3DPFIX: Improving Remote Novices' 3D Printing Troubleshooting through Human-AI Collaboration

NAHYUN KWON, Texas A&M University, USA
TONG STEVEN SUN, George Mason University, USA
YUYANG GAO, Home Depot, USA
LIANG ZHAO, Emory University, USA
XU WANG, University of Michigan, USA
JEEEUN KIM*, Texas A&M University, USA
SUNGSOO RAY HONG*, George Mason University, USA

The widespread consumer-grade 3D printers and learning resources online enable novices to self-train in remote settings. While troubleshooting plays an essential part of 3D printing, the process remains challenging for many remote novices even with the help of well-developed online sources, such as online troubleshooting archives and online community help. We conducted a formative study with 76 active 3D printing users to learn how remote novices leverage online resources in troubleshooting and their challenges. We found that remote novices cannot fully utilize online resources. For example, the online archives statically provide general information, making it hard to search and relate their unique cases with existing descriptions. Online communities can potentially ease their struggles by providing more targeted suggestions, but a helper who can provide custom help is rather scarce, making it hard to obtain timely assistance. We propose 3DPFIX, an interactive 3D troubleshooting system powered by the pipeline to facilitate Human-AI Collaboration, designed to improve novices' 3D printing experiences and thus help them easily accumulate their domain knowledge. We built 3DPFIX that supports automated diagnosis and solution-seeking. 3DPFIX was built upon shared dialogues about failure cases from Q&A discourses accumulated in online communities. We leverage social annotations (i.e., comments) to build an annotated failure image dataset for AI classifiers and extract a solution pool. Our summative study revealed that using 3DPFIX helped participants spend significantly less effort in diagnosing failures and finding a more accurate solution than relying on their common practice. We also found that 3DPFIX users learn about 3D printing domain-specific knowledge. We discuss the implications of leveraging community-driven data in developing future Human-AI Collaboration designs.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing systems and tools**; **Interactive systems and tools**; • **Computing methodologies** → *Computer vision.*

Additional Key Words and Phrases: Online Troubleshooting, 3D Printing, Remote Novice, Human-AI Collaboration, Community-augmented AI, AI-driven Troubleshooting, Schema Development

*Co-corresponding authors

Authors' addresses: Nahyun Kwon, nahyunkwon@tamu.edu, Texas A&M University, College Station, Texas, USA; Tong Steven Sun, tsun8@gmu.edu, George Mason University, Fairfax, Virginia, USA; Yuyang Gao, yuyang_gao@homedepot.com, Home Depot, Atlanta, Georgia, USA; Liang Zhao, liang.zhao@emory.edu, Emory University, Atlanta, Georgia, USA; Xu Wang, xwanghci@umich.edu, University of Michigan, Ann Arbor, Michigan, USA; Jeeeun Kim, jeeeun.kim@tamu.edu, Texas A&M University, College Station, Texas, USA; Sungsoo Ray Hong, shong31@gmu.edu, George Mason University, Fairfax, Virginia, USA.

ACM 2573-0142/2024/4-ART11
https://doi.org/10.1145/3637288

ACM Reference Format:
Nahyun Kwon, Tong Steven Sun, Yuyang Gao, Liang Zhao, Xu Wang, Jeeeun Kim, and Sungsoo Ray Hong. 2024. 3DPFIX: Improving Remote Novices' 3D Printing Troubleshooting through Human-AI Collaboration. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 11 (April 2024), 33 pages. https://doi.org/10.1145/3637288

1 INTRODUCTION

With the advent of technology and low-cost 3D printers [14, 78], 3D printing has become increasingly available to a broader group of novice users. Learning 3D printing means handling errors caused by several factors, such as machine settings, calibration, materials characteristics, etc. Handling failed printing attempts is an indispensable stepping stone to accumulating hands-on experience and being familiar with the 3D printing domain [6]. However, accommodating 3D printing and troubleshooting can be time-consuming and demanding, especially for *remote novices* who do not have in-person support from advanced users [2, 6, 37]. Since remote novices have no in-person support, they often utilize online resources predominantly categorized into online troubleshooting archives and online communities. Online archives are an online knowledge base that introduces comprehensive 3D printing failure types, viable potential solutions, and other online tutorials [73]) in a structured way (e.g., Simplify3D guide [62]). An alternative resource is online communities, such as Thingiverse forums [38], 3DPrinting subreddit [64], and Stack Overflow [1] where group members post questions to seek help from advanced users. While online archives present comprehensive failure types and solutions, it can be costly for novice users to build up their schematic understanding of 3D printing troubleshooting to digest information [40, 56]. Meanwhile, while online communities can provide a tailored answer that can fit the individual's specific condition, finding the right solution that can work for remote novices can be uncertain. Also, they can easily be intolerable when the suggested solution does not work [37]. These difficulties may, in turn, impose a barrier for remote novices to get into the 3D printing domain [6, 73].

