Research Statement

HE Shengfeng
School of Computing and Information Systems, Singapore Management University
Tel: (65) 6826-4973; Email: shengfenghe@smu.edu.sg
25 (Day) 12 (Month) 2023 (Year)

1. **Background**

Humans are avid consumers of visual content. Every day, people watch videos, play digital games and share photos on social media. However, there is an asymmetry – while everybody is able to consume visual data, only a chosen few are talented enough to effectively express themselves visually. For the rest of us, most attempts at creating or manipulating realistic visual content end up quickly "falling off" the manifold of natural images. My research goal is to understand human-centric visual properties and interpret generative models, which can drive the approaches for preserving visual realism while creating and manipulating photographs. Specifically, I focus on three research directions:

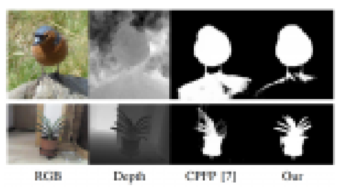
(1) I investigate new methods to enable machine understanding of multimedia content, structure, semantics, and the associate values.
(2) I design new generative models to help humans create visual content and synthetic training data more easily. Our models can synthesize photorealistic outputs (e.g., images, videos, 3D data, multimodal data) for downstream applications.
(3) I develop approaches for opening up the "black box" of generative models and interpret their latent semantics. Once the latent space is revealed, these pre-trained models can be re-used for synthesis, editing, and image-to-image translation tasks.

In the following, I will highlight my research contributions in these three themes. I will conclude with future research agenda.

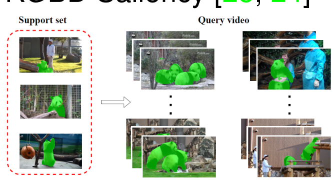
2. **Content Understanding**

To properly create a visual world, we need to understand it in advance. My group creates algorithms to understand the scenes in images and videos. Understanding the scenes and people's activity are fundamental steps toward building socially-aware agents, semantic image/video retrieval, captioning, and question-answering. Toward this goal, I focused on (1) exploring human visual saliency in the scene; (2) understanding object/regions in both image and video domains; and (3) learning from imperfect data.

2.1. **Visual Saliency.** My research focuses on studying how human perceives important objects/areas in the scene, i.e., simulating visual saliency in the scene. My early works explored an alternative flash/no-flash stimulus that better formulated the human visual attention [8]. Later I contributed one of the first methods to leverage deep convolutional features for saliency detection effectively [10]. To satisfy the need for practical applications, I also developed an extremely efficient saliency detection method that can run at 30 FPS on a CPU [31]. Humans can perceive depth and temporal information, and therefore my group studied saliency stimuli in different modalities, such as RGBD data [25, 24], video saliency [28, 33]. In particular, I cooperated with Tencent and Huya to apply our video saliency method [28] on intelligent bullet chatting. I also explored saliency in many other tasks, like top-down saliency [11], salient object subitizing [7], saliency in visual question answering [6], in which they explored the mechanism and effectiveness of saliency under various settings.



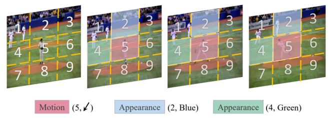
RGBD Saliency [25, 24]



Few-Shot Video Object Segmentation [2]

2.2. **Object/Scene Analysis.** Locating and segmenting objects/regions are one of the main themes of computer vision. I started my journey of computer vision by designing an object tracker [12] and it is enhanced with a gating mechanism [20]. Meanwhile, I developed the first orientation-aware class-agnostic object detector (i.e, object proposal) [9], and it was then extended to stereo and temporal domains [13, 14]. Besides detection, I also interested in designing application-specific algorithms to accurately segment object/regions, e.g., ultra high-resolution image [17], bird-view projection [46], glass segmentation [23], interactive matting [47], crowd scene [21], and curvilinear structure [34]. All these methods considered the professional domain knowledge of the applications, rather than designing an application-irrelevant deep network architecture. In the temporal domain, I proposed an efficient O(n) supervoxel method that faster than the existing one by 11x [32]. Meanwhile, my group designed a reciprocal method that integrated spatial and temporal information for video object segmentation methods [29], and studied the repetitive temporal patterns for action counting [49].

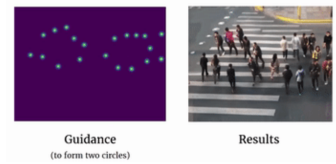
2.3. **Learning from Imperfect Data.** Humans are remarkable at learning and adapting to new tasks from very few examples (few-shot learning) and even samples not from the same domain (sketches). From the machine perspective, domain gap and data imperfection are the main barriers to computer vision algorithms becoming practical. I studied different vision problems learning from the data of different domains [22, 27, 39], from zero or a few samples [2, 42], and in self-supervised manners [36, 35]. In particular, I developed a new network distillation method that can extract "master" knowledge from a better but cross-domain model [27]; in [2], my group delved into the first many-to-many attention for few-shot video object segmentation; two consecutive works [36, 35] explored the learning of self-supervised representation from video using spatio-temporal statistics.



