RESEARCH STATEMENT

Fisher Yu

Intelligent systems with computer vision capabilities are poised to improve every aspect of our lives. Recent progress in computer vision research, especially with Deep Learning methods, offers promises of transformative automation in the foreseeable future. At home, appliance automation will make domestic lives more efficient and pleasant. On the road, self-driving vehicles will make transportation less expensive and more efficient. In retail, autonomous stores will streamline our shopping experiences and warehouse robots will make logistics more efficient. However impressive these recent advances are in simple cases, and excessively hyped in the popular press, advanced automation still face significant hurdles to reach the future of automation at planet scale.

The full promise of automated machine learning has yet to be realized. Current systems require extensive manual intervention and tuning. A state-of-the-art intelligent system requires a customized pipeline to understand domain knowledge and recognize domain-specific patterns. The development process typically has two stages: model design and data collection. Model design usually involves manual trial and error, or expensive exhaustive search. When an application of an existing method in a new domain appears, skilled research engineers has to redo most of the modeling. Furthermore, our understanding of modeling the dynamic world is still primitive compared to our achievements in image representation, which hinders the applicability of current methods in interactive, non-stationary environments.

While new algorithms are exciting, and a key focus of mine, performance and impact in the current era are often limited as much or more so by the quality of available data and associated annotations. Although they are often an after-thought for many researchers, I consider the systems issues in handling, labeling, and manipulating large-scale datasets generated by and for intelligent systems to be a first-class problem. For computer vision research to fully thrive, we not only need better core algorithms for perception and action, but we also need new algorithms and system frameworks to deploy the autonomous systems autonomously.

I have been pioneering frameworks to analyze scene structures and dynamics, and seek systematic solutions for processing and understanding truly planet-scale imagery. I believe there are three pillars of research necessary to support true AI: we must simultaneously tackle the fundamentals of deep image representation, model dynamic and interactive scenes, and develop the infrastructure and algorithms required for large-scale perceptual data management. I have pursued an aggressive research agenda in these areas, with high-impact publications in each area, which I will describe further in the following sections.

Unified Image Representation Framework

In industry, an increasing number of research engineers are engaged in tuning models for specific applications. Specialization becomes a serious issue when an intelligent system such as an autonomous vehicle needs to perform multiple tasks at the same time. It is a real requirement and challenge to develop robust image representation models that can directly solve multiple problems with a similar nature and generalize to new tasks. I take a bottom-up approach in pursuit of a unified representation framework by starting with understanding the requirements for previously isolated recognition problems. Towards this goal, my study begins with pixel-level prediction problems and connects them with broader applications.

Pixel-Level Prediction Most recognition tasks, such as semantic segmentation, boundary prediction, and detection, require pixel-level predictions. My first question about convolutional networks was whether we can design a network structure specialized to pixel-level prediction problems with the capability to transfer the representation learned from large-scale classification datasets to other tasks. This question led to furthering the understanding of dilated convolution [13], first proposed in wavelet transforms. Two properties of dilated convolution make it widely applicable: 1) It can "dilate" any image classification network to preserve learned connections and produce high-resolution feature maps; 2) It can aggregate long-range context information efficiently. This pixel-level representation was further improved by my work on Dilated Residual Networks [14]. The research community has found many exciting applications

of dilated networks, such as Google's audio generation method (WaveNet [6]), makeup transfer [2], 3D volume segmentation, and completion [5].

Unifying Basic Vision Problems Because other recognition problems can be considered as a form of pixel-level prediction, image classification becomes an interesting obstacle for connecting different types of recognition problems. Quantitative evaluation of Dilated Residual Networks (DRN) [14] on image classification shows that high-resolution feature maps for pixel-level prediction network can actually improve image-level prediction as well. Qualitative study of the results also lead to new insight into weakly supervised object localization and new network structures for more accurate segmentation. Although dilated convolution has become a basic building block for pixel-level prediction, skip connections with bottleneck layers are still popular due to their efficiency and higher resolution. My work on Deep Layer Aggregation (DLA) [16] combined the efficiency of skip connections and the semantics/resolution balance of dilated convolutions. The framework can construct practical solutions for a wide range of problems, from fine-grained classification to boundary prediction. The DLA representation is versatile enough to represent both image-level information and fine differences between adjacent pixels. Besides classification image recognition problems, DLA can also extract hierarchical features for dense correspondences [11] and scene prediction [4].

