Train the Spire: An ML-Driven Single Player GWAP for Image Annotation

Keno Nanninga*
University of Oldenburg
Oldenburg, Niedersachsen, Germany
keno.nanninga@uni-oldenburg.de

Abdulrahman Mohamed Selim*
Interactive Machine Learning
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
abdulrahman.mohamed@dfki.de

Sara-Jane Bittner
Interactive Machine Learning
German Research Center for Artificial Intelligence (DFKI)
Oldenburg, Germany
sara-jane.bittner@dfki.de

Pascal Lessel
German Research Center for Artificial Intelligence (DFKI), Saarland Informatics Campus
Saarbrücken, Germany
pascal.lessel@dfki.de

Michael Barz
Interactive Machine Learning
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
michael.barz@dfki.de

Daniel Sonntag
Interactive Machine Learning
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
Applied Artificial Intelligence
University of Oldenburg
Oldenburg, Germany
daniel.sonntag@dfki.de

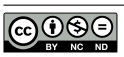
Abstract

Training image classification models requires large labelled datasets, which is particularly challenging in specialised domains where manual expert annotation remains the default, such as eye tracking. To address this challenge, we present *Train the Spire*, a single-player game with a purpose (GWAP) that embeds image annotation within turn-based card game mechanics for crowdsourcing data annotations. The system uses a few-shot deep learning classifier to validate player-generated labels, providing immediate feedback through rewards and penalties. The game incorporates elements, such as progression systems, a companion agent for system transparency, and balanced difficulty, to maintain player engagement while ensuring annotation accuracy. In this paper, we present the system design, implementation details, and evaluation study design for comparing *Train the Spire* against a baseline annotation tool using the *VISUS* mobile eye-tracking dataset in an online user study measuring effectiveness, usability, and enjoyment.

CCS Concepts

• **Human-centered computing** → **Computer supported cooperative work**; • **Information systems** → *Information retrieval*; *Collaborative and social computing systems and tools*; • **Computing methodologies** → *Supervised learning by classification*; *Neural networks*.

*These authors contributed equally to this research.



ETRA '26, Marrakesh, Morocco

ACM ISBN 979-8-4007-2519-7/26/06
https://doi.org/10.1145/3797246.3805861

ACM Reference Format:
Keno Nanninga, Abdulrahman Mohamed Selim, Sara-Jane Bittner, Pascal Lessel, Michael Barz, and Daniel Sonntag. 2026. Train the Spire: An ML-Driven Single Player GWAP for Image Annotation. In *2026 Symposium on Eye Tracking Research and Applications (ETRA '26), June 01–04, 2026, Marrakesh, Morocco.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3797246.3805861

1 Introduction

Training image classification models still relies on human-labelled data, which is expensive and time-consuming to obtain [Amin et al. 2023; Li et al. 2024]. Although Machine Learning (ML) methods such as active learning can improve efficiency by prioritising informative samples [Amin et al. 2023; Kurzhals 2021; Li et al. 2024], they do not remove the need for human annotation [Barz et al. 2025]. Crowdsourcing offers a scalable alternative by distributing tasks across many non-expert contributors and validating labels through mechanisms such as majority voting [Hammon and Hippner 2012]. This approach has supported large datasets, e.g., ImageNet [Deng et al. 2009] and MS COCO [Lin et al. 2014], but maintaining engagement during repetitive annotation remains difficult.

Game With a Purpose (GWAP)s address this challenge by embedding useful computation in enjoyable gameplay [Von Ahn and Dabbish 2008]. They can support intrinsic and extrinsic motivation through progression systems [Hallifax et al. 2023], scarcity [Hamari 2015], and unpredictable rewards [Hinterreiter et al. 2024]. Early GWAPs, e.g., *ESP-Game* showed that they can generate many labels [von Ahn and Dabbish 2004]. Recently, GWAPs have integrated ML into gameplay, e.g., through active learning and multiplayer validation [Barrington et al. 2012] or single-player annotation with real-time classifier feedback [Jesus et al. 2008]. However, existing ML-based GWAPs often require large initial datasets, depend on multiplayer synchronisation, or lack comparison with conventional annotation tools.