In this work, we improve remote novices' 3D printing troubleshooting experience through a novel human-AI collaboration design. Our design aims at helping remote novices to naturally establish their schematic understanding [56] about their specific printing failure types and general 3D printing knowledge through the "learning-by-using" approach. To determine the requirements of our new design, we conducted a formative study (S1) with 76 active 3D printing users in the online community through an online survey. Based on the former work that focuses on understanding how novices use walk-up-and-print services (e.g., maker spaces, fab labs, print shops) [6, 37], we conducted a formative study (S1) to more specifically learn how they troubleshoot using online resources, mostly online archives, and online communities. In particular, we sought to understand how users formulate their schematic understanding of the troubleshooting task, establish their strategies, and perceive challenges and their desire for future advanced tools in leveraging online sources. S1 found that users can access generic & common solutions listed in online archives, but novices had more difficulties in understanding how they can lexically explain their problem when searching. When matching their case with representative images provided for each printing failure type, their limited understanding of 3D printing and unfamiliar technical terms further constrained them. Meanwhile, we found novice users tend to rely on online communities to get custom solutions as they can inquire easily with plain languages (e.g., photos of a failed print and print settings). But there exists inevitable latency to get the answer mainly due to the unbalanced number of novices seeking help and experts who can help. Rather than doing their "homework" by reading the cumulative knowledge in the online community, we found users tend to upload the issues that have been explained in the past.

Based on the insights found in S1, we designed 3DPFIX, a system that provides an AI-driven "curated" troubleshooting workflow in detecting the failure type and finding a solution. In building our design, 3DPFIX transfers ***social annotation*** [40]—the knowledge accumulated in online communities contributed by their community members—into intelligent and interactive troubleshooting guidance powered by AIs to significantly reduce novices' effort for accomplishing troubleshooting. First, to leverage social annotation in designing 3DPFIX, we extracted information accumulated in the 3D printing community relevant for automating the Q&A process, such as user-posted photos showing the failed print and textual discussion threads to diagnose/resolve the failure and implement failure type classification models. This design decision was made to help remote novices articulate their cases more easily by simply uploading their photos, instead of relying on domain-specific terminologies to describe their cases for search. Especially in remote settings, using visual cues can lead users to self-investigate by helping them gain a deeper understanding of the issue (e.g., a certain setting can cause certain visual cues on print). This would be more useful in remote settings than just using current text-based search schemes. Second, 3DPFIX helps users better relate their image with possible failure types suggested by AIs, 3DPFIX leverages the XAI technique, generally known as a local explanation along with a structured explanation of visual features of the failure type collected from the online community. Third, upon the selection of the failure type, 3DPFIX suggests the feasible solutions, starting from the most generic solution and then advanced solutions based on a user's look-up.

To understand how the new design can improve remote novices' experience in troubleshooting, we conducted a summative study (S2) with 13 novice users in experimental settings and 6 users in more natural settings. By deploying 3DPFIX for a short period for qualitative feedback, we received comments from 6 active 3D Printing users. S2 found several positive signals of 3DPFIX in helping remote novices. For example, S2 found using 3DPFIX helped participants find a solution with significantly reduced effort and cognitive load. Their perceived efficiency in diagnosing the failure type was significantly lower than their common practice. The solutions the participants found were evaluated by 3D printing experts in a blind setting. As a result, expert reviews found that the solutions they found using 3DPFIX were rated significantly higher than the baseline scores. Finally, 3DPFIX significantly increased learnability-related performance than baseline.

