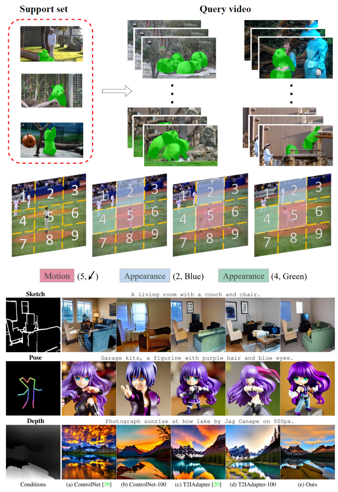
2. Content Understanding

Creating a visual world first requires a deep understanding of it. My group develops algorithms to interpret scenes in images and videos, which is fundamental for building socially aware agents, semantic image and video retrieval, captioning, and question-answering systems. Our work spans three major areas: (1) exploring human visual saliency, (2) understanding objects and regions in both image and video domains, and (3) learning from imperfect data.

2.1. **Visual Saliency.** My research investigates how humans perceive important objects or areas within scenes, simulating visual saliency. My early work introduced an alternative flash/no-flash stimulus that better represented human visual attention [8]. I then contributed to one of the first methods leveraging deep convolutional features for effective saliency detection [10]. To meet the demands of practical applications, I developed a highly efficient saliency detection method capable of running at 30 FPS on a CPU [32]. Recognizing that humans perceive depth and temporal information, my group explored saliency across modalities such as RGBD data [27, 26] and video saliency [29, 34]. We applied our video saliency methods to intelligent bullet chatting in partnership with Tencent and Huya [29], and explored various other applications of saliency, including top-down saliency [11], salient object subitizing [7], and saliency in visual question answering [6].

2.2. **Object and Scene Analysis.** Locating and segmenting objects and regions is a central theme in computer vision. My journey in this field began with the design of an object tracker [12], later enhanced with a gating mechanism [22]. I also developed the first orientation-aware, class-agnostic object detector (object proposal) [9], which was extended to stereo and temporal domains [13, 14]. My research includes the development of application-specific algorithms for high-precision segmentation in various contexts, such as ultra-high-resolution images [18], bird-view projection [47], glass surfaces [25], and curvilinear structures [35]. I also contributed to temporal segmentation, proposing an efficient O(n) supervoxel method that outperforms existing methods by 11x [33], and integrated spatial and temporal information for video object segmentation [30].

2.3. **Learning from Imperfect Data.** Humans excel at learning and adapting from limited or imperfect data, a capacity that remains a challenge for machine learning. I have addressed this by studying how to improve vision algorithms across domains [24, 28], from zero or few-shot samples [2, 42], and through self-supervised learning [37, 36]. My work includes network distillation methods for extracting knowledge from cross-domain models [28], few-shot learning for video object segmentation [2], and self-supervised representation learning from spatio-temporal statistics [37, 36]. Recent efforts focus on few-shot learning for complex scene object counting [52] and addressing domain gaps [44]. Our work also extends to conditional diffusion learning under limited supervision [49].

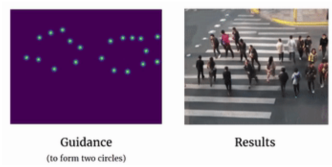


Few-Shot Learning [2, 36, 49]

3. Content Creation

Advances in machine modeling of the visual world offer exciting opportunities for creating, enhancing, and interacting with visual media. My research has made significant contributions to image synthesis and editing by integrating learning, vision, and graphics. Specifically, I have developed learning-based algorithms for (1) generating data from limited observations, (2) manipulating images to achieve artistic transformations, and (3) recovering corrupted or missing visual information.

3.1. **Visual Synthesis.** To address the data-hungry demands of deep learning, my group developed various synthesis methods for generating plausible data. For instance, I explored image reflection simulation beyond previous linear constraints [38], and developed algorithms to synthesize multi-view faces [43, 40]. In the area of crowd analysis, we created the first interactive crowd video synthesis method, CrowdGAN, which generates crowd behaviors with minimal user input [1]. Additionally, I pioneered an adversarial learning method that discovers multi-class attributes beyond binary ones, applied to both face and cartoon attributes [46]. Recently, we introduced a diffusion framework for synthesizing wide-angle 3D photographs [16].

