RESEARCH STATEMENT

Yong Jae Lee, Associate Professor, Computer Sciences Dept. UW-Madison

My research in computer vision and machine learning aims to develop intelligent systems capable of understanding our multimodal world, enabling them to serve as valuable and effective assistants to humans. My work has particularly focused on designing novel algorithms that require minimal human supervision. Since becoming a professor in 2014, I have made several significant contributions toward this goal over the past decade. Below, I highlight the most impactful achievements.

1 General-Purpose Multi-Modal AI Assistants

Creating general-purpose intelligent AI assistants that can reason about images and videos, and communicate in natural language to help humans with various tasks, is a central AI goal. We created one of the first such systems, LLaVA (Large Language and Vision Assistant) [10, 9]. Previous AI systems were limited to narrow tasks like image captioning, lacking the ability to robustly handle open-ended tasks (e.g., *What is unusual about this image? What can I do with the ingredients in the refrigerator shown in the picture?*). Achieving this required two breakthroughs: (i) a scalable approach for obtaining visual instruction understanding training data at large-scale and high quality, and (ii) a novel neural network architecture for efficiently and effectively processing both images and text. We addressed this with our *visual instruction tuning* pipeline, which reformats existing computer vision datasets to train a multimodal (vision and language) model to follow human instructions. It leverages text-only language models to automate converting caption and bounding box annotations into high-quality visual question answering training data, with minimal human input. The reformatted data is used to train a new multimodal neural network architecture to take in images and text, and produce meaningful text outputs. The architecture's simple design enables efficient training. LLaVA has had significant impact, and has been a **major catalyst in establishing a new subfield within AI focused on general-purpose multimodal AI assistants.**

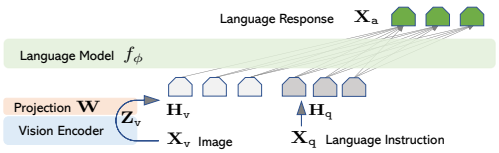

Figure 1: LLaVA [10, 9] is a general-purpose multi-modal vision-language assistant that can understand images and text instructions, and respond in natural language.

Despite exciting progress, there are many open questions in this space. Most notably, humans learn through diverse modalities (e.g., sound, vision, smell, touch, taste) that often supervise each other, beyond just language and vision. Moreover, it is not clear whether language should be the foundation upon which other modalities should be built upon. I am currently investigating these directions, including approaches that are more interactive [3], more personalized [12], and more efficient [4].

2 Efficient Object Instance Segmentation

Instance segmentation involves detecting, segmenting, and classifying each object instance in an image. Before 2019, leading methods used a two-stage pipeline: generating a set of object proposals and then classifying and segmenting them, which was computationally slow and far from realtime. We created YOLACT [1] a one-stage algorithm that skips the proposal stage, enabling real-time performance. YOLACT divides the task into two parallel subtasks: (1) generating full image size prototype masks, and (2) predicting per-instance mask coefficients. The final instance masks are produced by linearly combining prototypes with the mask coefficients. **YOLACT was the first real-time instance segmentation algorithm with highly competitive accuracy** on challenging benchmarks. It was selected for oral presentation at ICCV (a premier computer vision conference), an honor given to the top papers (<5% acceptance rate), and won the Most Innovative Award at the 2019 COCO Object Detection Challenge.

Today, most real-time convolution-based instance segmentation algorithms build on YOLACT's parallel processing approach. We also made further improvements to the approach [2], retaining its speed while increasing its accuracy, and adapted it for edge devices [11]. More recently, we have pioneered generalist segmentation models, X-Decoder [19] and SEEM [22], which can be prompted with text and visual inputs to perform a wide range of segmentation tasks (e.g., instance, panoptic, point-based, reference-based, text-based) with the same neural network architecture, unlike prior methods which designed separate specialist models for each segmentation setting.

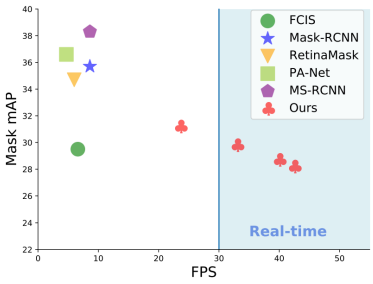

Figure 2: YOLACT [1] was the first real-time (>30 FPS) instance segmentation method with competitive accuracy on the COCO dataset.

3 Controllable Image Generation

Despite huge advances in image generation research, reliable systems for highly-controllable visual content creation that accurately match user specifications are still lacking. We created FineGAN [15] and MixNMatch [8], which were among **the first methods to yield structured, disentangled representations of background, object shape, color/texture, and pose for image generation with minimal supervision**. Our solution, based on information theory, automatically disentangles key visual factors without corresponding annotations. This allows users to modify specific factors (e.g., object color) while keeping others (e.g., object shape, pose, background) constant during image generation.

