Research Statement

Fanyi Xiao (fyxiao@ucdavis.edu)
University of California at Davis

1 Introduction

Humans perceive and interact with the visual world by processing an input "video stream" captured by our eyes. In addition to the RGB pixels presented in each "frame", our brain also effectively utilizes rich information like motion and audio signals presented in this video stream. Therefore, in order to enable human-like visual perception capability (such that we can enable applications like robot navigation in a fresh environment), it is vital to develop algorithms that can effectively understand videos. Moreover, according to Cisco Visual Networking Index, videos constitute the largest chuck of Internet traffic – 75% of all Internet traffic in 2017 and predicted to grow to 82% by 2022. Therefore, models that can understand rich semantics (e.g., recognizing actions/activities, localizing humans/objects, inferring intents and predicting future, etc.) in videos will have huge practical impact.

The computer vision community has witnessed great advances in the past few years on visual perception. The progress is in large propelled by the advent of large-scale datasets like ImageNet for image classification, MS-COCO for instance segmentation and ADE20K for scene understanding. Despite their contributions to enable large-scale supervised training, these datasets are mostly based on *static images* and therefore steer significant research efforts towards visual understanding on images. However, as mentioned before, humans do not perceive the visual world by parsing a set of independent snapshots, and this is true in almost all scenarios in which we would like an AI agent navigate and interact with the environment using its visual capability – it's almost certain that the agent would receive a video stream as input instead. Thus, an image-based algorithm would most likely deliver suboptimal results as 1) they ignore rich information like motion and audio completely; and 2) they suffer from a domain gap if trained on images and tested on videos (e.g., videos have specific challenges like motion blur which does not occur in images).

With this motivation, I have dedicated my PhD to researching effective video understanding algorithms and proposed several state-of-the-art approaches for various tasks ranging from video segmentation [1], action detection [2] (utilizing the idea of *progressive learning* to iteratively expand/refine results spatially and temporally), video object detection [3] (temporal propagation via an aligned spatial-temporal memory) and audiovisual video recognition [4] (hierarchical audiovisual fusion with training dynamics matching). Furthermore, as I accumulated more experiences working with videos, I realized the huge potential of videos to serve as a source of self-supervision to train visual models. For this research direction, I have done several work exploring different aspects of videos that could be utilized to provide supervision – using synchronization between audio and visual clips to learn video representations [4], learning from videos to disentangle visual factors for image generation [5] and training weakly-supervised object detectors using motion [6]. Finally, beyond videos, I'm also interested in exploring multimodal learning – vision + X, where X can be language, attributes, etc. – to exploit the structures in other modalities for better visual understanding [7, 8].

2 Video Understanding

Progressive learning in video understanding. Videos are sequence data in nature and therefore we can exploit its sequential structure to ease the task of learning algorithms. For example, the goal of video object segmentation is to produce a spatiotemporal mask tube of foreground objects (Fig. 1 left). A major challenge to achieve this is the large appearance variation of objects across time (e.g., rear view of a person changing to its side view). To overcome this, in [1], we propose to start from *key-frames* (i.e., frames on which it is easier to detect foreground objects) and *progressively* expand across time while iteratively updating the object's appearance model [1]. By progressively learning to segment foreground objects, our method achieved state-of-the-art performance on unsupervised video segmentation and this work has been used as the foundation for my other weakly-supervised object detection work [6].

Specifically, our method takes in an unlabeled video as input and outputs spatiotemporal mask tubes for foreground objects [1] (Fig. 1 left). I noticed that previous video segmentation methods either perform global clustering of foreground/background regions which is challenging due to the large appearance variation across time, or resort to tracking-based approach which is prone to the "drifting" problem (i.e., gradually accumulating localization errors and eventually miss the target object). By progressively detecting and segmenting objects, the proposed approach avoids challenges facing both global clustering based approach – as appearance variation between neighboring frames is minimized – as well as tracking based method – as we start from a diverse set of key-frames (this contrasts tracking which

Figure 1: **Progressive learning for video understanding**. Left: Inputs and outputs of video object segmentation [1]. Right: Progressively refine action tubelets both temporally and spatially for action detection [2].

always starts from the first frame).