This work offers the following contributions:

- **S1: Understanding Remote Novices' 3D Printing Troubleshooting Practice and Challenges**: We seek to more deeply understand how remote novices utilize online resources in their troubleshooting practice and what challenges they encounter in general. Based on our findings, we derive design requirements that can improve remote novices' troubleshooting experience.
- **3DPFIX**: We build 3DPFIX, a novel human-AI collaboration design devised based on S1 findings. Using the collective social annotations extracted from online archives and communities, 3DPFIX realizes a series of image classifiers for supporting remote novices' detection of their failure type and the mapped solutions.
- **S2: Effect of 3DPFIX**: We compare remote novices' use of 3DPFIX to their current practice to understand how 3DPFIX can positively impact novice users' task performance and 3D printing learnability in both behavioral and perceptual manners.
- **Implications for Design**: We reflect on the implications of considering social annotation in developing AI-driven troubleshooting systems to improve users' 3D printing troubleshooting experience and beyond.

2 RELATED WORK

We briefly cover empirical studies that deepen our understanding of how novices learn 3D printing techniques and knowledge in situ. Several previous studies have focused on 3D printing novices who have in-person support, which can largely help them to be familiar with the 3D printing domain [37]. As the makerspace concept became widespread through public print centers, fab labs, and creating hubs, several studies found that in-person help enables casual makers who lack prior knowledge in 3D printing to have advanced experiences in operating and maintaining machines [6]. In general, however, getting support from experts is not easy for casual makers as 3D experts are scanty [37]. Meanwhile, the blooming internet culture has enabled people in remote environments to get fast access to social contracts through online communities [53]. Online communities play a key role in letting 3D printing newcomers gradually adopt new concepts through social interaction with the more experienced practitioners [26]. However, printing 3D objects at home without any in-person support from experts can be finicky. One example can be adjusting print parameters on their machine from shared 3D objects online [47]. Since the novices lack prior knowledge to create 3D objects due to difficulty in learning 3D modeling software, they often download user-created 3D files from public repositories such as Thingiverse [2, 37, 38]. Since not every 3D modeler shares the detailed print settings, including the machine or material used [47] that are often critical for successful 3D printing, it costs more time for novices to find the optimal settings. A study also pointed out that it is common for the modelers not to update changes, and even if they do, updating modification is often done in the comments that could be easily overlooked [71].

Next, we review work in sensemaking and social annotation to characterize the nature of troubleshooting as an exploratory cognitive task. Sensemaking is the process of establishing one's internal representation of the target problem space through interacting with information resources [51, 56]. Several theories have been proposed to explain a user's behavior in sensemaking, including cost structure model [56], information foraging [50], and more [70]. At an early stage, sensemaking focused on an ecosystem between a single user and a single information system. As the literature in CSCW and social computing grows, sensemaking has influenced on developing social information foraging theory [49]. Social information foraging has largely influenced in designing CSCW applications, including collaborative information seeking [31, 32], collaborative data analytics [69], collaborative search [45], social network question and answering [46], and beyond [9, 70]. 3D printing troubleshooting is achieved largely by active interaction between a novice and an advanced users in a remote setting using the information repository built based on social annotation [40]. On top of the established social annotation, novices achieve their sensemaking goals by defining search keywords, reading articles, and refining keywords to follow up to gradually develop one's internal representation. Through their exploration, they may be acquainted with domain-specific knowledge and language, types of viable 3D printer models and their pros and cons, and more importantly, types of printing failures and plausible reasons about why they happen, how to fix them, and many more [2].

While social sensemaking and annotation have inspired numerous applications in CSCW, there have been relatively a scarce of approaches that leverage AI in facilitating novices' sensemaking process, e.g., [30, 34, 35, 58, 76] in troubleshooting. Rather, many focused on improving 3D printing computational pipelines, such as real-time failure detection systems with video cameras. A quality monitoring pipeline can compare the location or shape of the 3D object in the middle of printing using machine learning techniques [5, 13, 48]. In practitioners' fields, AI Build [72] caters an autonomous, large-scale 3D printing with real-time failure detection and correction in robot-arm-based printing, powered by computer vision technology. However, these approaches are scaffolded by additional hardware support that enables real-time monitoring and capturing failures,

which might not be an optimal solution for end-users [5, 13, 48, 72]. This monitoring technique with a camera also tends to be bound to the lighting condition, the specific printer they used for experiments [13], and calibration for an accurate camera positioning [5]. In general, previous AI-driven techniques aim at detecting machine-originated defects rather than supporting human operators' understanding.

