APSIPA Transactions on Signal and Information Processing, 2022, 11, e18

Forum Paper

The Future of Computer Vision

Jingjing Meng[1], Xilin Chen[2], Jürgen Gall[3], Chang-Su Kim[4], Zicheng Liu[5], Alessandro Piva[6] and Junsong Yuan[1]*

[1]*Department of Computer Science and Engineering, University at Buffalo, The State University of New York, Amherst, NY, USA*
[2]*Institute of Computing Technology, Chinese Academy of Sciences, China*
[3]*Department of Information Systems and Artificial Intelligence, University of Bonn, Germany*
[4]*School of Electrical Engineering, Korea University, Korea*
[5]*Microsoft Corporation, Redmond, WA, USA*
[6]*Department of Information Engineering, University of Florence, Italy*

ABSTRACT

This article summarizes the panel discussion on "The Future of Computer Vision," organized by the U.S. Local Chapter of APSIPA on August 22, 2021. This panel brings together some of the world's leading experts in Computer Vision (CV) to discuss "The Future of Computer Vision." The panel is composed of academic and industry leaders from three continents (Asia, Europe, and North America), who have made key contributions in various CV topics. The panel discussed and debated on different future issues, including but not limited to, the trend and impact of CV, the role of CV in combating social injustice, the impact of deep learning, emerging areas, and how to better prepare for a career in CV.

Keywords: Computer vision, reliability, deep learning

*Corresponding author: Jingjing Meng, jingjing.meng1@gmail.com.

Received 31 October 2021; Revised 18 March 2022
ISSN 2048-7703; DOI 10.1561/116.00000009

1 Introduction

[Jingjing] In recent years, we have witnessed great progress in computer vision (CV) research. Leading conferences in CV, such as the Conference on Computer Vision and Pattern Recognition (CVPR), and the International Conference on Computer Vision (ICCV), have grown from a few hundred attendees to around 10,000 in recent years. With strong government and industry support in this field, a growing number of participants are getting into CV. Many start to wonder whether we are heading in the right direction. What are the challenges we are facing in current CV research? How should we address these challenges? What is the role of academia in this aspect? What should we focus on next? How can we better prepare ourselves for the next wave?

To answer these burning questions, on August 22, 2021, the U.S. local chapter of the APSIPA brought together the world's leading experts in CV to discuss "The Future of Computer Vision." Given the large attendance and interest from the audience, we decided to produce an article summarizing the opinions of the experts. The members of the panel are the authors of this paper. Our panel discussions focused on four areas: (1) technology outlook, (2) challenges, (3) career advice, and (4) computer vision and society. The audience also had the opportunity to ask questions and interact with the panel at the end of the panel discussion. The panel discussion had close to 90 attendees. We summarize the main points in the conclusion section.

2 Technology Outlook

2.1 Robustness of CV Techniques

The first question posted to the panel is "Nowadays, CV technologies has been applied to mission critical applications, such as autonomous driving. However, in the past, CV was believed to be best suited for non-mission critical applications due to its robustness issue. Do you think this issue has been resolved and why?"

[Jürgen] More than 10 years ago, we were happy that things started to work, such as face detection [15] and pose estimation with Kinect cameras. But the robustness is still one of the big challenges and we have not really addressed it. Autonomous driving is a very good example. For instance, a car was driving in a construction site and the driver-assistance system of the car made a full break even though no other cars were around it. This actually shows that the robustness is still a huge problem. This topic is still underrepresented in the CV conferences. We are getting better results for some applications, but do things really work in 99.999% of the cases? In many cases we are still very far

away and achieve maybe 90 or 92 percent accuracy. Making systems reliable and closing the gap is still a big open problem.

[Zicheng] I agree. My team builds products using CV. Reliability is an issue not just for mission critical applications. Even for non-mission-critical products, it affects the value proposition. If something does not work very well, e.g., works 80% of the time, then for the other 20%, either customers are not happy, or we will need to have engineers to spend time to improve the system. That affects the cost. So reliability is a big issue, and we need to make it much better. That's why Jürgen said 99% is not enough when you really need 99.999%. It is actually not a ridiculous number, because some customers do want that number. They think that if you cannot achieve that number, it does not make a good value proposition for their application. For instance, in autonomous driving, near 100% is important, as it could mean the difference between life and death. So it is a very practical issue.

2.2 Explainability of Deep Learning based CV Techniques

The second question posted to the panel is "Is the lack of reliability and explainability of deep learning based CV tools worrying you? What's the role of CV researchers in this regard?"

[Junsong] To me, the reliability and explainability are highly coupled issues. The community has talked about them for a long time, but explainability has never been very clearly defined. We know better about reliability. Later on, I realized that actually these two have a strong correlation. If we have very good reliability, for example, for whatever reasons can achieve satisfactory performance, maybe explainability is less of a concern. However, if the accuracy is not satisfactory, then explainability becomes more critical, because AI may need to explain at what situations that errors will happen so that we can prepare ahead; Second, if AI is going to fail, it is better to explain why it fails, so that we may know how to improve it. To me, it is because of the not so reliable performance of the approach, we want to ask why, to understand how and when we can use this CV technology and how we can improve it [33, 36].

[Alessandro] The explainability is very sensitive in some applications [42], for example, multimedia forensics [40], where some tools for image and video integrity verification have been developed. They can help to identify if an image/video is authentic one or a modified/fake one. Such methods could be used as an example in front of a court to know if an image could be used to accuse some person. This is a very sensitive scenario for explainability. If you are not able to explain why we achieve a given result using the multimedia forensics method, we cannot use it in front of a court [7, 32]. On the contrary, if we use algorithms based on statistical models, then we can justify the results,

so explainability can be achieved. In many cases when we use deep learning, it is not yet possible.

[Zicheng] I think that the role is to make them better, make them more reliable, e.g., improving the precision and recall. It's kind of vague to say "making it explainable," despite a good intention. Right now, there is an area called *interpretable computing*, and people come up with different ways to show that their model is explainable [31, 42], (e.g., Figure 1). But that may not match what the users think of explainability. It is like a North Star to put it there, but the North Star is not clearly defined just yet. As CV researchers, we still have a lot of work to do, which is a good thing.