Progressive learning is a general idea that can be applied in many other video understanding tasks as well. Following up on [1], with collaborators we applied progressive learning in the domain of action detection [2] (Fig. 1 right). In this work, we not only progressively expand on the temporal axis, but also train our models to progressively refine spatial localization of persons. With a small number of initial proposals, we are able to refine them temporally and spatially to produce state-of-the-art results on action detection.

Learning temporal continuity for *better* video understanding. Compared to image understanding, video understanding approaches face unique challenges (e.g. motion blur and camera defocus). However, at the same time, videos present abundant opportunities for algorithms to produce *stronger* results with *faster* runtime. One of such opportunities comes from the utilization of *temporal continuity* (e.g., if a hamster occurs in a video frame, it's likely that it will still be there in the next frame, see Fig. 2 left). To effectively model temporal continuity in video object detection, we proposed the first end-to-end video detection architecture with an aligned spatial-temporal memory to propagate information across time [3]. Our detector achieves state-of-the-art performance on the standard ImageNet VID benchmark while maintaining a clean end-to-end runtime.

Despite being a fundamental challenge that lies in the heart of many video understanding tasks, I often found it hard to find an off-the-shelf video object detector that has a clean end-to-end architecture (i.e., no pre-processing to extract optical flows, etc.) while at the same time provides state-of-the-art accuracy. This motivated me to create a new video object detector with these considerations. To propagate information across time, previous work typically either performs temporal smoothing in an ad-hoc way (i.e., detect on each frame separately and link results together), or they apply temporal aggregation within a fixed window which does not keep any long-term information. In contrast, we proposed an aligned spatial-temporal memory for this purpose [3]. First, a carefully designed *spatial-temporal memory module* (STMM) is tasked to propagate memory maps. To account for objects' movement, we propose *MatchTrans* module which explicitly aligns STMM memory maps across frames. MatchTrans computes a lightweight (~1/10 computation of FlowNet) approximation of optical flow and warp memory maps accordingly. As shown in Fig. 2 (right), memory maps become much cleaner with explicit alignment, which subsequently leads to better detection results.

Learning temporal continuity for *faster* video understanding. Recently, together with my advisor, we have co-supervised an undergraduate research project on real-time instance segmentation (named "YOLACT") [9, 10]. YOLACT is first real-time ($>$ 30 FPS) instance segmentation method with $>$ 30 mAP on the challenging MS-COCO benchmark (Fig. 3 left). The key idea of YOLACT is to simultaneously produce a set of shared "prototype masks" as well as a set of "instance coefficients", which are then used to combine prototype masks into instance masks (Fig. 3 left). Since prototype masks are shared across all instances and also the instance assembly step is quite lightweight (linear combination), YOLACT is able to strike an impressive speed/accuracy trade-off. Examples of instance segmentation produced by YOLACT is shown in Fig. 3 (right).

The success of YOLACT inspired us to extend it to video domains. In addition to improving accuracy, temporal continuity also helps in accelerating video models. In an ongoing work, instead of computing full feature stack for every frame in the video, we combine a partially computed feature representation and a temporally propagated feature to perform instance segmentation. We are able to largely accelerate video instance segmentation while retain accuracy by exploiting this cross-frame feature redundancy.