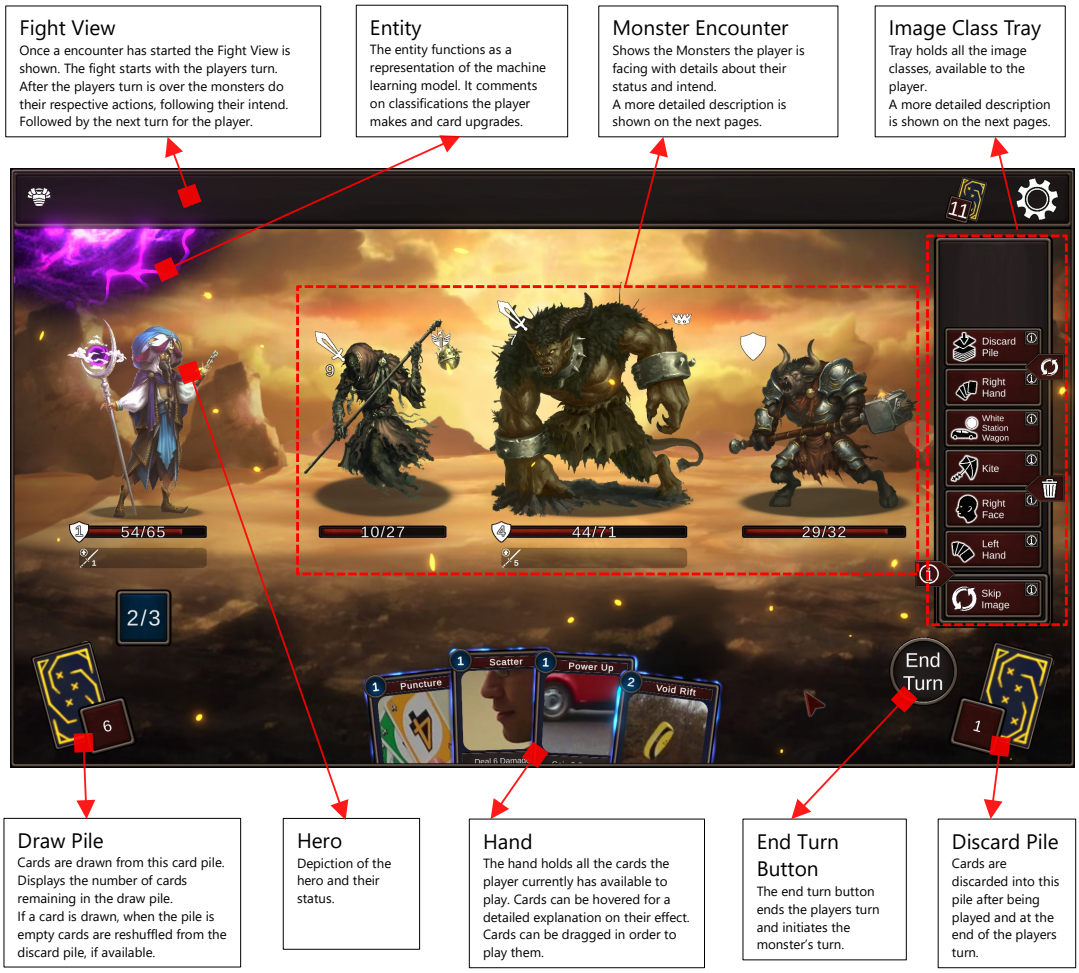


Figure 1: A detailed visualisation of the battle view in *Train the Spire*.

We present *Train the Spire*, a single-player turn-based card game that integrates image annotation into its core mechanics. The system uses a few-shot image classification model trained on six samples per class to verify player-generated labels for eye-tracking data. Its design combines gamification techniques with principles from Self-Determination Theory (SDT) [Vallerand 2000] to address key limitations of prior GWAPs. *Train the Spire* highlights a practical approach to scalable annotation for specialised domains such as eye tracking.