Self-supervised Learning [35]

3. Content Creation

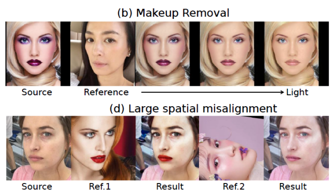
Machines' ability to model and recreate our visual world paves a promising path towards a deeper understanding of visual data. It also opens up fascinating opportunities for everyone to enhance, visualize, and interact with visual media. My research has substantially contributed to image synthesis and editing by drawing and integrating ideas from learning, vision, and graphics. Specifically, I developed learning-based algorithms for (1) creating additional data based on limited observations; (2) manipulating images to become another art form; and (3) recovering the corrupted/missing information of the image.



Interactive Crowd Synthesis [1]

3.1. **Image/Video Synthesis.** To satisfy the data-hungry nature of deep learning, my group developed different synthesis methods for creating plausible data. For example, image reflection lacks paired data for training, I studied the simulation of image reflection beyond the previous linear constraint [37]. In order to reveal the identity of a face from an arbitrary angle, I developed algorithms to synthesize multi-view faces [43, 40]. Crowd analysis relies heavily on diverse data. I proposed the first interactive crowd video synthesis method that can generate crowd behaviors with minimum user efforts [1], which can be beneficial for crowd counting, anomaly detection, and crowd video prediction.


Pixelization [5]



Makeup Transfer [3]

3.2. **Image/Video Manipulation.** Manipulating visual content to be another art form is entertaining and with a demand for social media. Motivated by the popularity of pixel art games, I developed the first deep learning-based pixelization method that can automatically transfer a clipart into a pixel art [5]. To mimic the cartoon styles from artists, I proposed a cartoonization that learns from line tracing data [18]. Changing makeup can be time-consuming, and my group developed the first spatially-invariant makeup transfer that is robust in live-streaming [3]. Existing video sharing services require storing a thumbnail and a short video for video preview, I cooperated with Tencent to design a video snapshot that can restore a short video based on a single image [50], which largely reduced the storage burden of video preview.

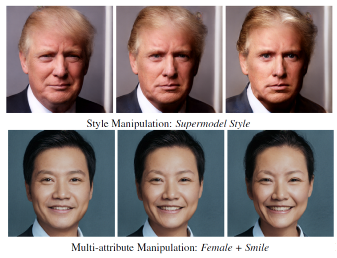
3.3. **Image/Video Restoration.** Recovering images from corrupted and missing information is a long-standing problem. For the tasks that deal with corrupted information, like image denoising, deblurring, and shadow removal, my studies focus on designing extra image priors [16, 30] or forming delicate learning structures [4, 15, 26]. For those required to recover the missing components, I developed a vehicle recovery method to visualize the invisible part for downstream segmentation [44], an L_0 regularized downscaling approach to maintain salient low-resolution features [19], and an example-based colorization method that can produce vivid colors for grayscale images [38].

4. **Interpretable Generative Models**

Deep neural network models are often criticized as being black boxes that lack interpretability, because of their millions of unexplained model parameters. Specifically, the training of generative models requires a massive amount of data and computational efforts, which limits the usage of complex models for wider AI applications. My recent research lies in interpreting the latent semantics of generative models, such that the pre-trained large-scale models can be easily re-used for other purposes. To understand the latent space of a GAN, I focus on three directions: 1) discovering the interpretable latent directions; 2) inverting a real image to the latent code; 3) re-using and extending a pre-trained GAN in other tasks.

Face Deblurring [30]

4.1. **Interpretable Generative Directions.** Although GAN is trained from noise to image, a properly trained GAN latent space shows semantically structured organization. As a result, I aim at finding those meaningful directions of a pre-trained GAN. However, previous methods were limited to discovering binary classes based on paired data. I developed an adversarial learning method that can discover more attributes beyond binary attributes like style [45]. I demonstrated the effectiveness of our discovered directions not only on face attributes but also on cartoon attributes.



Editing with Interpretable GAN Directions [45]

4.2. **GAN Inversion.** To enable the powerful editing ability of GAN to real images, we need to first convert the real image to a latent code. However, converting it to a latent code is not trivial to retain a faithful reconstruction. Based on the observation that the continuity brought by consecutive images can be used as an indicator to constrain the editability. I developed the first video-based GAN inversion method that maintained both reconstruction fidelity and editability of GAN [41].

Video GAN Inversion [41]

4.3. **Reusing a Generative Model.** Except editing the output of a pre-trained GAN, it contains the hidden potential for other purposes. Given a StyleGAN that can synthesize high-resolution random faces, I developed an extreme upscaling method (up to 64x) [48]. Specifically, it maps a low-resolution input to a latent code, which is optimized to produce as close as possible to the original high-resolution one in a progressive manner. It outperforms state-of-the-art upscaling methods by a large margin. Meanwhile, it sheds light on the encoding structure for other applications that re-use a pre-trained GAN.

