Bridging Physical World and Virtual Applications by Recreating, Processing, and Manipulating

Yifan Jiang, University of Texas at Austin

The central goal of my research is to build intelligent machines that can recreate our visual world and manipulate it further in virtual apps. I strive to achieve this by building a bridge to connect the visual contents between the physical world and virtual scenes. My research helps people better restore our visual world in the digital format and render artistic effects more easily. Especially, I study three important questions:

- *How can we design algorithms that restore and infer the physical world from corrupted measurements (noisy inputs, sparse view, and limited lighting condition, etc.)?*
- *How to build intelligent machines for recreating, editing, and manipulating photo-realistic visual contents (appearances, 3D shape/geometry, and animation)?*
- *How to accelerate the deployment of these intelligent machines for products in real scenarios (smartphones, FPGAs, and other edge devices)?*

I consider the learning-based algorithm to be a powerful tool and a central hub, in addressing those challenges tremendously faster and better. In the following, I will highlight my research contributions in these three themes and conclude with a future research agenda.

1 Inferring Physical World from Corrupted Signals

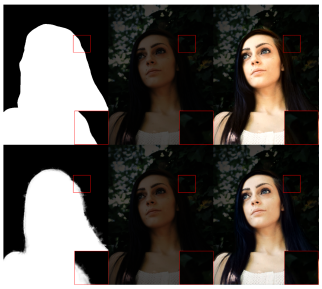
Physical world contains satisfactory visual contents and attracts our human beings to record it. However, it is only within the recent decade that digital photography has finally eclipsed film in sales, unit volume, and the number of pictures taken. Moreover, many of them are still hardly captured and saved in digital formats, due to the dark lighting condition, blurry artifacts, limited resolution, and so on. My research aims to leverage learning priors to restore the clean signal from these corrupted measurements.

Towards this goal, I have developed novel algorithms that (1) learn to enhance low-light images without paired supervision [1]; (2) learn to adopt user interaction for region-specified light enhancement [2]. (3) learn to efficiently recover clean images from noisy measurements using a faster dynamic filter [3]; (4) learn to jointly solve different distortions in a unified framework [4]; (5) learn to reconstruct a 3D world from a 2D image using neural radiance field [5]. The innovation from my research largely benefits from three key design philosophies: Data-efficiency (fewer annotations), Model-efficiency (one model to solve all tasks), and Computation-efficiency (faster inference time). Below I will outline each of my contributions.

Light Enhancement [1]

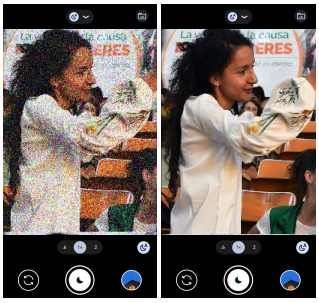
Unsupervised Low-light Enhancement Image captured in low-light conditions suffers from low contrast, poor visibility, and high ISO noise. Those issues challenge both human visual perception that prefers high visibility images, and numerous intelligent systems relying on computer vision algorithms, such as "Night Sight" mode in modern smartphone camera apps, all-day autonomous driving, and biometric recognition. The state-of-the-art approaches using deep learning techniques can properly handle these artifacts and help restore the clean signals using learning priors. However, they heavily rely on either synthesized or captured corrupted and clean image pairs to train. Instead, my TIP paper [1] developed an unsupervised method to enhance low-light image inputs. Our promising work (cited over **920** times) was implemented by popular third-party image manipulation software and open-source GNU Image Manipulation Program toolbox [GIMP-ML] and was also extensively adopted or modified by the related CV functionalities in mainstream Apps (TikTok and PicsArt). It was highlighted as one of the IEEE Signal Processing Society (SPS)'s top-25 most downloaded articles during 2021 - 2022.

Region-Controllable Light Enhancement While most existing enhancing algorithms are trained to enlighten a given image in a globally homogeneous way, they are neither capable of enhancing only local regions of interest ("where") nor producing outputs at a range of different illumination levels ("how much"). Those can significantly limit the prospect of flexible, customized, or even user-interactive enhancement. To address these gaps, my ACMMM'2022 work [2] allows users to directly specify "where" and "how much" they want to enhance from an input low-light image. Meanwhile, the proposed model will maintain an overall consistent visual appearance and plausible composition.



Controllable Enhancement [2]

Fast and Memory-friendly Kernels for Image Denoising Image denoising is fundamental to the study of computer vision. Recent advances in deep learning have sparked significant interest in learning end-to-end mappings directly from corrupted observations to the unobserved clean signal. These networks appear to learn a prior over the appearance of "ground truth" noiseless images in addition to the statistical properties of the noise present in the input images. However, these deep networks usually suffer from notoriously large computations. My ECCV'2022 paper [3] tackled this issue by proposing a fast and memory-friendly dynamic operator, which predicts spatially-varying kernels at low resolution and adopts a fast fused operator to jointly upsample and apply these kernels at full resolution.



