Bridging Virtual and Physical World by Recreating and Manipulating Visual Contents

Yifan Jiang, University of Texas at Austin

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

- *how can we design algorithms that restore and infer 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)?*

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 two themes and conclude with future research agenda.

1 Inferring Physical World from Corrupted Signals

Physical world contains satisfactory visual contents and attracts our human beings to record it. It is only within the recent decade that digital photography has finally eclipsed film in sales, unit volume, and the number of pictures taken. However, 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.

Unsupervised Low-light Enhancement Image captured in low-light conditions suffer 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 all-day autonomous driving and biometric recognition. The state-of-the-art approaches using deep learning techniques 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 **420** times) was implemented by popular third-party image manipulation software and open-source toolboxes, and was also extensively adopted or modified by the related CV functionalities in mainstream Apps (TikTok and PicsArt).

Light Enhancement [1]

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 an end-to-end mapping directly from corrupted observations to the unobserved clean signal, without explicit statistical modeling of signal corruptions. However, these deep networks usually suffer from notoriously large computation. My ECCV'2022 paper [2] tackled this issue by proposing a fast and memory-friendly dynamic kernels, which predicts spatially-varying kernels at low resolution and adopts a fast fused operator to jointly upsample and apply these kernels at full resolution.

Image Denoising [2]

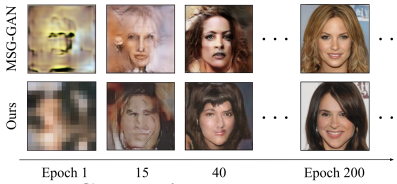
Single Image Novel-view Synthesis The task of novel view synthesis has recently seen dramatic progress as a result of using neural radiance field (NeRF). However, training NeRF requires dense captured views and the corresponding camera poses. Several recent attempts to train a NeRF using sparse views, in contrast to them, my ECCV'2022 work [3] pushed the setting of sparse views to the extreme, by training a neural radiance field on only one single view.

Novel-view Synthesis [3]

2 Creating, Editing, and Manipulating Visual Contents

While restoring physical world help us record memorial events, recreating photo-realistic visual contents can bring us more entertainments, 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.

Recreating Visual Contents via Generative Models Many applications for recreating visual contents 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 **AutoGA**N [5] architecture outperformed all existing handcrafted GAN models on the CIFAR-10 image generation benchmark by then, highlighted by **Synced AI Review** [link]. More recently, my NeurIPS'2021 paper accomplished a new piece of high-visibility work namely **TransGAN** [4]: 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 **150** citations and the open-sourced GitHub repository gained more than **1400** stars in less than one year.

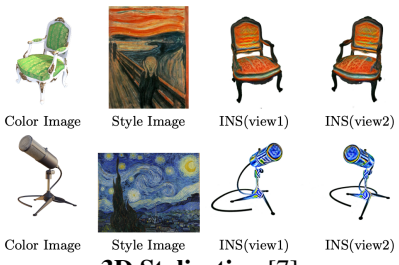


Generative Model [4]

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 [6] proposed the first self-supervised harmonization framework that needs neither human-annotated mask nor professionally created images for training.


Harmonization [6]

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 that how we can edit/manipulate these continuous representations. My ECCV'2021 paper [7] proposed a unified inplicit 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 [7]

3 Future Directions

My future works will explore how to leverage multiple modalities (e.g., lidar, thermal camera, etc.) to reconstruct more realistic visual content from the physical world, and how to make these learning-based algorithms more efficient in the training/inference stages.

References

[1] **Y. Jiang**, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, and Z. Wang. Enlightengan: Deep light enhancement without paired supervision. *IEEE Transactions on Image Processing*, 2021.

[2] **Y. Jiang**, B. Wronski, B. Mildenhall, J.T. Barron, Z. Wang, and T. Xue. Fast and high-quality image denoising via malleable convolutions. In *ECCV*, 2022.

[3] **Y. Jiang***, D. Xu*, P. Wang, Z. Fan, H. Shi, and Z. Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *ECCV*, 2022, [*] indicates equal contribution.

[4] **Y. Jiang**, S. Chang, and Z. Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. *NeurIPS*, 2021.

[5] X. Gong, S. Chang, **Y. Jiang**, and Z. Wang. Autogan: Neural architecture search for generative adversarial networks. In *ICCV*, 2019.

[6] **Y. Jiang**, H. Zhang, J. Zhang, Y. Wang, Z. Lin, K. Sunkavalli, S. Chen, S. Amirghodsi, S. Kong, and Z. Wang. Ssh: A self-supervised framework for image harmonization. In *ICCV*, 2021.

[7] **Y. Jiang***, Z. Fan*, P. Wang*, X. Gong, D. Xu, and Z. Wang. Unified implicit neural stylization. In *ECCV*, 2022, [*] indicates equal contribution.