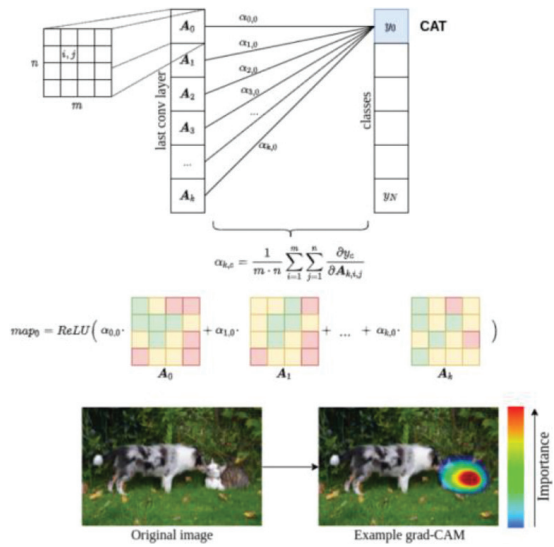

Figure 1: (Cited from [42]). Visual explanation of how grad-CAM works.

[Jürgen] I agree that explainability needs to address the question "explain to whom?" first. Earlier we had a very good example of explainability for law cases. Explainability needs to be studied in the context of specific user groups, otherwise it is meaningless. For most customers, explainability does not really matter so much. For instance, if your car breaks down every 1000 km and it tells you this component broke and that component broke, you are annoyed that your car does not work, but you do not care why it does not work. This information, however, is relevant for repairing the car. There are different types of explainability, e.g., as debugging tools for developers, or for the end users. I think one has to consider explainability to whom, and we may need to develop different tools as CV researchers [18, 19, 31, 36, 42].

2.3 Deep Learning based Approaches vs. Principled Approaches

The third question posted to the panel is "Considering the pros and cons of using deep learning based vs. principled/handcrafted tools, what should the CV community focus on next?"

[Zicheng] For industry, deep learning gives you a way to achieve certain usability in terms of precision and recall. On the other hand, it can be expensive to build a machine learning product because it requires so much labelled data. And if this doesn't work, you do not know why, and you are not sure when you are going to achieve the desired performance. So it's really a dilemma. Regarding what the CV community should focus on next, I think even right now many great researchers have been thinking about how we can do better, by maybe changing the current systems in a more dramatic way. I know that Jay has been doing really great work in this direction, trying to change the current system to make the learning process more explainable [20, 46]. It is a very different paradigm. I am sure that there are many other researchers who are thinking about it. I am sure that there will be people who continue the current progress. And there will be people who are thinking about more fundamental ways for machine learning.

[Chang-Su] I think that deep learning is very good in some sense but only for certain applications. Traditional techniques are more reliable, for example, in interactive segmentation. So combining deep learning with traditional techniques is also important in my opinion.

[Jürgen] Deep learning models are powerful in settings where we have very large datasets that we can use to train the networks. But just talk to people from other disciplines, they do not have so much data. You may have to wait an entire year to get a collection of data, and then you have to wait another year to get the next collection of data, for instance, in agriculture applications. So, you cannot easily collect 50 million images which are already available over the Internet. There is also an interesting discussion on how you can actually combine deep learning approaches with traditional approaches, for instance, how can you combine deep learning approaches with symbolic representations. The old-fashioned AI approaches, which are more based on symbolic representations, cannot be easily combined with deep learning, because deep learning has the constraint that everything has to be differentiable [7, 36]. This constraint, that has been imposed, has worked very well, but it is unclear if everything must be differentiable at the end of the story.

[Xilin] Although deep learning has been very successful in the industry and some typical tasks such as face recognition, it is still very hard for complex tasks such as service robots at home, which can do cooking, cleaning, or even baby-sitting. The reason is that deep learning is largely dependent on the

specific task, e.g., person detection, and large scale data collection for training. We cannot make a universal machine with vision system handling all different tasks together. In my opinion, deep learning should be the building block for future complex vision systems. So for CV researchers, we need to go back to focus our research on the mechanism of the vision system itself. Deep learning is only the building block for us, and the functional module for some basic tasks.

2.4 Intersections between CV and Other Disciplines

The fourth question posted to the panel is "Beyond machine learning, how would cognition, neuroscience and other disciplines assist in vision tasks? Or would they assist?"

[Junsong] I think those disciplines beyond machine learning, such as cognition, neuroscience are playing more and more critical roles [22, 43] (Figure 2). One of the reasons is that eventually we want the machine to be able to work together with a human. Human has eyes, so we can see the world and understand the world through our eyes. Nowadays, many machines have cameras, i.e., they have their own eyes. If you really wanted the machine to be intelligent, for example, for service robots that can help you in your daily activities, the machine also needs to understand what the human is doing and why they are doing this. Now lots of work has been focused on understanding what the human is doing [9, 28, 38, 41]. But on top of that, an even more critical question is why they are doing this, how they are doing this, before the robot can offer meaningful help. In terms of that, I think cognition is important to understand and to help bridge the gap so human and machine can work together in the future.

[Zicheng] I want to add that language will change CV in a fundamental way, because when we describe an image, we need language. Currently the vocabulary becomes a big problem in CV tasks, e.g., object detection. When people try to expand the vocabulary for object detection, the relationship between the words becomes a problem, for example, words have semantic overlaps and may conflict with each other. So I think that CV needs language to be an integral part of it, and language will bring new changes to CV methodologies.

[Xilin] In the early stage, CV was not focused on knowledge or understanding, but targeted to the problems such as categorization and 3D reconstruction, which are essential functions of CV and animal vision. Once we pay more attentions on understand the world, cognition and neuroscience are definitely important. For the purpose of measuring the degree of understanding the world, language is an important tool, especially for the high-level CV tasks.