Fast Image Denoising [3]

Joint Image Denoising and Enhancement Low-light images captured in the real world are inevitably corrupted by sensor noise. Such noise is spatially variant and highly dependent on the underlying pixel intensity, deviating from the oversimplified assumptions in conventional denoising. Existing light enhancement methods either overlook the important impact of real-world noise during enhancement, or treat noise removal as a separate pre- or post-processing step. My TIP work [4] seamlessly integrates light enhancement and noise suppression parts into a unified and physics-grounded optimization framework. We form these two parts into one principled plug-and-play optimization, and adopt the half quadratic splitting (HQS) algorithm to optimize it.

Joint Denoising & Light Enhancement [4]

Single Image Novel-view Synthesis The task of novel view synthesis has recently seen dramatic progress as a result of using neural radiance field (NeRF). However, training NeRF requires dense captured views and the corresponding camera poses. Several recent attempts to train a NeRF using sparse views, in contrast to them, my ECCV'2022 work [5] pushed the setting of sparse views to the extreme, by training a neural radiance field on only one single view.

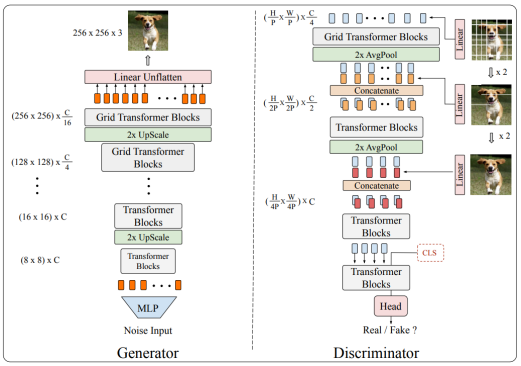
Novel-view Synthesis [5]

2 Creating, Editing, and Manipulating Visual Contents

While restoring the physical world helps us record memorial events, recreating photo-realistic visual contents can bring us more entertainment, from art to games. Instead of manually designing/creating visual objects, I build machine intelligence and learn to create various virtual contents and improve their visual quality. For example, (1) Recreating: I introduced advanced techniques (e.g., AutoML and Vision Transformer) into the architecture design of Generative Adversarial Networks (GANs). The resultant architectures, AutoGAN [6] and TransGAN [7], show strong abilities in producing visually pleasing images; (2) Processing: I developed a self-supervised algorithm (SSH [8]) that learns to adjust the appearance of foreground images according to the given backgrounds, which makes the compositing images in virtual applications more natural and realistic; (3) Manipulating:

I designed a unified framework (INS [9]) to stylize implicit representations, which enables direct editing on implicit representations and produces 3D stylization effects using the neural radiance field (NeRF). Details are shown in the following parts.

Recreating Visual Contents via Generative Models Many applications for recreating visual content are based on Generative Adversarial Networks (GANs), an active research area receiving explosive attention. Rather than just applying GANs out of the box, I have been actively innovating GAN model architectures, leading to several new GAN models now widely adopted by the community. For example, my ICCV'19 paper was the first to introduce AutoML to discover better GAN architectures and design principles. The resultant **AutoGAN** [6] architecture outperformed all existing handcrafted GAN models on the CIFAR-10 image generation benchmark by then, attracted significant attention from the research community as well as press coverage[1]. More recently, my NeurIPS'2021 paper accomplished a new piece of high-visibility work namely **TransGAN** [7]: it challenged the common wisdom that convolutions (with the strong inductive bias for natural images) are indispensable for high-quality image generation. Instead, we built the first strong GAN using only vanilla vision transformers aided with customized training recipes. It was later covered by **Quanta Magazine** [link], as one of the three well-known works picked to illustrate the vision transformer wave. The paper has attracted over **390** citations and the open-sourced GitHub repository gained more than **1500** stars in one year.

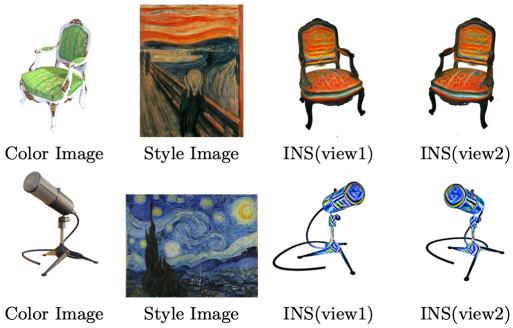


Generative Model [7]

Image Harmonization by Self-supervised Learning Image harmonization aims at adjusting (harmonizing) the appearance of a foreground object to better match the background image so that the resulting composite is more realistic. Existing methods train deep neural networks to address this problem, however, collecting high-quality paired harmonization data is time-consuming and laborious. For example, it requires an accurate mask of the foreground object in each image. My ICCV'2021 work [8] proposed the first self-supervised harmonization framework that needs neither human-annotated masks nor professionally created images for training.