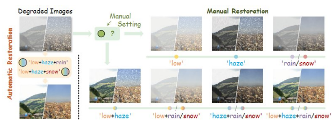


Interactive Crowd Synthesis [1]

3.2. **Visual Manipulation.** Transforming visual content into artistic styles has a wide range of applications in social media and entertainment. I developed the first deep learning-based pixelization method for transferring clipart into pixel art [5] and a cartoonization method that learns from line tracing data [19]. To reduce the time spent on makeup transfer in live streaming, my group created a robust, spatially-invariant makeup transfer algorithm [3]. We also collaborated with Tencent to design a video snapshot method that restores a short video from a single image [53], reducing storage costs for video previews. My work on GAN inversion [41] and extreme upscaling methods [48] highlights the broader applications of pre-trained GANs for high-resolution face synthesis and other use cases. Recently, we developed interactive tools for point-based image editing [20] and synchronized 3D dance animations with music beats [15].

Pixelization [5]

Interactive Editing [20]

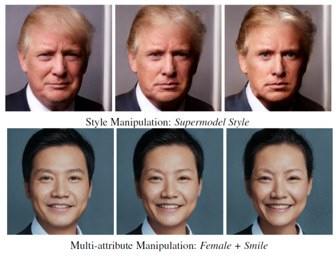


Universal Restoration [4]

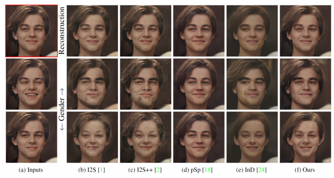
3.3. **Visual Restoration.** Restoring images from corrupted or missing information is a longstanding challenge in computer vision. My work addresses this through the design of learning structures and image priors [17, 31], as well as methods for vehicle recovery and downscaling [21, 45]. Recent efforts focus on integrating multiple types of image degradation into a single restoration framework [4] and restoring 3D content from 2D formats for efficient storage and processing [23].

4. **Interpretable Generative Models**

Deep neural network models are often criticized for being "black boxes" due to their millions of unexplained parameters, which hinder interpretability. This lack of transparency is particularly pronounced in generative models, where training demands massive datasets and substantial computational resources, limiting their broader applicability. My research addresses this gap by interpreting the latent semantics of generative models, enabling the reuse of pre-trained large-scale models for diverse applications. Specifically, I focus on three core directions: 1) discovering interpretable latent directions, 2) inverting real images into latent codes, and 3) reusing and extending generative priors for novel tasks.



Editing with Interpretable Directions [46]



Video Inversion [41]

Extreme Rescaling with generative priors [51]

4.1. **Interpretable Generative Directions.** Despite being trained to generate images from noise, the latent spaces of well-trained generative models exhibit semantically structured organization. My research aims to uncover these meaningful latent directions within pre-trained models. Unlike earlier approaches limited to binary attributes derived from paired data, I developed an adversarial learning method capable of discovering a broader spectrum of attributes, including style variations [46]. Building on this, I extended the exploration to interpretable subspaces, enabling the creation of 3D-aware and animatable artforms [50]. These discovered directions have proven effective not only in manipulating face attributes but also in enhancing the expressiveness of cartoon-style attributes.

4.2. **Generative Model Inversion.** To unlock the powerful editing capabilities of generative models for real images, it is essential to invert a real image into a corresponding latent code. However, achieving faithful reconstructions while preserving editability poses significant challenges. Observing that the continuity of consecutive frames in videos can guide better constraints, I introduced the first video-based inversion method, which ensures both reconstruction fidelity and editability in GANs [41].

In another line of work, I addressed the generative inversion problem as an extreme super-resolution task, mapping low-resolution inputs to high-resolution latent codes. By leveraging StyleGAN's ability to synthesize high-resolution random faces, I developed a progressive upscaling method that achieves up to 64x resolution enhancement [48]. This method optimizes the latent code to produce high-resolution outputs that closely match the originals, demonstrating the potential of generative inversion for real-world applications.

4.3. **Generative Priors.** Beyond editing pre-trained generative models, their latent spaces hold untapped potential for a variety of applications. To push the boundaries of reconstruction, I explored generative priors in an invertible rescaling framework, achieving superior reconstruction fidelity [51]. Additionally, I disentangled these priors into identity and style representations, significantly advancing applications in anime and cartoon generation [39]. These findings not only improve the flexibility and utility of generative models but also provide insights into their encoding structures, paving the way for novel applications leveraging generative priors.

