IEEE Computer

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

**Xiao Hu, Ziteng Jiao, Ayden Kocher, Zhenyu Wu[+], Junjie Liu[‡], James C. Davis, George K. Thiruvathukal[*], Yung-Hsiang Lu**
{hu440, jiao19, akocher, davisjam, yunglu}@purdue.edu - Purdue University.
[+]wuzhenyu_sjtu@tamu.edu. [‡]liujunjie10@meituan.com. [*]gkt@cs.luc.edu- Loyola University Chicago

***Abstract*—**
**Mobile and embedded devices are becoming ubiquitous. Applications such as rescue with autonomous robots and event analysis on traffic cameras rely on devices with limited power supply and computational sources. Thus, the demand for efficient computer vision algorithms increases. Since 2015, we have organized the IEEE Low-Power Computer Vision Challenge to advance the state of the art in low-power computer vision. We describe the competition organizing details including the challenge design, the reference solution, the dataset, the referee system, and the evolution of the solutions from two winning teams. We examine the winning teams' development patterns and design decisions, focusing on their techniques to balance power consumption and accuracy. We conclude that a successful competition needs a well-designed reference solution and automated referee system, and a solution with modularized components is more likely to win. We hope this paper provides guidelines for future organizers and contestants of computer vision 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 spacecrafts. The same positive influence of competitions applies to the field of computer vision. FERET from NIST [13] set up the standard of face recognition. ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [15] 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.

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 [1], [16]. 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