Extreme Upscaling with a Pre-trained GAN [48]

5. **Ongoing and Future Directions**

To summarize, my research goal is to develop algorithms that can understand and recreate the visual world. My research to date tackled the significant challenges via 1) exploiting internal data/latent structures and associations; 2) leveraging unlabeled or imperfect visual data; 3) creating data for task-specific augmentation. Moving forward, I am excited to explore the following research questions:

(1) **How can we synthesize the visual world using multiple modalities?** Babies/toddlers learn to perceive the world, not by memorizing millions of labeled training images. Instead, they learn by interacting with physical objects and exploring the world through multiple modalities (e.g., sound, touch, smell, taste, language, and vision). My research on learning with imperfect supervision (few-shot learning, zero-shot learning, and self-supervised learning) primarily focuses only on image data. Next, I would like to develop algorithms that can capitalize on these unlabeled but rich multi-modality signals to create robust generative models.
(2) **How can we recreate the visual world with style?** My dream is to manipulate visual data in different artistic styles. However, existing methods lack precise and robust stylish representations, making these methods not feasible in practice. I have begun to explore these problems from a professional artist perspective, and simulate the professional process for different art forms. By injecting professional experiences in the loop, the workflow can be easily integrated into the industrial process and fit practical necessities.
(3) **How can we extend the GAN interpretability to unseen tasks and environments?** One of the major problems in interpretable GAN is that a trained model's performance is often significantly degraded in new visual domains. As our world is continuously changing, the existing static training and testing paradigm in GAN inevitably does not lead to a promising path toward generalization. In the past, I have addressed the problem via unsupervised domain adaptation or self-supervised learning. However, such settings cannot be directly adopted when exploring the latent space of a GAN. In the future, I would like to formulate GAN interpretability as a class-agnostic plugin, or a continually revising process.

With my prior research contributions to the relevant problems, I am very excited to carry on my research trajectories with my students, colleagues, and collaborators. Apart from tackling the core research questions, I will also continue collaborating with researchers to tackle challenging cross-disciplinary tasks. Bringing together expertise across diverse fields will lead to out-of-box and practical solutions to impactful research questions.

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. 1, 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] Yong Du, Guoqiang Han, Yinjie Tan, Chufeng Xiao, and Shengfeng He. Blind image denoising via dynamic dual learning. *IEEE TMM*, 2020. 3

[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. 2

[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. 2

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

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

[15] Jianbo Jiao, Wei-Chih Tu, Ding Liu, Shengfeng He, Rynson WH Lau, and Thomas S Huang. Formnet: Formatted learning for image restoration. *IEEE TIP*, 29:6302–6314, 2020. 3

[16] 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. 3

[17] 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. 2

[18] 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
[19] Junjie Liu, Shengfeng He, and Rynson WH Lau. $l\_\{0\}$-regularized image downscaling. *IEEE TIP*, 27(3):1076–1085, 2017. 3
[20] 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. 2
[21] Yongtuo Liu, Qiang Wen, Haoxin Chen, Wenxi Liu, Jing Qin, Guoqiang Han, and Shengfeng He. Crowd counting via cross-stage refinement networks. *IEEE TIP*, 29:6800–6812, 2020. 2
[22] Jianming Lv, Kaijie Liu, and Shengfeng He. Differentiated learning for multi-modal domain adaptation. In *ACM MM*, pages 1322–1330, 2021. 2
[23] 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. 2
[24] 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
[25] 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
[26] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson WH Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *CVPR*, pages 4067–4075, 2017. 3
[27] 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
[28] 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
[29] 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. 2
[30] 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. 3
[31] 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
[32] 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. 2
[33] 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
[34] 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. 2
[35] 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
[36] 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
[37] 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
[38] Chufeng Xiao, Chu Han, Zhuming Zhang, Jing Qin, Tien-Tsin Wong, Guoqiang Han, and Shengfeng He. Example-based colourization via dense encoding pyramids. In *Computer Graphics Forum*, volume 39, pages 20–33, 2020. 3
[39] Xuemiao Xu, Hai He, Huaidong Zhang, Yangyang Xu, and Shengfeng He. Unsupervised domain adaptation via importance sampling. *IEEE TCSVT*, 30(12):4688–4699, 2019. 2
[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. 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] 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. 3
[45] 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. 3
[46] 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. 2
[47] Xin Yang, Ke Xu, Shaozhe Chen, Shengfeng He, Baocai Yin Yin, and Rynson Lau. Active matting. *NeurIPS*, 31:4590–4600, 2018. 2
[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. 3
[49] Huaidong Zhang, Xuemiao Xu, Guoqiang Han, and Shengfeng He. Context-aware and scale-insensitive temporal repetition counting. In *CVPR*, pages 670–678, 2020. 2
[50] 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. 3