Evolution of Winning Solutions in the 2021 Low-Power Computer Vision Challenge

Xiao Hu, **Ziteng Jiao**, and **Ayden Kocher**, Purdue University
Zhenyu Wu and **Junjie Liu**, Loyola University Chicago
James C. Davis, Purdue University
George K. Thiruvathukal, Loyola University Chicago
Yung-Hsiang Lu, Purdue University

We organized the 2021 IEEE Low-Power Computer Vision Challenge to advance state-of-the-art solutions in low-power computer vision. Here, we examine the winning teams' development patterns and design decisions, focusing on their techniques to balance power consumption and accuracy to provide guidelines for future competitions.

Competitions drive innovation and promote creativity. The DARPA Grand Challenge opened the era of autonomous driving; the Ansari X Prize opened the era of reusable spacecraft. The same positive influence of competition applies to the field of computer vision. The Face Recognition Technology

Digital Object Identifier 10.1109/MC.2023.3250246
Date of current version: 26 July 2023

program from the National Institute of Standards and Technology[1] set up the standard of facial recognition. The ImageNet Large-Scale Visual Recognition Challenge[2] established deep learning as the mainstream approach for computer vision. These competitions created an incentive of surpassing the existing solutions and provided a platform for researchers to benchmark their solutions. We believe the next challenge in computer vision is to achieve state-of-the-art performance on resource-constrained devices.

To take further advantage of competitions, the IEEE Annual International Low-Power Computer Vision Challenge (LPCVC) has been held to identify energy-efficient computer vision solutions since 2015.[3,4] These solutions may apply to energy-constrained systems equipped with digital cameras, such as mobile phones, aerial robots, and automobiles. From 2015 to 2017, LPCVC competitions were held on site at large conferences (the Design Automation Conference in 2015–2016 and the International Conference on Computer Vision and Pattern Recognition in 2017–2018). On-site competitions allowed contestants to bring their own hardware, including experimental boards, mobile phones, tablets, field-programmable gate arrays, and desktops. To encourage more participation, the competition was hybrid in 2018: contestants could bring their own hardware on site, and a separate track allowed contestants to submit their code online using the same hardware. Since 2019, the competitions have been entirely online.

In 2021 LPCVC, 53 teams from four different countries submitted 366 solutions during the submission window (1 August–1 September) (Figure 1). A public leaderboard ranked all submitted solutions during the month. A total of 138 solutions from 17 teams outperformed our open source reference solution.