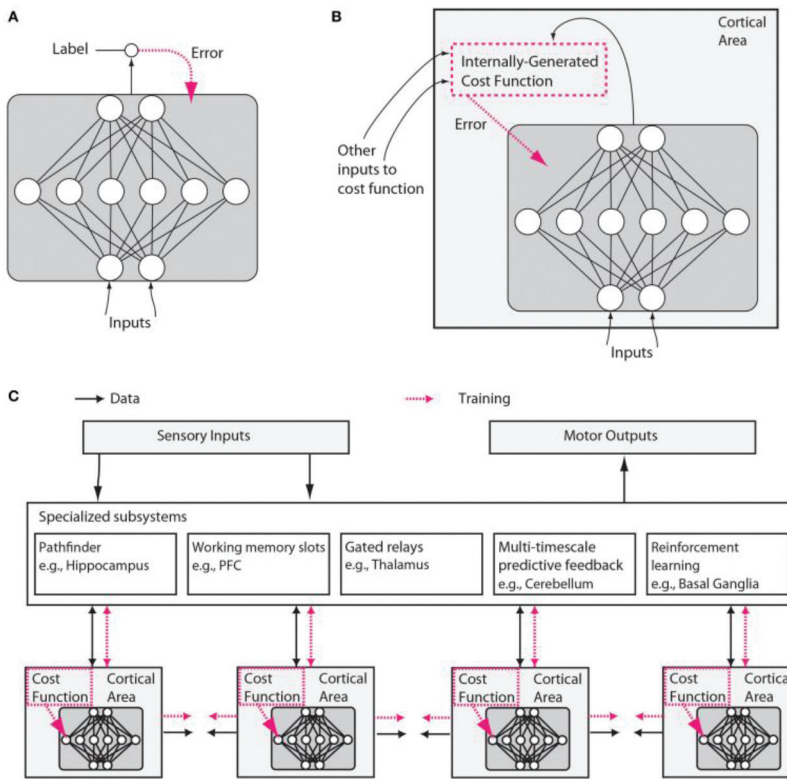

Figure 2: (Cited from [22]). Putative differences between conventional and brain-like neural network designs. (A) Supervised training via conventional deep learning (B) supervised training of networks in the brain. (C) Internally generated cost functions and error-driven training of cortical deep networks. More details can be found in [22].

Therefore, it is very necessary to make these two parts working closely, for instance, in visual-question-answering (VQA) tasks, trying to understand what happened in an image or a video [2, 35, 39]. That's why in recent years, we see lots of papers in CVPR, ICCV on related topics.

2.5 The Ultimate Goal of CV

The fifth question posted to the panel is "Should the ultimate goal of CV be to mimic human vision or to create super vision that is beyond human capabilities?"

[Xilin] In general, the answer is YES. A CV system with higher temporal and spatial resolution, broader spectrum, capturing signals beyond 2D is definitely

the dream of CV researchers. Meanwhile, it's possible to reach such a goal of super vision system as we have more sensors to utilize, which give us a much wider spectrum of sensory input. For instance, Microsoft Kinect has changed CV dramatically by providing RGBD images instead of RGB only. So that's where super vision is needed. Another example is for healthcare, where enhanced information is leveraged by different sensors to obtain more detail signals about the human being, which cannot be directly obtained from the RGB images. Again, it depends on your area.

[Junsong] If you look at the autonomous driving scenario, there were debates on whether we should only use the optical camera to mimic human vision and drive, or we should use other sensors together to learn driving, such as camera, Lidar, Radar. Today the more reliable autonomous vehicles use different kinds of sensors including RGB camaras [3, 5]. I think that is one of the advantages of using super vision, i.e., to achieve better reliability. However, I still believe that if the task is for the camera to interact with the human, then mimicking human vision, including understanding human languages [14], will become necessary.

[Zicheng] I always see computers as a tool to human beings, not to mimic a human. So as a tool, you will never replace human beings. I don't believe that computers will be the same as a brain, because brains are biological, while computers are silicon. The two materials are fundamentally different. So that may be my personal bias, but I view computers as a tool to humans, to be the best assistant to human beings. To this end, as long as the computer does useful things for the human society, it does not matter whether it is super vision or human vision.

[Jürgen] There is a more general question. Currently, we build very specific tools, which is very different to humans, who can do lots of different things [6, 10] at the same time. From the economic perspective, we have specialized tools that work very well, but also tools that can be used for very different tasks. The question is, from an economic point of view, whether there is actually a need to have more universal approaches, which are not limited to solving a single task. From a research perspective, having a more universal tool is quite attractive. However, from the industrial point of view, I am a little sceptical whether this is actually something desirable. For specific tasks, machines are doing a better job than humans. This is obviously the goal you want to achieve if it is for productivity, because machines do not get tired, you can use additional sensor technology and get a higher accuracy. There is, however, the research question whether it is possible to build a universal CV tool, given that we have developed for the past 20 years specific tools which are used in different applications.

3 Challenges

3.1 Possible Challenges for CV Research

The first question posted to the panel is "Despite the successes of CV applications, do you see any challenges CV research is facing?"

[Junsong] Some challenges from my personal perspective is the way deep learning has influenced our way of thinking. Deep learning is a great tool to help machine learning models to achieve good performance, but on the other hand, if you look at the papers, for example in CVPR, 10 years ago and today, you are going to see a big difference in the way we approach CV tasks. For instance, if you do 3D vision, in the past you need to first study multi-view geometry instead of machine learning [13]. While today, regardless of the CV task, from object recognition, detection/tracking, scene segmentation, to depth estimation, optical flow, and view synthesis, the way we approach the problems becomes similar, e.g., through data-driven approach [4, 44]. Give me training data first to build the regression model, if the performance is not good enough, can we get more data, e.g., through data augmentation, to fine tune? I have no objection to this approach, but if for all different CV applications we apply the same pipeline without questioning if there is any other alternatives, that could be an issue. It is the lack of the diversity of the approaches that worries me. And we tend to believe that deep learning is the only way to go. If this trend continues, I do not know how it is going to impact CV in the long term.

[Zicheng] Despite the fact that right now the CV conferences are very big with a lot of papers, I see that there is a trend of CV becoming an application of machine learning. That is a danger. If the trend continues, there will be no CV discipline anymore because one can just apply machine learning and that's it. So that is really a big challenge to the entire CV research community. I do not think that the current way of deep learning is going to solve all CV problems, so I do believe that at some point people will come up with ideas to change the current learning paradigm, maybe that's going to be the renaissance of CV. Currently behind the booming of CV there is a big danger. I don't have a solution, but I feel that at least we need to inject the traditional model-based thinking there. Traditionally we describe objects using geometry, motion, etc., basically try to model objects first [1, 27, 29]. Nowadays we extract the features first. So they are very different approaches. Integrating the currently prevalent data-driven approaches with the traditional model based approaches may help save some effort on the data-demanding aspect. I believe that new learning paradigms will emerge.

[Chang-Su] Nowadays about 10K papers are submitted to CVPR every year and 25% papers are accepted. We can safely assume that not all of them are

good papers. It is okay that a not-so-good paper is accepted, but it is not okay that an important idea is rejected. When there are too many papers, it breaks down the review system. We need to find a way to select good papers and good ideas from so many papers. Another thing I worry about is that since it is becoming too competitive, people care about state-of-the-art performance only, which suffocates creative ideas. We spent so much time in fine-tuning algorithms, but so little time in developing new ideas, analyzing the hidden mechanism, and formulating the problem. In academia, we need to teach how to develop ideas and how to think analytically. I feel that we are not doing very well in this aspect right now.