Dynamic Network Structure While convolutional networks have static structure at inference time, it has been shown that humans can dynamically adapt to the difficulty of the situation, with the resulting intuition system making decisions quickly and unconsciously. The key is the ability to decide when to pass the decision making to a deliberate thinking process. SkipNet [7] explored this idea: The network can learn how many layers to use for each input image and try to use less computation for simpler images without sacrificing overall prediction accuracy. The idea is extended to make the computation decision at the channel level in our Task-Aware Fearture Embedding Network (TAFE-Net) [8] method. TAFE-Net uses a separate controller to predict network layer weights based on its inputs, which can be either task encoding, word2vec description or the input image itself. Besides computation reduction, a dynamic network can also bring generalizability beyond input-level supervision, e.g. zero- and few-shot learning tasks, by predicting a new image representation network for each task and learning a joint embedding for the task network output and image representation. Dynamic networks deserve further investigation as a solution for incremental and life-long learning.

Dynamic Scene Understanding and Interaction

The world is in constant motion. To comprehend it, we need to go beyond images and analyze the motion and temporal information hidden in image sequences. Video analysis has even more diverse methods and domains than image analysis. In general, video tasks can be categorized by the targets of the investigation. For example, the goal can be recognition of human activity or static 3D geometry. My research interest here lies in dynamic scene understanding for autonomous driving, including tracking and prediction. Driving scenarios contain a diverse set of scenes and actors due to the variation of locations, weather, time of the day, etc. Interactions between moving objects pose a challenging policy learning problem. Autonomous driving requires a holistic understanding of scene structure and dynamics; I study these problems at these levels: scene, object, and pixel.

Ego-motion Prediction from Scene Structure When we are driving at an intersection, we know that our choices are either to make a turn or go straight. If the car in front of us stops, we know we should stop as well to avoid a collision. Humans learn these reactions from a large amount of prior experience. This intuition inspired me to investigate whether we can learn the same reactions from large-scale driving video datasets recording both human actions and visual inputs [10]. In that work, we aim to find that a convolutional network, with an LSTM for aggregating temporal information, can learn reasonable driving behavior from large-scale driving data. The trained agent understands how to drive straight, follow curvy lanes, make turns, and stop for traffic. Most of the time we drive and react to scenes subconsciously, however, we also do careful planning for more challenging maneuvers such as lane changing. Humans go through a loop, alternating between observing and planning. Accordingly, I led a team to develop a

model for predictive control [4]. Given the previous, current observations, and a sequence of actions into the future, the model predicts scene structure evolution, represented by semantic segmentation. Then the model will select the action which would generate the best expected sequence of events. This model can be paired with the reactive model designed in Xu *et al.* [10] as an affordance model for more efficient action sampling. The predictive, affordance and temporal models are jointly learned at training time with online-exploration. The final model can learn faster than existing Reinforcement Learning models and acquire human-like driving behavior by self-exploring in simulated environments without pre-training.

3D Object Tracking Although scene structure prediction can give us a holistic understanding of the scene dynamics, individual objects still require our attention because of their diverse behaviors. Complicated object motion makes scenes hard to comprehend, requiring a large amount of training data to learn end-to-end. Therefore, I investigated whether it would make policy learning more data efficient and explainable if we modeled the object motion separately. Consequently, I led another team to work on a full pipeline for 3D vehicle tracking from a monocular camera [3]. In this work, we aim to estimate the complete 3D information of a moving vehicle, including position, dimensions, and orientations. The resulting pipeline shows us that we can not only predict 3D vehicles from sequences of monocular images, but also improve 2D tracking information by understanding the relationship with 3D occlusion. We obtained state-of-the-art results on KITTI, an established benchmark for 2D vehicle tracking, even though our pipeline was designed for estimating full 3D information.

