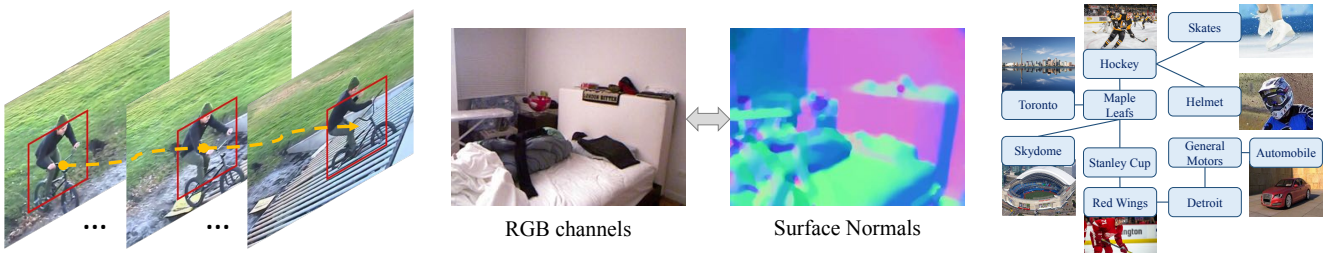
Research Statement – Xiaolong Wang

xiaolonw@cs.cmu.edu – http://www.cs.cmu.edu/~xiaolonw/

In recent years, the field of computer vision has been completely transformed by the success of deep neural networks. A key ingredient behind this success is a large amount of human supervision for training deep networks. Unfortunately, this key ingredient also turns out to be the biggest bottleneck: the number of labels is limited by the high cost of human labor. As computer vision works towards more difficult and structured AI tasks, it becomes more challenging for humans to provide training supervision. Given this situation, we ask the question: is there any information in the data we have not fully utilized to guide learning?



(a) Structure in Video Data (b) Structure in RGB+Depth (RGBD) Data (c) Structure in Semantic Labels

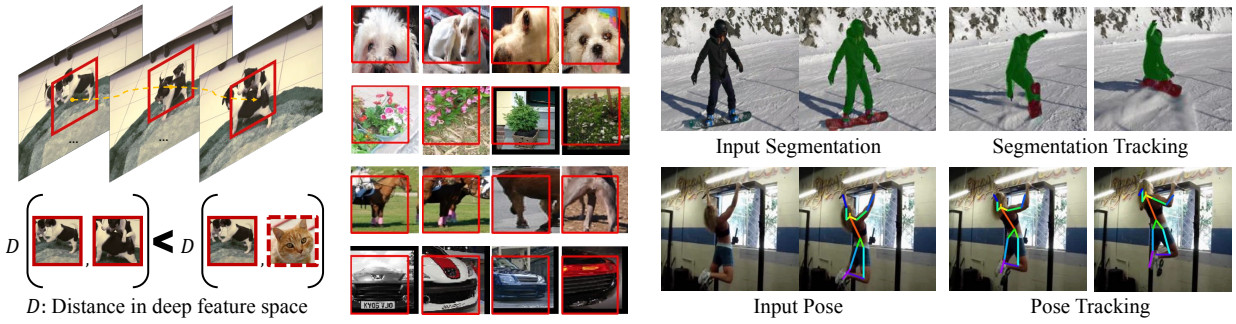
Figure 1: Both visual data and semantic labels are highly structured: (a) The same objects transform smoothly through different frames in a video; (b) RGB channels and surface normals (computed from depth, we use green for horizontal surface, blue for facing right and red for facing left) are highly correlated in RGBD data; (c) In a knowledge graph, different semantic labels can have shared components and appearance (e.g., Helmet and Hockey).

Interestingly, most of the visual world is highly structured. There is a lot of redundant information in the visual data and semantic labels. For instance, the information in a video is smaller than the set of its pixels because the same objects appear over time (Fig. 1 (a)). Similarly, if we look at RGB+Depth (RGBD) data from the Kinect (Fig. 1 (b)), the RGB channels and the surface normal map (computed from depth channel) are highly correlated as they share the same underlying image structure. In addition to visual data, there is also structure in semantic or label space. For example, there might be different concepts or labels related to "Hockey" in a knowledge base (Fig. 1 (c)), and yet many of them share similar visual components and appearance.