[Jürgen] We also mentioned previously the robustness issue, for instance, how do we deal with an object that the network has never seen [12, 11]. At the moment, the networks provide some confidence for a closed set, but how about open set recognition? When the deep learning network sees for the first time an elephant that it has never seen before, the network will give you a confidence score, but this confidence is not really reliable. How to deal with this kind of less restrictive open set problems is also a difficult challenge.

3.2 Combining Different Sensors to CV Research

The second question posted to the panel is "What are the challenges and opportunities of combining optical sensors with other sensors for CV research?"

[Zicheng] Obviously the depth sensors are very important, especially the long range ones. Current version of Microsoft Kinect has a very short range, which is a typical limitation of current consumer depth cameras. I think that the Lidar sensors are really good, as they can see objects in the distance [21]. I'd love to see the Lidar sensors become cheaper. Obviously if you can see 10 m or 150 m away, it is going to change the field dramatically. So good sensors are really important. RGB cameras are not sufficient for lots of applications that require 3D understanding. Therefore, I hope that depth sensors eventually will have longer range, larger field of view, higher frame rate and become cheaper.

[Alessandro] There are other applications where we have many frequency bands, such as remote sensing applications [48], or multispectral image analysis for diagnosis and restoration of paintings and other cultural heritage goods [26]; there we have a lot of frequency bands, so we have a huge number of images that we have to combine together; if we need to use deep learning, we have to collect a very huge amount of information, so I think that in these particular cases, it is challenging to use deep learning approaches.

3.3 The Role of Journals in CV

The third question posted to the panel is "In CV research, as publications in conferences such as CVPR and ICCV are popular nowadays, in this situation, what will be the role of journals – for example, JVCI – in the field of CV?"

[Zicheng] I am sure that many researchers have this question. First of all, some papers are not suited for publication in conferences. You may have seen in CVPR type of conferences where the reviewers may give you feedback that your topic is not well suited for the conference, even though the criteria are not well defined. Definitely, there are some works better suited for journals, instead of for conferences. Second, because of the way conferences run, the turnaround time from submission to acceptance decision is very short. Even though there is the rebuttal period for most conferences, the short time makes it difficult to change the fate of many papers. On the contrary, journals in general allows a paper to be revised and improved multiple times, which is an advantage. In addition, I think that another aspect of journals is the archiving service they provide, meaning that after 20 years, you can still find a journal paper, versus finding the conference papers. The journal publishers usually organize or maintain journal papers better than conference proceedings. So, I think that journals will continue to play a role.

[Jürgen] In conferences you can present your ideas in front of the audience directly, although it is not the case at the moment as conferences are virtual due to the pandemic. As Zicheng mentioned, one difference is in the review process. For journals you can revise your paper after receiving the feedback from the reviewers and it will be reviewed again by the same reviewers. Another aspect is that you do not have a strict page limit for journals. Often you propose your idea first in conferences, and when your idea becomes more mature, you submit an extension as a journal version. I think that this is also important for young students. At the end of your PhD study, you will have this full-fledged version of your idea with a detailed analysis in a long journal publication. The importance of journals is also evident in that the longstanding works, which have been presented in conferences, usually have the extension published in journals.

[Chang-Su] I agree with Zicheng and Jürgen. I want to add that the role of Associate Editors is becoming more important in journals. Some journal reviewers treat journal submissions as conference papers, for instance, asking you to compare with the latest state-of-the-art (SOTA). Then after a round of revision, new SOTA comes out, so you need to add more experiments. Therefore, the AEs should watch out for such reviews, otherwise the journal papers will lose their merits.

3.4 Reviewing Process in CV

The fourth question posted to the panel is "Because of the popularity of CV, there are too many to review. How can we improve or revolutionize the review process to select good papers from so many papers?"

[Zicheng] One radical idea: maybe for conferences, we do not need reviewers. Just publish all papers on arxiv, we are doing it anyway nowadays, let people read them. People will have a way to find good papers, cite them, and talk about them. I think that this is an automatic filtering process. Based on the ratings, or download numbers, the good ones can be identified, then presented at conferences. These papers will go through a more rigorous review process if the authors want to extend the work to a journal publication. Right now, each CVPR paper is reviewed by three reviewers and there are some random factors in reviewer selection. Three reviewers are not enough to decide the quality of a paper, especially when the reviewers have vastly different opinions. Having many more people read the paper and comment on it will provide more statistically meaningful reviews. Of course we would need to prevent fraud, as this system can be easily gamed.

[Junsong] Some of the ML papers are doing this, e.g., OpenReview. There's being a long-time debate on this. There are certainly pros and cons. On the positive side, OpenReview tries to promote openness in scientific communication, particularly the open peer review process.

[Xilin] For ICML and other machine learning conferences, the major motivation for OpenReview is to ensure that the reviewers are responsible as everyone can read the comments from the reviewer. Publication of papers is different from the publication of books, where the publisher takes full responsibility for the quality of the book. To publish a book, the publisher will ask the author(s) to write one or more chapters and outline of the proposed book at beginning. Then the publisher will quickly work with the author(s) to review the outline and sample chapters, only after this stage the author(s) will get to write the whole books. For the conferences right now, I do not think that the review procedure is a big deal. Even we have 1000+ submissions, it means that thousands of researchers are working in this area. That is why some conference such as NeurIPS just recruits all the authors as their reviewers. To keep the quality of the reviewers, these conferences prefer to post all reviews in OpenReview to make the reviewers more responsible. On a related note, for journals, I think that SAEs and AEs should take more responsibility for desk reject, as many paper submitted are not even qualified for reviews.

[Zicheng] I agree. I have seen people submitting papers way below the bar for either conferences or journals. Even back in the 90's, we were discussing

how to prevent people from submitting obviously low-quality papers. Basically that is a waste of the whole community's time. There was an idea of charging a fee for each submission, to penalize those who submit low-quality papers. It could be a factor in terms of how to improve the system.

4 Career Advice

The question posted to the panel is "How should the course content of CV be adapted to meet future needs? Is it sufficient for students to just learn how to use pytorch, tensorflow? Are traditional foundational courses on image processing still important? What other suggestions do you have for students who are interested in pursuing a career in CV?"

[Jürgen] I think that it is not enough to be able to write a data loader. You should have very good programming skills. In industry, it is not enough to have some experience with Python. It is also important to have a broad background knowledge, because, at the end, many ideas are actually taken from traditional approaches. The neural networks use a lot of basic components from CV and image processing, such as how can we enhance or speed up the way how convolutions work, or geometric transformations. I think it is very important to have an overview of different ideas.