Integrated audiovisual learning for video recognition. Audio is a largely under-explored modality despite its effectiveness in recognizing actions and activities in videos. Considering the action "playing saxophone", one would expect

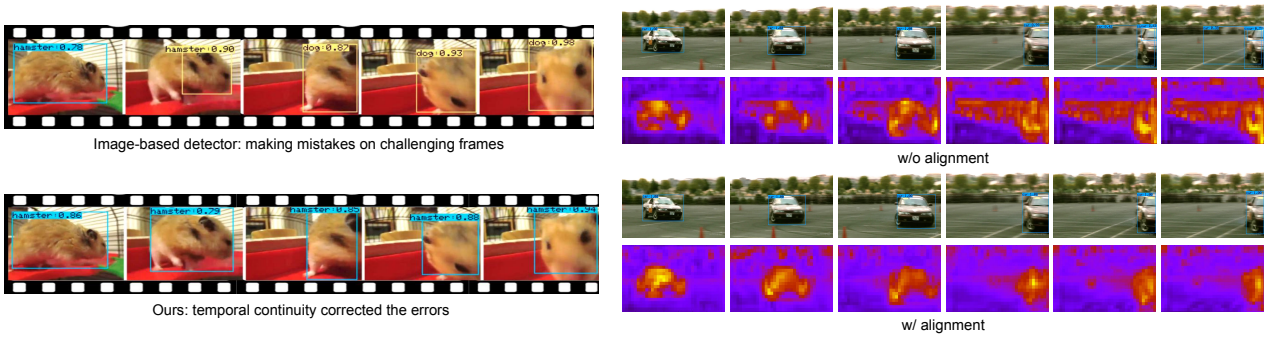

Figure 2: **Video object detection with aligned spatial-temporal memory** [3]. Left: Comparing video and image object detectors (blue/yellow correspond to true/false positives). Right: Effects of explicitly aligning memory.

Figure 3: **YOLACT: Real-time Instance Segmentation** [9, 10]. Left: YOLACT/YOLACT++ strikes an impressive accuracy/speed trade-off compared to previous methods. Right: instance segmentation examples produced by YOLACT++.

that the unique sound signature could significantly facilitate recognizing the class. Furthermore, visually subtle classes such as "whistling", where the action itself can be difficult to see in video frames, can be much easier to recognize with the aid of audio signals. Previous work on audiovisual video recognition typically adopts the "late-fusion" paradigm which refers to processing audio and visual inputs separately and fuse their outputs at the end. Despite theoretical appealingness of integrated audiovisual modeling (i.e., early/intermediate-level audiovisual fusion demonstrated in McGurk effect[1]), previous approaches are haunted by the "training dynamics mismatch" problem that we identify (Fig. 4 left) – audio subnetwork trains much faster than its visual counterpart (~1/3 training iterations compared to visual subnetwork). This mismatch leads to severe overfitting on audio inputs and thus poor generalization. To address this, we propose a simple yet surprisingly effective training strategy to randomly drop audio inputs with probability P during training (dubbed as *DropPathway*). With DropPathway, we are able to train our integrated Audiovisual SlowFast Networks [4] with hierarchical audiovisual fusion and demonstrate superior performance compared to late-fusion approaches. AVSlowFast is the first integrated audiovisual modeling approach with state-of-the-art performance on all major benchmarks (Kinetics, Charades, EPIC-kitchen for action recognition and AVA for action detection). Also, since audio inputs are one-dimensional, they are much cheaper to process compared to visual inputs. With modest computation increase (in some case as low as 2%), we are able to effectively use audio for better video understanding.

3 Learning from Videos with Minimal Supervision

While pushing the boundary of video understanding, a key issue arises with the current state-of-the-art video understanding paradigm – it requires a large amount of labeled data to train models. As the task becomes more finer-grained (i.e., classification to detection, which further requires localization), the cost of labeling such dataset increases tremendously (image-level class labels to bounding box labels). This motivated me to work on algorithms that could learn with minimal human supervision. Also, as I work more and more with videos, I realize the huge potential of