5. **Ongoing and Future Directions**

In summary, my research goal is to develop algorithms that can both understand and recreate the visual world. To date, I have addressed key challenges by: (1) exploiting internal data structures and latent associations, (2) leveraging unlabeled or imperfect visual data, and (3) generating data for task-specific augmentation. Looking ahead, I am excited to explore the following research questions:

(1) **How can we synthesize the visual world using multiple modalities?** Humans, especially babies and toddlers, learn to perceive the world not by memorizing millions of labeled images but through interaction with their environment, engaging multiple sensory modalities—vision, sound, touch, smell, and language. While my research in imperfect supervision (such as few-shot, zero-shot, and self-supervised learning) has primarily focused on image data, I aim to

extend this by developing algorithms that leverage rich, unlabeled multimodal signals. The goal is to create robust generative models capable of synthesizing the visual world by integrating diverse sensory data.

(2) **How can we recreate the visual world with artistic styles?** My long-term vision is to enable the manipulation of visual data in various artistic styles. However, current methods lack precise, robust stylistic representations, limiting their practical use. I plan to address these issues by simulating the creative processes of professional artists and embedding professional experiences into the workflow. This approach will bridge the gap between theoretical models and real-world artistic practices, enabling the seamless integration of style transfer and artistic manipulation into industrial workflows.

(3) **How can we embed generative models in robots for open-environment tasks?** A key area of interest for me is the intersection of generative models and robotics. Specifically, I aim to design a drawing robot capable of creating visual art on any surface in open environments. This robot would integrate generative models with real-time sensory input, allowing it to adapt its artistic output to different surfaces and contexts. Such a system would require robust algorithms for real-world visual recognition, spatial understanding, and surface adaptability, paving the way for practical applications in interactive art, education, and industrial design.

Building on my prior contributions, I am excited to continue exploring these research directions in collaboration with my students, colleagues, and partners. In addition to addressing the core questions outlined above, I look forward to cross-disciplinary collaborations that can lead to innovative and practical solutions to pressing research challenges.

References

[1] Liangyu Chai, Yongtuo Liu, Wenxi Liu, Guoqiang Han, and Shengfeng He. Crowdgan: Identity-free interactive crowd video generation and beyond. *IEEE TPAMI*, 2020. 2

[2] Haoxin Chen, Hanjie Wu, Nanxuan Zhao, Sucheng Ren, and Shengfeng He. Delving deep into many-to-many attention for few-shot video object segmentation. In *CVPR*, pages 14040–14049, 2021. 2

[3] Han Deng, Chu Han, Hongmin Cai, Guoqiang Han, and Shengfeng He. Spatially-invariant style-codes controlled makeup transfer. In *CVPR*, pages 6549–6557, 2021. 2

[4] Yu Guo, Yuan Gao, Yuxu Lu, Huilin Zhu, Ryan Wen Liu, and Shengfeng He. Onerestore: A universal restoration framework for composite degradation. In *ECCV*, 2024. 2

[5] Chu Han, Qiang Wen, Shengfeng He, Qianshu Zhu, Yinjie Tan, Guoqiang Han, and Tien-Tsin Wong. Deep unsupervised pixelization. *ACM TOG*, 37(6):1–11, 2018. 2

[6] Shengfeng He, Chu Han, Guoqiang Han, and Jing Qin. Exploring duality in visual question-driven top-down saliency. *IEEE TNNLS*, 31(7):2672–2679, 2019. 1

[7] Shengfeng He, Jianbo Jiao, Xiaodan Zhang, Guoqiang Han, and Rynson WH Lau. Delving into salient object subitizing and detection. In *ICCV*, pages 1059–1067, 2017. 1

[8] Shengfeng He and Rynson WH Lau. Saliency detection with flash and no-flash image pairs. In *ECCV*, pages 110–124, 2014. 1

[9] Shengfeng He and Rynson WH Lau. Oriented object proposals. In *ICCV*, pages 280–288, 2015. 1

[10] Shengfeng He, Rynson WH Lau, Wenxi Liu, Zhe Huang, and Qingxiong Yang. Supercnn: A superpixelwise convolutional neural network for salient object detection. *IJCV*, 115(3):330–344, 2015. 1

[11] Shengfeng He, Rynson WH Lau, and Qingxiong Yang. Exemplar-driven top-down saliency detection via deep association. In *CVPR*, pages 5723–5732, 2016. 1

[12] Shengfeng He, Qingxiong Yang, Rynson WH Lau, Jiang Wang, and Ming-Hsuan Yang. Visual tracking via locality sensitive histograms. In *CVPR*, pages 2427–2434, 2013. 1

[13] Shao Huang, Weiqiang Wang, Shengfeng He, and Rynson WH Lau. Stereo object proposals. *IEEE TIP*, 26(2):671–683, 2016. 1