2 Game Design & Implementation

We derived five requirements from the literature to guide the design and implementation of our game:

(1) **Ensure Data Quality:** The game must produce accurate annotations and deter cheating, which can severely reduce computation quality [Von Ahn and Dabbish 2008]. This can be supported by rules that promote correct outputs, e.g., behaviour monitoring and accuracy-based rewards.

(2) **Address the Cold Start Problem:** Multiplayer GWAPs face cold start problems when too few players are available for pairing [Šimko et al. 2013]. Recorded gameplay can simulate partners, but still requires many initial recordings [Šimko et al. 2013]. Single-player games that use ML to verify annotations reduce reliance on other players [Jesus et al. 2008], but still require an initial labelled dataset. Few-shot learning can reduce this requirement while maintaining sufficient accuracy for predictable winning conditions and reduced frustration [Von Ahn and Dabbish 2008].

(3) **Provide a Robust ML Model:** Rewarding agreement with model predictions may encourage players to imitate the model rather than give honest labels, which can undermine autonomous engagement according to SDT [Vallerand 2000]. The model must therefore improve through iterative retraining on player-generated labels. The initial validation dataset monitors accuracy, prevents deployment of compromised models, and preserves annotation integrity.

(4) **Ensure Enjoyable Gameplay:** The game should balance difficulty with player skill to support flow [Nakamura and

Csikszentmihalyi 2009] and sustain engagement through progressive challenge. Gamification elements such as progression systems, scarcity, and unpredictable rewards can strengthen extrinsic motivation. Time pressure should remain limited to protect annotation quality. A companion agent may further increase engagement by making the ML model more relatable [McQuiggan and Lester 2007].

(5) **Ensure Low Implementation Effort and High Accessibility:** Development effort should be low, and the game should be accessible without external software, complex installation, registration, or login. Mechanics and controls should be intuitive and need little tutorial explanation.

2.1 Train the Spire

Train the Spire is a turn-based single-player digital card game inspired by *Slay the Spire*[1]. Its single-player design addresses the cold start problem (Requirement 2), while its round-based structure provides repeated opportunities for label generation. It also includes rogue-like elements[2], i.e., players lose all progress upon death. Combined with randomised challenges and rewards, this supports high replayability with low implementation effort (Requirements 4 and 5). However, difficulty must be balanced carefully, as failure forces players to restart from the beginning. Figure 1 shows the main elements of *Train the Spire*.

Players ascend the spire by fighting increasingly powerful monsters in turn-based combat. This gradual increase in difficulty matches improving player skill, consistent with flow theory [Nakamura and Csikszentmihalyi 2009] (Requirement 4). Players also build and optimise their deck using cards with effects, e.g., dealing damage or applying status effects. These effects are strengthened through a labelling mechanic in which each card shows an image to classify, as shown in Figure 2. Class labels appear in a tray beside the playing field, and players classify images by dragging labels onto cards. When matched, both are consumed. The tray cycles over time, with labels removed from the bottom and new ones added at the top as cards are played. Cards gain experience from correct classifications and level up to strengthen their effects, which can increase player motivation [Segundo Díaz et al. 2022] (Requirement 4). The experience required increases with each level. Incorrect classifications reduce card experience, discouraging careless annotations and supporting data quality (Requirement 1). Players may skip images when no suitable label is available, refreshing the image without consuming labels and reducing guessing. Once the tray is full, however, they must annotate before playing more cards.

2.2 Machine Learning for Label Verification

A few-shot ML model validates player-generated labels by checking prediction confidence (Requirement 2). As the sole label verifier, it must be accurate enough to motivate players and prevent exploitation (Requirement 3). We implemented few-shot learning using Matching Networks [Vinyals et al. 2017], which perform well with small training sets through attention and memory mechanisms, enabling rapid concept acquisition and strong generalisation. This reduces initial dataset requirements while maintaining competitive classification performance.

