Research Statement

Chi Keung Tang

During the first decade of my academic career (2000–2009) the focus was *vision and graphics* when SIGGRAPH published some of the most impactful papers to date. Relevant research momentum was fueled by high grossing Pixar/Disney movies and computer games in the coming years, where computer vision found its instrumental role in enabling high-quality special efforts which were used to be costly to produce. My SIGGRAPH papers *Lazy Snapping* and *Poisson Matting* published during this time continue to inspire researchers and practitioners, with the patented technologies applied in TV/movie post-production. Notably, they were HKUST's *first SIGGRAPH papers* as well as the first time a Hong Kong-based researcher publishing more than one in the same year. After these impactful papers, we witnessed a sea change in the landscape of graphics research in Hong Kong not limited to HKUST: faculty were invited to serve on SIGGRAPH committees, students and faculty (many are HKUST PhD graduates) have been making consistent presence on the SIGGRAPH map and recognized as leaders in 2D and 3D computer graphics. These first SIGGRAPH papers have gathered over 2500 citations combined and become "must-read" papers in graduate courses in computational photography around the world. The papers are centered around the main theme of *interactive computer vision*, where one focus is to "bring the user into the loop" to achieve computer vision tasks that have been traditionally very difficult to automate. By accepting a few simple and loose hints, such as strokes or scribbles the user casually marks on an image, the automatic system can output segmentation results of pixel or even subpixel accuracy.

Following the revolution brought by deep convolutional neural network in machine learning, *deep learning for computer vision* was the focus during the second decade of my academic career (2010–2019). Over the past decade, I have focused on publishing high-impact work in top computer vision and machine learning venues *CVPR, ICCV, ECCV, NeurIPS*, where in particular *CVPR* is ranked #4 of all publications in h-5 index, only after *Nature*, *The New England Journal of Medicine*, and *Science*. During this time, *Network Trimming*, an arXiv report which was rejected by NeurIPS'16, have become one of the seminal works for network pruning to achieve efficient deep architectures. Our recent work on *Few Shot Learning*, a series of CVPR and ECCV papers published since 2020, contribute to image and video object detection and semantic segmentation. We proposed novel deep architectures for video object segmentation, video instance segmentation, multiple object tracking and segmentation, and occlusion-aware video object inpainting.

Stepping into the third decade (2020–) of my academic career, given the lightning rate of computer vision and machine learning research which hold the key to success in the next-generation high-impact applications, my recent works focus on

Neural Radiance Fields (NeRFs). NeRFs are one of the hottest emerging topics in computer vision, computer graphics and machine learning due to their impressive 3D results and high potential in augmented reality, storytelling or even "storyliving" technologies leading to high-quality metaverse applications for immersive user experience. To enable such relevant important applications, our fundamental NeRF contributions are essential on unsupervised multi-view object segmentation (NeurIPS'22), region proposal network for general object detection (*NeRF-RPN*), and direct 3D landmark detection on NeRFs (FLNeRF). In *NeRF-RPN*. Given a pre-trained NeRF model, NeRF-RPN detects all bounding boxes of objects in a scene. By exploiting a novel voxel representation that incorporates multi-scale 3D neural volumetric features, we demonstrate it is possible to regress the 3D bounding boxes of objects in NeRF directly without rendering the NeRF at any viewpoint. Our 3D coarse-to-fine FLNeRF model efficiently samples from the NeRF on the whole face with individual facial features for accurate landmarks. To mitigate the limited number of facial expressions in the available data, nonlinear augmentations are applied at facial features in fine scale to simulate large emotions range, including exaggerated facial expressions (e.g., blowing cheek, wide opening mouth, blinking eyes), for training FLNeRF. Consequently, an animator can easily edit, control and even transfer emotion from another face NeRF. With precise landmarks on facial features, exaggerated facial expressions with wide grinning, mouth wide open, cheek blowing can be readily achieved. NeRF-RPN and FLNeRF are argably the first significant

work on NeRFs in their respective areas, which are fundamental and essential in enabling future important 3D NeRF-based applications. Demo links: NeRF-RPN, FLNeRF.

Unsupervised Unimodal Pre-training for Boosting Zero-shot Vision-Language Tasks. Large-scale joint training of multimodal models, e.g., CLIP, have demonstrated great performance in many vision-language tasks. However, image-text pairs for pre-training are restricted to the intersection of images and texts, limiting their ability to cover a large distribution of real-world data, where noise can also be introduced as misaligned pairs during pre-processing. Conversely, unimodal models trained on text or image data alone through unsupervised techniques can achieve broader coverage of diverse real-world data and are not constrained by the requirement of simultaneous presence of image and text. In this paper, we demonstrate that using large-scale unsupervised unimodal models as pre-training can enhance the zero-shot performance of image-text pair models. Our thorough studies validate that models pre-trained as such can learn rich representations of both modalities, improving their ability to understand how images and text relate to each other. Our experiments show that unimodal pre-training outperforms state-of-the-art CLIP-based models by 6.5% (52.3% $\rightarrow$ 58.8%) on PASCAL-5^i and 6.2% (27.2% $\rightarrow$ 33.4%) on COCO-20^i semantic segmentation under zero-shot setting respectively. By learning representations of both modalities, unimodal pre-training offers broader coverage, reduced misalignment errors, and the ability to capture more complex features and patterns in the real-world data resulting in better performance especially for zero-shot vision-language tasks.

Segment Anything in High Quality. The recent Segment Anything Model (SAM) represents a big leap in scaling up segmentation models, allowing for powerful zero-shot capabilities and flexible prompting. Despite being trained with 1.1 billion masks, SAM's mask prediction quality falls short in many cases, particularly when dealing with objects that have intricate structures. We propose HQ-SAM, equipping SAM with the ability to accurately segment any object, while maintaining SAM's original promptable design, efficiency, and zero-shot generalizability. Our careful design reuses and preserves the pre-trained model weights of SAM, while only introducing minimal additional parameters and computation. We design a learnable High-Quality Output Token, which is injected into SAM's mask decoder and is responsible for predicting the high-quality mask. Instead of only applying it on mask-decoder features, we first fuse them with early and final ViT features for improved mask details. To train our introduced learnable parameters, we compose a dataset of 44K fine-grained masks from several sources. HQ-SAM is only trained on the introduced detaset of 44k masks, which takes only 4 hours on 8 GPUs. We show the efficacy of HQ-SAM in a suite of 9 diverse segmentation datasets across different downstream tasks, where 7 out of them are evaluated in a zero-shot transfer protocol. Our code and models will be released.

Efficient Dataset Distillation via Critical Samples Selection. Data-efficient learning has drawn significant attention, especially given the current trend of large multi-modal models, where dataset distillation can be an effective solution. However, the dataset distillation process itself is still very inefficient. In this work, we model the distillation problem with reference to information theory. Observing that severe data redundancy exists in dataset distillation, we argue to put more emphasis on the utility of the training samples. We propose a family of methods to exploit the most valuable samples, which is validated by our comprehensive analysis of the optimal data selection. The new strategy significantly reduces the training cost and extends a variety of existing distillation algorithms to larger and more diversified datasets, e.g., in some cases only **0.04%** training data is sufficient for comparable distillation performance. Moreover, our strategy consistently enhances the performance, which may open up new analyses on the dynamics of distillation and networks. Our method is able to extend the distillation algorithms to much larger-scale datasets and more heterogeneous datasets, e.g., ImageNet-1K and Kinetics-400. Our code will be made publicly available.