[Alessandro] I agree that just knowing how to use pytorch or tensorflow is not enough. I am not an expert on it, so probably have a bias on this direction. But if you do not know enough about your data collected for training and testing, there could be a problem. As your data can be biased, even if you obtain very good results, it could be due to the fact that your network is learning this particular bias and not indeed doing a good job. So I think that you should have some basic know-how about the source of your data, why your data are done in this way, and from where they come.

[Junsong] I think that image processing is still a very relevant course to CV. I teach a course called CV and image processing this fall. It was offered as two separate courses in the past, but we eventually combined them. It looks like students are more excited about CV than image processing. But I want to give you two reasons why image processing is still very relevant to today's CV. First, if we look at convolutional neural network (CNN), which is the popular tool we use today, the neural network and back propagation are invented by image processing community, but the convolution has a deep root in signal processing and image processing. Even with today's deep learning models, I believe it is still quite valuable to understand Fourier Transform, the relationship with convolution, and knowing the spectral analysis of signals. Also, if you look at the recent new successful deep learning tools, for example, graph neural

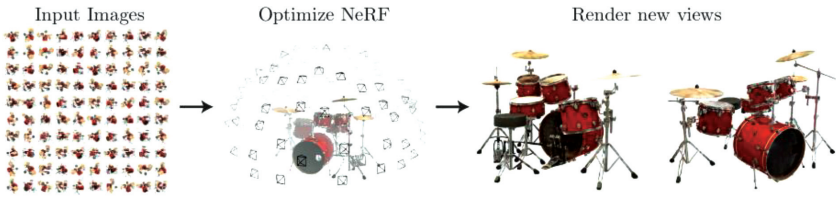

Figure 3: (Cited from [17]). Graph Convolutional Network.

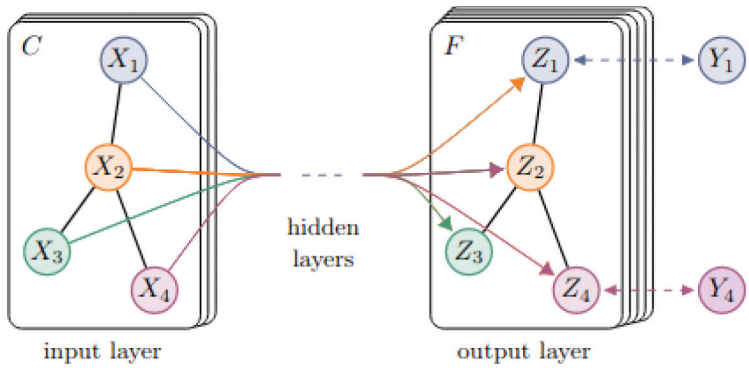

Figure 4: (Cited from [24]). Optimization for a continuous 5D neural radiance field representation.

networks, which is highly relevant to the graph signal processing [17, 34] (Figure 3). If you do not understand graph spectrum, you will have problem understanding how graph neural network works. Also a recent work for view synthesis, NeRF [23, 24, 37] (Figure 4), for synthesizing a new view from existing multi-view image collections, incorporates positional encoding using Fourier features to improve the high frequency details. From my interactions with my students, who are working on this task, those who had the background in image processing are much easier to appreciate the algorithm. So I believe that image processing is still very relevant.

[Alessandro] I want to add another note because I also teach image processing. I think that image processing courses need to be updated. For example, if we take a book on image processing, you can still read that the in camera processing just includes basic operations like color filter array demosaicking, gamma correction, and white balancing. But if you just take any smartphone nowadays, inside its camera, there are so many different processes that we do not even know. As teachers, we should know what happens inside the camera device, otherwise, we are not able to explain the processes that have produced a given digital image. So we also need to update the image processing part.

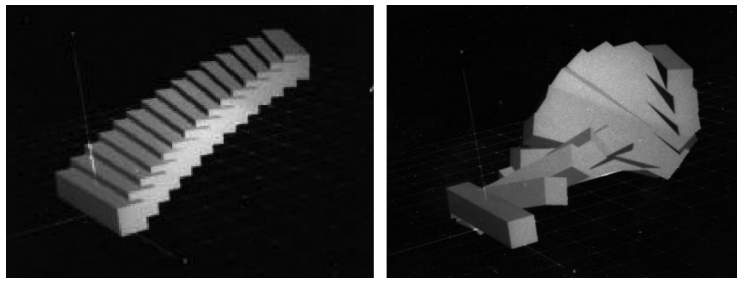

Figure 5: (Cited from [30]). Screw motion (left) versus linear interpolation of pose parameters (right).

[Zicheng] One quick comment. I hope that CV courses will continue to teach 3D vision. Traditionally in CV, the main problem was to solve the structure from motion problem [8, 30] (Figure 5). That was a classical CV problem, and so far it still has not been solved. But I see that many students in CV nowadays have never learned anything about it. They do not know how to use two views to estimate depth. I heard that many universities do not even teach 3D vision at all, because people think that it is not as useful. However, in industry, many of the applications that we have do require 3D understanding. So I hope that the curriculum should include 3D related concepts and techniques in CV.

[Xilin] Yes, I think that geometry may be the most beautiful parts in CV. I really like it. Today when we talk about 3D, yes, you can get 3D just from data. However, if you try to understand the underlying rules just from data, you would need a lot of data. On the contrary, if you have the geometric model or some physical model, it is become easier. Right now in the deep learning era, we can model the geometry or some physical as data constraints. For instance, this year we have a work published in CVPR, trying to get the 3D information using the geometry as a constraint [25]. It does make everything quite easy. That is why I think geometry is still a very important part. Also, as Junsong mentioned, a lot of people think that CNN is hard to understand. Once you know that the basic convolution comes from signal processing, and that we just change from human hand-crafted filters to data-trained filters, everything becomes easy to understand. So we need to understand the basic principles behind the data.

5 CV and Society

5.1 CV for Combating Social Injustice

The first question posted to the panel is "What's the role of CV in combating social injustice (or is it making it worse, e.g., adversary attack, recognizing people of color less accurately)?"

[Chang-Su] Like any other technology, CV can be used in bad ways as well as in good ways. It is a critical problem that face recognition methods less accurately recognize people of different colors. Technically, we can solve the problem by constructing well-balanced datasets. The fact is that even good tools can be put into bad use. As engineers, we should cooperate with other members of the society to prevent it, such as teaching them that these tools have limitations and there are no error-free methods.