Building on these observations, *my research focuses on exploiting different forms of structure in data for learning visual representations.* The structure here includes the spatial-temporal structure in videos, the 3D structure in RGBD data, and the semantic structure in knowledge bases. There are two principal directions I have explored: first, to use the structure information from the data itself as a supervisory signal for learning visual representations (i.e., self-supervised learning), eliminating the need for manual labels; second, to explicitly model the structure in data via relationship reasoning for visual recognition and interaction. I will explain my research on these two directions as follows.

(1) Self-supervised learning for generalization. To break the limitations of human supervision, I have been working on utilizing the structure and redundant information in data as supervisory signals to train deep networks. In this way, we can scale the algorithm to an unlimited amount of unlabeled data and potentially generalize better in various tasks. One important supervisory signal is time. Following time in a video, we can observe the same object changes smoothly in a sequence. By tracking a sequence of views of the object, it gives us the signal to learn view-invariant representations [1, 2] (Fig. 1 (a)), as well as find correspondence between views [3]. Another important signal is the 3D structure of the world. For example, in RGBD data (Fig. 1 (b)), I have trained deep networks to reason about the correlations between the RGB image and the corresponding surface normals [4, 5] for 3D scene understanding.

(2) Relationship reasoning for recognition and interaction. Most current approaches for visual tasks feed all the raw data into a simple deep network and attempt to learn everything (including the structure) implicitly. While this has brought us many successes, my work has shown that explicit reasoning through the structure in data leads to improved performance and new abilities. Specifically, I have worked on modeling relationships between repeating and correlating patterns in space and time to improve video recognition [6, 7, 8, 9]. Besides visual data, I have also utilized the structure in label space (Fig. 1 (c)) to improve recognition. By reasoning about the relationships between semantic categories, I have shown the possibility to generate tens of thousands classifiers without seeing any training images for them [10]. Going beyond recognition, relationship reasoning in semantics also facilitates interaction. I have created an agent which uses a semantic knowledge base to help it navigate and actively interact with scenes [11].



(a) Unsupervised Tracking (b) Top Activations (c) Video Propagation with Self-Supervised Representation

Figure 2: Self-supervised learning from time: (a) Learning from tracking through time; (b) Each row shows top 4 patches that maximally activate one neuron in the self-supervised network with (a); (c) Tracking results given segmentation or pose skeleton on the first frame.

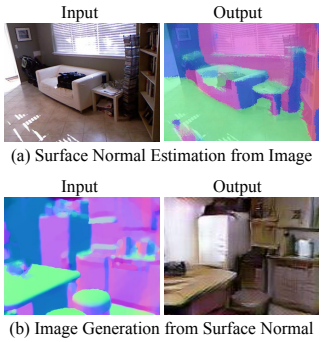
1 Self-supervised learning for generalization

Training Convolutional Neural Networks (ConvNets) with large-scale semantic category labels has offered researchers a generic representation that can be applied to different vision tasks. However, there are two shortcomings in the supervised training paradigm: First, we cannot annotate every example in the world, which limits the generalization ability of the representation. Second, detailed understandings of objects and scenes (e.g., human motion and 3D layout) are hard to be directly derived from semantics. In order to obtain representations that generalize better, we must search for other forms of supervision besides semantic labels. Towards this goal, I have proposed and worked on two types of supervision based on the implicit structure in videos and 3D data, and trained ConvNets in a self-supervised manner without manual annotations.

Learning from tracking through time. I have explored the use of temporal structure in videos to train visual representations (ICCV 2015 [1]). The key idea is to use visual tracking: two patches in the same track should have similar deep feature representations given that they might be the same object with different views or deformations (Fig. 2 (a)). I designed a system to extract millions of tracks from 100K YouTube videos for training. Surprisingly, even without any labels, semantics emerge after training. As shown in Fig. 2 (b), each row shows the top 4 patches that maximally activate one neuron in our network. This is one of the *first* works showing that we can train a ConvNet with a standard architecture in a self-supervised manner. Moving ahead, I have also worked on combining the tracking signal with another self-supervised signal in training (ICCV 2017 [2]). My latest results show that the learned representation transfers better to 3D understanding tasks than the representation trained with large-scale semantic labels. These observations suggest the great potential of the generalization ability of self-supervised representations.