 © XXXX IEEE 

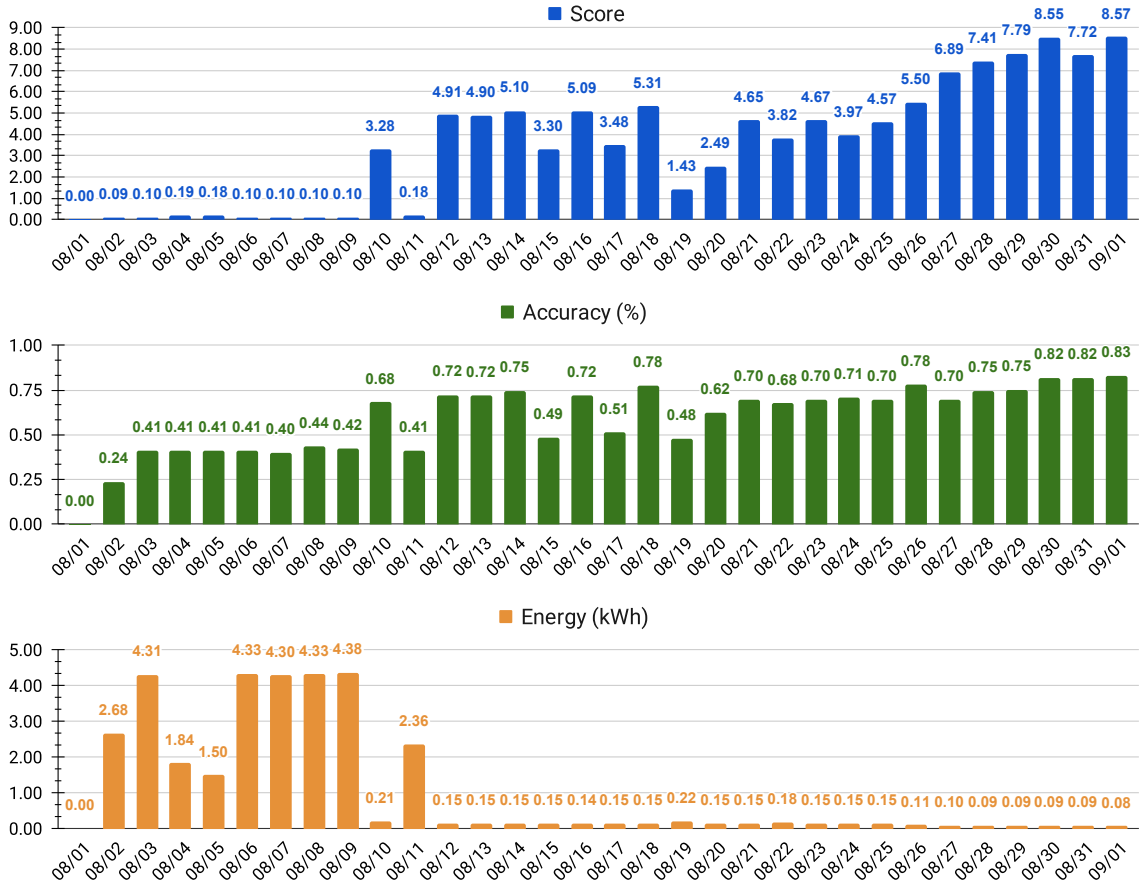


Figure 1: The highest score, highest accuracy, and lowest energy on each day during August 2021.

competitions were held on-site at large conferences (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, FPGAs, 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 the 2021 LPCVC, 53 teams from 4 different countries submitted 366 solutions during the submission window (08/01 - 09/01) (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. Compared with the reference solution, the best solution improved accuracy by 3.43 times (343%) using only 4.0% (96% reduction) of the energy. This paper analyzes all submissions from the top two teams and presents their important design decisions. This paper aims to help organizers design future competitions and to help contestants explore design space and win competitions.

## 2021 IEEE Low-Power Computer Vision Challenge (Video Track)

Multi-Object Tracking (MOT) is a challenging problem in computer vision [4], [10]. MOT aims to determine the identities and trajectories of multiple moving objects in a video. 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 the tracking accuracy. Although some application scenarios can address this with an array of cameras, others envision following the

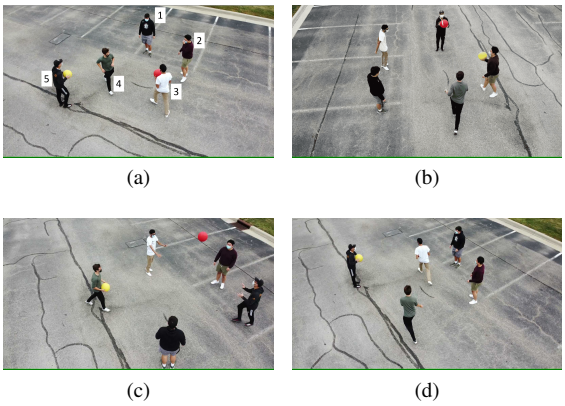

Figure 2: Four frames in one sample video for multiple object tracking. Each person is labeled a number between 1 to 5. Balls have different colors. The balls, the people, and the cameras may move simultaneously. Occlusion may occur: in (d) the red ball is occluded by the person with white shirt.

objects of interest using of Unmanned Aerial Vehicles (UAVs, also called drones). 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 has 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 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 were required to perform Multi-Class (balls and humans) Multi-Object Tracking on a series of 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 was provided; contestants could use any training data.

## Referee System

Figure 3 shows the architecture and how information flows through the automated referee system. 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 CSV (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 provided an open-source reference solution on GitHub [6] as a baseline for contestants to create better solutions. The purpose is to help participants understand the submission format while encouraging creativity. From our experience in the previous competitions, the reference solution is used as an example to present the submission formatting but not limiting innovative designs. It also serves as the qualification: a submitted solution is disqualified if it is inferior to the

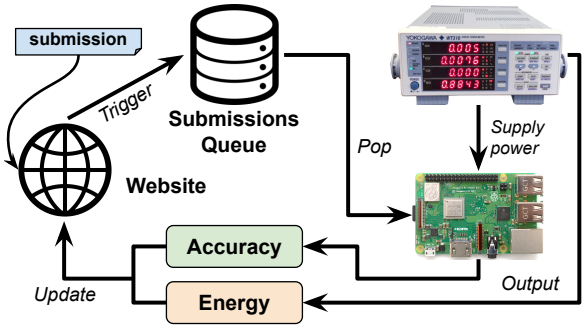


Figure 3: The automated referee system.

| 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 |
| ... | ... | ... | ... | ... | ... | ... | ... |

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 center of each bounding box with width and height. The bottom table is an example of the expected output format. The last column (Meaning) helps human interpret the information and is not included in the file.