[1] https://www.youtube.com/watch?v=G-lN8vWm3m0

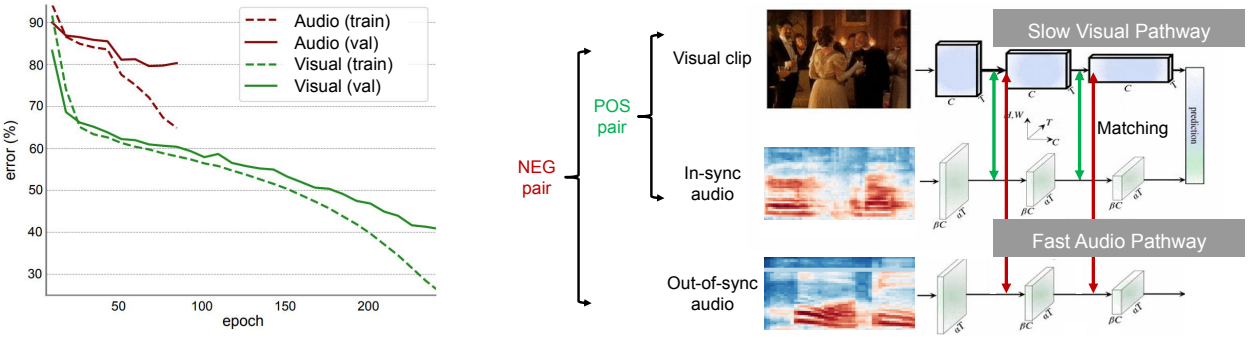

Figure 4: **AVSlowFast Networks** [4]. Left: Illustration of the training dynamics mismatch between audio and visual models. Right: Training self-supervised audiovisual synchronization prediction with AVSlowFast.

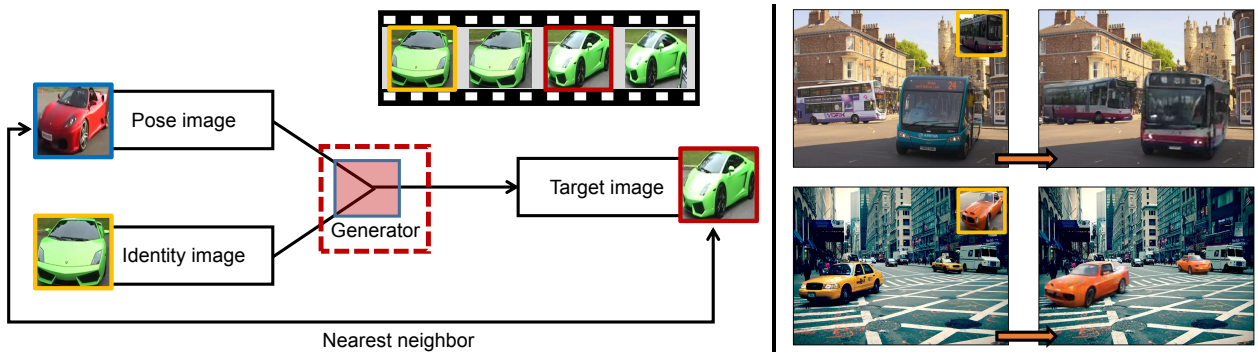

Figure 5: **Disentanglement** [5]. Left: Harvesting pseudo ground-truths from videos to learn disentanglement for image generation. Right: Generating images by combining identity (top-right) and pose from different objects.

unlabeled/weakly-labeled videos to serve as a source of supervision to train visual models.

Self-supervised video representation learning. In addition to supervised audiovisual video recognition, I am also interested in applying AVSlowFast Networks [4] to learn video representations without human annotation (i.e., self-supervised representation learning). As demonstrated in previous work [11], synchronization between audio and visual signals can be used to learn video representations. Specifically, instead of training AVSlowFast model to classify action labels of each video (i.e., supervised training), we train AVSlowFast to predict whether a pair of audio and visual clips are in-sync or not (Fig. 4 right). Positive pairs are those in which audio and visual clips are tightly synchronized, whereas negative pairs are either from different videos or from same video but with shifted start/end point. We train AVSlowFast with this audiovisual synchronization prediction task on large-scale dataset like Kinetics. To test the quality of the learned representations, we transfer the representation to perform video classification on target datasets like UCF and HMDB. With this simple audiovisual training objective, our AVSlowFast model learned a state-of-the-art video representation that outperforms previous methods by a large margin.

