Yu-Wing TAI

Research Statement

Google Scholar: https://scholar.google.com/citations?user=nFhLmFkAAAAJ&hl=en

Research Interests and Past Experiences

My research interests are in Deep Learning and Computer Vision. In my early academic career, I focused my research in image processing and computational photography, a technique to understand the image formation process and use algorithms to enhance the ability of cameras. I was also interested in 3D capturing of the real-world using computational photography techniques. After I moved to a company and witnessed the impact of deep learning, I found that recognition, detection, segmentation, and video understanding have far more impacts and application values. Because of my job nature, I shifted my research interests from low level vision to high level vision.

Using my experiences in the current company as an example, when a user uploaded a video to a short-video platform, it has to pass through multiple steps before it is ready to be transmitted to another user for watching. In short, one has to first apply video analyses to filter out improper contents such as porn/violence recognition, and/or face recognition to remove politically sensitive videos. Next, video classification is applied to group videos into several big categories to enhance video recompression. A very fast video saliency detection algorithm is applied to protect region-of-interests during the video recompression. Some of the low-quality videos might also need to go through enhancement processes to improve the quality of uploaded videos. The next step is video tagging by applying various algorithms for video scene partitioning, object and action recognition, ocr, and even speech and nlp recognition, resulting in a very large feature vector describing the contents of a video. With that, a recommendation system can learn a user profile that keeps track of a user's interests based on the videos a user has watched before. A video search engine can also be developed according to the large feature vector of video contents. Based on the recommendation system and search engine, a company can make profits from advertisements to connect service providers and/or merchandise sellers with their customers. Note that all of the above processes are fully automated and the video tagging is dynamic which evolves with user uploaded videos. While this is not 100% identical for different companies, we see that this is fundamentally the business model of youtube, meta, and tiktok. We can see that deep learning and computer vision play very important roles in the above processes.

Recently, I am particular interested in solving problems that can bridge the gap between papers and real-world applications. I category my recent works into two big topics: **Solving data hungry problem in real-world applications** and **Towards 3D understanding of our physical world**

Solving data hungry problem in real-world applications

Over the past few years, we have witnessed the success of deep learning in image recognition thanks to the availability of large-scale human-annotated datasets such as PASCAL VOC, ImageNet, and COCO. Although these datasets have covered a wide range of object categories, there are still a significant number of objects that are not included. Indeed, real-world object categories follow a longtail distributions as shown in Fig. 1. Can we perform the same task without a lot of human annotations? This problem motivates my recent works on Few Shot Learning, a series of papers [1-16] published since 2020, contribute to image and video object detection, segmentation, pose estimation and action recognition.

Figure 1. Left: Real-world object categories follow a longtail distributions. Right: Given different objects as supports (top corners above), our approach [1] can detect all objects in the same categories in the given query image.

In my CVPR'20 paper [1], my coauthors and I have developed a new few-shot object detection (FSOD) network that aims at detecting objects of unseen categories with only a few annotated examples. Central to our method are our Attention-RPN, Multi-Relation Detector and Contrastive Training strategy, which exploit the similarity between the

few shot support set and query set to detect novel objects while suppressing false detection in the background. To train our network, we contribute a new dataset that contains 1000 categories of various objects with high-quality annotations. To the best of our knowledge, this is one of the first datasets specifically designed for few-shot object detection. Once our few-shot network is trained, it can detect objects of unseen categories without further training or finetuning.

In another CVPR'20 paper [2], we take a step forward to study the few shot segmentation (FSS) problem. The FSS paper is one of the first paper that aims at evaluating few shot segmentation algorithms at large scale. Similar to FSOD, we built a benchmark dataset consists of 1000 object classes with pixelwise annotation of ground-truth segmentation. Unique in FSS-1000, this dataset contains significant number of objects that have never been seen or annotated in previous datasets, such as tiny daily objects, merchandise, cartoon characters, etc. We build our baseline model using standard relational network. To our surprise, we found that training our model from scratch using FSS-1000 achieves comparable and even better results than training with weights pre-trained by ImageNet which is more than 100 times larger than FSS-1000. This finding echoes our results in FSOD where increasing number of object categories is far more important than increasing number of samples of each class in training data.