Learning dense correspondence from time. The value of self-supervised learning also lies in scaling up learning for tasks where human annotations are hard to obtain. In collaboration with Prof. Alexei Efros's group, we developed a self-supervised framework for learning deep spatial features that enable finding dense correspondence between frames separated far in time [3] (i.e., long-range flow). Note that this is extremely difficult for human to label, because annotating which pixel in one frame corresponds to a pixel in another frame is labor intensive. Our model learns to track spatial features back and forth through time, relying on the inherent cycle-consistency of events in time for self-supervision (i.e., consistency of the starting query and ending result). The acquired representations can be applied across a wide range of visual correspondence tasks (e.g., tracking segmentation and pose in Fig. 2 (c)) *without* any further training on the target dataset. These results show the encouraging generalization ability of self-supervised learning on tasks where human labels are hard to obtain.

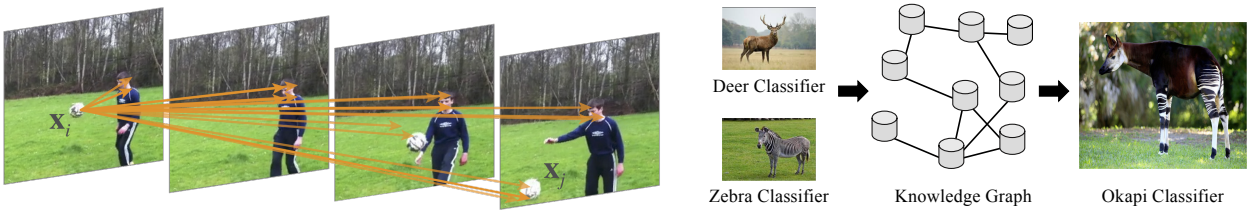
Learning from 3D. Aside from temporal signals, I have also investigated using depth extracted from sensors such as Kinect to learn visual representations. Given the RGBD data, we can split it into RGB channels and the depth channel. I explored the correlations between the RGB channels and the corresponding depth map by learning a mapping from RGB images to 3D surface normals (Fig. 3 (a)). Inspired by early work in geometric reasoning, I proposed a two-stage ConvNet which incorporated mid-level and high-level constraints for surface normal prediction. This work (CVPR 2015 [4]) is one of the *first* attempts of training a ConvNet for surface normal estimation and 3D scene understanding. Besides, I have also worked on generating RGB images conditioned on surface normals, using conditional Generative Adversarial Networks (ECCV 2016 [5]). This work is one of the early works in image-to-image translations. I have not only shown better results in image generation (Fig. 3 (b)), but also obtained RGBD representations for recognition tasks in 3D.



(a) Surface Normal Estimation from Image

(b) Image Generation from Surface Normal

Figure 3: Learning from 3D.



(a) Spacetime Non-local Operation (b) Generate Novel Classifier with Knowledge Graph

Figure 4: (a) Non-local Neural Networks for human action recognition; (b) Zero-shot recognition with knowledge graphs.

2 Relationship reasoning for recognition and interaction

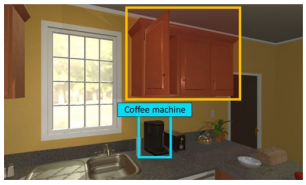
Modeling the structure of the data also plays an essential role in visual semantic tasks where human annotations are required. I have designed network modules to explicitly model and reason about the correlated information in the data. By adding this explicit reasoning instead of hoping that reasoning will happen implicitly in a simple deep network, the model can utilize the data much more efficiently, which can lead to better performance and even help low-shot or zero-shot learning (where training data is very limited). I have conducted research on modeling the relationships between repeating patterns across space and time in videos, as well as semantic relationships in knowledge space. I organized these relationships in graph structure and performed reasoning via graph neural networks.