Probabilistic Dense Correspondence Among dynamic scene understanding topics, point matching, including optical flow and stereo, has the longest research history. It is a fundamental tool in both low-level and high-level video understanding. Modeling uncertainty in this setting is a key challenge, We recently proposed the first deep probabilistic framework for learned image correspondence representation. The probabilistic framework can capture large-support match density between pixels of two frames, and it is applicable for both optical flow and stereo matching. The primary challenge is the prohibitive computational cost of the underlying high-resolution 4D cost volume. I proposed Hierarchical Discrete Distribution Decomposition (HD^3) [11] to overcome this problem. The critical observation is that we can actually decompose the large 4D volume into smaller ones in a method similar to a quadtree and learn the decomposed distribution sequentially. The resulting framework can be applied on existing feature pyramids extracted from convolutional networks and can estimate the actual match density, which gives arg max as optical flow, max as uncertainty, and probability mass transfer as region propagation. Despite its simplicity and versatility, the framework achieves competitive results on established benchmarks.

Perceptual Data: Infrastructure and Algorithms

The grand challenge of a computer vision system is not just enabling autonomous systems to perceive the scene, objects, and their behaviors, but to also digest and discover the knowledge learned from the planet-scale visual data. I approach this problem from the data and system infrastructure perspective, and develop algorithms that can process the data both at large-scale and interactively with the human in the loop.

Comprehensive Vision Benchmark Advances in algorithms are always achieved in tandem of releases of new challenging benchmarks. I have worked with others to build large-scale visual datasets for both 2D and 3D worlds. We built ModelNet [9] and ShapeNet [1] for 3D object shape analysis and recognition. We then moved from analyzing objects to whole scenes in Song *et al.* [5]. I am also interested in exploring the boundaries of algorithm evaluation at real-world scale and diversity. In LSUN [15], I downloaded 1 billion images from the Internet and released 70 million images with high annotation accuracy. Each category in this dataset can have millions of images, as opposed to the thousands found in the previous datasets. The images have inspired the machine learning community to study how to learn the manifold of natural images. The scale and density of the images in LSUN enabled this research opportunity for the first time. At UC Berkeley, I led the team to work on BDD100K, the largest driving video dataset in the research community, with video clips in diverse conditions including various weather conditions, time of the day, and types of scenes. The challenges from BDD100K have attracted hundreds of teams

to compete to develop the best perception algorithms.

Interactive Data Processing System For any model to learn, there need be clues about supervision. Babies start learning by observing and interacting with the world to train their motor control and perception. Children more actively interact with the world to acquire new knowledge. However, most existing machine learning systems and frameworks are either model agnostic or assume a static model. The real-time and dynamic architecture requirement for an interactive system is not well explored, even though it will have a significant impact on human-machine collaboration and robot exploration. I am leading a team to work on Scalabel [12], a web-based system for exploring interactive system design issues. It is already open-sourced, and several academic and industrial groups use it to label their data today. We are endeavoring to promote awareness of interactive data processing and to provide the community with a robust infrastructure foundation.

Planet-Scale Visual Understanding Finally, I extend my research to tackle the system and algorithm challenges of processing planet-scale image datasets and to understand the knowledge hidden in them. This is challenging for both existing system and algorithm development. For this purpose, an active learning pipeline was created to label the data more efficiently. The model can label 40 images per one human labeled image without compromising overall labeling accuracy. The computation and storage systems are designed based on the active learning algorithm, so that they are efficient enough to process 1 billion images under academic compute settings. The resulting dataset, called LSUN [15], has been used extensively in studying impact of visual data density and generative models, because of the sheer amount of data for each visual concept. I am expanding the work to scale up the data volume further at UC Berkeley and the new initiative is called Visual Factor Graph. It will not only label the images with human in the loop, but also look for the relationships between images in a vast image database to find breakthroughs in scene understanding, low shot learning, and active learning. Scalabel [12] is supporting this exploration in the system side and seeking new designs for dealing with planet-scale data.