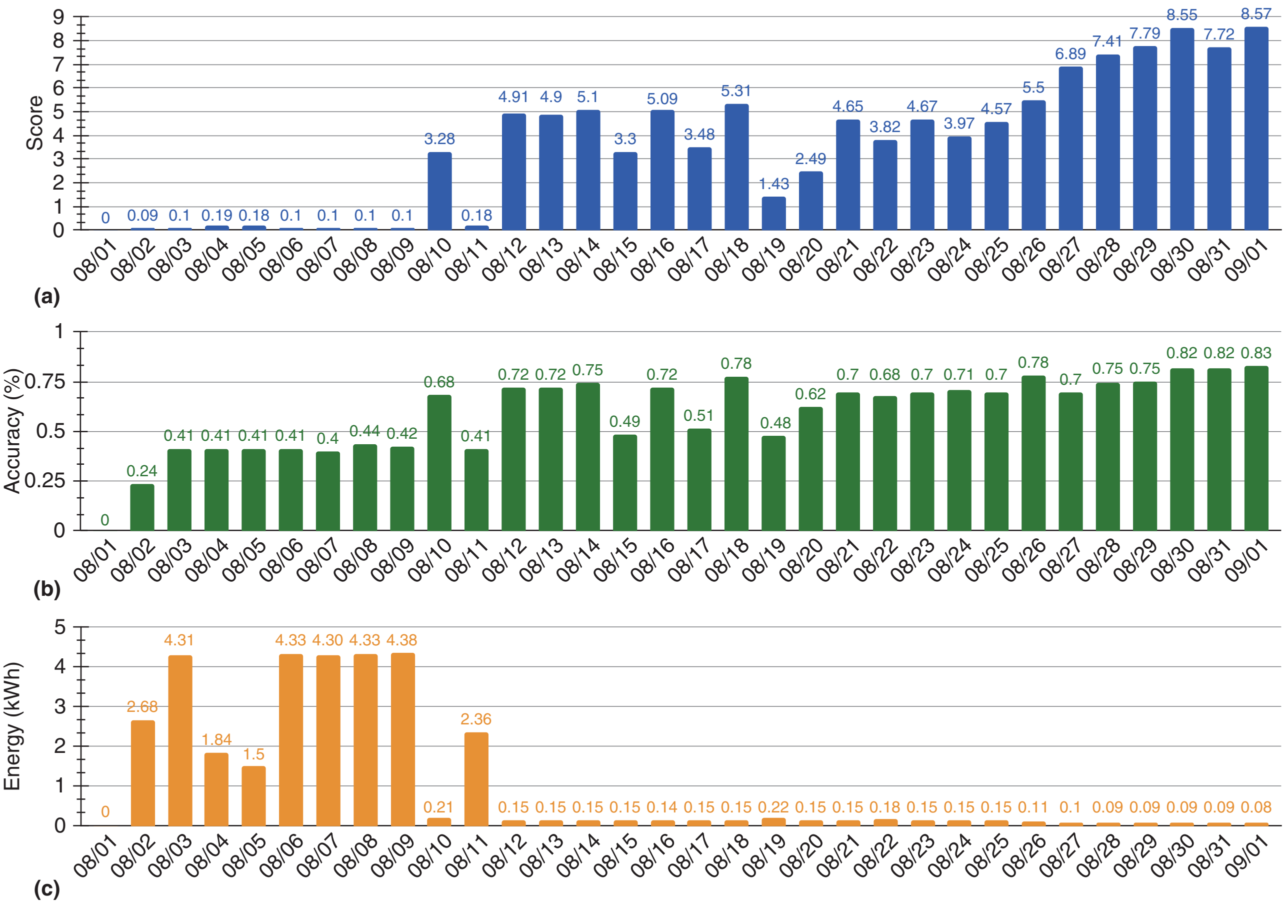


FIGURE 1. (a) The highest score, (b) the highest accuracy, and (c) the lowest energy on each day during August 2021.

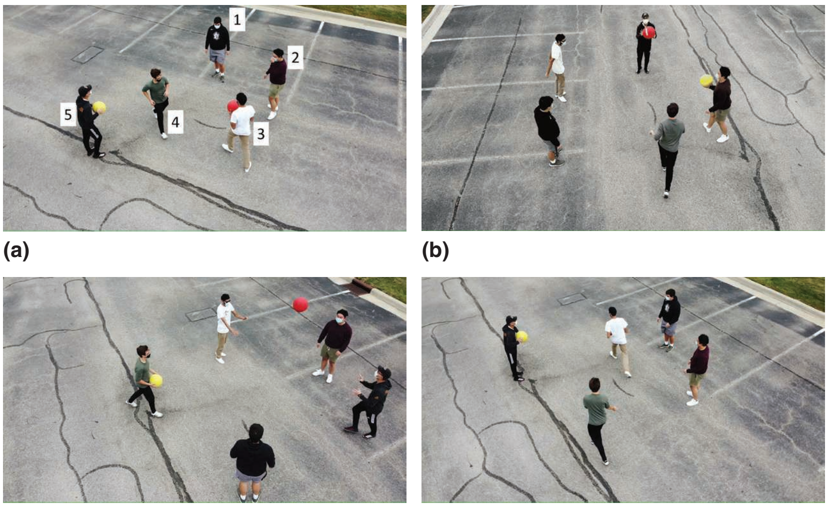
Compared with the reference solution, the best solution improved accuracy by 3.43 times (343%) using only 4% (a 96% reduction) of the energy. This article analyzes all submissions from the top two teams and presents their important design decisions. This article aims to help organizers design future competitions and help contestants explore design space and win competitions.

2021 IEEE LOW-POWER COMPUTER VISION CHALLENGE (VIDEO TRACK)

In the video track, contestants are required to solve an instance of the multiobject tracking (MOT) problem. MOT is a challenging problem in computer vision.[5,6] It aims to determine the identities and trajectories of multiple moving objects in a video. However, MOT is limited by input frames—if the input frames come from a stationary camera, tracking can only happen within the frame, and the occlusions interfere with the tracking accuracy. Although some application scenarios can address this with an array of cameras, others envision following the objects of interest using unmanned aerial vehicles (UAVs, also called *drones*). Therefore, UAVs have received increasing attention in research and industry communities for their flexibility. From video surveillance to crowd behavior analysis, many application scenarios can benefit from analyzing drone-captured video with MOT solutions.

MOT on UAVs exhibits two major challenges: 1) the dynamic background makes tracking more difficult, and 2) the solutions need to be low power since the UAVs have limited energy from their onboard batteries. Although these constraints are not unique to UAVs, and many battery-powered systems need fast and energy-efficient solutions, most computer vision competitions focus exclusively on accuracy. To fill this gap, the 2021 LPCVC introduced a track that measured vision solutions in both accuracy and energy consumption.

