Statement of Purpose: Siddhartha Gairola *webpage:* https://sidgairo18.github.io/

Research Interests and Background: My research interests span Machine Learning (ML) and Computer Vision (CV). Building robust systems that model visual perception and understanding find important applications in medicine, photography, robotics, autonomous vehicles, and much more. I am particularly interested in building reliable systems that model visual perception with limited supervision. My prior research contributions have tackled key tasks in computer vision including (1) image compositing, (2) unsupervised representation learning, (3) few-shot image segmentation, and (4) AI-based vision tools for healthcare. These have resulted in publications at top conferences like **IJCAI, WACV, SIGIR, EMBC,** and **IMWUT/Ubicomp**. I wish to earn a Ph.D. in Computer Science to tackle core computer vision, graphics, and machine learning problems.

Computer Vision and Machine Learning: I delved into research in my sophomore year, when I started working with **Prof. P. J. Narayanan** at the Center for Visual Information Technology (CVIT) at IIIT Hyderabad. Initially, I explored problems in the space of computational photography and structure-from-motion. For my first research project, I proposed a data-driven approach to perform color-consistent sky replacement in outdoor images. I received the **Dean's Research Award** for this work, and it led to a publication at **MMM 2018**.

While working on these initial projects, I realized the utility of robust image features for visual tasks like image recognition, retrieval, style transfer, etc. This got me interested in **representation learning** to model the style and content in images. Style recognition and similarity are important measures for understanding abstract concepts like art, fashion, and design. However, the definition of style is contextual and vague. Deep neural networks (DNNs) perform well on image understanding tasks but require densely annotated large-scale datasets. Obtaining such annotations is expensive (in terms of time and money) and often infeasible. I developed a DNN based framework for learning a neural embedding that captures the 'look and feel' of an image. Unlike previous supervised learning methods, the proposed framework is **unsupervised** – it does not use categorical labels but uses a gram matrix feature-based clustering to get proxy labels for forming triplets of anchor, positive, and negative images. These triplets are used to train a siamese network with a triplet loss to learn an embedding useful for style-based search and retrieval. Despite being 16 times more compact than traditional representations, the embeddings achieved state-of-the-art results for style-based image retrieval and recognition on six datasets with different notions of style. Thus, the proposed method provides a general framework to learn style representations and can also be combined with other proxy measures for style-aware grouping. This work was published at **IEEE WACV 2020**.

At CVIT, a few of my lab mates and I had also been exploring **multimodal machine learning**. We participated in the Clickbait Detection Challenge co-organized by Google and the Webis Research Group. The goal was to build a system that automatically identifies clickbait posts (given its content and images) used to lure users into clicking on articles that fail to fulfill the post description. We proposed a siamese network that fuses visual and textual data to learn robust neural embeddings that are used to classify a social media post as clickbait or not. I took the lead on developing the visual feature extractor and the siamese network module. We secured 3^{rd} place globally, and published an extension of this work at **ACM SIGIR 2018**.

My research journey at IIIT-Hyderabad culminated in the form of my Master's thesis. Working at CVIT introduced me to the fascinating world of research, where I developed a solid foundation in computer vision and machine learning, and skills to collaborate effectively and conduct principled research. The brainstorming sessions with my advisor and senior graduate students taught me to communicate well, ask questions, and develop methods for answering them. I left CVIT feeling optimistic about research and wished to explore similar problems further.

After completing my thesis, I joined the **Media and Data Science Research** lab at **Adobe** as an intern. Here I worked on building an efficient solution to detect and segment unseen (not seen during training) object classes in images. Image segmentation is crucial for visual understanding and is a common use case for software like Adobe Experience Manager – a digital asset management solution for brands like Audi, Nike, etc., to maintain their product images. However, supervised methods fall short because of the limited availability of good-quality annotated data. An automated solution to segment 'novel' object classes with minimal supervision can save a lot of time and effort. During my internship, I carried out an extensive literature survey of the domain and narrowed our focus to the **Few-Shot Image Segmentation** problem, *i.e.,* learning to segment a query image given only a few samples (1-5) as support for each unseen object class. I proposed a novel method that improves the similarity propagation between the support and query image features to get accurate segmentation. The method introduces two key components which provide a non-trivial improvement ($> 6\%$ mean-IoU) over prior methods. First, we jointly predict the support mask and the query mask to enforce self-similarity. Secondly, we introduce a novel foreground-background attentive fusion mechanism to utilize similarities in the background regions of the query and support images. The method achieved state-of-the-art performance on the one-shot and five-shot segmentation benchmarks. The work was published at **IJCAI 2020**, and a **US patent** has been filed for the same.