Through our review, we identify the two gaps we can investigate further to facilitate remote novices' learning of 3D printing through their troubleshooting experience. First, while several studies have provided useful insights regarding how 3D printing users motivate their learning through in-person setting [15, 37, 53], we are motivated to further understand how novices use online resources in coping with troubleshooting in remote settings. Therefore, we further investigate practices and challenges of the remote novices in utilizing popular online troubleshooting guides and soliciting support from advanced users to identify new design opportunities. Second, while social information foraging and social annotation have provided a useful framework that designers can use to transfer their insights into new applications, developing an interactive human-AI collaboration design for 3D printing remote novices has not been seriously considered in the past literature. To better support our target users, we adopt a user-centered design to develop a viable system that motivates lowering barriers for 3D printing troubleshooting.

3 S1: FORMATIVE STUDY

Taking one step further from the previous work that focused on the novices who have in-person support [6, 37], Study 1 (S1) aims at two main objectives:

- Seeking to understand *remote* novices' practices, challenges, and future desire in handling 3D printing troubleshooting.
- Establishing design requirements for support tools that can potentially improve remote novices' troubleshooting processes.

As a starter, we crawled 26,894 posts from the FixMyPrint subreddit [66] to understand how 3D printer users communicate for troubleshooting in online communities. Next, we conducted an online survey to understand novices' strategies, challenges, and desires in utilizing existing online resources for troubleshooting compared to user groups with advanced knowledge.

3.1 Preliminary Observation

In this phase, we observed behaviors of members in FixMyPrint subreddit [66], one of the popular online 3D printing troubleshooting communities with over 100k members, to understand how they share 3D printing issues and interact with others to solve them using a board discussion. We also looked into FixMyPrint's policy & announcements to see how they foster effective troubleshooting-related communication. Our observation revealed the growing need for targeted solutions for novices, continuously growing posts seeking answers to fix their issues (see Figure 1, (a)). Such needs seemed to be aligned with the active participation of motivated 3D printing enthusiasts and experts' altruistic motivations to benefit 3D printing novices as a community.

Unfortunately, a nontrivial amount of posts are eventually left unanswered. One viable reason is that questions are repetitively overflowing the boards. Many of them share commonalities thus existing solutions can be easily found in prior posts by searching and finding similar cases, which is a practice that is not well-used by novices. The announcement on the main page recommends reading online web archives first before posting questions, especially Simplify3D guide [62] first, along with other options such as Rigid.ink [63] Matterhackers [43], and Reprap [54] if desired. AutoModerator in this community immediately sends an automatic comment for every new post saying *"Most common print quality issues can be found in the Simplify3D print quality guide"*. Many

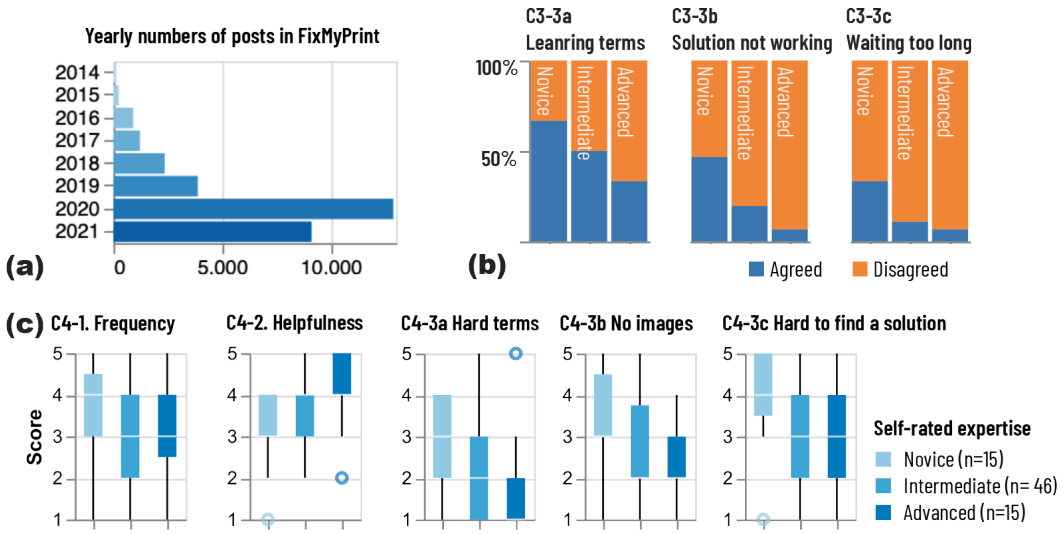


Fig. 1. (a) Yearly count of Q&A posts in FixMyPrint (2021 counts posts between Jan. and June), (b) Survey results of C3-3, users' changing perception regarding online communities depending on their self-rated expertise, (c) Survey results of C4-3, users' changing perception regarding online archives depending on their self-rated expertise

online communities' internal recommendation system pulls popular posts onto the main page, further marginalizing non-busy posts and depreciating ones with only a little to no discussion. In our informal conversation with community moderators, they seemed to be aware of this issue, encouraging users to re-post when initial attempts were left unresolved due to a lack of attention.

