Research Statement
Huaizu Jiang

Overview

My research is driven by the desire to build intelligent agents that can interact with their surroundings and other peers. To this grand end, I have mainly focused on teaching machines to develop visual intelligence, an essential ability an organism has for understanding and interpreting the dynamic visual world. In the long term, my research aims to teach machines to develop visual intelligence in a manner analogous to humans by drawing inspiration from humans’ visual perception capabilities. In the short term, my research goal is to create smart visual perception tools to improve people’s life experiences of using cameras.

Human beings have the remarkable capability to learn from limited data, with partial or little annotation, in sharp contrast to computational vision models that rely on large-scale, manually labeled data. This capability to learn from sparse data may stem in part from our ability to interpret the dynamic visual world holistically. When interpreting a scene, people simultaneously infer many properties such as depth, motion, location, and the semantic categories of different elements. Solving one task is often helpful in solving others; for example, a car’s motion is likely to be rigid while a person’s movement is non-rigid. Thus, motion could provide a cue for object identification, or conversely, knowledge of the object could help us interpret the motion. Computational models of perception, however, tend to focus on one task at a time, and benchmark datasets are typically designed to evaluate individual tasks. Therefore, one of my major research directions is to investigate how to teach machines to perceive and interpret our dynamic visual world holistically in a manner analogous to humans [1, 4, 6, 7].

Even with a quick glance, we humans can describe a visual scene in great detail, which is still far beyond the capability of state-of-the-art computational perception models. Human language is not only an informative tool for communications about observations of the dynamic visual world but also provides labels for visual perception tasks. Correctly describing a visual scene or answering questions about the content requires a set of visual perception capabilities. For example, action recognition is needed in order to answer the question, “what is the person doing?” Moreover, unlike closed vocabularies used in most existing benchmark datasets, free-form descriptions and question answering of the visual scene, consisting of nouns and verbs, provide enormous class labels for a great variety of visual concepts. Thus, another one of my major research directions is to study how to teach machines to develop visual intelligence using natural language as a scaffold [3, 9].

In addition to addressing basic research questions, my research is also motivated by real applications. Advanced visual perception models have been already deployed to serve millions of people each day, like detecting faces when we take photos on a cell phone. I am passionate about such applied research that provides opportunities to improve our daily lives in the near future. In specific, I am enthusiastic about building tools that help people more easily record and share their daily life experiences [5].

Such multi-disciplinary research spans the areas of computer vision, natural language processing, computational photography, and machine learning. In the following sections, I will introduce my recent advances in these directions and also briefly present my future research plans, highlighting the key challenges.

Understanding the Dynamic Visual World: From Motion to Semantics

Reliance on strongly supervised models with manually labeled data inherently prohibits us from modeling the dynamic visual world, as manual annotations are tedious, expensive, and
not scalable, especially if we would like to solve multiple scene understanding tasks at the same time. Even worse, in some cases, manual annotations are completely infeasible, such as the motion vector of each pixel (*i.e.*, optical flow) since humans cannot reliably produce these types of labeling. In some other cases, the data we use to train a model is quite different from the data on which we are going to use the model. Consider running a face detector trained using images from the web on surveillance videos, for example. Such domain gaps often yield inferior results.

Motion information contained in real-world videos, as a result of moving camera, independently moving objects, and scene geometry, consists of abundant information, revealing the structure and complexity of our dynamic visual world. My work has shown that using unlabeled videos, in either a self-supervised or semi-supervised manner, dramatically reduces reliance on manual annotations, leading to a more precise understanding of the dynamic visual world.

**Self-Supervised Relative Depth Learning.**

When moving through the world, it is natural for an intelligent agent to associate image patterns with the magnitude of their displacement over time: as the agent moves, far away mountains don’t move much; nearby trees move a lot. This natural relationship between the appearance of objects and their apparent motion is a rich source of information about the relationship between the distance of objects and their appearance in images. In [1], I designed a pretext task of estimating the relative depth of elements of a scene (*i.e.*, ordering the pixels in an image according to distance from the viewer), as shown in Fig. 1. The goal of this pretext task was to induce useful feature representations in deep Convolutional Neural Networks (CNNs). These induced representations, using 1.1 million video frames crawled from YouTube within one hour without any manual labeling, provide valuable starting features for the training of neural networks for downstream tasks. It is promising to match or even surpass what ImageNet pre-training gives us today, which needs a huge amount of manual labeling, on tasks such as semantic image segmentation as all of our training data comes almost for free.