The contestants are required to perform multiclass (balls and persons) MOT on videos captured by UAVs. Figure 2 shows four example frames from one video. The solutions should determine when the balls change hands by indicating the frame number and the ball possessor. Sample test data were provided; however, contestants can use any training data.



(a) (b) (c) (d)

FIGURE 2. Four frames in one sample video for MOT. Each person is labeled a number between 1 and 5. Balls have different colors. The balls, the people, and the cameras move simultaneously. Occlusion may occur during these movements: in (d) the red ball is occluded by the person with a white shirt.

Referee system

Figure 3 shows the architecture of the automated referee system and how information flows through it. A contestant uploads a solution to the competition website: https://lpcv.ai. These solutions enter a queue to be evaluated by the referee system. To process a submission, the referee system resets the evaluation board to a clean state and then executes the submission. Power measurement starts when a submitted solution starts running. After a submission completes, the referee system calculates the score and updates the public leaderboard on the website. Online submissions require a common hardware platform for comparing the speed—we used a Raspberry Pi 3B+ because it is a popular platform for embedded systems.

A submitted solution receives two input files: a testing video and a calibration file. The expected output is a comma-separated value file storing the frame when a ball changes hands. Table 1 shows the expected format of the output file. A submission program is disqualified if it cannot be executed or generates the wrong output format.

Reference solution

We provide an open source reference solution on GitHub[7] as a baseline for contestants to create better solutions. From our experience in the previous competitions, the reference solution is used as an example to present the submission formatting but not limit innovative designs. It also serves as the qualification: a submitted solution is disqualified if it is inferior to the reference solution.

To encourage innovation, the reference solution provides a sample adopting the conventional multiclass MOT paradigm using "tracking by detection" (Figure 4). YOLOv5,[8] an advanced version of the You Only Look Once (YOLO) object detector,[9] is the detector of our choice because of its flexibility in training and high inference speed. DeepSORT[10] is used to track the moving object because it contains multiple dimensions of features to track the instance across frames and has been widely used in many MOT projects. The reference solution ranked No. 2 on the fourth day of the challenge, 4 August 2021. When the challenge concluded on 1 September 2021, the same reference solution (two versions) ranked 139 and 147 among 158 valid submissions.

Evaluation metrics

The evaluation metrics are designed to balance multiple factors. First, the organizers did not wish to use the per-frame annotations commonly adopted in conventional MOT datasets. Creating such annotations requires significant effort from the organizers. Also, comparing the submitted solutions with the ground truth, frame by frame, will require significant computation on the referee system and delay posting the scores on the leaderboard. Second, the main purpose of this tracking problem is to detect when the balls change hands and who holds which ball. The event of capturing a ball is more important than the duration of holding a ball. The accuracy is determined by detecting when a ball is caught using two major components of a MOT solution: object detection and reidentification. A catch is defined as the moment a thrown ball touches a person's hand. Reidentification determines which person catches the ball.

When a submitted solution reports a catch, the index frame can belong to one of three categories:

1. *True positive* (*TP*): A catch is caught correctly. Suppose a ball

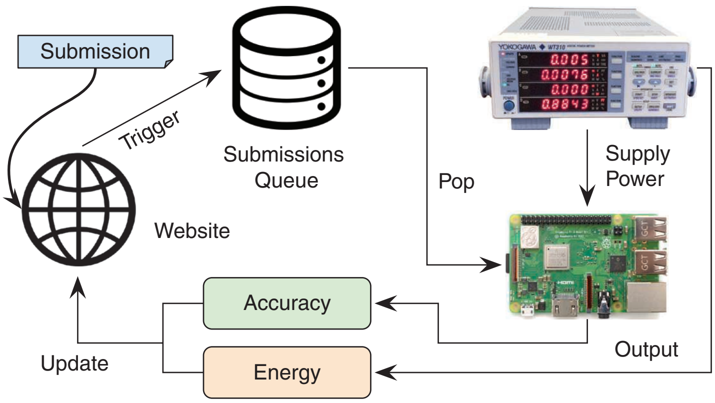


FIGURE 3. The workflow of our automated referee system.

TABLE 1. The top table is an example of the input file provided with the test video. Class 0 is a person and 1 is a ball. Following the YOLO annotation format, *X* and *Y* are the absolute centers of each bounding box with width and height. The bottom table is an example of the expected output format. The last column (Meaning) helps interpret the information and is not included in the file.