As Figure 1 (a) shows, demand for seeking a targeted answer is soaring while experts' availability to read and comment remains flat, resulting in an imbalance. Among $\approx 27k$ posts that we collected from the FixMyPrint subreddit, 15% of the posts ($\approx 6k$) have never been answered and nearly 50% have 3 comments or fewer. Given the depth of discussion needed to reach an eventual solution, our data acknowledge substantial posts ending up being unsuccessful attempts. For posts with at least one comment, the average time to get the first comment was about 789 minutes, which indicates the wait time for getting attention is nearly 13 hours on average, which does not even guarantee resolving the issue. We assume such an imbalance in supply and demand can hinder novices from finding valid solutions and attaining 3D printing domain knowledge promptly. Through our observation, we were motivated to deeply understand the challenges of novices, as well as to identify further how the current online archives impose more challenges.

3.2 Online Survey

3.2.1 Methodology. As a way to understand the challenges of 3D printing users in troubleshooting failures more deeply, we designed a survey and reached out to active 3D printing practitioners in online communities. To recruit 3D printing users online with different levels of expertise, we posted invitation flyers to popular Reddit 3D printing communities, including FixMyPrint [66], Ender3 [65], 3DPrinting [64], and Prusa3D [67] that had at least 40k members. Then, we reached out to 842 members of the chosen subreddits who recently posted inquiries through direct messages. We also sent an invitation flyer to popular 3D printing-related discord channels with more than 2k members, titled Print Everything [11], Creality 3D Printers [10], and Prusa3D [55]. Our invitation resulted in 76 responses being returned.

In the survey, we first asked participants to self-evaluate their expertise in 3 levels: novice (G_{nov}), intermediate (G_{int}), and advanced (G_{adv}). We used self-assessed expertise instead of more quantitative measures such as years of experience. Unlike other domains, such as education, which can fairly reflect people's expertise using quantitative metrics, 3D printing is highly arbitrary in defining the expertise. For example, many hours of taking workshops and being tightly committed and engaged in makerspaces may intensively increase their expertise in a short time. Among 76 respondents, 15 self-assessed themselves as novice, while 46 assessed as intermediate and 15 assessed as advanced, respectively. Then we asked the 9 questions of the 5 categories below:

C1. Printing environment: What type of filaments and printers do you use?
C2. Common strategies for troubleshooting: Please indicate your "strategies" to use varying online resources when handling troubleshooting.
C3. About using online communities
- C3-1. When using online communities, what do you usually do to resolve issues?
- C3-2. Which aspects of online communities help you the most?
- C3-3. Please indicate if you agree with the statements as follows: When using online communities, (a) I learned technical terms and several useful tips through discussion; (b) Solutions suggested did not really solve my issue; (c) I needed to wait too long to get others' responses.

C4. About using online archives
- C4-1. How often do you read online archives?
- C4-2. To what extent is reading online archives helpful in resolving your issues?
- C4-3. Please indicate how much you agree with the following aspects in applying information learned from online archives: When using online archives, (a) It is hard to understand technical terms; (b) None of the example images in the archives are similar to my case; (c) It is hard to find a specific solution that applies to my own printing issue

C5. Further remark: Open-ended questions that they can freely describe their experience regarding troubleshooting using online resources.

For questions in C1-3, we provided multiple choice options that respondents can choose, including 'other' to describe further answers that do not fall under any options.

3.2.2 Results. About the common troubleshooting strategies (C2), reading well-known online archives and the most popular suggestion (N = 61, 79% of participants), followed by searching previous online community posts (64%), and posting problems on online communities' (51%). Relatively fewer respondents relied on video sources such as YouTube (11%) or search Google (9%).