While working on this project, I observed that class-conditional similarity matching could only match pixels with a similar class mix between the query and the support images. Hence, I wish to explore introducing inductive biases via representations that capture the geometric structure and physics of the real world to improve image understanding. My internship experience enhanced my skills as an engineer and researcher. I learned to write well-documented and optimized code, and build applications to integrate my research project as a software tool.

Applying ML, Vision, HCI to Healthcare: To further my research experience, I joined the **Microsoft Research (MSR)** Lab as a **Pre-doctoral Research Fellow** (residency program). I am currently working with **Dr. Mohit Jain** and **Dr. Nipun Kwatra** on developing AI-based low-cost diagnostic solutions. Over the past year, I have worked on building a smartphone-based corneal topographer to diagnose keratoconus – a severe eye disease. Such a device can be highly useful in rural and remote locations where modern medical facilities are inaccessible.

A commercial corneal topographer is an expensive (∼$10,000) device with proprietary hardware and software. It projects a concentric ring pattern on the eye and reconstructs the corneal surface using ray-optics and 3D geometry principles. The generated topography maps highlight any deformities in the cornea. I directed this project from end to end, and after months of consistent efforts we built a working prototype. Building a low-cost device with the same fidelity as a commercial topographer required several key design and technical innovations. First, I designed a 3D printed conical attachment to project a concentric ring pattern on the eye. Second, I developed an AI-assisted smartphone app to ensure proper alignment and image quality with a hand-held device. Third, to process the captured images I implemented an analysis pipeline that involved image processing and a 3D reconstruction algorithm to generate the corneal topography. Further, due to the lack of hardware sensors on the smartphone, I developed an approximate vision-based algorithm to estimate the working distance (between camera and eye), which is required for reconstruction. And to automatically detect keratoconus, we trained a CNN-based classifier on the output maps. We conducted an extensive clinical evaluation of our system (at Sankara Eye Hospital) and found that it performed at par with the gold-standard medical device 'Optikon Keratron'. This project led to a publication at **ACM IMWUT/Ubicomp 2021**. The proposed attachment costs **$33** (**303x** less than Keratron) and weighs just **140 grams**, making it ideal for use by community health workers. We are now working towards deploying the device at Sankara clinics across India for mass screening of keratoconus, through which we envision early interventions and timely treatment.

My experience at MSR has broadened my perspective on how technology can have an impact beyond the realm of computer science as a field. I also noticed that there exists a huge gap between research and deployment. Once the system is deployed, many unforeseen scenarios can arise – system crashes, unexpected data (due to changing environment), and edge cases. These unexpected scenarios are not accounted for when working in a controlled lab setting. This emphasized to me the necessity of building robust and generalizable systems.

Ph.D. and Future Goals: After completing my Ph.D., I wish to continue in academia. I have found learning about new problems, conducting research, collaborating with fellow researchers, and teaching – fulfilling and enriching. A Ph.D. degree would help me build a strong foundation and depth to achieve this goal. This decision is informed by my research background and positive experience working as a teaching assistant for graduate and undergraduate level courses. My experience as a TA improved my understanding of multiple core subjects, and it was gratifying to see students indulge in critical thinking and ask challenging questions.

During my Ph.D. I wish to continue working on fundamental problems in Computer Vision and Machine Learning. I am particularly interested in deep learning with minimal supervision, and developing robust knowledge representations that model visual perception. My experience with deploying deep learning (DL) based systems exposed an important practical issue that arises when moving from a research setting to deployment. The DL model works well on fixed test data but fails to generalize on real-world data, where the distribution of images and the number of classes may differ. This is quite common and it is impractical to get densely annotated large-scale data for each use case. I wish to solve this by investigating techniques like transfer learning, domain adaptation, self-supervised learning to make machine learning systems accessible and easy to deploy.

At **UPenn**, I am interested in working with **Prof. Mark Yatskar** on research projects at the intersection of computer vision and machine learning. I am especially excited about deep learning for visually grounded representations to model machine intelligence. **Prof. Dinesh Jayaraman's** group engages with many vital problems in visual understanding along with applications to robotics that interest me, and I would love to continue my research with him. I am also interested in working with **Prof. Kostas Daniilidis** on his work at the intersection of 3D computer vision, graphics, and deep learning. My experience in computer vision and machine learning projects will add value to such research. Additionally, I would also like to collaborate with **Prof. Mingmin Zhao** on machine learning and health sensing. My work at MSR on developing smartphone-based diagnostic solutions aligns well with his line of research. UPenn has a strong and diverse faculty body, close collaborations across research groups, and hence, would cater well to my aspirations of conducting fundamental and interdisciplinary research. I believe my strong background in research, software engineering experience, and alignment of interests with faculty makes me a good fit for UPenn's graduate program.