[14] Shao Huang, Weiqiang Wang, Shengfeng He, and Rynson WH Lau. Egocentric temporal action proposals. *IEEE TIP*, 27(2):764–777, 2017. 1

[15] Zikai Huang, Xuemiao Xu, Cheng Xu, Huaidong Zhang, Chenxi Zheng, Jing Qin, and Shengfeng He. Beat-it: Beat-synchronized multi-condition 3d dance generation. In *ECCV*, 2024. 2

[16] Yutao Jiang, Yang Zhou, Yuan Liang, Wenxi Liu, Jianbo Jiao, Yuhui Quan, and Shengfeng He. Diffuse3d: Wide-angle 3d photography via bilateral diffusion. In *ICCV*, 2023. 2

[17] Jianbo Jiao, Qingxiong Yang, Shengfeng He, Shuhang Gu, Lei Zhang, and Rynson WH Lau. Joint image denoising and disparity estimation via stereo structure pca and noise-tolerant cost. *IJCV*, 124(2):204–222, 2017. 2

[18] Qi Li, Weixiang Yang, Wenxi Liu, Yuanlong Yu, and Shengfeng He. From contexts to locality: Ultra-high resolution image segmentation via locality-aware contextual correlation. In *ICCV*, pages 7252–7261, 2021. 1

[19] Simin Li, Qiang Wen, Shuang Zhao, Zixun Sun, and Shengfeng He. Two-stage photograph cartoonization via line tracing. In *Computer Graphics Forum*, volume 39, pages 587–599, 2020. 2

[20] Haofeng Liu, Chenshu Xu, Yifei Yang, Lihua Zeng, and Shengfeng He. Drag your noise: Interactive point-based editing via diffusion semantic propagation. In *CVPR*, 2024. 2