Train the Spire was developed as a dataset-independent annotation game, i.e., its mechanics and interface support straightforward substitution of datasets without changing the core architecture. In our implementation, we used the *VISUS* mobile eye-tracking dataset [Kurzhals et al. 2014]. The training and validation sets each contained 66 images (6 per class across 11 classes), while the labelling set comprised 11,989 images from which player annotations were drawn. The model remained static during the study to ensure consistency across players, though future versions may iteratively retrain on player-generated labels to guide annotation behaviour and further reduce initial training data requirements. The model is also represented as a companion agent (Requirement 4), narrating gameplay and providing encouragement [McQuiggan and Lester 2007]. It congratulates players on correct annotations, announces level changes, and expresses uncertainty by questioning its own suggestions when players select incorrect labels, thereby improving system transparency [Sartor and Lagioia 2020]. The model also populates the labelling tray by offering contextually appropriate class labels for datasets with many classes.

3 Planned Evaluation

We evaluate *Train the Spire* against a simple web-based baseline annotation tool in an unmoderated online user study to address the following research questions:

- RQ1 How effective is our single-player GWAP *Train the Spire* in generating image annotations compared to a traditional annotation tool for mobile eye-tracking data?
- RQ2 How does *Train the Spire* affect usability and user enjoyment compared to the baseline tool?
- RQ3 How does post-hoc retraining the few-shot classification model on labels collected through *Train the Spire* affect classification performance?[3]

3.1 Study Procedure

After accepting the data usage agreement, participants complete an initial questionnaire on demographics and gaming experience. They then enter an open-play phase in which they can use either tool freely. Usage logs record time spent, annotation rate, and accuracy for both systems. After using at least one tool, participants complete a second questionnaire and enter a prize raffle. The evaluation questionnaire includes System Usability Scale (SUS) [Brooke 1996] for both tools, the full Player Experience Inventory (PXI) [Abeele et al. 2020] for the game, and three enjoyment items adapted from the PXI for both tools. Open-ended questions collect qualitative feedback. Participants may revisit both tools and revise their questionnaire responses multiple times.

4 Conclusion

We introduced *Train the Spire*, a single-player GWAP that generates image labels verified by a few-shot ML model. By using a classifier trained on only six examples per class, the game addresses key challenges in prior GWAPs, especially the cold start problem and

[1] https://www.megacrit.com/ (Accessed January 20, 2026)

[2] For definition: https://www.roguebasin.com/index.php/Berlin_Interpretation (Accessed January 20, 2026)

[3] RQ3 is addressed after study completion because the model remains static during the user study to ensure consistent conditions across participants.

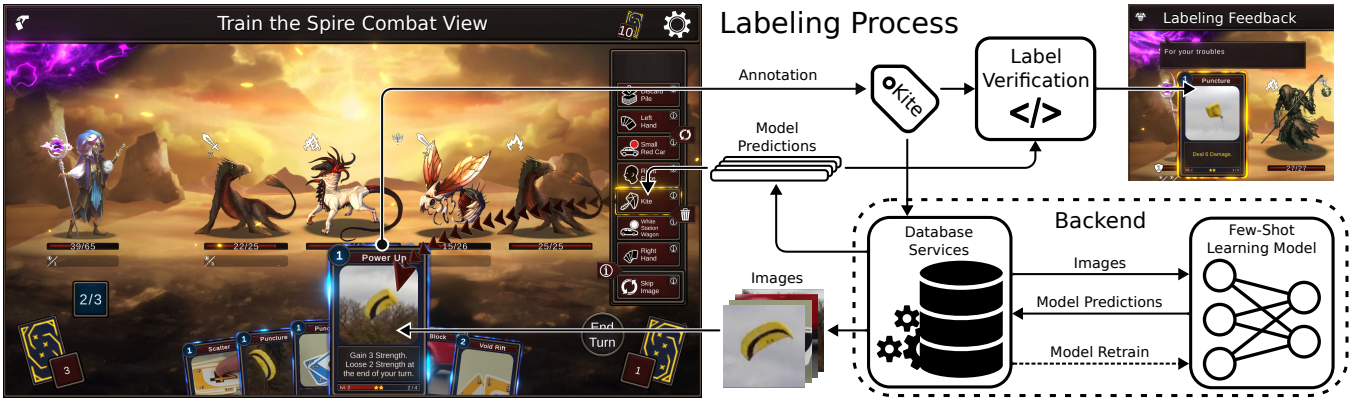


Figure 2: Visualisation of label-related information flow in *Train the Spire*. The backend provides image-prediction pairs. Players match model-suggested labels to cards via drag-and-drop. Model predictions are then used to verify the generated labels, and players are rewarded or punished based on label accuracy.