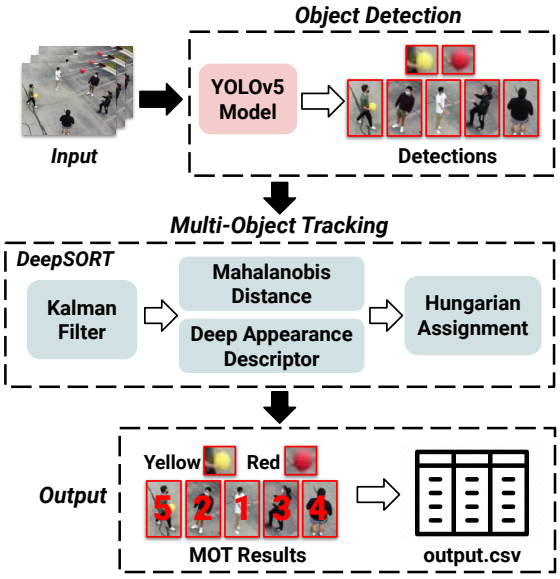


Figure 4: The workflow of the reference solution. The Multi-Object Tracking block follows the object association architecture listed in DeepSORT.

reference solution.

To encourage innovation, the reference solution provides a sample adopting the conventional multi-class multi-object tracking paradigm using "tracking-by-detection" (Figure 4). YOLOv5 [8], an advanced version of the YOLO object detector [14], is the detector of our choice because of its flexibility in training and high inference speed. DeepSORT [19] 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, 2021/08/04. When the challenge concluded on 2021/09/01, 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 per-frame annotations, commonly adopted in conventional multi-object tracking datasets. Creating such annotations require significant efforts 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 an MOT solution: object detection and re-identification. A catch is defined as the moment a thrown ball touches a person's hand. Re-identification 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 is caught at frame $t$ in the ground truth, the reference system accepts the answer within $\pm 10$ 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.

$F_1$-score is commonly used as evaluation metrics 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 in Equation (1).

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

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 groundtruth $total_i$ (Equation (2)). $i$ is the index frame and $n$ is the total number of balls in the input video.

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

The original numerator $TP$ in Equation (1) is replaced by $score_{TP}$. Since $TP$ represents the frames 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 with Equation (3).

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

In the example showed at Table 2: frame 31 and 95 in the output are within $\pm$ 10 frames from the ground truth frame 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 frames in the ground truth, therefore it is classified as $FP$. The final $accuracy$ is: $\frac{1+2/3}{2+\frac{1}{2}(1+2)} = 0.48$.

| Frame | Red | Blue | Green | Result |
|---|---|---|---|---|
| *Groundtruth* | | | | |
| 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 |

Table 2: Example output and ground truth for one input video.

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

## Evolution of Winners' Solutions

To better understand the design decisions of the participants, this paper 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 WormpexAI. The second award belongs to the baseSlim team from Meituan. The accuracy and energy difference between each submission from both teams are showed in Figure 5 and Figure 6. Important submissions are divided into sections on the figures.