Image Harmonization [8]

3D Stylization via Implicit Representation While implicit neural representation (INR) reveals multiple advantages compared to conventional discrete signals on 3D scene reconstruction, it is still unknown how we can edit/manipulate these continuous representations. My ECCV'2021 paper [9] proposed a unified implicit neural stylization framework. which can edit the appearance of both 2D (SIREN) and 3D (SDF/NeRF) continuous representations. I developed a novel self-distillation geometry consistency loss which preserves the geometry fidelity of the stylized scenes.



3D Stylization [9]

[1]For example "AutoML + GAN = AutoGAN! AI Can Now Design Better GAN Models Than Humans", Synced AI Technology & Industry Review (Aug. 2019). Also featured on Towards Data Science (Sep. 2019).

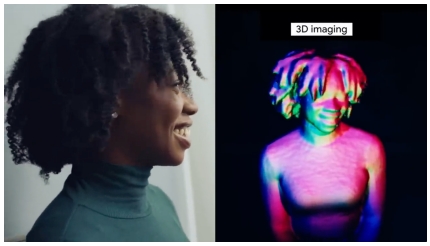
3 Efficient Deployment on Edge-devices

Recent breakthroughs in deep neural networks (NNs) have fueled a growing demand for intelligent edge devices. However, applications for inferring/recreating/manipulating require real-time inference and in-situ learning. For example, image denoising algorithms are expected to be deployed on modern smartphones. The limited computing and energy resources available at the edge stand at odds with the massive and growing learning costs for state-of-the-art deep neural networks. My research also engaged extensive efforts in building efficient training and inference pipelines for computer vision algorithms. Many of them are powered by interpretability [10], Automated machine learning [11, 12] (AutoML), and hardware-software co-design [13].

4 Future Directions

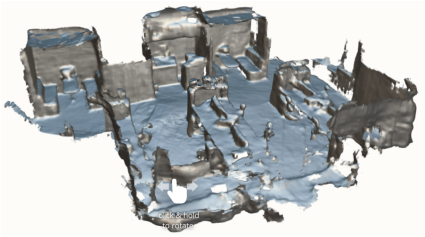
My future works will explore how to leverage machine learning techniques to further enhance the experience of photographs capturing and reconstructing more realistic visual content from the physical world, as well as how to make these learning-based algorithms more efficient in the edge. Below I will outline each proposed project in detail.

Generating Front-facing Views in Facetime Video App. Online video apps (E.g., Facetime) are crucial tools to connect family and friends. However, existing devices (iPhone, iPad, Macbook, and etc.) are not able to put the front camera in the center of screens, and thus the human faces in the captured views are not front-facing. Although [14] renders different novel views of captured faces to solve this issue, it highly relies on multiple cameras with different angles. Therefore, rendering a front-face view using a single camera is still challenging. My ECCV'2022 work [5] proposes a potential solution to render novel views of a single in-the-wild image, but requires a per-scene optimization procedure for each scene. In this project, we consider the human face as a strong prior and adopt neural rendering techniques [15, 16] to train a feed-forward network to solve this problem. For each face, we will adopt off-the-shelf algorithms [17] to get 3D Face geometry. With the help of 3D geometry and learning priors from training data, we are expected to learn a model that is able to render the front-facing view of the captured face.



Front-facing View Synthesis [14]

Real-time 3D Reconstruction on Mobile Devices with Lidar Guidance. Modeling the geometry of a scene is crucial to VR/AR applications. For example, portable Augmented Reality (AR) devices (e.g., the Magic Leap One) reconstruct the scene geometry and localize users further. The recently released ARkit[2] API from Apple provides Lidar depth (estimated depth map using sparse lidar rays) to developers and enables various downstream AR applications, where one of them is 3D geometry reconstruction. Existing methods [18] mainly adopt multi-view geometry to estimate 3D geometry, which is computationally expensive. With the help of lidar depth, we are able to adopt low-quality depth maps as geometry priors to accelerate the 3D reconstruction algorithm and enable real-time estimation on the edge devices (E.g., iPhone 14 Pro).



