Research Statement

Jia-Bin Huang, Assistant Professor, Virginia Tech

The central goal of my research is to build intelligent machines that can *understand* and *recreate* the visual world around us. While recent years have witnessed impressive progress in data-driven computer vision, the state-of-the-art vision systems today rely on millions of *manually labeled* training examples. The dependency on diverse, high-quality training datasets substantially limits the applicability to problem domains where large-scale human annotations are expensive, scarce, or ambiguous. My research aims to overcome these limitations by leveraging a vast amount of *unlabeled* visual data that are readily available on the web. Specifically, my recent work has focused on addressing the following three key research questions:

1. How can we create algorithms that learn/adapt to new tasks and environments with limited supervision?
2. How can we model and recreate the visual world by using the structures of visual data?
3. How can we learn to understand temporal dynamics and complex human activities?

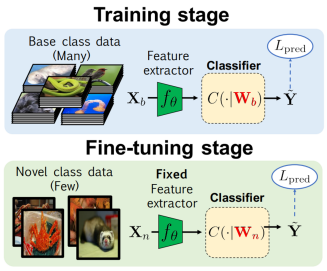
In the following, I will highlight my research contributions in these three themes and then describe my research work's projection onto cross-disciplinary collaboration. I will conclude with future research agenda.

1. Learning to See with Limited Supervision

Unlike current algorithms that require millions of manually labeled training examples, humans learn to perceive our visual world through observations, goal-oriented movements, and embodied interaction with the physical world. Alleviating the reliance on labeled examples allows us to capitalize on the vast amount of unlabeled images and videos (and the associated modalities or meta-data) to *learn to see* with limited manual supervision. Enabling machines to learn/adapt to new tasks is one of the most crucial steps to democratize AI technology.

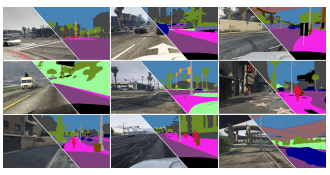
Toward this goal, I have developed novel algorithms that (1) learn novel concepts using *few examples* [3, 46, 22], (2) learn to solve structured prediction problems using *synthetic data* [19, 5, 25, 14], (3) learn to localize objects in images using *weak supervision* (e.g., with only image-level annotations) [6, 4, 30, 29], and (4) learn visual representations using *videos* [49, 48, 35]. This research theme is being supported by my NSF CRII award and the 3M non-tenured faculty award. Below I will outline each of my contributions.

Learning from few examples. Humans are remarkable at learning and adapting to new tasks from very few examples (few-shot learning). While significant progress has been made, mainly using meta-learning algorithms, the differences in implementation and experimental details of various methods make a fair comparison difficult. My work [3] established the first consistent comparative analysis and introduced a surprisingly simple yet effective baseline method that achieves competitive performance with sophisticated meta-learning algorithms. Building on [3], I created an algorithm for few-shot learning *under domain shifts* [46]. The core technical idea is to design a *learned feature-based augmentation* to explicitly optimize generalization (validation) error. I recently applied the ideas of feature-based augmentation to semi-supervised learning and showed substantial improvement over prior arts [22].



Few-shot learning [3]

Learning from synthetic data. For high-level problems, one may still be able to collect many labeled training examples at the cost of time and money. However, this is often not feasible for most low-level vision problems where dense annotations are required. To tackle this limitation, I explored learning with *synthetic training data* (e.g., rendered frames from game engines) and designed algorithms to adapt the trained model to the real domain. My research along this direction demonstrated substantially improved results on optical flow [25], multi-view stereo [19], and general dense prediction problems [5]. In addition to using existing synthetic datasets, I also created a new one that enables the community to study new problems (semantic amodal video objection segmentation) [14].

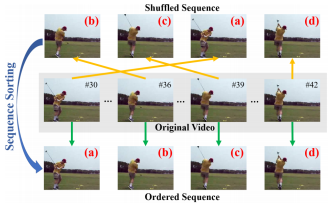


Syn2Real transfer [5]

Learning from weak supervision. Object detection/segmentation is a building block for many high-impact downstream applications (e.g., self-driving cars). However, curating large-scale *instance-level* annotations (bounding box or segmentation) is costly and error-prone. My work instead used *image-level* supervision indicating the presence of the objects [30, 29]. Image-level annotations are much easier to obtain (e.g., many are readily available through text tags, GPS tags, and image search queries). Pushing toward finer-grained localization, I designed novel weakly supervised algorithms for learning *pixel-level* semantic segmentation [51] and joint matching and co-segmentation [6, 4].