**Holistic Scene Flow Estimation.** As we humans look around, we do not solve a single vision task at a time. Instead, we perceive our surroundings in a holistic manner, doing visual understanding using all visual cues jointly. By simultaneously solving multiple tasks together, one task can influence another. In [4], I proposed a neural network architecture, called SENSE, which shares common feature representations among four closely-related tasks: optical flow estimation, disparity estimation from stereo, occlusion estimation, and semantic segmentation, as displayed in Fig. 2. The key insight was that sharing features makes the network more compact and induces better feature representations. For real-world data, however, not all annotations for four

![Figure 1: RGB images and their estimated relative depth maps](image1.png)

![Figure 2: In [4], given video frames from two stereo cameras, four closely-related tasks were solved simultaneously: optical flow estimation, disparity estimation from stereo, occlusion estimation, and semantic segmentation.](image2.png)
tasks mentioned above are always available at the same time. To this end, I designed loss functions, which exploit interactions of different tasks and do not need manual annotations, to better handle partially labeled data in a semi-supervised manner, leading to superior understanding performance of the dynamic visual world. This work was selected as an Oral presentation at ICCV 2019, where the acceptance rate was 4.6%.

Improved Object Detection From Unlabeled Videos.
Modern object detectors can localize objects from different categories in a single image, such as human faces [2]. Yet detectors still make mistakes even with high confidence, which are called hard examples. Moreover, the domain gap between training data and testing data often leads to inferior detection results. How can we improve the object detector? Instead of manually annotating more data to re-train the model, which is expensive and tedious, my colleagues and I used temporal consistency in video frames to identify those hard examples automatically in [6], as shown in Fig. 3. The object detector is then fine-tuned using automatically mined hard examples. Better object detection performance can be achieved on faces, pedestrians, and other categories [6], even when the testing data is significantly different from the training data [7].

Emergent Visual Perception via Linguistic Explanations and Descriptions
Natural language, as an effective tool, is often used to explain and describe what people perceive about the visual world. When explaining visual scenes in terms of question-answer pairs (i.e., Visual Question Answering) and describing a visual scene using natural language, various visual perception capabilities are needed, such as human action recognition and object localization as displayed in Fig. 4. My research has demonstrated that visual perception skills emerge when training an agent to explain and describe visual scenes using natural language.

Rethinking Bottom-Up Attention for Visual Question Answering. For Visual Question Answering (VQA), when asking questions about a scene, people tend to capture unique and salient parts. Bounding boxes (or regions) provide a natural way to identify such salient parts in the input image. Consequently, region-based visual features often referred to as bottom-up attention, have dominated joint vision and language understanding tasks like VQA. However, it requires expensive bounding box-level annotations to train an object detector in advance. In [3], I found that grid convolution features work as well as region features under the same pre-training budget (using region-level annotations), indicating learned feature representations, instead of the feature format (using regions or not), is the key for better VQA accuracy. The value of using grid features for VQA is that it opens a new gate of discarding expensive region-level annotations for pre-training of VQA models. Using large-scale, weakly supervised web images with noisy hashtags for pre-training achieved comparable VQA accuracy to an object detector pre-trained with region-level annotations. Remarkably, without using any region-level annotations, visual perception capabilities naturally emerged during training of the model merely with the supervision of VQA accuracy. For in-
stance, as shown in Fig. 4, the VQA model learned to correctly localize human actions and objects without having access to such fine-grained annotations.

**Describing Fine-Grained Visual Differences.** Short sentences using natural language (a.k.a., attribute phrases) provides an intuitive way to describe differences of a unique semantic visual property between two objects of the same category, which are otherwise expensive to annotate, such as using paired bounding boxes. In [9], my colleagues and I investigated using attribute phrases to describe visual differences between paired instances from different models of aircraft, as shown in Fig. 5. In such a setting, reasoning about attribute phrases is a practical way of discovering and grounding describable attributes for fine-grained categories. As a result, more discriminative representations emerged for fine-grained visual recognition tasks, leading to better accuracy than hand-crafted annotations such as a set of pre-defined attributes.

**SuperSloMo: Synthesizing Slow-Motion Videos with a Plain Camera**

There are many moments in our lives that we might want to record with a camera in slow-motion because they are either memorable or otherwise hard to see clearly with our naked eyes: the first time a baby walks, a difficult skateboard trick, a dog catching a ball, etc. While it is possible to take 240 frame-per-second (fps) videos with a cell phone nowadays, professional high-speed cameras are still required for higher frame rates. Besides, many of the moments we would like to slow down are unpredictable, and as a result, are recorded at standard frame rates. Recording everything at high frame rates is impractical—it requires large memories and is power-intensive for mobile devices.