Frame	Class	ID	*X*	*Y*	Width	Height
0	0	1	50.41015	0.39583	0.02031	0.03425
0	0	2	0.36835	0.61990	0.04557	0.18055
0	1	3	0.41015	0.39583	0.03593	0.16296
...	...	...	...	...	...	...

Frame	Yellow	Orange	Red	Purple	Blue	Green	Meaning
0	0	1	5	2	3	0	Initial setting
5	0	1	5	2	4	0	Person 4 catches blue ball
30	0	3	5	2	4	0	Person 3 catches orange ball
60	0	3	1	2	4	0	Person 1 catches red ball
...	...	...	...	...	...	...	...

is caught at frame *t* in the ground truth; the reference system accepts the answer within ±10 frames from the ground truth frame. If multiple output frames are within the range, the earliest frame is selected, so more accurate output is encouraged.
2. *False positive* (*FP*): A catch is falsely detected. This reduces the scores of the solutions that output too many irrelevant frames.
3. *False negative* (*FN*): The solution fails to detect a catch.

The F_1-score is commonly used as an evaluation metric in machine learning as it elegantly sums up the predictive performance of a model by combining two otherwise competing metrics: precision and recall.[11] The conventional F_1-score is represented here in

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}. \quad (1)$$

For this competition, *TP* is not uniform in all cases. If *TP* only counts the frame that has a correct detection, other attributes within the detection (how many pairs of balls/person within the frames are correctly detected) will be neglected. Thus, we have $score_{TP}$ for each *TP* frame, which is calculated by dividing the number of correct ball/person values $correct_i$ over the total catches in the ground truth $total_i$

$$score_{TP} = \sum_{i=0}^{n} \frac{correct_i}{total_i} \quad (2)$$

where *i* is the index frame and *n* is the total number of balls in the input video.

The original numerator *TP* in (1) is replaced by $score_{TP}$. Since *TP* represents the frame of correct detections, and $score_{TP}$ gives the accuracy within the correct detection, this gives a better evaluation of the performance for the entire solution. Finally, the *accuracy* is calculated based on

$$accuracy = \frac{score_{TP}}{TP + \frac{1}{2}(FP + FN)}. \quad (3)$$

In the example shown in Table 2, frames 31 and 95 in the output are within ±10 frames from the ground truth frames 30 and 90; therefore, they are classified as *TP* with corresponding $score_{TP}$; frame 60 and frame 115 are missing in the output, so *FN* is 2; frame 48 is not within any range of the

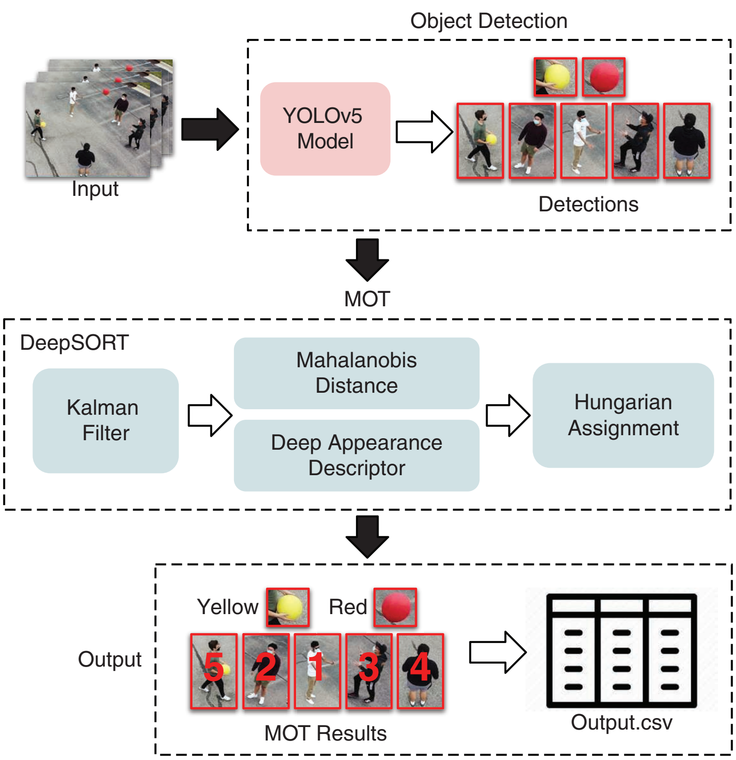


FIGURE 4. The workflow of the reference solution. The MOT block follows the object association architecture listed in DeepSORT.