Online Communities: difficulty in searching and obtaining timely help. We designed C3 to confirm the usage patterns we found in the preliminary online community observation and to hear more about barriers and desires. It seemed that users post their own questions since searching the key articles often fails, regardless of their level of expertise. P32 (intermediate) mentioned: *"First of all, [I] look for previous posts with similar issues, then ask those that answered those posts, and if nothing worked, posting my own with details on settings [...] and where it failed with a close-up on the failure"*. The most common troubleshooting strategy (C3-1) was searching for posts that handled a similar problem from the previous discussion board (69.2%), followed by uploading posts with photos of failed prints (62.8%). Many understood how useful well-articulated discussion records with details are to obtain more accurate suggestions, as well as the fact that they assist by building assets and a healthy community, attracting more members in the longer term. Many respondents indicated that their initial posts include sufficient details: *"I have found giving as much detail along with posting images is a good way to not only get help myself but provide people in the future a resource as well (P14, intermediate)"*.

Nonetheless, several novice participants found it hard to search for the right discussion record that matches their case. For instance, P19 (novice) pointed to the problem of Reddit's search function: *"Reddit's decentralized model limits search functionality because of lack of tags or use of key terms"*. P17 (novice) mentioned that *"Searching doesn't always help, but I try it anyway just to do due diligence"*. Not knowing the terminology to articulate what is happening, novices' success rate to hit the right post discussing similar cases through text-based query has no choice but to be limited.

Several respondents found getting comments from others in online communities hard (C3-3), while the majority of respondents found immediate support is the key. 61.5% of respondents noticed that the benefit of online communities significantly reduces when they cannot have "synchronous feedback" or "instant support". *"The only real issue is that it can take quite a while to get just one response, and sometimes posts end up not getting answered at all"* (P26, intermediate).

As found (Section 3.1), the core value of online communities seemed to be custom feedback obtained from experienced users without domain-specific language. 61.5% of respondents loved using online communities for troubleshooting owing to the possibility of getting their issues diagnosed using photos of a failed print. Meanwhile, people with lower expertise felt they learned more about 3D printing language than people with higher expertise did (see Figure 1, (b) C3-3a). On the other hand, the risk of using online communities is the supply-and-demand gap in Q&A, which inevitably disables synchronous feedback and lowers the chance of getting a timely solution. Interestingly, we found novice users perceive the waiting time as longer than more advanced groups (see Figure 1 (b) C3-3c). For instance, P10 (intermediate) drew attention to the high demand of novice users requesting help in comparison with the relatively low number of experienced users giving useful suggestions, stating *"There are too many users requiring help compared to those who can actually help. That makes it harder for me to get a response"*. Even though the perceived waiting time of the advanced user groups is relatively shorter than the novice group, they are aware of the issue; P13 (advanced) criticizing Reddit's availability noted; *"Reddit is not an appropriate forum for troubleshooting. Troubleshooting should really be done with live [and more interactive] feedback. Posting a bunch of pictures and waiting doesn't lead to good results"*.

In sum, we found discourses of online communities and custom suggestions from experienced users benefit remote novices. They learn domain-specific languages and experience useful tips to fortify their 3D printing knowledge by obtaining suggestions from others, trying recommended solutions, and asking for further help if they are not resolved. However, we also found experienced users in online communities scarce, making it hard for novices to elicit suggestions.

Online archives: unfamiliar technical terms, difficulty in finding applicable problem & solution. C4's main goal is to understand how and why users' perceptions about using online archives vary, and whether there exists differences depending on their expertise. We conducted Spearman's rank correlation for every C4 question with the null hypothesis that respondents' Likert scale rates will not vary between groups with different expertise levels (H_0). Regarding the frequency of using online archives (C4-1), Spearman's rank correlation test found that the frequency didn't vary significantly depending on their level of expertise (G_{nov}: M = 3.53, SD = 1.20, G_{int}: M = 3.27, SD = 1.18, and G_{adv}: M = 3.27, SD = 1.18, r(74) = 0.036, p = 0.754, see Figure 1 (c)). On the other hand, we found that people with lower expertise tend to perceive online archives significantly less useful than the group(s) with more experience (G_{nov}: M = 3.07, SD = 0.85, G_{int}: M = 3.93, SD = 0.77, and G_{adv}: M = 4.13, SD = 0.81, r(74) = 0.449, p = 0.000, see Figure 1 (c)). We especially noticed the gap between G_{nov} and G_{int} to be way larger than the gap between G_{int} and G_{adv}. These results indicate that the level of expertise does not affect the frequency of use, but novices may be unlikely to obtain useful information compared to advanced users.