![Figure 6: SuperSloMo converts a plain video into slow-motion by synthesizing multiple intermediate frames (two are shown here enclosed with red color). More demo videos are available at https://youtu.be/MjViy6kyiqa and https://tinyurl.com/vlrp8br.](image)

In [5], I developed a system called SuperSloMo. It takes two consecutive frames of a plain video (e.g., 30-fps) to synthesize multiple intermediate frames to convert it into a slow-motion version with arbitrary higher frame rates, such as 240-fps or 1080-fps, in a single run without human intervention. An example is shown in Fig. 6. Such a technique helps compress and transmit videos (only keeping and sending key video frames and discarding the rest as they can be later synthesized). It is potentially useful for virtual reality (VR) devices as well to generate smooth view transitions as well. **SuperSloMo was accepted as a Spotlight presentation at CVPR 2018, where the acceptance rate was 6.7%**. It has received a lot of media coverage and was recognized as one of the ten coolest papers in CVPR 2018.¹ It has also attracted attention from industry, including companies like ESPN, Adobe, Google, etc. A third-party implementation has received 2000 stars on GitHub.²

²[https://github.com/avinashpaliwal/Super-SloMo](https://github.com/avinashpaliwal/Super-SloMo)
Future Research Plan

In the future, to fulfill the desire to create intelligent agents, I plan to continue explorations of a holistic understanding of dynamic visual world, joint reasoning of visual and linguistic data, and enhancing user experiences of recording and sharing life experiences. I am also passionate about investigating new directions. There are several key challenges to be addressed, summarized as follows.

**Holistic 3D Scene Understanding of Dynamics and Interaction.** We live in a 3D visual world, which is constantly in motion. My goal is to build a holistic 3D scene understanding model unifying appearance, geometry, motion, and semantics. Such holistic model will not only allow us to infer various scene dynamics, including depth, ego-motion, object motion, semantic parsing, etc, but also incorporate interactions of agents in the scene. Interactions of agents are crucial for developing intelligence. An important task for kids to learn in kindergarten, for example, is how to play together with others. Investigating agents’ interactions is beneficial for applications as well. In an urban traffic scene, An autonomous driving vehicle must predict intent and motion trend of other (autonomous) vehicles, pedestrians, and bicyclists/motorcyclists. Such a 3D holistic scene understanding model will have a huge impact on the advance and safety of next-generation intelligent systems.

**Enhancing Photography Experiences in the Mobile Era.** Taking photos has never been so easy with smartphones. Recording VLOGs (video blogs) and sharing them on social media is a fashion living style nowadays. It imposes great challenges as well as opportunities for researchers to improve users’ photography experiences on cell phones. On the one hand, photos and videos casually shot by non-expert users are usually less visually appealing. Photos or videos may be tilted and look too dark or blurry. On the other hand, the latest smartphones are usually equipped with powerful computational resources and cameras, which allow running of powerful visual perception models to provide instant assistant to users. Following the effort of synthesizing slow-motion videos [5], and from a broader perspective of deep generative models, I am passionate about developing algorithms that enhance users’ photography experiences to more easily record and share their daily lives, for example, via VLOGs.

**Beyond Perception: Visual Reasoning with Mined Knowledge from Text.** As a hallmark of human intelligence, visual reasoning leads to decisions by connecting evidence obtained from perception, which is crucial for the development of visually intelligent agents that can be useful in dynamic and novel environments. The acquirement of *commonsense knowledge* plays a critically important role in visual reasoning. We use commonsense knowledge every single day in our daily lives. Suppose we are in an unfamiliar Airbnb apartment, as shown in Fig. 7. We are in the hallway and want to find a TV. Which directions should we choose? Commonsense tells us that a TV is more likely to be near a couch, so we might decide to turn right that leads us to find a TV quickly. In [8], my colleagues and I demonstrated such commonsense co-occurrences knowledge can be mined from unlabeled images. Alternatively, human language, in the form of text like books, for example, also conveys abundant background
knowledge. In fact, such prior knowledge from visual and linguistic data are complementary: it is easier to learn physical commonsense knowledge from visual data such as a coin can not stand up on its own; linguistic data provide factual knowledge (e.g., lemons are sour) and knowledge of social protocols (e.g., people usually bring presents to a birthday party). Investigating how to harvest knowledge jointly from visual and linguistic data, either from structured data like knowledge bases or unstructured data such as (unlabeled) images and text corpus, is a promising direction that I would like to explore in the future.

**Toward Understanding Pre-training of Deep CNNs.** My research [1] has demonstrated that pre-training of deep CNNs, either self-supervised or supervised, leads to better accuracy on a downstream task than a random initialization, as shown in Fig. 8. Particularly, when the number of training samples in a downstream task is limited, such pre-training is critical. We know training of deep CNNs often leads to a reasonably good and stable local minimum. Little is known so far, however, about the dynamics of the optimization process given different initialization. I am interested in analyzing the behaviors of such optimizations and furthermore taking inspirations to design better initialization schemes. After all, pre-trained weights are essentially equivalent to a better initialization of the CNNs once the pre-training is finished. It is impactful on broad fields, including machine learning, computer vision, and natural language processing.

**References**