TABLE 2. Example output and ground truth for one input video.

Frame	Red	Blue	Green	Result
Ground truth				
30	1	2	3	
60	1	3	4	
90	2	1	3	
115	4	2	1	
Example output				
31	1	4	3	TP, 2/3
48	5	3	4	FP
95	2	1	3	TP, 1

frames in the ground truth; therefore, it is classified as *FP*. The final *accuracy* is $(1+2/3)/(2+(1/2)(1+2)) = 0.48$

$$score = \frac{accuracy}{energy}. \tag{4}$$

EVOLUTION OF WINNERS' SOLUTIONS

To better understand the design decisions of the participants, this article analyzes the solutions submitted by the top two winning teams (see Table 3). The champion is the VITA team from the University of Texas and Wormpex AI. The second award belongs to the baseSlim team from Meituan. The accuracy and energy differences between each submission from both teams are shown in Figures 5 and 6. Important submissions are divided into sections on the figures.

The baseSlim team

Section A. The baseSlim team's first submission used a combination of NanoDet[12] and JDETracker,[13] but the program produced no output. In the second submission (Section A of Figure 5), the team replaced JDETracker with the DeepSORT used in the reference solution. The resultant score was eight times better than the reference solution, given the low-power profile of NanoDet.

Section B. The fifth submission made significant progress by updating the structure to NanoDet as the detector and DeepSORT as the tracker. The solution also has an improved feature extractor for the reidentification module in DeepSORT by retraining the tracking pretrained weights. The fifth submission obtained a score of 2.26. The team further improved the accuracy by pruning the DeepSORT weights in the sixth submission. This improvement in accuracy also increased the energy consumption. The sixth submission replaced NanoDet with YOLOx and tuned the pretrained weights of the Visual Object Classes dataset. The eighth submission reduced energy consumption with nearly no change in

TABLE 3. Final scores of the top two teams and the reference solution. Energy is in kilowatt hours, and accuracy is in percentages. VITA has lower energy consumption; baseSlim has higher accuracy.

Team	Energy	Accuracy	Score	Count
VITA	0.09	0.79	8.57	22
baseSlim	0.1	0.83	8.56	14
Reference	2.26	0.23	0.11	2

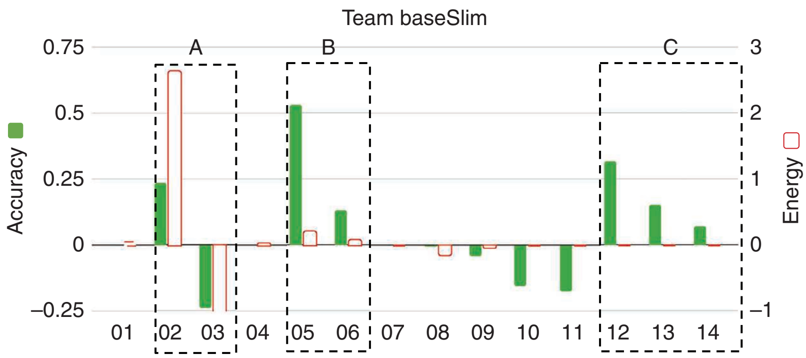


FIGURE 5. Changes in accuracy (in percentages) and energy consumption (in kilowatt hours) over the solutions from the baseSlim team. The first pair, labeled 01, shows the scores from the first submission. Higher accuracy (positive) and lower energy (negative) are preferred.

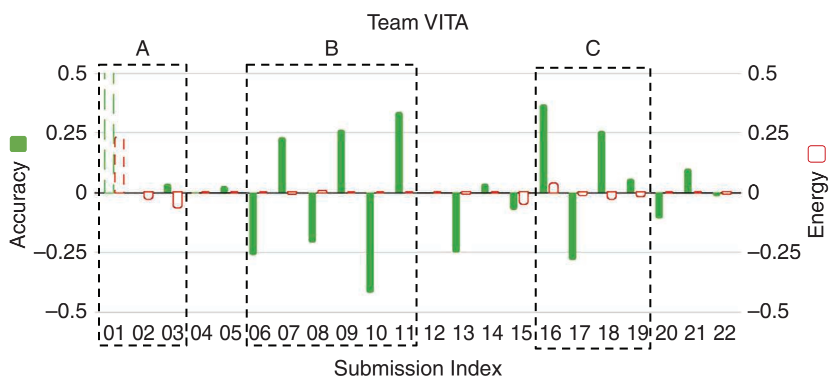


FIGURE 6. The difference of accuracy (in percentages) and energy consumption (in kilowatt hours) between the submissions from the VITA team. The first submission has the actual data instead of the difference.

accuracy. The ninth and 10th submissions attempted to accelerate execution, but the accuracy decreased. The slight reduction in energy consumption was accompanied by a significant reduction in accuracy (10th and 11th submissions).