[Jürgen] I think there is also a related discussion in the AI community, not only in CV. Twenty years ago, we were so happy to achieve at least something which is close to human performance. I am not sure how this exactly started, but some also believe, that there is this super AI, which solves all the problems of the society. Obviously, if the training data comes from human annotations, the network will also be biased in a certain direction. I think the expectation is a bit too high, as it is not that straightforward to have the AI completely free of any kind of bias. You have to really get rid of all the bias inherent in the data, which is human since humans always have a bias. However, as Chang-Su has mentioned before, it is important that the awareness is there. When we also teach people to be aware of the bias, this has a positive effect. CV can be very helpful for many critical tasks, but it can also be abused. So, we should be aware of it. For nearly any technology from computer science, the technology can be used in a positive and negative way.

5.2 Government/Industry Supports for CV Research

The second question posted to the panel is "What kind of government/industry supports for CV are available in the three continents where our panelists are from?"

[Chang-Su] In South Korea, there are a lot of research funds for AI research from government. CV is regarded as one of the most important applications of AI. From the industry side, Samsung and LG are strong supporters. Recently Hyundai Motor and other major companies in Korea joined the force as well. So in terms of funds, it is a good time for CV research in South Korea.

[Xilin] In China, we have a lot of researchers working on different tasks in the industry, from autonomous driving, vision-based products assessment, to medical imaging. Therefore the job market for CV attracts a lot of students trying to get into this glory area.

[Jürgen] Under the large umbrella of AI, there are many worldwide programs, including Europe and Germany. There are so-called competence centers for machine learning in Germany. You can find these opportunities when the government makes an announcement on an AI program. It is interesting to

see how many groups are doing AI right now, even though they used to be in other disciplines, which is a good thing.

[Alessandro] I would say the same. AI is one of the key priorities in Europe right now, so there is a strong interest in it. Also, there is an interest in what is called trustworthy AI.

[Junsong] I think that the US has also started to invest much more in AI. For example, NSF funded a number of AI institutes recently, with diverse topics in AI theory and applications, which cover from agriculture, healthcare and manufacturing etc.

6 Future Directions of CV

The last question posted to the panel is "Where do you see the CV research going next 5–10 years?"

[Zicheng] Five to ten years is a long time. In 2011 few people had predicted that deep learning was going to be so successful. It is really hard to predict what is going to happen in the next 5 years, but I do think that right now people start to feel the pain of data labelling and are thinking of ways to overcome this data problem, i.e., learning without that much data [45]. Another possibility is to inject some of physics into the learning system. People talk about common sense, and physics is a type of common sense. Moving forward, we may want to inject physical models, such as geometry and motion, into the learning system.

[Jürgen] It is always difficult to predict the future. At the time before we had the ImageNet competition, I had never seen such a huge shift in such a short time in terms of the methodology. Maybe we will have another shift in 5–10 years, but I think that the neural networks will definitely still be there, at least as building blocks. Looking forward, we might get a little bit away from the current dataset driven approach. Also, right now we focus on content analysis like detecting objects or recognizing actions, but there is not so much work on real understanding. We have approaches that generate captions for images, but this is not yet a real understanding. Maybe in 10 years, we can actually move on to have this high-level understanding as humans do. It is not just having a sentence, but to really get a clear structure of what is going on. There may be a holistic approach that from whatever kind of image data generates a full set of descriptions, instead of a few labels.

[Chang-Su] I think that it is very difficult to predict the future, but I hope that the smart people here in CV will come up with more elegant theories to explain the experimental success of deep learning.

[Junsong] I also hope that in the next 10 years, even if we still use deep learning, deep learning will become the tools of our CV research, rather than humans become the tools of deep learning, e.g., data annotation and cleaning.

7 Question from the Audience

Due to the limit of time, we only had time for one question from the audience, which is below from Prof. Jay Kuo:

[Jay] I see two areas that used to be very hot and now becomes cold. One is computer graphics. I remember those years when SIGGRAPH was really popular, it attracted all the attention and gradually became dominated by industry. As the bar becomes higher and higher, people from academia started to feel that their work cannot compete with the quality of works from industry. Consequently, researchers from academia gradually left computer graphics field. Gradually, the conference becomes mostly beautiful contents without much novel ideas, and SIGGRAPH does not retain its prestigious status anymore.

Another field is coding and compression. Image and video coding was very hot for a period of time, but then gradually became more and more specialized, higher and higher complexity. Those labor-intensive work may be suitable for industry, but is not suitable for academia. So academia people left. Very few professors wanted to do image/video coding, and the field started to lack young talent. I am worried that similar things will happen to CV, as the bar becomes higher and higher for the data and the GPU requirements. If you want to get good performance, you will need much more resource, which the academia people cannot afford. Although there are a lot of papers, even with academia, most are internship works, i.e., students had internship with some company and then bring back the work. Few works are really from a university environment, because the university cannot afford similar kind of computing environments and data as industry. I think that once academia people start to be pissed off by this kind of harsh environment, they will leave. Again, conferences attract people because of innovation, not just beautiful products and polished products. If you do not give a space for academia people, if they do not join, it will eventually ruin the field. That is my observation, and I would like to know whether the panelists have some comments along this line.

7.1 The Role of the Academia

[Zicheng] I want to echo what Jay said. I feel like that, right now, deep learning has converted CV into engineering problems. You need resources such as data and GPUs to get the best performance. It is not a fair competition between academia and industry. But I personally think that the role of

academia is even more important than before, which is not really to compete on the engineering side, but to innovate. I think this is a great time for professors if you think differently, if you think in a disruptive way instead of following the industry trend, e.g., data-efficient training [16, 47]. Jay's team has done some really impressive work in a totally different direction against the trend. We need more thinkers like Jay, because only people in academia have this kind of freedom to think differently. I have to say that doing CV using deep learning shouldn't be the future. There will be better ways to learn without so much data, as that is how humans learn. Although nobody knows how to get there yet, we cannot just keep following the industry, which could lead us to the end of CV.

7.2 Challenges for the Academia

[Jürgen] I also agree that this is a major concern. There have always been differences in the available resources between academia and industry. But if the required resources to do research become very high, it becomes uninteresting for academia. The resource issue is amplified by the focus on accuracy. If I have more hardware, I can train my network more often than someone from academia to get better hyperparameters and better results. That means that it is more likely that the paper will get accepted. This is problematic and needs to be somehow addressed in the review process. Accuracy is important, but it should not only be about accuracy. Otherwise, many groups in academia will get squeezed out because they do not have the resources to run so many trials for optimizing just the hyperparameters. The other issue is recruiting of postdocs. As student you can join industry with some freedom to do research while getting much better paid. There is therefore a lack of postdocs in academia. In the long term, if there are no postdocs, there will be no professorships.