Future Work

In the future, I hope to continue exploring new models for image representation, holistic understanding of scene dynamics, and automatic knowledge discovery in planet-scale data, with strong system support for delivering the models to the real world. I hope that eventually, we can develop an algorithm and system design that can digest and learn new knowledge by itself.

Joint Model Selection and Data Annotation Recently, automatic architecture search for image classification problems has attracted much research attention. One of my immediate goals is to investigate how we can automatically design network for general recognition problems, which will make such techniques more useful and lead to new insights. At a higher level, model design usually only takes a small amount of time in the whole life cycle of a model. A majority of the engineering time is spent on data collection and annotation. In the current literature, model selection and active data annotation are studied separately, though they are deeply coupled together. The model selection depends on the dataset it is trained on, and data annotation based on active learning will be affected by the bias of the underlying model. I want to study how to optimize the two procedures jointly. The selected model should be more robust for different samples of data, and the annotated data should be useful for a wide range of models.

Dynamic 3D Scene Understanding for Tracking and Prediction I aim to obtain a holistic view of the 3D scene, including appearance, geometry, motion, and context. I will try to combine the dynamics at different levels together to infer static structure, ego-motion, object motion, and region propagation jointly by aggregating info from different sensors and prior knowledge. This will become the meta representation of the dynamic 3D scenes and provide sufficient information for perception and subsequent decisions. There is still a long way to go, but the resulting understanding can have a high impact on the advance, safety, security, and interpretability of future intelligent systems.

Cloud Brain for Planet-Scale Intelligent Systems Unlike humans, machines can easily store and exchange data. The aggregated data may create knowledge superseding human capabilities, but compre-

hending it requires systems that can share that information efficiently and algorithms that can extract knowledge from the data. LSUN [15] was my first step towards solving this problem. As mentioned above, my next step is towards Visual Factor Graph models, which connect each image to a large network of visual factors defined by object names, scene types, actions, interactions, and prototypical examples. This will be an efficient learning problem backed by an existing vast knowledge base. The primary barrier is to scale up such systems to use fewer samples and less computation, similar to how humans learn new concepts.

References

[1] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Yi Li, and Fisher Yu. ShapeNet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[2] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. PairedCycleGan: Asymmetric style transfer for applying and removing makeup. In *CVPR*, 2018.

[3] Hou-Ning Hu, Qizhi Cai, Dequan Wang, Ji Lin, Min Sun, Philipp Krähenbühl, Trevor Darrell, and Fisher Yu. Joint monocular 3d vehicle detection and tracking. *arXiv preprint*, 2018.

[4] Xinlei Pan, Xiangyu Chen, Qizhi Cai, John Canny, and Fisher Yu. Semantic predictive control. *arXiv preprint*, 2018.

[5] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 190–198. IEEE, 2017.

[6] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *SSW*, page 125, 2016.

[7] Xin Wang, Fisher Yu, Zi-Yi Dou, and Joseph E Gonzalez. SkipNet: Learning dynamic routing in convolutional networks. In *ECCV*, 2018.

[8] Xin Wang, Fisher Yu, Ruth Wang, Yi-An Ma, Azalia Mirhoseini, Trevor Darrell, and Joseph E Gonzalez. TAFE-Net: Task-aware feature embeddings for efficient learning and inference. *arXiv preprint arXiv:1806.01531*, 2018.

[9] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d ShapeNets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[10] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. *CVPR*, 2017.

[11] Zhichao Yin, Trevor Darrell, and Fisher Yu. Hierarchical discrete distribution decomposition for match density estimation. *arXiv preprint*, 2018.

[12] Fisher Yu. Scalabel: A Web-based Annotation System for Low-Latency Distributed Human-Machine Collaboration. `https://www.scalabel.ai`, 2018.

[13] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.

[14] Fisher Yu, Vladlen Koltun, and Thomas A Funkhouser. Dilated residual networks. In *CVPR*, volume 2, page 3, 2017.

[15] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[16] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *CVPR*, 2018.