Regarding the analyses to understand possible factors that make the online archives hard to use, Spearman's rank correlation rejected every null hypothesis in this category. First, technical

terms can be a bigger barrier to using online archives for novices than the experienced (G_{nov}: M = 3.40, SD = 1.20, G_{int}: M = 2.65, SD = 1.07, and G_{adv}: M = 2.47, SD = 0.96, r(74) = -0.250, p = 0.030, see Figure 1 (c), C4-3a). Next, Spearman's rank correlation test found that relating example images used in the online archive to a user's specific problem becomes harder if their expertise levels are lower (G_{nov}: M = 2.87, SD = 1.09, G_{int}: M = 2.11, SD = 1.07, and G_{adv}: M = 1.80, SD = 1.11, r(74) = -0.313, p = 0.006, see Figure 1, C4-3b). Finally, the lower self-rated expertise levels are, the more difficulties in applying a general level of solution description to their specific case (G_{nov}: M = 4.00, SD = 1.10, G_{int}: M = 2.93, SD = 0.92, and G_{adv}: M = 3.07, SD = 1.18, r(74) = -0.269, p = 0.019, see Figure 1 (c), C4-3c). These findings could explain why online archives are perceived as less useful by users with lower-level expertise particularly. Novices may need to invest more time and effort to understand domain-specific language to fully digest the articles. Limited knowledge about terminologies may also negatively affect the way that novices search posts due to the text-based service. While images are more direct and intuitive to deliver the context, representative images displayed may be unfamiliar, making it also hard to locate similarities, given that every single sign and symptom could be unique in different users. Even if they were able to find an article relevant, novices may also face barriers in selecting one among several suggested solutions.

3.3 Design Requirements for Remote Novices

Remote novices' desire seems straightforward–(1) being capable of easily identifying their problems, (2) using an easy inquiry, and (3) obtaining the targeted solution to resolve the issue directly, without circumvention or prediction. "Remote experts" are a big help since they can immediately notice issues that are shown with minimal textual/visual information, and give straightforward answers to seemingly-unique printing issues than examining existing posts. However, the overwhelmingly increased volume of posts makes the experts an uncertain asset.

To mitigate the gap we identified in S1, we propose a novel human-AI collaboration system leveraging social annotation obtained from existing online communities that have accumulated remote experts' knowledge over time, which can advance remote novices' 3D printing troubleshooting experience. Using social annotation (comments), we seek to expand the role of AI as a curator who can balance the load for online experts. The role of these remote experts includes guiding novices to identify the type of problem, match feasible solutions in plain language, enabling access to targeted solutions similar to the human experts' custom solutions. Specific design requirements we learned from S1 are shown as follows.

DR1. Easy articulation of the problem: Troubleshooting often starts by investigating noteworthy visual features from a failed print to understand the printing failure type. Visually investigating the problem is an intuitive and straightforward way to diagnose a failure type. To facilitate easy diagnosis, our first DR is to use photos as input. By enabling an image-based search, remote novices can avoid describing unfamiliar domain terms. Considering text-based communication appeared in both online archives and communities, we hypothesize that an automatic diagnosis powered by vision-based models can better assist remote novices.

DR2. Reliable relation of a user-uploaded image with a failure type: Diagnosing 3D printing failures type by referring to a single representative image for each type could not be complete, because failures visual traits can be different from one case to another. To help remote novices better identify the most viable failures shown in the image, the system should help users get the visual reasoning of the failure types (e.g., [8, 23, 24]). Through the system, users should be able to understand which visual features belong to a certain failure type.

DR3. An unobstructed flow of thoughts without being interrupted by unfamiliar technical terms: The use of synonyms without explicit clarification can confuse users (e.g., stringing - wisping, vibration - ghosting/rippling/ringing), the new design should provide an explanation of

newer technical terms quickly. Being on-demand is also essential, considering the different levels of knowledge that user groups might possess.