Space and time relationships for video recognition. In neural networks, both convolutional and recurrent operations can only process one local neighborhood at a time. In order to overcome this, I have introduced non-local operations as neural network modules (CVPR 2018 [6]), which capture the long-range dependencies between repeating, correlated patterns in videos. Specifically, this non-local operation updates the feature at a position as a weighted sum of the related or similar features at all positions across space and time (Fig. 4 (a)). By embedding the non-local operations in neural networks (namely, Non-local Neural Networks), the model achieved state-of-the-art performance in the task of human action recognition. Since I released the code to the public, researchers have been applying the method to different tasks including image recognition, image synthesis, defense against adversarial attack, and reinforcement learning. Motivated by this work, I have also built a method that uses a space-time region graph, which takes object regions as nodes and connects them across space and time (ECCV 2018 [7]). By reasoning with graph neural networks on object level concepts in an explicit way, it not only leads to better performance but also makes the model more explainable.

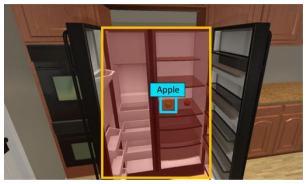
Knowledge graph reasoning for large-scale image recognition. In addition to images and videos, semantic labels are also highly structured and correlated among each other. For example, the animal okapi can be described as "a zebra-striped four-legged animal with a brown torso and a deer-like face." Given this description, even if we have not seen an okapi, we can still recognize it. I mentored a Master's student on a project where we used both semantic embeddings and the categorical relationships in knowledge graphs to describe categories (CVPR 2018 [10]). Given existing classifiers (e.g., deer and zebra) with training examples, we can generate novel visual classifiers (e.g., okapi) in a zero-shot setting (Fig. 4 (b)). This allows us to scale up the classification to tens of thousands of classes without requiring new training examples.

Knowledge graph reasoning for interaction. Going beyond recognition, knowledge-based reasoning also has great potential in visual interaction. I mentored a Ph.D. student on a project that used a knowledge graph to guide an agent to perform semantic navigation (i.e., finding a semantic object by navigating to it) in a virtual indoor environment [11]. The provided knowledge includes the relative positions and co-occurrence between semantic objects. For example, in Fig. 5 (a), even without seeing a mug, thanks to the reasoning in the knowledge graph, the agent knows that the likely location in which to find a mug is the cabinet near the coffee machine. Similarly, in Fig. 5 (b), knowing that apples are usually stored inside of a refrigerator, the agent can find the apple easily even though it is small and hidden.

In our paper [11], knowledge reasoning is performed via a graph neural network, trained together with the policy decision network with reinforcement learning. As the agent navigates in the room, the graph network helps the agent update its understanding of the observations and make decisions. We showed that navigation policies can generalize to novel scenes and unseen targets with knowledge reasoning, which has rarely been addressed in reinforcement learning. This is my first step towards using knowledge to guide interactions with the environment.



(a) Semantic target: "Mug"



(b) Semantic target: "Apple"

Figure 5: Semantic navigation.

(a) Training Data Collection (b) Predicted pose from the trained model

Figure 6: We design a system to observe how human interacts with the repeating scene in TV series.

3 Future Research Agenda

My aim for research is to build an AI system that can scale up its learning ability beyond human supervision, acquire common sense knowledge, and interact with the world using the knowledge it has learned. Importantly, more knowledge can be acquired throughout the interaction. There is still a long way to go before I reach this goal, and below I outline three topics that I plan to pursue.

Self-supervised learning. Going beyond the limits of semantic supervision, I believe self-supervised learning has a great potential to produce richer representation and allow for learning at a much larger scale. I see two possible goals for self-supervised learning.

- Obtain a *universal representation*. We can utilize multiple sources of structure information in data as signals to learn a *universal representation*, which can be generalized to every task. One potential way is to combine different signals from low levels (e.g., motion and boundary) to high levels (e.g., physical and functional properties) with curriculum learning. Intuitively, the high-level structure would be easier to learn given the emergence of low-level representation.
- Obtain *task specialized representations*. The goal is to learn representation for tasks where human annotations are hard to come by. There are already encouraging results in this direction (e.g., my work on finding dense correspondence in videos [3]). However, the applications are still restricted to a few areas. I plan to explore task-specific, self-supervised learning on a large range of visual problems. Across these tasks, I believe there will be a shared principle in algorithm design.