Table 3: Final scores of the top 2 teams and the reference solution. Energy is in kWh and accuracy is in %. The VITA team has lower energy consumption; baseSlim, higher accuracy.

| Team | Energy | Accuracy | Score | Count |
|---|---|---|---|---|
| VITA | 0.09 | 0.79 | 8.57 | 22 |
| baseSlim | 0.10 | 0.83 | 8.56 | 14 |
| Reference | 2.26 | 0.23 | 0.11 | 2 |

### The baseSlim Team

**Section A** The baseSlim team's first submission used a combination of NanoDet [12] and JDETracker [18], but the program produced no output. In the second submission (Section A), the team replaced JDETracker with the DeepSORT used in the reference solution. The resultant score was 8 times better than the reference solution given the low-power profile of NanoDet.

**Section B** The 5th 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 re-identification module in the DeepSORT by retraining the tracking pre-trained weights. The 5th submission obtained a score of 2.26. The team further improved the accuracy by pruning the DeepSORT weights in the 6th submission. This improvement in accuracy also increased energy consumption. The 6th submission replaced NanoDet by YOLOx and tuned the pre-trained weights of the VOC dataset. The eighth submission reduced energy consumption with nearly no change in accuracy. The 9th 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 pre-trained 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 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 accuracy by 0.1533. The final (14th) submission used better pre-trained weights. This submission achieved an accuracy of 0.83 at energy usage of 0.097 — 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 [17], [9].

### 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 quantiza-

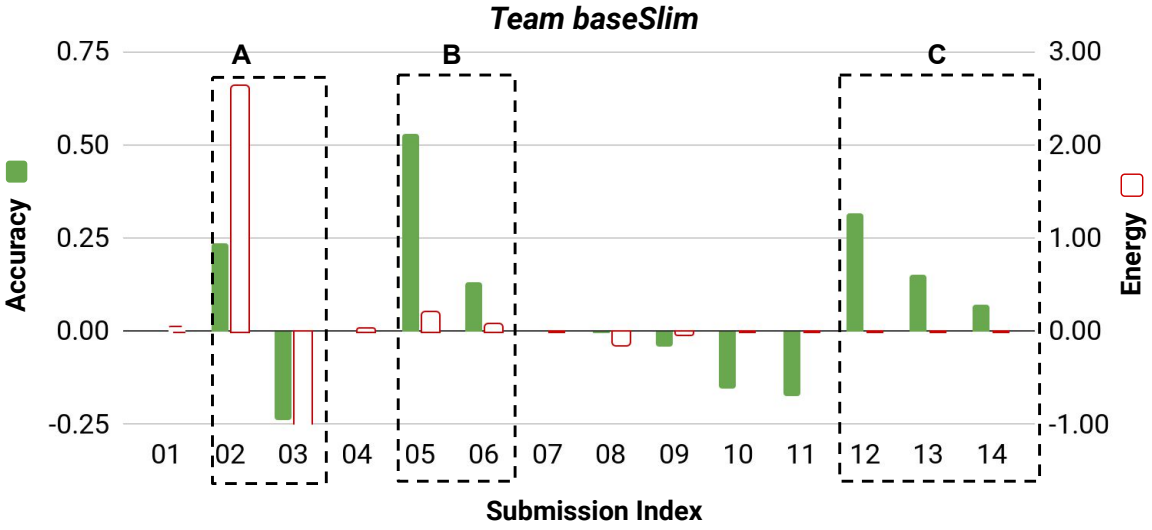


Figure 5: Changes in accuracy (%) and energy consumption (kWh) over the solutions from the baseSlim team. The first pair, labeled 01, shows the scores from the first submission. Higher accuracy (positive) and lower energy (negative) are preferred.