8 Conclusion

[Jingjing] Overall, the panel is positive about the future of CV. In all the three continents where the panelists come from, currently governments and industries are very supportive of research in CV (often under the big AI umbrella). The research grants and job markets are promising. Our panelists acknowledge the successes of deep learning based approaches in recent years, and agree that deep learning will be here to stay. However, they also caution that long-standing challenges such as the reliability and explainability issues in CV have not been addressed yet despite the current hype. Therefore, there are still much work to be done for CV researchers. Moving forward, they foresee new approaches beyond deep learning will come out. Possible

directions suggested by the panelists include exploring other disciplines such as cognitive science and language, understanding fundamental models of CV, and bringing in common sense knowledge and physical models. Career-wise, the panelists suggest that students who are interested in pursuing a career in CV do not solely focus on learning the deep learning tools. Courses in image processing, 3D vision and geometry will help build a solid foundation in the long term and should not be overlooked. As for any other technologies, CV techniques may be a double-edged sword. Therefore, CV researchers should bear in mind the inherent bias from data or learning algorithms that could reinforce social injustice. And it is part of their responsibilities to educate the public about the algorithm's societal implications. In 5–10 years, the panel foresees that researchers in CV will come up with solutions to overcome the data-intensive problem with deep learning, better understand deep learning models, incorporating common sense knowledge into learning. The hope is that even if deep learning continues to play a role in CV research, they will be regarded as one of the tools, not the only way to solve CV problems. At the end, the audience and panelists discussed the current trend of industry overwhelming CV conferences due to the resource-inequality between industry and academia and performance-focused review outcomes. The panel suggests that the CV community should be mindful of its potential adverse impact on academia research and the CV field itself.

Acknowledgement

We had close to 90 participations on Zoom. The authors acknowledge the support of the APSIPA U.S. Local Chapter, the active participations of the audience, APSIPA and many individuals/groups who have helped to promote the panel. Special thanks go to Prof. Kenneth K. M. Lam at the Hong Kong Polytechnic University for hosting the panel recording on the APSIPA server, Nan Xi and Jialian Wu at the University at Buffalo for technical and reference assistance.

Biographies

Jingjing Meng received the B. Eng. degree in electronic and information engineering from Huazhong University of Science and Technology, China, the M.S. degree in computer science from Vanderbilt University, U.S., and the Ph.D. degree in electrical and electronic engineering from Nanyang Technological University, Singapore. She is currently an Assistant Professor of Teaching and Research at the University at Buffalo, U.S. She is a senior member of IEEE.

Her main research interests are in video/image content analysis, visual search and discovery.

Xilin Chen is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS). He has authored one book and more than 200 papers in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multimodal interfaces. He is a fellow of ACM, IAPR, and CCF. He was a recipient of several awards, including China's State Natural Science Award in 2015, and China's State S&T Progress Award in 2000, 2003, 2005, and 2012 for his research work.

Jürgen Gall received the B.Sc. degree in mathematics from the University of Wales Swansea, in 2004, the master's degree in mathematics from the University of Mannheim, in 2005, and the Ph.D. degree in computer science from the Saarland University and the Max Planck Institut für Informatik, in 2009. He was a postdoctoral researcher with the Computer Vision Laboratory, ETH Zurich, from 2009 until 2012 and senior research scientist with the Max Planck Institute for Intelligent Systems in Tübingen from 2012 until 2013. Since 2013, he is professor with the University of Bonn and head of the Computer Vision Group. He is a member of the IEEE.

Chang-Su Kim received the Ph.D. degree in electrical engineering from Seoul National University (SNU), in 2000. From 2000 to 2001, he was a Visiting Scholar with the Signal and Image Processing Institute, University of Southern California, Los Angeles. From 2001 to 2003, he coordinated the 3D Data Compression Group at the National Research Laboratory for 3D Visual Information Processing, SNU. From 2003 to 2005, he was an Assistant Professor with the Department of Information Engineering, The Chinese University of Hong Kong. In September 2005, he joined the School of Electrical Engineering, Korea University, where he is a Professor.

Zicheng Liu is currently a partner research manager with Microsoft, Redmond, WA, USA. Before joining Microsoft Research, he was with Silicon Graphics Inc. His current research interests include computer vision, vision-language learning, and machine learning. He received a B.S. degree from Huazhong Normal University, Wuhan, China, in 1984, a M.S. degree from the Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing, China, in 1989, and a Ph.D. degree in computer science from Princeton University, Princeton, NJ, USA, in 1996. He is a fellow of the IEEE.

Alessandro Piva received his Ph.D. degree in "Computer Science and Telecommunications Engineering" from the University of Florence on 1999. From 2002 until 2004 he was Research Scientist at the National Inter-university Consortium for Telecommunications (CNIT). Since 2005 he was Assistant Professor at the University of Florence. He is actually Associate Professor at the same

Department, and lecturer for the course "Image Processing and Protection" of the Laurea Degree in Telecommunications Engineering of the University of Florence. His research interests lie in the areas of Information Forensics and Security, and of Image and Video Processing.

Junsong Yuan is Professor and Director of Visual Computing Lab at Department of Computer Science and Engineering (CSE), State University of New York at Buffalo (UB), USA. Before joining SUNY Buffalo, he was Associate Professor (2015–2018) and Nanyang Assistant Professor (2009–2015) at Nanyang Technological University (NTU), Singapore. He obtained his Ph.D. from Northwestern University in 2009, M.Eng. from National University of Singapore in 2005, and B.Eng. from Huazhong University of Science Technology (HUST) in 2002. His research interests include computer vision, pattern recognition, video analytics, human action and gesture analysis, large-scale visual search and mining. He is a Fellow of IEEE and IAPR.

References

[1] A. M. Andrew, "Multiple View Geometry in Computer Vision," *Kybernetes*, 2001.

[2] S. Antol *et al.*, "Vqa: Visual Question Answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[3] S. A. Bagloee *et al.*, "Autonomous Vehicles: Challenges, Opportunities, and Future Implications for Transportation Policies," *Journal of Modern Transportation*, 24(4), 2016, 284–303.

[4] Y. Cai *et al.*, "3D Hand Pose Estimation Using Synthetic Data and Weakly Labeled RGB Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[5] S. Campbell *et al.*, "Sensor Technology in Autonomous Vehicles: A Review," *2018 29th Irish Signals and Systems Conference (ISSC). IEEE*, 2018.

[6] X.-W. Chen and X. Lin, "Big Data Deep Learning: Challenges and Perspectives," *IEEE Access*, 2, 2014, 514–25.