Weak supervision [6]

Learning from videos. I am excited about learning from videos because videos provide much richer information than still images. First, analyzing a video reveals object motions, occlusion/dis-occlusions, camera poses, and scene structures. Using these cues, I developed *self-supervised* algorithms for learning monocular depth estimation and optical flow [49], visual odometry [35], and video depth [35]. Second, videos offer observations of how our world works (e.g., dropping a pen, it will fall on the floor). Specifically, I used natural chronology order of videos for representation learning [26] and domain adaptation [9, 8].

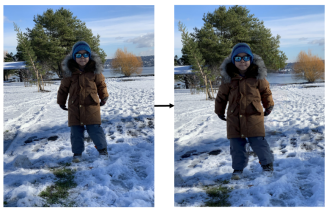


Self-sup. learning [26]

2. Learning to Recreate Our Visual World

Machines' ability to model and recreate our visual world paves a promising path towards a deeper understanding of visual data. It also opens up fascinating opportunities for everyone to enhance, visualize, and interact with visual media. My research has substantially contributed to computational photography by drawing and integrating ideas from learning, vision, and graphics. Specifically, I developed learning-based algorithms for (1) creating 3D photos or videos using only monocular images/videos, (2) recovering missing information from partial observations, and (3) translating visual contents across domains. This research theme is supported via two research grants funded by Denso and research gifts from Google, Facebook, and Adobe.

3D photos. 3D photography—capturing views of the world with a camera and using image-based rendering techniques for novel view synthesis—is a fascinating way to record and reproduce visual perception. It provides a dramatically more immersive experience than old 2D photography: almost lifelike in Virtual Reality, and even to some degree on regular flat displays when displayed with parallax. I developed a method for converting a single input image into a 3D photo, a multi-layer representation that contains hallucinated color and depth structures in regions occluded in the original view [39]. Along with the paper, I created demo videos and open-sourced the code. My work reached millions of people and HBO documentary filmmakers who adopted it to tell more compelling visual stories.

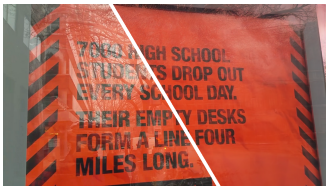


3D photography [39]

3D videos. I further continue to work on creating *3D video* aiming to enable free viewpoint and time navigation of causally captured videos. I tackled the first challenge for creating 3D videos: reconstructing dense, geometrically consistent depth for all pixels in a monocular video [35]. I showed that our method enables dynamic scene reconstruction and advanced video-based visual effects

Consistent depth [35]

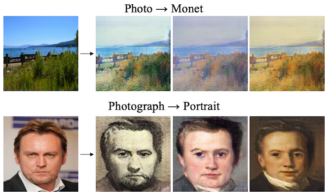
Seeing the unseen. Traditional imaging systems have limitations in optics, sensors and cannot reproduce or manipulate the rich visual perception. My work contributed to overcoming these limitations, allowing us to "seeing the unseen". Specifically, I developed algorithms for filling plausible missing contents in images [43] or videos [44, 10], recovering high spatial frequency details [45, 23, 24], inferring the high-dynamic range contents [33], color information [40], propagating details from guidance [32, 31], and seeing through obstructions such as re-



Obstruction removal [34]

flections or fences with layer decomposition [34]. Unlike existing approaches that treat the observed images as plain 2D signals, the unifying ingredient in my work is to model and interpret visual signals in terms of physical surface, object, scene, and image formation process. This perspective allows my methods to learn from data effectively and generalize well to unseen test cases.

Image-to-image generation Many of the methods above required *paired* training data and can produce one *single* output (ignoring the multi-modal nature of the output in many problems). In light of this, I created an algorithm that can learn diverse image-to-image translation models from unpaired datasets using disentangled representations [27, 28]. In my follow-up research, I extended it to handle guidance signals [1] and go beyond generating plain 2D images [18].