tion, the YOLOv5s model was only 1.29 MB (the released YOLOv5 model was 13.9 MB). These changes led to 2.78 times better accuracy than the reference solution. Their first solution also improved the DeepSORT tracker by replacing the original backbone Wide Residual Network (WRN) [20] with ResNet18 [5]. With the new backbone, the VITA team trained a tracking model of size 2.81 MB through pruning, 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 [7]. With the help of the action detector, the 2nd submission reduced energy by 0.23 kWh. The 3rd submission compressed the tracking model even more, from 2.81 MB to 0.31 MB through pruning. As a result, the 3rd 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 6th 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 26.00%. The 7th submission was similar to the 5th submission. The 8th submission attempted to improve the action detector but the accuracy dropped by 201.00% again. In the 9th submission, the team used the DeepSORT tracking which improved 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 41.70%. The 11th submission removed the calibration and used a smaller pre-trained 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 36.70%. In this submission, the team learned the lessons from all the components that did not help improve their submissions and finalized their 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. Due to this increase, the score of the 16th submission was lower than some of their previous submissions. The team implemented a correction strategy in their action detector. The max number of balls and persons were marked at the beginning of the video based on the given annotation files. When the query reached the max number but the detector detects a new ball or person is in the video, the detector will first try to re-identify again to see if the new object could be linked with any existing profiles. This strategy helped the team to reduce much time of correcting themselves, and an accuracy increase

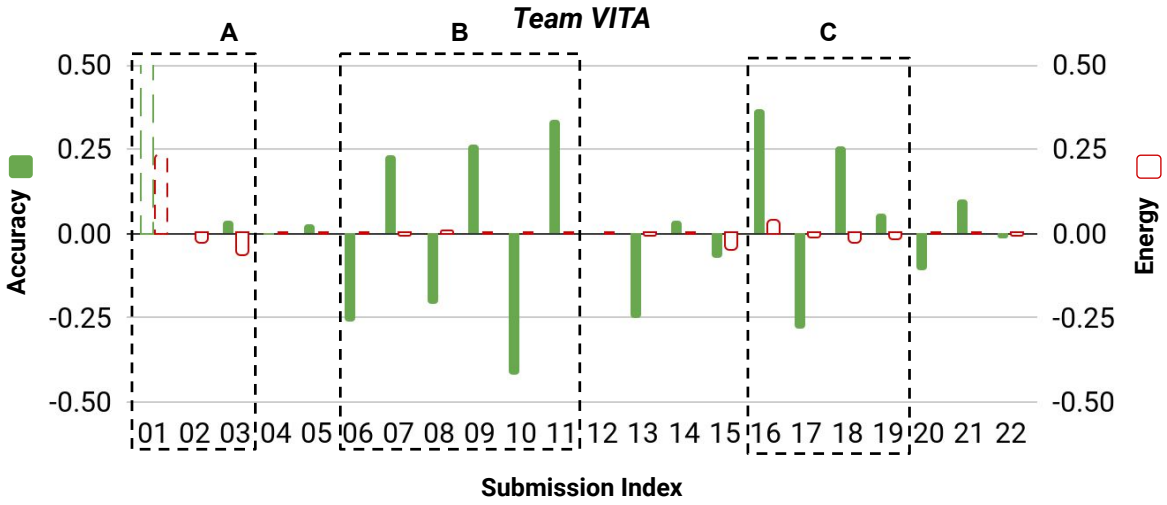


Figure 6: The difference of accuracy (%) and energy consumption (kWh) between the submissions from the VITA team. The first submission has the actual data instead of the difference.

of 25.67% and an energy usage decrease of 0.026 kWh in the 18th submission. Finally, the team reached the highest accuracy at 81.30% in the 19th submission.

Later submissions explored the trade-offs between accuracy and energy usage. With all the previous lessons, the VITA team reached the highest score among all submissions in LPCVC 2021 at 8.57 with accuracy at 79.00% 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 paper [7].

## 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 to reduce energy consumption. Both teams found success by tuning individual modules while sticking within 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 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 can 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 they did not explore new designs. In the future, we will explore whether providing zero, one, or multiple reference solutions promotes greater design diversity.