Video representation learning. While deep learning has made great strides in image recognition, it is still struggling in video understanding. In my experience in human action recognition [6, 7, 8, 12], I have observed two problems in the field: (i) first, the supervision from classification is not necessarily correct in capturing temporal information–for example, to classify the action "swimming," the model does not need to capture the motion but recognize the water; (ii) second, the scale of training examples is much smaller than image datasets, due to the cost of labels. These problems require us to search for other supervisions. One direction in which I plan to work is looking for training signals for learning spatial-temporal representation in a self-supervised manner, which has been rarely explored so far. The other direction is to re-define the problem itself. I plan to design a dataset which requires the model to reason about the cause and effect of the action through time to solve the task. One possible direction is to disentangle the appearance from the target tasks in the dataset. In this way, the network will be forced to learn better abstraction beyond semantics.

Common sense knowledge and interaction. I would like the AI system to have a deeper understanding of how objects can be used and the functionality of scenes and how humans interact with them (i.e., affordance). As a first step, I created a system which goes through different TV series and observes how humans interact with the scenes (CVPR 2017 [13]). As illustrated in Fig. 6 (a), the system performed human pose estimation on every frame, and collected all the human poses in the same scene together. With this data, I trained a deep network which can predict how humans interact with a scene (Fig. 6 (b)). Inspired by this, I plan to make the system observe how humans interact with each other in the videos. My goal is to obtain the common sense, including the intentions of the actions and even the social rules behind the behaviors. I believe my experience in video understanding will also help towards this goal.

However, the process of acquiring common sense knowledge should not only be in a passive manner. I plan to create an agent to move around and perform interactions. By reasoning with the interactions, new knowledge can be absorbed. I have already shown navigation with semantic knowledge base in [11]. Moving forward, I have three goals in this direction: (i) First, design agents which not only rely on semantic knowledge but also on richer information including affordances and the common sense of human behaviors, which can be acquired from videos. (ii) After interacting with the environment and observing the changes, we will allow the agent to update its knowledge base. Potentially, the updated knowledge base can then offer better guidance for future interactions. I plan to integrate the knowledge updates and the evolution of actions into an iterative learning procedure. (iii) Most importantly, my ultimate goal is to make this agent work in the real world beyond simulations, where I see many opportunities for collaboration with Robotics and NLP.

References

[1] X. Wang and A. Gupta, "Unsupervised learning of visual representations using videos," in *ICCV*, 2015.

[2] X. Wang, K. He, and A. Gupta, "Transitive invariance for self-supervised visual representation learning," in *ICCV*, 2017.

[3] X. Wang*, A. Jabri*, and A. A. Efros, "Learning correspondence from the cycle-consistency of time," in *Submission to CVPR, 2019*.

[4] X. Wang, D. F. Fouhey, and A. Gupta, "Designing deep networks for surface normal estimation," in *CVPR*, 2015.

[5] X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," in *ECCV*, 2016.

[6] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *CVPR*, 2018.

[7] X. Wang and A. Gupta, "Videos as space-time region graphs," in *ECCV*, 2018.

[8] X. Wang, A. Farhadi, and A. Gupta, "Actions $\sim$ Transformations," in *CVPR*, 2016.

[9] Y. Yuan, X. Liang, X. Wang, D.-Y. Yeung, and A. Gupta, "Temporal dynamic graph lstm for action-driven video object detection," in *ICCV*, 2017.

[10] X. Wang*, Y. Ye*, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," in *CVPR*, 2018.

[11] W. Yang, X. Wang, A. Farhadi, A. Gupta, and R. Mottaghi, "Visual semantic navigation using scene priors," in *Submission to ICLR, 2019*.

[12] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *ECCV*, 2016.

[13] X. Wang*, R. Girdhar*, and A. Gupta, "Binge watching: Scaling affordance learning from sitcoms," in *CVPR*, 2017.