the need for large initial training datasets. Its dataset-independent architecture also supports straightforward substitution of image classification datasets without changing the core mechanics. *Train the Spire* shows how game mechanics and ML-based verification can be combined to create an engaging annotation interface. Guided by five design requirements from the literature, the system balances data quality, usability, and enjoyment while keeping implementation effort low. Its implementation on the *VISUS* mobile eye-tracking dataset [Kurzhals et al. 2014] illustrates its relevance to specialised domains where expert labelling is costly and time-consuming. At the same time, systems such as *Train the Spire* may raise privacy concerns or influence annotation behaviour when applied to sensitive data or higher-stakes settings. Although these risks are limited in our present study context, future deployments should consider appropriate consent, secure data handling, and potential misuse. Overall, *Train the Spire* provides a practical approach to scalable crowdsourced annotation and a basis for studying how game mechanics can improve engagement and throughput in human-in-the-loop annotation systems.

Acknowledgments

This work was funded by the European Union under grant number 101093079 (MASTER), the Federal Ministry of Research, Technology and Space (BMFTR) under grant number 16IW23002 (No-IDLE) and grant number 16IW24006 (NoIDLEChatGPT), the Lower Saxony Ministry of Science and Culture (MWK) in the zukunft.niedersachsen program, and the Endowed Chair of AAI at the University of Oldenburg.

References

Vero Vanden Abeele, Katta Spiel, Lennart Nacke, Daniel Johnson, and Kathrin Gerling. 2020. Development and validation of the player experience inventory: A scale to measure player experiences at the level of functional and psychosocial consequences. *International Journal of Human-Computer Studies* 135 (March 2020), 102370. https://doi.org/10.1016/j.ijhcs.2019.102370

Sareer Ul Amin, Adnan Hussain, Bumsoo Kim, and Sanghyun Seo. 2023. Deep learning based active learning technique for data annotation and improve the overall performance of classification models. *Expert Systems with Applications* 228 (Oct. 2023), 120391. https://doi.org/10.1016/j.eswa.2023.120391

Luke Barrington, Douglas Turnbull, and Gert Lanckriet. 2012. Game-powered machine learning. *Proceedings of the National Academy of Sciences* 109, 17 (April 2012), 6411–6416. https://doi.org/10.1073/pnas.1014748109 Publisher: Proceedings of the National Academy of Sciences.

Michael Barz, Omair Shahzad Bhatti, Hasan Md Tusfiqur Alam, Duy Minh Ho Nguyen, Kristin Altmeyer, Sarah Malone, and Daniel Sonntag. 2025. eyeNotate: Interactive Annotation of Mobile Eye Tracking Data Based on Few-Shot Image Classification. *Journal of Eye Movement Research* 18, 4 (Aug. 2025), 27. https://doi.org/10.3390/jemr18040027 Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

john Brooke. 1996. SUS: A 'Quick and Dirty' Usability Scale. In *Usability Evaluation In Industry*. CRC Press. Num Pages: 6.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Miami, FL, USA, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Stuart Hallifax, Maximilian Altmeyer, Kristina Kölln, Maria Rauschenberger, and Lennart E. Nacke. 2023. From Points to Progression: A Scoping Review of Game Elements in Gamification Research with a Content Analysis of 280 Research Papers. *Proc. ACM Hum.-Comput. Interact.* 7, CHI PLAY (Oct. 2023), 402:748–402:768. https://doi.org/10.1145/3611048

Juho Hamari. 2015. *Gamification: Motivations & Effects*. Aalto University.