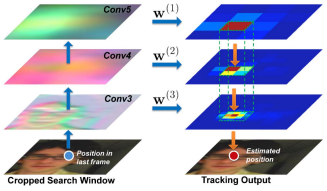


Diverse translation [27]

3. Learning to Understand People in Images and Videos

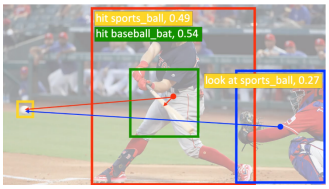
Another exciting research direction I have taken is to create algorithms to understand people in images and videos. Understanding human activity is a fundamental step toward building socially-aware agents, semantic image/video retrieval, captioning, and question-answering. Toward this goal, I focused on (1) understanding temporal dynamics and (2) recognizing human actions across visual domains. This research theme is supported by a Samsung Global Research Outreach Award, a Google Faculty Award, and a Ford Motor research grant.

Video tracking and segmentation. Establishing long-range correspondence in dynamic videos is a core computer vision problem with numerous applications, e.g., video surveillance, autonomous driving, and understanding human activities. My early work [36, 37] contributed one of the first methods to leverage deep convolutional features for tracking effectively. I also developed trackers for specific target objects such as faces [47] or eye gazes [41]. Building upon these foundations, I further developed methods for pixel-wise tracking and segmentation [15, 17, 16], amodal segmentation [14], and geometrically consistent 3D reconstruction of people [50].



Visual tracking [36, 37]

Human activity understanding. One challenge in activity understanding is recognizing activities where we do not access many training examples. My research specifically aims to address the generalization problem via debiasing [7], unsupervised domain adaptation [9, 8], and exploiting contextual information in the scene [12, 11]. It enables transferring knowledge from one domain to another and recognizes rare (or unseen) activities during the training.



Human-object inter. [11, 12]

4. Cross-Disciplinary Research

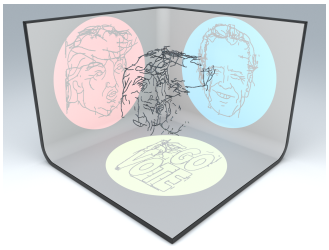
At Virginia Tech, I enjoy collaboration with researchers from different fields to tackle cross-disciplinary research challenges, including health care [21, 38], computational ecology [42], security [20], and art [2, 13].

Semi-automated rehabilitation systems in the home. With the aging of the US population, there is an increasing need for effective and accessible rehabilitation services for debilitating illnesses and injuries such as stroke and arthritis. Together with researchers in the Biomedical Engineering, Mechanical Engineering, Computer Science, and Carilion Clinic, I designed and implemented a vision-based system for semi-automated rehabilitation (providing immediate feedback based on patients' movement) [21, 38]. Our promising preliminary work led to an NSF Smart and Connected Health grant.

Rehabilitation system [21]

Computation-aware autonomy. As we enter an autonomy-driven future, it is critical to facilitate the safe and reliable coordination of diverse cyber-physical systems (CPS). I am currently working with experts in robotics, computer architecture, and real-time systems to develop *computation-aware autonomy* that seeks to integrate algorithms for decision-making, visual sensing, motion, and underlying computational resources. This research effort is currently funded via an NSF CPS medium grant.

Vision and art. I enjoyed applying 3D vision algorithms for interdisciplinary collaborations on creating new forms of artistic expression. For examples, I developed *multi-view wire art* [13] — a 3D sculptural art with multiple interpretations when perceived at different viewpoints. Our sculptures were exhibited in dining halls across the campus for several months. Another example is the *Source Form* project. Teaming with the School of Visual Arts and Mechanical Engineering, I contributed to building an end-to-end system that takes user queries, produces 3D models from crowd-sourced photos, and 3D prints physical form [2].