3D Reconstruction [18]

Rendering 3D Photo Effects using Wide and Ultra-wide Camera Stereos. While capturing static images helps people save memorial moments in their lives, Apple's Live Photos can further store frames before and after capturing to bring more entertainment. Recently, a new application [19] called "3D photo" enables users to see captured images in a 3D view and provide immersive experiences. Their main pipeline firstly estimates the depth map of the given image and then generates multi-plane

3D Photo [19]

[2]https://developer.apple.com/augmented-reality/arkit/

images. Meanwhile, a pre-trained image inpainting model will inpaint the missing holes in the novel view of multi-plane images and render complete photos. Although their 3D effects are promising, the performance is highly affected by the depth estimator and the inpainted texture could be fake in some cases. In this project, we propose to use the wide and ultra-wide camera in iPhone 14 to render a more photorealistic 3D photo. Since these two cameras are placed in different positions, the captured views can be adopted to build image stereos and help enhance the quality of the estimated depth map. Moreover, with the help of another captured image, we are able to render smooth interpolation between these two angles using neural rendering techniques [16] and make the rendered 3D photo satisfying.

References

[1] **Yifan Jiang**, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *Transcation on Image Processing (**TIP**)*, 2021.

[2] Dejia Xu, Hayk Poghosyan, Shant Navasardyan, **Yifan Jiang**, Humphrey Shi, and Zhangyang Wang. Recoro: Region-controllable robust light enhancement by user-specified imprecise masks. *ACM Multimedia (**MM**)*, 2022.

[3] **Yifan Jiang**, Bartlomiej Wronski, Ben Mildenhall, Jonathan T. Barron, Zhangyang Wang, and Tianfan Xue. Fast and high-quality image denoising via malleable convolutions. In *European Conference on Computer Vsion (**ECCV**)*, 2022.

[4] Zeyuan Chen, **Yifan Jiang**, Liu Dong, and Zhangyang Wang. Cerl: A unified optimization framework for light enhancement with realistic noise. *Transcation on Image Processing (**TIP**)*, 2022.

[5] **Yifan Jiang***, Dejia Xu*, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *European Conference on Computer Vsion (**ECCV**)*, 2022, [*] indicates equal contribution.

[6] Xinyu Gong, Shiyu Chang, **Yifan Jiang**, and Zhangyang Wang. Autogan: Neural architecture search for generative adversarial networks. In *International Conference on Computer Vision (**ICCV**)*, 2019.

[7] **Yifan Jiang**, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Thirty-fifth Conference on Neural Information Processing Systems (**NeurIPS**)*, 2021.

[8] **Yifan Jiang**, Zhang He, Jianming Zhang, Yilin Wang, Zhe Lin, Kalyan Sunkavalli, Simon Chen, Sohrab Amirghodsi, Sarah Kong, and Zhangyang Wang. Ssh: A self-supervised framework for image harmonization. In *International Conference on Computer Vision (**ICCV**)*, 2021.

[9] **Yifan Jiang***, Zhiwen Fan*, Peihao Wang*, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Unified implicit neural stylization. In *European Conference on Computer Vsion (**ECCV**)*, 2022, [*] indicates equal contribution.

[10] Bowen Pan, Rameswar Panda, **Yifan Jiang**, Zhangyang Wang, Rogerio Feris, and Aude Oliva. Ia-red^2: Interpretability-aware redundancy reduction for vision transformers. *Thirty-fifth Conference on Neural Information Processing Systems (**NeurIPS**)*, 2021.

[11] **Yifan Jiang**, Xinyu Gong, Junru Wu, Honghui Shi, Zhicheng Yan, and Zhangyang Wang. Autox3d: Searching ultra-efficient architecture for video understanding. *Winter Conference on Applications of Computer Vision (**WACV**)*, 2022.

[12] Yonggan Fu, Zhongzhi Yu, Yongan Zhang, **Yifan Jiang**, Chaojian Li, Yongyuan Liang, Mingchao Jiang, Zhangyang Wang, and Yingyan Lin. Instantnet: Automated generation and deployment of instantaneously switchable-precision networks. *The 58th Design Automation Conference (**DAC**)*, 2021.

[13] Mengshu Sun, Haoyu Ma, Guoliang Kang, **Yifan Jiang**, Tianlong Chen, Xiaolong Ma, Zhangyang Wang, and Yanzhi Wang. Vaqf: Fully automatic software-hardware co-design framework for low-bit vision transformer. *arXiv preprint arXiv:2201.06618*, 2022.

[14] Goolge AI Blog. https://blog.google/technology/research/project-starline/.

[15] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[16] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021.

[17] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4):1–13, 2021.

[18] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. NeuralRecon: Real-time coherent 3D reconstruction from monocular video. *CVPR*, 2021.

[19] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8028–8038, 2020.