## Conclusion

In this paper, we present the preparation process of organizing the 2021 Low-Power Computer Vision Challenge UAV video track and the evolution of top two winning teams' solutions. We summarize 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 many design choices throughout their submissions to achieve balance between accuracy and energy consumption. The success of 2021 LPCVC, along with the previous competitions, help to shift the computer vision competition focus from accuracy only to both accuracy and power efficiency. The application scenario of computer vision on UAV paved the way for two follow-up competitions: the 2023 IEEE Autonomous UAV Chase Challenge and the 2023 LPCVC UAV Segmentation track. More evaluation criterias such as fairness [2] and robustness [3] may be considered in future challenges. We hope this paper benefits future competition organizers as well as participants, promoting our goal of advancing innovation in Computer Vision.

## Acknowledgments

The 2021 Low-Power Computer Vision Challenge was supported by Facebook, Xilinx, Elan Microelectronics, the IEEE Computer Society, the National Science Foundation (CNS-2120430, OAC-2107230, OAC-2104709). 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.

## Authors

**Xiao Hu,** obtained his MSc from the Elmore Family School of Electrical and Computer Engineering at Purdue University. His research broadly connects to the fields of human-computer interaction, low-power computer vision, fairness in artificial intelligence, and machine learning. He is a software engineer at Qualcomm.

**Ziteng Jiao** obtained his Bachelor's degree from the College of Science at Purdue University. His research interests include Energy-Efficient Computing and Operating Systems Design.

**Ayden Kocher** obtained his Bachelor's degree from the Elmore Family School of Electrical and Computer Engineering at Purdue University. His research interests include computer vision on edge devices, optical character recognition, and inference throughput.

**Zhenyu Wu,** is a Researcher in Wormpex AI Research. He obtained his Ph.D. from VITA Lab, advised by Prof. Zhangyang (Atlas) Wang. His research interests lie in visual privacy/fairness, object detection, and model compression. Wu is the leader of the VITA team.

**Junjie Liu,** is a senior research engineer in Meituan working on efficient computer vision models for embedded device and UAV. His research topics include computer vision, model compression, and embedded systems. Liu is the leader of the baseSlim team.

**James C. Davis,** is an assistant professor in the Elmore Family School of Electrical and Computer Engineering at Purdue University. His research topics include software engineering and cybersecurity. He is a Member of the ACM and a Senior Member of the IEEE.

**George K. Thiruvathukal,** is Professor and Chairperson in the Department of Computer Science at Loyola University Chicago and Visiting Computer Scientist at Argonne National Laboratory in the Leadership Computing Facility. His research topics include parallel and distributed systems, software engineering, computer science, and embedded systems. He is a Senior Member of the IEEE.

**Yung-Hsiang Lu,** is a professor in the Elmore Family School of Electrical and Computer Engineering at Purdue University. His research topics include computer systems, computer vision, and embedded systems. He is a Fellow of the IEEE and Distinguished Scientist of the ACM.

## REFERENCES

1. S. Alyamkin, M. Ardi, A. C. Berg, A. Brighton, B. Chen, Y. Chen, H.-P. Cheng, Z. Fan, C. Feng, B. Fu, K. Gauen, A. Goel, A. Goncharenko, X. Guo, S. Ha, A. Howard, X. Hu, Y. Huang, D. Kang, J. Kim, J. G. Ko, A. Kondratyev, J. Lee, S. Lee, S. Lee, Z. Li, Z. Liang, J. Liu, X. Liu, Y. Lu, Y.-H. Lu, D. Malik, H. H. Nguyen, E. Park, D. Repin, L. Shen, T. Sheng, F. Sun, D. Svitov, G. K. Thiruvathukal, B. Zhang, J. Zhang, X. Zhang, and S. Zhuo. Low-power computer vision: Status, challenges, and opportunities. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(2):411–421, 2019.
2. J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018.
3. G. Chen, W. Wang, Z. He, L. Wang, Y. Yuan, D. Zhang, J. Zhang, P. Zhu, L. Van Gool, J. Han, S. Hoi, Q. Hu, M. Liu, A. Sciarrone, C. Sun, C. Garibotto, D. N.-N. Tran, F. Lavagetto, H. Haleem, H. Motorcu, H. F. Ateş, H.-H. Nguyen, H.-J. Jeon, I. Bisio, J. W. Jeon, J. Li, L. H. Pham, M. Jeon, Q. Feng, S. Li, T. H.-P. Tran, X. Pan, Y.-M. Song, Y. Yao, Y. Du, Z. Xu, and Z. Luo. Visdrone-mot2021: The vision meets drone multiple object tracking challenge results. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 2839–2846, 2021.
4. G. Ciaparrone, F. L. Sánchez, S. Tabik, L. Troiano, R. Tagliaferri, and F. Herrera. Deep learning in video multi-object tracking: A survey. *CoRR*, abs/1907.12740, 2019.
5. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual

learning for image recognition. *CoRR*, abs/1512.03385, 2015.

6. X. Hu, A. Kocher, and Z. Jiao. 21LPCVC-UAV_VIdeo_Track-Sample-Solution. https://github.com/lpcvai/21LPCVC-UAV_VIdeo_Track-Sample-Solution, 2021.
7. X. Hu, Z. Wu, H.-Y. Miao, S. Fan, T. Long, Z. Hu, P. Pi, Y. Wu, Z. Ren, Z. Wang, and G. Hua. $E^2TAD$: An energy-efficient tracking-based action detector, 2022.
8. G. Jocher, A. Stoken, J. Borovec, NanoCode012, ChristopherSTAN, L. Changyu, Laughing, tkianai, A. Hogan, lorenzomammana, yxNONG, AlexWang1900, L. Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, F. Ingham, Frederik, Guilhen, Hatovix, J. Poznanski, J. Fang, L. Yu, changyu98, M. Wang, N. Gupta, O. Akhtar, PetrDvoracek, and P. Rai. ultralytics/yolov5: v3.1 - SOTA Realtime Instance Segmentation, Oct. 2020.
9. J. Liu, D. Wen, H. Gao, W. Tao, T.-W. Chen, K. Osa, and M. Kato. Knowledge representing: Efficient, sparse representation of prior knowledge for knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.
10. W. Luo, X. Zhao, and T. Kim. Multiple object tracking: A review. *CoRR*, abs/1409.7618, 2014.
11. D. Powers. Evaluation: From precision, recall and f-factor to roc, informedness, markedness correlation. *Mach. Learn. Technol.*, 2, 01 2008.
12. RangiLyu. Nanodet-plus: Super fast and high accuracy lightweight anchor-free object detection model. https://github.com/RangiLyu/nanodet, 2021.
13. P. J. Rauss, J. Phillips, M. K. Hamilton, and A. T. DePersia. FERET (Face Recognition Technology) program. In D. H. Schaefer and E. F. Williams, editors, *25th AIPR Workshop: Emerging Applications of Computer Vision*, volume 2962, pages 253 – 263. International Society for Optics and Photonics, SPIE, 1997.
14. J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.
15. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014.
16. G. K. Thiruvathukal, Y.-H. Lu, J. Kim, Y. Chen, and B. Chen, editors. *Low-Power Computer Vision: Improve the Efficiency of Artificial Intelligence*. Chapman and Hall/CRC, ISBN 9780367744700, Feb. 2022.
17. H. Wang, J. Liu, X. Ma, Y. Yong, Z. Chai, and J. Wu. Compressing models with few samples: Mimicking then replacing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 701–710, June 2022.
18. Z. Wang, L. Zheng, Y. Liu, and S. Wang. Towards real-time multi-object tracking. *CoRR*, abs/1909.12605, 2019.
19. N. Wojke, A. Bewley, and D. Paulus. Simple online and realtime tracking with a deep association metric. *CoRR*, abs/1703.07402, 2017.
20. S. Zagoruyko and N. Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016.