Wire art [13]

5. Ongoing and Future Directions

To summarize, my research goal is to develop algorithms that can understand and recreate the visual world. My research to date tackled the significant challenges via leveraging unlabeled or weakly labeled visual data and exploiting structures such as geometry, compositionality, and image formation. Moving forward, I am excited to explore the following research questions:

- ***How can we train the vision models using multiple modalities?*** Babies/toddlers learn to perceive the world, not by memorizing millions of labeled training images. Instead, they learn by interacting with physical objects and exploring the world through multiple modalities (e.g., sound, touch, smell, taste, language, and vision). My research on learning with limited supervision (few-shot learning, self-supervised learning, and semi-supervised learning) primarily focuses only on image data. Next, I would like to develop algorithms that can capitalize on these unlabeled but rich multi-modality signals to create robust perception models.

- ***How can we recreate the visual world with photorealism?*** My dream is to bring our visual memory to life by modeling and recreating visual data. However, existing 3D representations (e.g., explicit depth estimation, mesh reconstruction used in my work on 3D photos/videos) remain not sufficiently flexible for creating photorealistic rendering. I have begun to explore rendering using *neural implicit representation* to model the complex visual phenomenon (e.g., view-dependent effects, semi-transparent surfaces). In contrast to existing work in this field that focuses on *memorization*, I will instead strike for *generalization*. I will also tackle the challenge of creating representations that support intuitive editing and manipulation (e.g., lighting, viewpoints, time).

- ***How can we create models that adapt to unseen tasks and environments?*** One of the major problems in deep learning is that a trained model's performance is often significantly degraded in new visual domains. As our world is continuously changing, the existing *static* training and testing paradigm in machine learning inevitably does not lead to a promising path toward generalization. In the past, I have addressed the problem via unsupervised domain adaptation or self-supervised learning. However, such settings remain artificial (e.g., unsupervised domain adaptation assumes that we have access to unlabeled data in the target domain). In the future, I would like to formulate visual learning as a continual, life-long process, and incrementally revise its belief in our visual world.

With my prior research contributions to the relevant problems, I am very excited to carry on my research trajectories with my students, colleagues, and collaborators. Apart from tackling the core research questions, I will also continue collaborating with researchers to tackle challenging cross-disciplinary tasks. Bringing together expertise across diverse fields will lead to out-of-box and practical solutions to impactful research questions.

References

[1] B. AlBahar and **J.-B. Huang**. Guided image-to-image translation with bi-directional feature transformation. In *ICCV*, 2019. 3

[2] S. Blanchard, **J.-B. Huang**, C. B. Williams, V. Meenakshisundaram, J. Kubalak, and S. Lokegaonkar. Source form an automated crowdsourced object generator. In *ACM SIGGRAPH 2019 Studio*. 2019. 3, 4

[3] W.-Y. Chen, Y.-C. Liu, Z. Kira, Y.-C. F. Wang, and **J.-B. Huang**. A closer look at few-shot classification. In *ICLR*, 2019. 1

[4] Y.-C. Chen, P.-H. Huang, L.-Y. Yu, **J.-B. Huang**, M.-H. Yang, and Y.-Y. Lin. Deep semantic matching with foreground detection and cycle-consistency. In *Asia Conference on Computer Vision*, 2018. 1, 2

[5] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and **J.-B. Huang**. CrDoCo: Pixel-level domain transfer with cross-domain consistency. In *CVPR*, 2019. 1

[6] Y.-C. Chen, Y.-Y. Lin, M.-H. Yang, and **J.-B. Huang**. Show, match and segment: Joint weakly supervised learning of semantic matching and object co-segmentation. *TPAMI*, 2020. 1, 2

[7] J. Choi, C. Gao, J. C. Messou, and **J.-B. Huang**. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *NeurIPS*, 2019. 3

[8] J. Choi, G. Sharma, M. Chandraker, and **J.-B. Huang**. Unsupervised and semi-supervised domain adaptation for action recognition from drones. 2020. 2, 3

[9] J. Choi, G. Sharma, S. Schulter, and **J.-B. Huang**. Shuffle and Attend: Video domain adaptation. In *ECCV*, 2020. 2, 3

[10] C. Gao, A. Saraf, **J.-B. Huang**, and J. Kopf. Flow-edge guided video completion. In *ECCV*, 2020. 2