Larissa Hammon and Hajo Hippner. 2012. Crowdsourcing. *Business & Information Systems Engineering* 4, 3 (June 2012), 163–166. https://doi.org/10.1007/s12599-012-0215-7

Smi Hinterreiter, Timo Spinde, Sebastian Oberdörfer, Isao Echizen, and Marc Erich Latoschik. 2024. News Ninja: Gamified Annotation of Linguistic Bias in Online News. *Proc. ACM Hum.-Comput. Interact.* 8, CHI PLAY (Oct. 2024), 327:1–327:29. https://doi.org/10.1145/3677092

Rui Jesus, Duarte Goncalves, Arnaldo J. Abrantes, and Nuno Correia. 2008. Playing Games as a Way to Improve Automatic Image Annotation. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, Anchorage, AK, USA, 1–8. https://doi.org/10.1109/CVPRW.2008.4563045

Kuno Kurzhals. 2021. Image-Based Projection Labeling for Mobile Eye Tracking. In *ACM Symposium on Eye Tracking Research and Applications (ETRA '21 Full Papers)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3448017.3457382

Kuno Kurzhals, Cyrill Fabian Bopp, Jochen Bässler, Felix Ebinger, and Daniel Weiskopf. 2014. Benchmark Data for Evaluating Visualization and Analysis Techniques for Eye Tracking for Video Stimuli. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization (BELIV '14)*. Association for Computing Machinery, New York, NY, USA, 54–60. https://doi.org/10.1145/2669557.2669558

Xiongquan Li, Xukang Wang, Xuhesheng Chen, Yao Lu, Hongpeng Fu, and Ying Cheng Wu. 2024. Unlabeled data selection for active learning in image classification. *Scientific Reports* 14, 1 (Jan. 2024), 424. https://doi.org/10.1038/s41598-023-50598-z Publisher: Nature Publishing Group.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014 (Lecture Notes in Computer Science)*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Scott W. McQuiggan and James C. Lester. 2007. Modeling and evaluating empathy in embodied companion agents. *International Journal of Human-Computer Studies* 65, 4 (April 2007), 348–360. https://doi.org/10.1016/j.ijhcs.2006.11.015

Jeanne Nakamura and Mihaly Csikszentmihalyi. 2009. Flow Theory and Research. In *The Oxford Handbook of Positive Psychology*, Shane J. Lopez and C. R. Snyder (Eds.). Oxford University Press, 195—-206. https://doi.org/10.1093/oxfordhb/9780195187243.013.0018

Giovanni Sartor and Francesca Lagioia. 2020. *The impact of the General Data Protection Regulation (GDPR) on artificial intelligence.* Publications Office of The European Union. https://doi.org/10.2861/293

Rosa Lilia Segundo Díaz, Gustavo Rovelo Ruiz, Miriam Bouzouita, and Karin Coninx. 2022. Building blocks for creating enjoyable games—A systematic literature review. *International Journal of Human-Computer Studies* 159 (March 2022), 102758. https://doi.org/10.1016/j.ijhcs.2021.102758

Robert J. Vallerand. 2000. Deci and Ryan's Self-Determination Theory: A View from the Hierarchical Model of Intrinsic and Extrinsic Motivation. *Psychological Inquiry* 11, 4 (2000), 312–318. Publisher: Taylor & Francis, Ltd..

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2017. Matching Networks for One Shot Learning. https://doi.org/10.48550/arXiv.1606.04080 arXiv:1606.04080 [cs, stat]

Luis von Ahn and Laura Dabbish. 2004. Labeling Images with a Computer Game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '04)*. Association for Computing Machinery, New York, NY, USA, 319–326. https://doi.org/10.1145/985692.985733

Luis Von Ahn and Laura Dabbish. 2008. Designing Games With A Purpose. *Commun. ACM* 51, 8 (2008), 58–67.

Jakub Šimko, Michal Tvarožek, and Mária Bieliková. 2013. Human computation: Image metadata acquisition based on a single-player annotation game. *International Journal of Human-Computer Studies* 71, 10 (Oct. 2013), 933–945. https://doi.org/10.1016/j.ijhcs.2013.05.002