Section C. The last three submissions achieved much better accuracy with negligible impacts on energy consumption. Up to the 11th submission, the team used pretrained weights stored in `.pth` format; this is the default format for models trained with PyTorch. In their 11th submission, the team converted the `.pth` weights into the `.jit` format. This reduced the model size to only 21.82% of the previous submissions. The just-in-time (JIT) compiler takes a PyTorch model and rewrites it to run at a higher efficiency. The team came back to the YOLOx model from NanoDet on submissions 12–14 and made great improvements in accuracy. The 13th submission replaced YOLOx with SPGNet and stored all color codes in a NumPy array. These changes increased the accuracy by 0.1533. The final (14th) submission used better pretrained weights. This submission achieved an accuracy of 0.83 at an energy usage of 0.097, for a score of 8.56. This is 77.9 times better than the reference solution. More details on model compression techniques used in the solution are reported elsewhere.[14,15]

The VITA team

Section A. VITA team's first submission used YOLOv5s as the detector, which required only 8.3% operations compared to the YOLOv5 model used in the reference solution. Through quantization, the YOLOv5s model was only 1.29 MB (the released YOLOv5 model was 13.9 MB). These changes led to a 2.78-times better accuracy than the reference solution. The team's first solution also improved the DeepSORT tracker by replacing the original backbone Wide Residual Network[16] with ResNet18.[17] With the new backbone, the VITA team trained a tracking model of size 2.81 MB through pruning; this was only 6.47% the size of the reference model. For inference, the team designed an action detector that dynamically classified and selected useful actions in the input video to minimize the frames that needed to be processed.[18] With the help of the action detector, the second submission reduced the energy by 0.23 kWh. The third submission compressed the tracking model even more, from 2.81 MB to 0.31 MB through pruning. As a result, the third submission decreased the energy consumption by 0.06 kWh, with a slight increase in accuracy.

Section B. The following submissions had wide fluctuations in accuracy, while the energy consumption remained nearly unchanged. The sixth submission attempted to improve the action detector by estimating the proximity of the balls and the people. However, this did not perform well, and the accuracy dropped by 26%. The seventh submission was similar to the fifth submission. The eighth submission attempted to improve the action detector, but the accuracy dropped by 201% again. In the ninth submission, the team used the DeepSORT tracking, which improved the accuracy to 77.67%. The 10th submission added calibration to the action detector and bounding boxes to make the tracking more precise, but the accuracy dropped by 41.7%. The 11th submission removed the calibration and used a smaller pretrained YOLOv5 model (from 1.29 MB to 0.93MB). The accuracy improved by 33.67%.

Section C. The VITA team had the highest increase in accuracy in their 16th submission at 36.7%.

In this submission, the team learned the lessons from all of the components that did not help improve its submissions and finalized its action detector by adding more cases to handle the different situations in the input video. What came with higher accuracy was more energy usage. A longer execution time was needed to complete the 15th submission, leading to an increase of 0.04 kWh. Because of this increase, the score of the 16th submission was lower than some of the team's previous submissions. The team implemented a correction strategy in its action detector. The maximum numbers of balls and persons were marked at the beginning of the video based on the given annotation files. When the query reaches the maximum number, but the detector detects a new ball or person in the video, the detector will first try to reidentify again to see if the new object could be linked with any existing profiles. This strategy helped the team greatly reduce the time of correcting itself, and an accuracy increase of 25.67% and an energy usage decrease of 0.026 kWh appeared in the 18th submission. Finally, the team reached the highest accuracy at 81.3% in the 19th submission.

Later submissions explored the tradeoffs between accuracy and energy usage. With all of the previous lessons, the VITA team reached the highest score among all submissions in LPCVC 2021 at 8.57 with the accuracy at 79% and energy usage at 0.09 kWh. More details of the development process, including model compression techniques and training, can be found in the VITA team's article.[18]

OBSERVATIONS AND SUGGESTIONS FOR FUTURE CHALLENGES

As shown in Figures 5 and 6, the winning teams' solutions did not achieve monotonic improvements. Instead, both teams experimented with different methods to improve accuracy and reduce energy consumption. Both teams found success by tuning individual modules while sticking to the same general modular design they started with. The teams' approaches suggest that winning solutions should be designed and implemented in modules so that replacing components can be easy.