[11] C. Gao, J. Xu, Y. Zou, and **J.-B. Huang**. DRG: Dual relation graph for human-object interaction detection. In *ECCV*, 2020. 3

[12] C. Gao, Y. Zou, and **J.-B. Huang**. iCAN: Instance-centric attention network for human-object interaction detection. In *BMVC*, 2018. 3

[13] K.-W. Hsiao, **J.-B. Huang**, and H.-K. Chu. Multi-view wire art. *ACM TOG (Proc. SIGGRAPH Asia)*, 37(6):242–1, 2018. 3, 4

[14] Y.-T. Hu, H.-S. Chen, K. Hui, **J.-B. Huang**, and A. G. Schwing. SAIL-VOS: Semantic amodal instance level video object segmentation-a synthetic dataset and baselines. In *CVPR*, 2019. 1, 3

[15] Y.-T. Hu, **J.-B. Huang**, and A. Schwing. MaskRNN: Instance level video object segmentation. In *NeurIPS*, 2017. 3

[16] Y.-T. Hu, **J.-B. Huang**, and A. G. Schwing. Unsupervised video object segmentation using motion saliency-guided spatio-temporal propagation. In *ECCV*, 2018. 3

[17] Y.-T. Hu, **J.-B. Huang**, and A. G. Schwing. VideoMatch: Matching based video object segmentation. In *ECCV*, 2018. 3

[18] H.-P. Huang, H.-Y. Tseng, H.-Y. Lee, and **J.-B. Huang**. Semantic view synthesis. In *ECCV*, 2020. 3

[19] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and **J.-B. Huang**. DeepMVS: Learning multi-view stereopsis. In *CVPR*, 2018. 1

[20] S. T. Jan, J. Messou, Y.-C. Lin, **J.-B. Huang**, and G. Wang. Connecting the digital and physical world: Improving the robustness of adversarial attacks. In *AAAI*, 2019. 3

[21] A. Kelliher, J. Choi, **J.-B. Huang**, T. Rikakis, and K. Kitani. Homer: An interactive system for home based stroke rehabilitation. In *International ACM SIGACCESS Conference on Computers and Accessibility*, 2017. 3

[22] C.-W. Kuo, C.-Y. Ma, **J.-B. Huang**, and Z. Kira. FeatMatch: Feature-based augmentation for semi-supervised learning. In *ECCV*, 2020. 1

[23] W.-S. Lai, **J.-B. Huang**, N. Ahuja, and M.-H. Yang. Deep Laplacian pyramid networks for fast and accurate super-resolution. In *CVPR*, 2017. 2

[24] W.-S. Lai, **J.-B. Huang**, N. Ahuja, and M.-H. Yang. Fast and accurate image super-resolution with deep laplacian pyramid networks. 41(11):2599–2613, 2018. 2

[25] W.-S. Lai, **J.-B. Huang**, and M.-H. Yang. Semi-supervised learning for optical flow with generative adversarial networks. In *NeurIPS*, 2017. 1

[26] H.-Y. Lee, J.-B. Huang, M. Singh, and M.-H. Yang. Unsupervised representation learning by sorting sequences. In *ICCV*, 2017. 2

[27] H.-Y. Lee, H.-Y. Tseng, Q. Mao, **J.-B. Huang**, Y.-D. Lu, M. Singh, and M.-H. Yang. DRIT++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 2020. 3

[28] H.-Y. Lee, H.-Y. Tseng, **J.-B. Huang**, M. Singh, and M.-H. Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 3

[29] D. Li, **J.-B. Huang**, Y. Li, S. Wang, and M.-H. Yang. Weakly supervised object localization with progressive domain adaptation. In *CVPR*, 2016. 1, 2

[30] D. Li, **J.-B. Huang**, Y. Li, S. Wang, and M.-H. Yang. Progressive representation adaptation for weakly supervised object localization. 42(6):1424–1438, 2019. 1, 2

[31] Y. Li, **J.-B. Huang**, N. Ahuja, and M.-H. Yang. Deep joint image filtering. In *ECCV*, 2016. 2

[32] Y. Li, **J.-B. Huang**, N. Ahuja, and M.-H. Yang. Joint image filtering with deep convolutional networks. 41(8):1909–1923, 2019. 2