[7] J. Choo and S. Liu, "Visual Analytics for Explainable Deep Learning," *IEEE Computer Graphics and Applications*, 38(4), 2018, 84–92.

[8] M. R. M. Crespo da Silva, "Equations for Nonlinear Analysis of 3D Motions of Beams," 1991, S51–9.

[9] E. De Schutter, "Deep Learning and Computational Neuroscience," *Neuroinformatics*, 16(1), 2018, 1–2.

[10] R. Fjelland, "Why General Artificial Intelligence will not be Realized," *Humanities and Social Sciences Communications*, 7(1), 2020, 1–9.

[11] J. Gall *et al.*, "Global Stochastic Optimization for Robust and Accurate Human Motion Capture," 2007.

[12] J. Gall, B. Rosenhahn, and H.-P. Seidel, "Robust Pose Estimation with 3D Textured Models," in *Pacific-Rim Symposium on Image and Video Technology*, Springer, Berlin, Heidelberg, 2006.

[13] G. Grubb *et al.*, "3D Vision Sensing for Improved Pedestrian Safety," *IEEE Intelligent Vehicles Symposium, 2004. IEEE*, 2004.

[14] ICCV21 CLVL, https://sites.google.com/view/iccv21clvl/home.

[15] H. Jiang and E. Learned-Miller, "Face Detection with the Faster R-CNN," *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017). IEEE*, 2017.

[16] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training Generative Adversarial Networks with Limited Data," NeurIPS-2020.

[17] T. N. Kipf and M. Welling, "Semi-supervised Classification with Graph Convolutional Networks," 2016, arXiv preprint arXiv:1609.02907.

[18] C.-C. J. Kuo and Y. Chen, "On Data-driven Saak Transform," *Journal of Visual Communication and Image Representation*, 50, 2018, 237–46.

[19] C.-C. J. Kuo, M. Zhang, S. Li, J. Duan, and Y. Chen, "Interpretable Convolutional Neural Networks via Feedforward Design," *Journal of Visual Communication and Image Representation*, 2019.

[20] X. Lei *et al.*, "TGHop: An Explainable, Efficient, and Lightweight Method for Texture Generation," *APSIPA Transactions on Signal and Information Processing*, 10, 2021.

[21] S. Liu *et al.*, "Computer Architectures for Autonomous Driving," *Computer*, 50(8), 2017, 18–25.

[22] A. H. Marblestone, G. Wayne, and K. P. Kording, "Toward an Integration of Deep Learning and Neuroscience," *Frontiers in Computational Neuroscience*, 10, 2016, 94.

[23] R. Martin-Brualla *et al.*, "Nerf in the Wild: Neural Radiance Fields for Unconstrained Photo Collections," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[24] B. Mildenhall *et al.*, "Nerf: Representing Scenes as Neural Radiance Fields for View Synthesis," in *European Conference on Computer Vision*, Springer, Cham, 2020.

[25] Y. Min *et al.*, "An Efficient Pointlstm for Point Clouds based Gesture Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[26] A. Pelagotti, A. D. Mastio, A. D. Rosa, and A. Piva, "Multispectral Imaging of Paintings," *IEEE Signal Processing Magazine*, 25(4), 2008, 27–36.

[27] J. F. Peters, "Foundations of Computer Vision: Computational Geometry, Visual Image Structures and Object Shape Detection," 124(Springer), 2017.

[28] B. A. Richards *et al.*, "A Deep Learning Framework for Neuroscience," *Nature Neuroscience*, 22(11), 2019, 1761–70.

[29] B. M. H. Romeny, "Geometry-driven Diffusion in Computer Vision," *Springer Science & Business Media*, 1, 2013.

[30] J. R. Rossignac and J. J. Kim, "Computing and Visualizing Pose-interpolating 3d Motions," *Computer-Aided Design*, 33(4), 2001, 279–91.

[31] W. Samek *et al.*, "Explainable AI: Interpreting, Explaining and Visualizing Deep Learning," *Springer Nature*, 11700, 2019.

[32] W. Samek and K.-R. Müller, "Towards Explainable Artificial Intelligence," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer, Cham, 2019, 5–22.

[33] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," 2017, arXiv preprint arXiv:1708.08296.

[34] F. Scarselli *et al.*, "The Graph Neural Network Model," *IEEE Transactions on Neural Networks*, 20(1), 2008, 61–80.

[35] K. J. Shih, S. Singh, and D. Hoiem, "Where to Look: Focus Regions for Visual Question Answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[36] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable Deep Learning Models in Medical Image Analysis," *Journal of Imaging*, 6(6), 2020, 52.

[37] L. Song *et al.*, "Stacked Homography Transformations for Multi-View Pedestrian Detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

[38] K. R. Storrs and N. Kriegeskorte, "Deep Learning for Cognitive Neuroscience," 2019, arXiv preprint arXiv:1903.01458.

[39] D. Teney *et al.*, "Tips and Tricks for Visual Question Answering: Learnings from the 2017 Challenge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[40] L. Verdoliva, "Media Forensics and DeepFakes: An Overview," *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 2020, 910–32.

[41] N. Vogt, "Machine Learning in Neuroscience," *Nature Methods*, 15(1), 2018, 33.

[42] N. Xie *et al.*, "Explainable Deep Learning: A Field Guide for the Uninitiated," 2020, arXiv preprint arXiv:2004.14545.

[43] Y.-H. Yiu *et al.*, "DeepVOG: Open-source Pupil Segmentation and Gaze Estimation in Neuroscience Using Deep Learning," *Journal of Neuroscience Methods*, 324, 2019, 108307.

[44] T. Yu *et al.*, "3D Object Representation Learning: A Set-to-set Matching Perspective," *IEEE Transactions on Image Processing*, 30, 2021, 2168–79.
[45] J. Zhang *et al.*, "Boosting Positive and Unlabeled Learning for Anomaly Detection with Multi-features," *IEEE Transactions on Multimedia*, 2019.
[46] M. Zhang *et al.*, "Pointhop: An Explainable Machine Learning Method for Point Cloud Classification," *IEEE Transactions on Multimedia*, 22(7), 2020, 1744–55.
[47] S. Zhao, Z. Liu, J. Lin, J.-Y. Zhu, and S. Han, "Differentiable Augmentation for Data-Efficient GAN Training," NeurIPS-2020.
[48] Q. Zhu *et al.*, "3d Lidar Point Cloud based Intersection Recognition for Autonomous Driving," *2012 IEEE Intelligent Vehicles Symposium. IEEE*, 2012.