We sent a survey to all participants from all of the different tracks of the 2021 LPCVC competition to collect their feedback. Based on this feedback, here are several suggestions for organizers of future challenges.

- An up-to-date leaderboard encourages innovations. Figure 1 shows that the best daily scores improved substantially over the month. It is possible to update the leaderboard quickly because the referee system was automated (shown in Figure 3). The UAV video track did not have any execution-related failures from the automated referee system.
- An open source scoring system helps participants understand how to optimize. Our referee system was open source, and contestants could fully understand how scores are calculated. An interesting insight from the survey is that the UAV video track received a 3.8/5 satisfaction score on the scoreboard. Since the UAV video track was the only one equipped with the automated referee system, it suggests that our approach benefited the participants by providing constant and reliable scoreboard updates.
- A reference solution is valuable. A reference solution serves multiple purposes. 1) It helps contestants understand the input and output formats. 2) It sets a minimum standard for qualification. 3) If it is well structured, it encourages contestants to experiment by replacing the components. Our survey results show a score of 4.4/5 on satisfaction with the reference solution. One potential disadvantage is that it may discourage participants' creativity in using drastically different approaches. We acknowledge that even the winning teams innovated only *within* the modular design of the reference solution—they improved components but did not explore new designs. In the future, we will explore whether zero or multiple reference solutions promote greater design diversity.

In this article, we presented the preparation process for organizing the 2021 LPCVC UAV video track and the evolution of the top two winning teams' solutions. We summarized that the key to a successful competition consists of a well-designed reference solution, an automated referee system, and a timely scoreboard. In the analysis of the evolution of the winning solutions, both teams experimented with many design choices throughout their submissions to achieve the delicate balance between accuracy and energy consumption. The success of the 2021 LPCVC, along with the previous competitions, helped to shift the computer vision competition focus from accuracy only to both accuracy and power efficiency. The application scenario of computer vision on UAVs paved the way for the following competitions: the 2023 IEEE Autonomous UAV Chase Challenge and the 2023 LPCVC UAV Segmentation track. More evaluation criteria, such as fairness[19] and robustness,[20] may be considered in future challenges. We hope that this report will be beneficial for both future competition organizers and participants to continue advancing innovation in computer vision. ■

ACKNOWLEDGMENT

The 2021 LPCVC was supported by Facebook, Xilinx, ELAN Microelectronics, the IEEE Computer Society, and the National Science Foundation (CNS-2120430, OAC-2107230, OAC-2104709, and OAC-2107020). Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and may not reflect the views of the sponsors.

REFERENCES

1. P. J. Rauss, J. Phillips, M. K. Hamilton, and A. Trent DePersia, "FERET (Face Recognition Technology) program," in *Proc. 25th AIPR Workshop, Emerg. Appl. Comput. Vis.*, D. H. Schaefer and E. F. Williams, Eds. Bellingham, WA, USA: SPIE, 1997, vol. 2962, pp. 253–263.
2. O. Russakovsky et al., "Imagenet large scale visual recognition challenge," 2014. [Online]. Available: https://arxiv.org/abs/1409.0575
3. S. Alyamkin et al., "Low-power computer vision: Status, challenges, and opportunities," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 9, no. 2, pp. 411–421, Jun. 2019, doi: 10.1109/JETCAS.2019.2911899.
4. G. K. Thiruvathukal, Y.-H. Lu, J. Kim, Y. Chen, and B. Chen, Eds., *Low-Power Computer Vision: Improve the Efficiency of Artificial Intelligence*. London, U.K.: Chapman & Hall, Feb. 2022.
5. G. Ciaparrone, F. L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera, "Deep learning in video

ABOUT THE AUTHORS

XIAO HU is a software engineer at Qualcomm, San Diego, CA 92124 USA. His research interests include low-power computer vision, fairness in artificial intelligence, and machine learning. Hu received a master's degree in electrical and computer engineering from Purdue University. Contact him at hu440@purdue.edu and https://xiaohu.info.

ZITENG JIAO is a student at Yale University, New Haven, CT 06520 USA. His research interests include energy-efficient computing and operating systems design. Jiao received a bachelor's degree in computer science from Purdue University. Contact him at jiao19@purdue.edu.

AYDEN KOCHER is a computer vision engineer at Vimaan, San Jose, CA 95131 USA. His research interests include computer vision on edge devices, optical character recognition, and optimizing inference throughput. Kocher received a bachelor's degree in electrical and computer engineering from Purdue University. Contact him at akocher@purdue.edu.