Supervise visual models using videos. As videos provide different views of the same object, we can learn from videos to disentangle visual factors like identity and pose *without explicit supervision* [5]. Unlike previous approaches which resort to brittle cyclic constraints (similar to the cycle constraints used in CycleGAN), we directly harvest pseudo ground-truth training tuples ($I_{identity}$, I_{pose}, I_{target}) from videos and train our generative model in a supervised manner (Fig. 5 left). Specifically, we first obtain $I_{identity}$ and I_{target} by randomly sampling a pair of frames from a video clip – such that we make sure that $I_{identity}$ and I_{target} share the same identity – and then obtain I_{pose} by retrieving a nearest neighbor for I_{target}. Models trained with this video supervision proves to produce better disentanglement than cyclic-constraint based methods, and thus leads to better image generation results (Fig. 5 right).

Videos can also be used as supervision to train object detectors. For most previous work on weakly-supervised object detection, they apply discriminative data mining in an image collection with only image-level labels (e.g., an image is labeled as "car" but no bounding boxes are given). It then searches for image patches that frequently occur in "car" images, but rarely occur in "not car" images. This leads to the mis-localization problem as it usually outputs those patches that are most discriminative but not necessarily corresponding to the full extent of objects. For example,

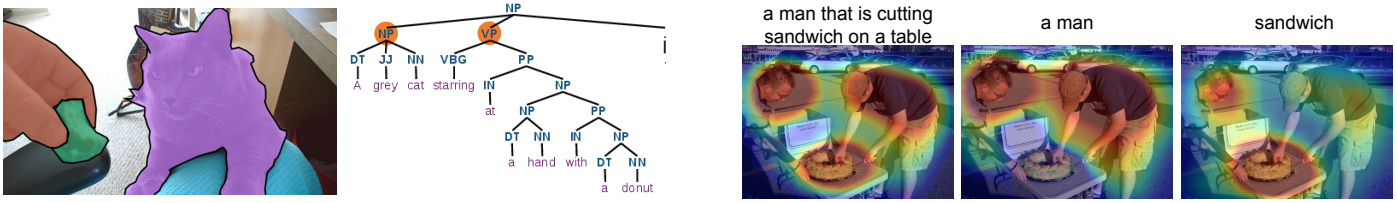

Figure 6: **Weakly-supervised Visual Grounding of Phrases with Linguistic Structures** [7]. Left: An input image and the parse tree corresponding to its caption. Right: Visual groundings of different phrases for an input image.

when mining in "car" images, one often obtain patches that correspond to "wheels" as they are the most discriminative parts of a car. To tackle this, we propose to use videos, which has motion information, to help improve localizations of patches [6]. Specifically, we first apply my previous work [1] to obtain a set of object proposal tubes in videos. Then, we match the discriminatively mined patches to those tubes, which could in turn refine the localization of patches. Finally, we transfer back refined patches to images and use them to train object detectors. This approach leads to much better localization for weakly-supervised detectors thanks to the use of motion in videos.