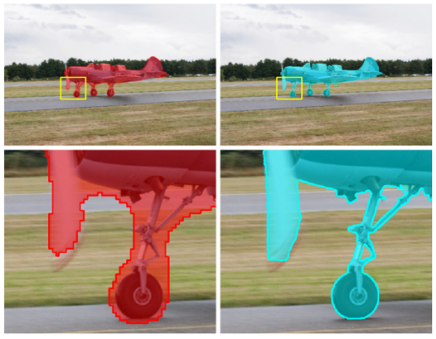

Figure 2. Segmentation of a high-resolution image (3492×2328). Left: Produced by Deeplab V3+. Right: Refined by our algorithm [3].

As human annotation for accurate boundary segmentation is very labor intensive, we developed the first deep learning-based Class-Agnostic Segmentation Refinement framework to speed up our annotation process (Fig. 2). This work was also published in CVPR'20 [3]. Similarly, we have also figured out the data hungry problem in 3D human pose annotations and developed a method to synthesize 2D-to-3D training data pairs to tackle the problem [4]. The same story happens to action recognition, where we tried to reformulate the problem of action recognition into atomic action recognition [5] which is a better definition of action to eliminate ambiguities in data label. We have demonstrated the efficacy of HAA500 where action recognition can be greatly benefited from a clean, highly diversified, class-balanced finegrained atomic action. On top of HAA500, we have also empirically investigated several important factors that can affect the performance of action recognition.

Video object segmentation can also be considered as a few shot problem where annotations are limited to a few frames, and the segmentation has to be propagated across the entire video sequences. In order to distinguish the problem from existing tracking-by-detection framework which require a large training dataset for learning the object recognition part, we consider interactive video object segmentation. The CVPR'21 paper [6] describes how to decouple interaction-to-mask and mask propagation in a fast and efficient manner, allowing for higher generalizability and better performance. The NeurIPS'21 paper [8] discusses the problem of previous STM module for segmentation propagation and propose the STCN module for even faster (20x faster than STM) and accurate propagation of segmentation masks across video frames. Note that STCN also win the second place of YouTubeVOS challenges (first place for unseen classes) in 2021.

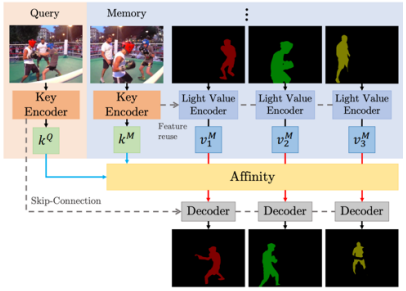

Figure 3. Our proposed Space-Time Correspondence Networks (STCN) [8]. We use Siamese key encoders to compute affinity directly from RGB images, making it more robust and efficient.

With the above mentioned ``baseline'' models, my recent works [7,9-16] improve the aforementioned papers from different perspective including an even more accurate instance segmentation algorithm by instance matting [11], boundary aware large scale instance segmentation by transfiner [10] and video transfiner [15], using self-support query to improve few-shot segmentation [13], extending FSOD to FDVOD [14], reducing bias in model pre-training using calibrations [12], and improving model generality by reducing domain gap of data using normalization perturbation [16]. Note that all these works are not just highly cited in google scholar, some of them have already transferred to real world products of my company in either better annotation tools and/or new interactive segmentation tools for image/video editing. Recognized novel objects in a video are also used for merchandise recommendations and/or video searching and grouping.

Towards 3D understanding of our physical world

Our physical world is 3D and we perceive and understand this physical world with two eyes that implicitly encode 3D information in our brain. Yet, most deep learning algorithms regard images as flat 2D arrays of pixels where pixel groupings form object instances. This representation ignores the fact that images are 2D projections of the 3D world. Discarding one of the dimensions gives rise to fundamental ambiguities and challenges to the task resulting from partial or total occlusion, especially for single-image methods. I believe 3D computer vision holds the key to success in the next-generation high-impact researches and applications.