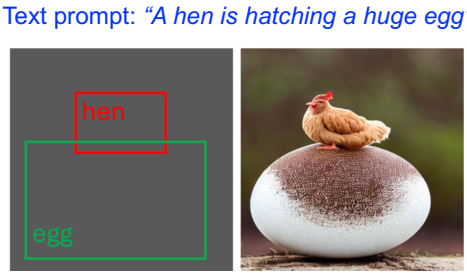

Figure 3: GLIGEN [7] efficiently converts a pre-trained text-to-image generation model into a grounded model where e.g., bounding boxes can control object location and size.

Our more recent work, GLIGEN (Open-set Grounded-Language-to-Image Generation) [7], extends this control to text-to-image diffusion models like StableDiffusion. Unlike standard models that rely on a single text prompt (e.g., "A brown dog"), GLIGEN enables conditioning on additional grounding inputs, including bounding boxes, keypoints, edge maps, and segmentation maps. It is **one of the first works to show how to efficiently and effectively add new functionalities to pre-trained diffusion models**.

4 Improving Robustness in Visual Recognition Models

Although deep networks have achieved remarkable breakthroughs in visual recognition, speech recognition, and natural language processing, they are prone to silly mistakes. For example, a tiny shift in the input image can cause drastic changes in the predicted label, due to aliasing from downsampling in convolutional networks. In [21, 20], we introduced a method to build anti-aliased convolutional networks. Unlike previous approaches that apply the same Gaussian blur across all spatial locations and feature channels, we made the key observation that the blur should vary based on content (e.g., edges vs smooth regions). We developed a content-aware anti-aliasing module, which adaptively predicts low-pass filter weights for different regions and feature channels, leading to significant improvements in image classification, segmentation, and domain generalization. **This work was awarded the Best Paper Award at BMVC 2020, selected as the top paper of the conference**, and we published an invited journal article for best papers in IJCV 2022 with further improvements.

5 Learning with Weak Human Supervision

Detecting and segmenting objects in images is a core problem in computer vision, but today's algorithms rely heavily on costly and error-prone bounding box or pixel-level annotations. For instance, building the popular MS COCO dataset required more than 70,000 hours to annotate 328,000 images for just 80 object categories. Clearly, this approach is not scalable to the vast range of visual concepts humans recognize. To address this, we developed novel algorithms for detecting and segmenting objects using only image-level tag annotations.

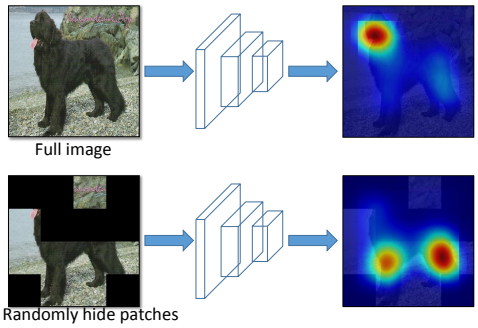

Figure 4: Hide-and-Seek [16] randomly hides patches in each training image (bottom), which forces the image classifier to go beyond just the most discriminative part (top) and instead learn to focus on all parts of the object.

My most impactful contribution in this space is the *Hide-and-Seek* approach [16]. The idea is simple yet highly-effective: random patches of images are hidden when training an image classification model. This forces the model to focus on different parts of the objects across images, which leads to it learning the full object representation (e.g., a whole dog) rather than just the most discriminative part (e.g., dog's face) like prior methods. This idea has also proven to be useful as a data augmentation technique for training deep networks, improving the state-of-the-art on a variety of tasks including image classification, segmentation, face recognition, and person re-identification [17].

6 Ongoing and Future Directions

In addition to the above themes, I am interested in exploring questions such as:

- *Can we develop AI systems that can learn from multiple modalities, beyond vision and language?* As mentioned previously, we humans learn about our world through signals acquired from multiple sources (e.g., sound, vision, smell, touch, taste), which often supervise each other. However, today's systems largely rely on vision and language only, and leverage language as the foundation. I believe that expanding to other modalities will be especially critical for creating systems that can learn to act in the physical world.

- *How can we create unbiased and secure AI algorithms?* As AI technology is becoming more integrated into our daily lives, addressing ethics, bias/fairness, and privacy/security questions are more important than ever. I have initial work in studying ways to ensure the privacy and security of users in the visual data that the algorithms process [14, 6, 5], to mitigate undesirable biases [18], to improve the robustness of deep networks [21, 20], and to create universal fake image detectors [13].

- *How can we better understand our AI systems?* While AI systems are advancing at an extraordinary pace, our understanding of them has not kept up. There are many open questions, ranging from whether language should be the foundation upon which other modalities should be built upon as in today's LLM-based systems, to whether today's models can reason or instead *pretend* to reason by mimicking reasoning patterns in the training data. I am interested in investigating such questions.

Over the next decades, I will strive to continue to be on the forefront in creating robust and useful AI systems that can learn with minimal human supervision. I am passionate about asking the *right* (meaningful and impactful) research questions, and proposing innovative and effective solutions to those questions. I am excited about the prospects of working towards these challenges with collaborators in vision and learning, and related fields including graphics, robotics, neuroscience, and cognitive science.