[21] Junjie Liu, Shengfeng He, and Rynson WH Lau. $l\_\{0\}$-regularized image downscaling. *IEEE TIP*, 27(3):1076–1085, 2017. 2
[22] Wenxi Liu, Yibing Song, Dengsheng Chen, Shengfeng He, Yuanlong Yu, Tao Yan, Gehard P Hancke, and Rynson WH Lau. Deformable object tracking with gated fusion. *IEEE TIP*, 28(8):3766–3777, 2019. 1
[23] Yuqin Lu, Bailin Deng, Zhixuan Zhong, Tianle Zhang, Yuhui Quan, Hongmin Cai, and Shengfeng He. 3d snapshot: Invertible embedding of 3d neural representations in a single image. *IEEE TPAMI*, 2024. 2
[24] Jianming Lv, Kaijie Liu, and Shengfeng He. Differentiated learning for multi-modal domain adaptation. In *ACM MM*, pages 1322–1330, 2021. 2
[25] Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiang Zhang, Xiaopeng Wei, and Rynson WH Lau. Don't hit me! glass detection in real-world scenes. In *CVPR*, pages 3687–3696, 2020. 1
[26] Yuzhen Niu, Guanchao Long, Wenxi Liu, Wenzhong Guo, and Shengfeng He. Boundary-aware rgbd salient object detection with cross-modal feature sampling. *IEEE TIP*, 29:9496–9507, 2020. 1
[27] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. Rgbd salient object detection via deep fusion. *IEEE TIP*, 26(5):2274–2285, 2017. 1
[28] Sucheng Ren, Yong Du, Jianming Lv, Guoqiang Han, and Shengfeng He. Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In *CVPR*, pages 13325–13333, 2021. 2
[29] Sucheng Ren, Chu Han, Xin Yang, Guoqiang Han, and Shengfeng He. Tenet: Triple excitation network for video salient object detection. In *ECCV*, pages 212–228, 2020. 1
[30] Sucheng Ren, Wenxi Liu, Yongtuo Liu, Haoxin Chen, Guoqiang Han, and Shengfeng He. Reciprocal transformations for unsupervised video object segmentation. In *CVPR*, pages 15455–15464, 2021. 1
[31] Yibing Song, Jiawei Zhang, Lijun Gong, Shengfeng He, Linchao Bao, Jinshan Pan, Qingxiong Yang, and Ming-Hsuan Yang. Joint face hallucination and deblurring via structure generation and detail enhancement. *IJCV*, 127(6):785–800, 2019. 2
[32] Wei-Chih Tu, Shengfeng He, Qingxiong Yang, and Shao-Yi Chien. Real-time salient object detection with a minimum spanning tree. In *CVPR*, pages 2334–2342, 2016. 1
[33] Bo Wang, Yiliang Chen, Wenxi Liu, Jing Qin, Yong Du, Guoqiang Han, and Shengfeng He. Real-time hierarchical supervoxel segmentation via a minimum spanning tree. *IEEE TIP*, 29:9665–9677, 2020. 1
[34] Bo Wang, Wenxi Liu, Guoqiang Han, and Shengfeng He. Learning long-term structural dependencies for video salient object detection. *IEEE TIP*, 29:9017–9031, 2020. 1
[35] Feigege Wang, Yue Gu, Wenxi Liu, Yuanlong Yu, Shengfeng He, and Jia Pan. Context-aware spatio-recurrent curvilinear structure segmentation. In *CVPR*, pages 12648–12657, 2019. 1
[36] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Wei Liu, and Yun-Hui Liu. Self-supervised video representation learning by uncovering spatio-temporal statistics. *IEEE TPAMI*, 2021. 2
[37] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *CVPR*, pages 4006–4015, 2019. 2
[38] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *CVPR*, pages 3771–3779, 2019. 2
[39] Chenshu Xu, Yangyang Xu, Huaidong Zhang, Xuemiao Xu, and Shengfeng He. Dreamanime: Learning style-identity textual disentanglement for anime and beyond. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 3
[40] Xuemiao Xu, Keke Li, Cheng Xu, and Shengfeng He. Gdface: Gated deformation for multi-view face image synthesis. In *AAAI*, volume 34, pages 12532–12540, 2020. 2
[41] Yangyang Xu, Yong Du, Wenpeng Xiao, Xuemiao Xu, and Shengfeng He. From continuity to editability: Inverting gans with consecutive images. In *ICCV*, pages 13910–13918, 2021. 2, 3
[42] Yangyang Xu, Chu Han, Jing Qin, Xuemiao Xu, Guoqiang Han, and Shengfeng He. Transductive zero-shot action recognition via visually connected graph convolutional networks. *IEEE TNNLS*, 2020. 2
[43] Yangyang Xu, Xuemiao Xu, Jianbo Jiao, Keke Li, Cheng Xu, and Shengfeng He. Multi-view face synthesis via progressive face flow. *IEEE TIP*, 30:6024–6035, 2021. 2
[44] Yingjie Xu, Bangzhen Liu, Hao Tang, Bailin Deng, and Shengfeng He. Learning with unreliability: Fast few-shot voxel radiance fields with relative geometric consistency. In *CVPR*, 2024. 2
[45] Xiaosheng Yan, Feigege Wang, Wenxi Liu, Yuanlong Yu, Shengfeng He, and Jia Pan. Visualizing the invisible: Occluded vehicle segmentation and recovery. In *ICCV*, pages 7618–7627, 2019. 2
[46] Huiting Yang, Liangyu Chai, Qiang Wen, Shuang Zhao, Zixun Sun, and Shengfeng He. Discovering interpretable latent space directions of gans beyond binary attributes. In *CVPR*, pages 12177–12185, 2021. 2, 3
[47] Weixiang Yang, Qi Li, Wenxi Liu, Yuanlong Yu, Yuexin Ma, Shengfeng He, and Jia Pan. Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation. In *CVPR*, pages 15536–15545, 2021. 1
[48] Zhou Yang, Yangyang Xu, Yong Du, Qiang Wen, , and Shengfeng He. Pro-pulse: Learning progressive encoders of latent semantics in gans for photo upsampling. *IEEE TIP*, 2022. 2, 3
[49] Yuyang Yu, Bangzhen Liu, Chenxi Zheng, Xuemiao Xu, Huaidong Zhang, and Shengfeng He. Beyond textual constraints: Learning novel diffusion conditions with fewer examples. In *CVPR*, 2024. 2
[50] Chenxi Zheng, Bangzhen Liu, Xuemiao Xu, Huaidong Zhang, and Shengfeng He. Learning an interpretable stylized subspace for 3d-aware animatable artforms. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 3
[51] Zhixuan Zhong, Liangyu Chai, Yang Zhou, Bailin Deng, Jia Pan, and Shengfeng He. Faithful extreme rescaling via generative prior reciprocated invertible representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5708–5717, 2022. 3
[52] Huilin Zhu, Jingling Yuan, Zhengwei Yang, Yu Guo, Xian Zhong, Zheng Wang, and Shengfeng He. Zero-shot object counting with good exemplars. In *ECCV*, 2024. 2
[53] Qianshu Zhu, Chu Han, Guoqiang Han, Tien-Tsin Wong, and Shengfeng He. Video snapshot: Single image motion expansion via invertible motion embedding. *IEEE TPAMI*, 2021. 2