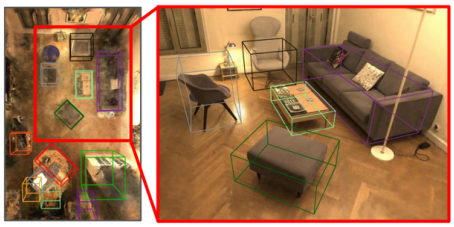

Figure 4. Region proposal results on a NeRF. Top 12 proposals in eight orientations with highest confidence are visualized. The NeRF is trained from the Living Room scene from INRIA.

My recent computer vision works focus on Neural Radiance Fields (NeRFs) which is one of the hottest emerging topics in computer vision, computer graphics and machine learning since it was first published in ECCV'20. Majority of recent works on NeRF are still focusing on improving the capturing quality and speeding up the rendering processes of NeRF. Our fundamental NeRF contributions, on the other hand, trying to understand the 3D world represented by NeRF. Up to now, we have contributed one of the first algorithms for unsupervised multi-view object segmentation (NeurIPS'22 [17]), region proposal network for general object detection (NeRF-RPN) [18]. In NeRF-RPN, given a pre-trained NeRF model, NeRF-RPN detects all bounding boxes of objects in a scene (Fig. 4). 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 method is general and can be applied to detect objects without class labels. To facilitate future research in object detection for NeRF, we also built a new benchmark dataset which consists of both synthetic and real-world data with careful labeling and clean up. Please watch the video for visualizing the 3D region proposals by our NeRF-RPN (Demo link: https://www.youtube.com/watch?v=M8_4Ih1CJjE).

I believe 3D understanding of our physical world based on NeRF representation is just the beginning of a new era in deep learning, computer vision, and computer graphics. There are still a lot of fruitful topics and applications waiting for us to explore and to develop. Without any doubt, this research direction also has the data hunger problem, and I believe my rich experiences in few shot learning and related applications would allow me to make even bigger impacts to research and to the society.

Future Research Plan

"We tend to overestimate the effect of technology in the short run, and underestimate the effect in the long run." – Amara's Law

The above statement has been quoted many times to describe the trend of development of many new technologies. This is especially true for deep learning when Deepmind first invented AlphaGo that beats the best professional player of Go over the world. People believed that one day AI could replace humans for many tasks but this day hasn't come yet.

As a computer vision and deep learning researcher, I believe many algorithms and models we have developed to understand our physical world based on 2D inputs are deficiency. In contrast, I believe 3D understanding is the future but it has a very long way to go. Can we capture and digitalize our physical world? How are we going to interact with an agent in the digitalized physical world? How can we distinguish AI generated contents (AIGC) in the digitalized physical world from our real physical world?

All these questions pointed to the next generation of applications related to metaverse, AIGC, autonomous driving and embodied AI. Without any doubt, they are not going to come in the short run, but I believe they will come one day in the long run. As a researcher who should have a concrete research plan, in the short run, I am planning to build a **3D capturing system** that would allow efficient data collection of human, objects, and indoor room environments. Such 3D data is very valuable and they are the driving force of my research towards 3D understanding. The corresponding published papers would also make a big impact to the research community and society. Next, I

would be interested in studying how to generate new contents in the digitalized 3D world (**3D AIGC**) governed by the captured 3D data. New contents can be a new object with different appearance, a virtual human that mimics activities of a person and/or a room or city with different configurations that would allow an AI agent to learn and interact inside the digitalized world. With that, one can imagine these technologies can revolutionize the existing *short-video clip industry* and even drama/movie industry where the rendered **3D AIGC contents** are as high quality as the videos captured in our real physical world. The technology can also be used to *protect people privacy* through generating virtual identity of a person. Lastly, the real autonomous driving and embodied AI would come.

References:

1. Qi Fan, Wei Zhuo, Chi-Keung Tang, **Yu-Wing Tai**, *Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector*, IEEE Conference on Computer Vision and Pattern Recognition (***CVPR***), 2020.
2. Xiang Li, Tianhan Wei, Yau Pun Chen, Chi-Keung Tang, **Yu-Wing Tai**, *FSS-1000: A 1000-Class Dataset for Few-Shot Segmentation*, IEEE Conference on Computer Vision and Pattern Recognition (***CVPR***), 2020.
3. Ho Kei Cheng, Jihoon Chung, **Yu-Wing Tai**, Chi-Keung Tang, *CascadePSP: Toward Class-Agnostic and Very High-Resolution Segmentation via Global and Local Refinement*, IEEE Conference on Computer Vision and Pattern Recognition (***CVPR***), 2020.
4. Shichao Li, Lei Ke, Kevin Pratama, **Yu-Wing Tai**, Chi-Keung Tang, Kwang-Ting Cheng, *Cascaded Deep Monocular 3D Human Pose Estimation with Evolutionary Training Data*, IEEE Conference on Computer Vision and Pattern Recognition (***CVPR***), 2020.
5. Jihoon Chung, Cheng-hsin Wuu, Hsuan-ru Yang, **Yu-Wing Tai**, Chi-Keung Tang, *HAA500: Human-Centric Atomic Action Dataset with Curated Videos*, IEEE International Conference on Computer Vision (***ICCV***), 2021
6. Ho Kei Cheng, **Yu-Wing Tai**, Chi-Keung Tang, *Modular Interactive Video Object Segmentation: Interaction-to-Mask, Propagation and Difference-Aware Fusion*, IEEE Conference on Computer Vision and Pattern Recognition (***CVPR***), 2021
7. Lei Ke, **Yu-Wing Tai**, Chi-Keung Tang, *Deep Occlusion-Aware Instance Segmentation with Overlapping BiLayers*, IEEE Conference on Computer Vision and Pattern Recognition (***CVPR***), 2021
8. Ho Kei Cheng, **Yu-Wing Tai**, Chi-Keung Tang, *Rethinking Space-Time Networks with Improved Memory Coverage for Efficient Video Object Segmentation*, Thirty-fifth Conference on Neural Information Processing Systems (***NeurIPS***), 2021
9. Lei Ke, Xia Li, Martin Danelljan, **Yu-Wing Tai**, Chi-Keung Tang, Fisher Yu, *Prototypical Cross-Attention Networks for Multiple Object Tracking and Segmentation*, Thirty-fifth Conference on Neural Information Processing Systems (***NeurIPS***), 2021
10. Lei Ke, Martin Danelljan, Xia Li, **Yu-Wing Tai**, Chi-Keung Tang, Fisher Yu, *Mask Transfiner for High-Quality Instance Segmentation*, IEEE Conference on Computer Vision and Pattern Recognition (***CVPR***), 2022
11. Yanan Sun, Chi-Keung Tang, **Yu-Wing Tai**, *Human Instance Matting via Mutual Guidance and Multi-Instance Refinement*, IEEE Conference on Computer Vision and Pattern Recognition (***CVPR***), 2022
12. Qi Fan, Chi-Keung Tang, **Yu-Wing Tai**, *Few-Shot Object Detection with Model Calibration,* European Conference on Computer Vision (***ECCV***), 2022
13. Qi Fan, Wenjie Pei, **Yu-Wing Tai**, Chi-Keung Tang, *Self-support few-shot semantic segmentation,* European Conference on Computer Vision (***ECCV***), 2022
14. Qi Fan, Chi-Keung Tang, **Yu-Wing Tai**, *Few-shot video object detection,* European Conference on Computer Vision (***ECCV***), 2022
15. Lei Ke, Henghui Ding, Martin Danelljan, **Yu-Wing Tai**, Chi-Keung Tang, Fisher Yu, *Video Mask Transfiner for High-Quality Video Instance Segmentation,* European Conference on Computer Vision (***ECCV***), 2022
16. Qi Fan, Mattia Segu, **Yu-Wing Tai**, Fisher Yu, Chi-Keung Tang, Bernt Schiele, Dengxin Dai, *Towards Robust Object Detection Invariant to Real-World Domain Shifts*, International Conference on Learning Representations (***ICLR***), 2023
17. Xinhang Liu, Jiaben Chen, Huai Yu, **Yu-Wing Tai**, Chi-Keung Tang, *Unsupervised Multi-View Object Segmentation Using Radiance Field Propagation*, Thirty-sixth Conference on Neural Information Processing Systems (***NeurIPS***), 2022
18. Benran Hu, Junkai Huang, Yichen Liu, **Yu-Wing Tai**, Chi-Keung Tang, *NeRF-RPN: A general framework for object detection in NeRFs*, arXiv preprint arXiv:2211.11646, 2022 (Accepted in CVPR'23)