ZHENYU WU is a researcher at Wormpex AI, Bellevue, WA 98004 USA. His research interests include video understanding, efficient vision on embedded systems, and adversarial machine learning. Wu received a Ph.D. in computer science from Texas A&M University. Contact him at wuzhenyu_sjtu@tamu.edu and https://wuzhenyusjtu.github.io.

JUNJIE LIU is a senior research engineer in Meituan, Beijing 100102, China. His research interests include computer vision, model compression, and embedded systems. Liu received a master's degree in computer science from the City University of Hong Kong. Contact him at liujunjie10@meituan.com.

JAMES C. DAVIS is an assistant professor in the Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47906 USA. His interests include software engineering and cybersecurity. Davis received a Ph.D. in computer science from Virginia Tech. He is a member of the Association for Computing Machinery and a Senior Member of IEEE. Contact him at davisjam@purdue.edu.

GEORGE K. THIRUVATHUKAL is a professor and the chairperson in the Department of Computer Science, Loyola University Chicago, Chicago, IL 60660 USA, and a visiting computer scientist at Argonne National Laboratory in the Leadership Computing Facility, Lemont, IL 60439, USA. His research topics include parallel and distributed systems, software engineering, and embedded systems. Thiruvathukal received a Ph.D. in computer science from the Illinois Institute of Technology. He is a Senior Member of IEEE. Contact him at gkt@cs.luc.edu.

YUNG-HSIANG LU is a professor at the Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47906 USA. His research interests include computer systems, computer vision, and embedded systems. Lu received a Ph.D. in electrical engineering from Stanford University. He is a Fellow of IEEE and a Distinguished Scientist of the Association for Computing Machinery. Contact him at yunglu@purdue.edu.

multi-object tracking: A survey," 2019. [Online]. Available: https://arxiv.org/abs/1907.12740
6. W. Luo, X. Zhao, and T.-K. Kim, "Multiple object tracking: A review," 2014. [Online]. Available: https://arxiv.org/abs/1409.7618
7. X. Hu, A. Kocher, and Z. Jiao. "21LPCVC-UAV_VIdeo_Track-Sample-Solution." GitHub. Accessed: Aug. 2, 2021. [Online]. Available: https://github.com/lpcvai/21LPCVC-UAV_VIdeo_Track-Sample-Solution
8. G. Jocher et al., "Ultralytics/yolov5: V 3.1 - SOTA realtime instance segmentation," Oct. 2020. [Online]. Available: https://github.com/ultralytics/yolov5
9. J. Redmon, S. Kumar Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2015. [Online]. Available: https://arxiv.org/abs/1506.02640
10. N. Wojke, A. Bewley, and D. Paulus, "Simple online and realtime tracking with a deep association metric," 2017. [Online]. Available: https://arxiv.org/abs/1703.07402
11. D. Powers, "Evaluation: From precision, recall and F-factor to ROC, informedness, markedness & correlation," *Mach. Learn. Technol.*, vol. 2, pp. 37–63, Jan. 2008.
12. R. Lyu. "Nanodet-plus: Super fast and high accuracy lightweight anchor-free

object detection model." GitHub. Accessed: Jul. 31, 2021. [Online]. Available: https://github.com/RangiLyu/nanodet
13. Z. Wang, L. Zheng, Y. Liu, and S. Wang, "Towards real-time multi-object tracking," 2019. [Online]. Available: https://arxiv.org/abs/1909.12605
14. H. Wang, J. Liu, X. Ma, Y. Yong, Z. Chai, and J. Wu, "Compressing models with few samples: Mimicking then replacing," in Proc. *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 701–710.
15. J. Liu et al., "Knowledge representing: Efficient, sparse representation of prior knowledge for knowledge distillation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2019, pp. 638–646, doi: 10.1109/CVPRW.2019.00090.
16. S. Zagoruyko and N. Komodakis, "Wide residual networks," 2016. [Online]. Available: https://arxiv.org/abs/1605.07146
17. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385
18. X. Hu et al., "E^2TAD: An energy-efficient tracking-based action detector," 2022. [Online]. Available: https://arxiv.org/abs/2204.04416
19. J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. 1st Conf. Fairness, Accountability Transparency*, PMLR, S. A. Friedler and C. Wilson, Eds. Feb. 2018, vol. 81, pp. 77–91.
20. G. Chen et al., "VisDrone-MOT2021: The vision meets drone multiple object tracking challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, 2021, pp. 2839–2846, doi: 10.1109/ICCVW54120.2021.00318.

Digital Object Identifier 10.1109/MC.2023.3292664