[33] Y.-L. Liu, W.-S. Lai, Y.-S. Chen, Y.-L. Kao, M.-H. Yang, Y.-Y. Chuang, and **J.-B. Huang**. Single-image hdr reconstruction by learning to reverse the camera pipeline. In *CVPR*, 2020. 2

[34] Y.-L. Liu, W.-S. Lai, M.-H. Yang, Y.-Y. Chuang, and **J.-B. Huang**. Learning to see through obstructions. In *CVPR*, 2020. 2, 3

[35] X. Luo, **J.-B. Huang**, R. Szeliski, K. Matzen, and J. Kopf. Consistent video depth estimation. 39(4), 2020. 1, 2

[36] C. Ma, **J.-B. Huang**, X. Yang, and M.-H. Yang. Hierarchical convolutional features for visual tracking. In *ICCV*, 2015. 3

[37] C. Ma, **J.-B. Huang**, X. Yang, and M.-H. Yang. Robust visual tracking via hierarchical convolutional features. *TPAMI*, 41(11):2709–2723, 2018. 3

[38] T. Rikakis, A. Kelliher, J. Choi, **J.-B. Huang**, K. Kitani, S. Zilevu, and S. L. Wolf. Semi-automated home-based therapy for the upper extremity of stroke survivors. In *Proceedings of the PErvasive Technologies Related to Assistive Environments Conference*, 2018. 3

[39] M.-L. Shih, S.-Y. Su, J. Kopf, and **J.-B. Huang**. 3D photography using context-aware layered depth inpainting. In *CVPR*, 2020. 2

[40] J.-W. Su, H.-K. Chu, and **J.-B. Huang**. Instance-aware image colorization. In *CVPR*, 2020. 2

[41] **J.-B. Huang**, Q. Cai, Z. Liu, N. Ahuja, and Z. Zhang. Towards accurate and robust cross-ratio based gaze trackers through learning from simulation. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, 2014. 3

[42] **J.-B. Huang**, R. Caruana, A. Farnsworth, S. Kelling, and N. Ahuja. Detecting migrating birds at night. In *CVPR*, 2016. 3

[43] **J.-B. Huang**, S. B. Kang, N. Ahuja, and J. Kopf. Image completion using planar structure guidance. *ACM TOG (Proc. SIGGRAPH)*, 33(4):1–10, 2014. 2

[44] **J.-B. Huang**, S. B. Kang, N. Ahuja, and J. Kopf. Temporally coherent completion of dynamic video. *ACM Transactions on Graphics (TOG)*, 35(6):1–11, 2016. 2

[45] **J.-B. Huang**, A. Singh, and N. Ahuja. Single image super-resolution from transformed self-exemplars. In *CVPR*, 2015. 2

[46] H.-Y. Tseng, H.-Y. Lee, **J.-B. Huang**, and M.-H. Yang. Cross-domain few-shot classification via learned feature-wise transformation. In *ICLR*, 2020. 1

[47] S. Zhang, **J.-B. Huang**, J. Lim, Y. Gong, J. Wang, N. Ahuja, and M.-H. Yang. Tracking persons-of-interest via unsupervised representation adaptation. *IJCV*, 128(1):96–120, 2020. 3

[48] Y. Zou, P. Ji, Q.-H. Tran, **J.-B. Huang**, and M. Chandraker. Learning monocular visual odometry via self-supervised long-term modeling. In *ECCV*, 2020. 1

[49] Y. Zou, Z. Luo, and **J.-B. Huang**. DF-Net: Unsupervised joint learning of depth and flow using cross-task consistency. In *ECCV*, 2018. 1, 2

[50] Y. Zou, J. Yang, D. Ceylan, J. Zhang, F. Perazzi, and **J.-B. Huang**. Reducing footskate in human motion reconstruction with ground contact constraints. 2020. 3

[51] Y. Zou, Z. Zhang, H. Zhang, C.-L. Li, X. Bian, **J.-B. Huang**, and T. Pfister. PseudoSeg: Designing pseudo labels for semantic segmentation. *arXiv preprint arXiv:2010.09713*, 2020. 2