4 Multimodal Learning

In addition to audio/motion that is prevalent in videos, I am also interested in learning from various other modalities in combination with visual inputs. In particular, I'm interested in exploiting the intrinsic structures presented in other modalities. For example, unlike visual input which has pixels as its atomic units, language is composed of a hierarchy of words and phrases, which are artificially created objects of high-level semantics and rich structures. Therefore, we propose to make use of the linguistic structure presented in image captions to visually ground (i.e., localize) arbitrary linguistic phrases (in the form of spatial attention masks) in a weakly-supervised manner [7]. Specifically, our model is trained with images and their associated image-level captions, without any explicit region-to-phrase correspondence annotations. To this end, we introduce an end-to-end model with two types of carefully designed loss functions. In addition to the standard discriminative loss, which enforces that attended image regions and phrases are consistently encoded, we propose a novel structural loss which makes use of the parse tree structures induced by the sentences. In particular, we ensure complementarity among the attention masks that correspond to sibling noun phrases (e.g., 'a hand' and 'with a donut' in Fig. 6 left), and compositionality of attention masks among the children and parent phrases ('a hand' and 'a hand with a donut'), as defined by the sentence parse tree. We validate the effectiveness of our approach on the MS-COCO and Visual Genome datasets. Some visual grounding examples are shown in Fig. 6 (right).

Other than language, I have also worked on learning from vision + attributes [8] to discover the spatial extent of relative attributes. To accomplish this, we first develop a novel formulation that combines a detector with local smoothness to discover a set of coherent visual chains across the image collection. We then introduce an efficient way to generate additional chains anchored on the initial discovered ones. Finally, we automatically identify the most relevant visual chains, and create an ensemble image representation to model the attribute.

5 Future Work

In the future, I would like to continue pushing the boundary of self-supervised learning of visual representations. The promising results obtained in AVSlowFast work [4] makes me believe that audiovisual representation learning has the potential to bridge the gap between supervised and self-supervised video representations. Also, as most self-supervised learning work focus on learning semantic representation from 2D pixels (or spatiotemporal voxels), self-supervised learning of 3D geometric understanding is largely under-explored. Since videos provide multiple views of the same object (implicit geometry), I would like to explore learning geometric understanding from unlabeled videos. Furthermore, as self-actuated movement is proven to be an essential part of acquiring visual perception (e.g., "two kitten experiment" by Held and Hein), I would like to explore this analogy in the context of visual representation learning and locomotor skills. A concrete direction that I'm particularly interested is to develop a close-loop between "move smarter" and "see better" (i.e., train agents to "move better" in order to "see better" and vice versa).

Meanwhile, I would also like to explore *efficient* video understanding pipelines. In addition to continue researching on video models with efficient runtime, I believe it is also important to explore training strategies (e.g., RL to search

for cost-effective schedule) that can accelerate video model training. Otherwise it is not sustainable from a economical/environmental perspective and also hinders the progress of research.

Finally, with my experiences working on learning from multiple modalities, I feel strong about going towards learning from multiple modalities. This could potentially include language, audio, touch and symbolic knowledge graph, which in my view are all required ingredients for human-level intelligence in the long run.

References

[1] Fanyi Xiao and Yong Jae Lee. Track and segment: An iterative unsupervised approach for video object proposals. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 5

[2] Xitong Yang, Xiaodong Yang, Ming-Yu Liu, Fanyi Xiao, Larry S Davis, and Jan Kautz. Step: Spatio-temporal progressive learning for video action detection. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[3] Fanyi Xiao and Yong Jae Lee. Video object detection with an aligned spatial-temporal memory. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 3

[4] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2019. 1, 3, 4, 5

[5] Fanyi Xiao, Haotian Liu, and Yong Jae Lee. Identity from here, pose from there: Self-supervised disentanglement and generation of objects using unlabeled videos. In *International Conference on Computer Vision (ICCV)*, 2019. 1, 4

[6] Krishna Kumar Singh, Fanyi Xiao, and Yong Jae Lee. Track and transfer: Watching videos to simulate strong human supervision for weakly-supervised object detection. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 5

[7] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 5

[8] Fanyi Xiao and Yong Jae Lee. Discovering the spatial extent of relative attributes. In *International Conference on Computer Vision (ICCV)*, 2015. 1, 5

[9] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. YOLACT: Real-time Instance Segmentation. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[10] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. YOLACT++: Better Real-time Instance Segmentation. *arXiv preprint arXiv:1912.06218*, 2019. 2, 3

[11] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *European Conference on Computer Vision (ECCV)*, 2018. 4