References

[1] D. Bolya, C. Zhou, F. Xiao, and **Yong Jae Lee**. YOLACT: Real-time Instance Segmentation. In *IEEE International Conference on Computer Vision* (ICCV), 2019. (**oral presentation**).

[2] D. Bolya, C. Zhou, F. Xiao, and **Yong Jae Lee**. YOLACT++: Better Real-time Instance Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (TPAMI), 44(2), 2020.

[3] M. Cai, H. Liu, S. K. Mustikovela, G. P. Meyer, Y. Chai, D. Park, and **Yong Jae Lee**. Making Large Multimodal Models Understand Arbitrary Visual Prompts. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2024.

[4] M. Cai, J. Yang, J. Gao, and **Yong Jae Lee**. Matryoshka Multimodal Models. *arXiv preprint arXiv:2405.17430*, 2024.

[5] Z. A. Din, H. Venugopalan, J. Park, A. Li, W. Yin, H. Mai, **Yong Jae Lee**, S. Liu, and S. T. King. Boxer: Preventing Fraud by Scanning Credit Cards. In *USENIX Security Symposium* (USENIX Security), 2020.

[6] X. Gu, W. Luo, M. Ryoo, and **Yong Jae Lee**. Password-conditioned Anonymization and Deanonymization with Face Identity Transformers. In *European Conference on Computer Vision* (ECCV), 2020.

[7] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and **Yong Jae Lee**. GLIGEN: Open-Set Grounded Text-to-Image Generation. *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2023.

[8] Y. Li, K. K. Singh, U. Ojha, and **Yong Jae Lee**. MixNMatch: Multifactor Disentanglement and Encoding for Conditional Image Generation. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2020.

[9] H. Liu, C. Li, Y. Li, and **Yong Jae Lee**. Improved Baselines with Visual Instruction Tuning. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2024. (**highlight presentation**).

[10] H. Liu, C. Li, Q. Wu, and **Yong Jae Lee**. Visual Instruction Tuning. In *Neural Information Processing Systems* (NeurIPS), 2023. (**oral presentation**).

[11] H. Liu, R. A. R. Soto, F. Xiao, and **Yong Jae Lee**. YolactEdge: Real-time Instance Segmentation on the Edge. In *IEEE International Conference on Robotics and Automation* (ICRA), 2021.

[12] T. Nguyen, H. Liu, Y. Li, M. Cai, U. Ojha, and **Yong Jae Lee**. Yo'LLaVA: Your Personalized Language and Vision Assistant. In *Neural Information Processing Systems* (NeurIPS), 2024.

[13] U. Ojha, Y. Li, and **Yong Jae Lee**. Towards Universal Fake Image Detectors that Generalize Across Generative Models. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2023.

[14] Z. Ren, **Yong Jae Lee**, and M. Ryoo. Learning to Anonymize Faces for Privacy Preserving Action Detection. In *European Conference on Computer Vision* (ECCV), 2018.

[15] K. Singh, U. Ojha, and **Yong Jae Lee**. FineGAN: Unsupervised Hierarchical Disentanglement for Fine-Grained Object Generation and Discovery. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2019. (**oral presentation**).

[16] K. Singh and **Yong Jae Lee**. Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-supervised Object and Action Localization. In *IEEE International Conference on Computer Vision* (ICCV), 2017.

[17] K. Singh, H. Yu, A. Sarmasi, G. Pradeep, and **Yong Jae Lee**. Hide-and-Seek: A Data Augmentation Technique for Weakly-Supervised Localization and Beyond. In *arXiv*, 2018.

[18] K. K. Singh, D. Mahajan, K. Grauman, **Yong Jae Lee**, M. Feiszli, and D. Ghadiyaram. Don't Judge an Object by Its Context: Learning to Overcome Contextual Bias. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2020. (**oral presentation**).

[19] X. Zou, Z.-Y. Dou, J. Yang, Z. Gan, L. Li, C. Li, X. Dai, J. Wang, L. Yuan, N. Peng, L. Wang, **Yong Jae Lee**, and J. Gao. Generalized Decoding for Pixel, Image and Language. In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2023.

[20] X. Zou, F. Xiao, Z. Yu, Y. Li, and **Yong Jae Lee**. Delving Deeper into Anti-aliasing in ConvNets. *International Journal of Computer Vision* (IJCV), 131:67–81, 2022. (**invited article for best papers of BMVC 2020**).

[21] X. Zou, F. Xiao, Z. Yu, and **Yong Jae Lee**. Delving Deeper into Anti-aliasing in ConvNets. In *British Machine Vision Conference* (BMVC), 2020. (**Best Paper Award**).

[22] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, and **Yong Jae Lee**. Segment Everything Everywhere All at Once. In *Neural Information Processing Systems* (NeurIPS), 2023.