DR4. Generic solutions first, case-specific solutions on demand: Since several different reasons underlie the seen features of printing failures (e.g., too low printing temperature and/or high printing speed can cause under-extrusion), there could be a diverse set of how-to suggestions. Also, searching for previous discussions is a common behavior to find useful dialogues, but one-size-fits-all does not apply to this case. A built-in search function is limited to retrieving the applicable posts, and a lack of novices' expertise may result in the wrong choice of suggestions. Improving the capability to find the applicable post among existing ones is thus essential, and to help users select the highly applicable solution which is worth trying first among many alternatives will eventually help reduce the time for trial and error. To address challenges in using two online resources, we apply the visual information-seeking mantra, "overview first, zoom and filter, and details on demand" [60] in providing feasible solutions. Online archives list generic solutions along with the description of the failure itself and common reasons that cause such problems. Also, as the online communities tend to encourage novices to append context (e.g., printer and filament types and settings) for accurate diagnosis and case-specific solutions, we are to provide a detailed description of the diagnosed failure and the generic solutions first (overview). If desired, users can proceed to open more case-specific solutions that do not fall into the generic cases (details on demand). Also, if available for individual solutions, the system should provide conditions or clues that further narrow down the search for options (zoom and filter).

4 3DPFIX: AI-DRIVEN DESIGN FOR AUTOMATED 3D PRINTING TROUBLESHOOTING

Online communities have accumulated ample real-world image data on 3D printing failures and discussions as a form of social annotation. While such resources provide a significant amount of textual records about the issues showing in the images, i.e., *social annotations*, remote novices' process to get the solution is not without uncertainty, as per our S1 findings. For instance, many of their postings may not receive any attention at all, they may have to wait until other experienced users respond, or they may overlook the existing posts and ask the same question. Using static online archives is also challenging due to unfamiliar technical terms, difficulties in relating their case with the description, and selecting suggestions to try. To provide a better troubleshooting experience to remote novices, 3DPFIX provides interactive features built based on the DRs we derived in S1. In particular, it provides an automatic diagnosis of the 3D printing failure type suggested by AI and helps users find corresponding solutions based on social annotations that we extracted from 3D printing online archives and communities.

4.1 3DPFIX Design Overview & Implementation

Following the four design requirements defined in S1, 3DPFIX is equipped with four main features: (1) a diagnosis of failure types from the user-uploaded photo (**DR1**), (2) visual reasoning through the grayscale saliency attention map to provide a local explanation of the CNN model's decision and representative sample images (**DR2**), (3) a detailed explanation of technical term with a hover-over window on demand (**DR3**), and (4) generic & case-specific solutions with filter conditions to narrow down the search (**DR4**).

#1. Image-based diagnosis for the easy articulation of the problem (DR1): Visual characteristics of failed prints play a key role in recognizing 3D printing failure types; automated diagnosis through user-uploaded photos is the most intuitive and simple way for novices. Users can start by uploading photos of their failed prints using our web-based interface. 3DPFIX examines if it contains any type of 3D printing failure by computing its probability of presence using all CNN classification models step by step. 3DPFIX shows results within a few seconds and lets users directly recognize

in synchronous settings. 3DPFIX does not display prediction probabilities directly to the users as presenting numeric values can confuse users, possibly due to an objective interpretation [75]. Instead, 3DPFIX presents abstracted labels of prediction results in four levels: Highly Likely (75% - 100%), Likely (50% - 75%), Unlikely (25% - 50%), and Highly Unlikely (0% - 25%). As in Figure 2, the 'Show AI's Best Guess' is activated, which shows the 'Highly Likely' failure type visible only by default. Upon preference, users can also enable 'Seeing all failure types' to check all the failure types that 3DPFIX supports.

#2. Visual reasoning of diagnosis for reliable relation between user-uploaded photos and diagnosed failure types (DR2): In addition to the diagnosis, 3DPFIX provides visual reasoning on the diagnosed results, which can support users to not only better appreciate classification models' decisions, but also learn various visual features to recognize different types of failure. As a lower subsection of Figure 2 shows, our tool also provides the grayscale saliency map generated by Grad-CAM [59] for each type diagnosed. Saliency maps present the region used for prediction, highlighting the point of interest. This accentuation helps users focus on the problematic parts in the photo, as well as contribute to the visual explanation of CNN models where it is used to make the decision [52]. Upon clicking each failure type, 3DPFIX provides the detailed descriptions under the *'What's this problem?'* tab. This tab contains several representative images with different visual features, extracted from our image dataset (See Figure 6), which will provide another visual reasoning for users to understand why such failure types were diagnosed. These images assist users to discern common visual characteristics, facilitating learning about print failures with their appearances. In addition, users can also read textual explanations for each failure type, such as what

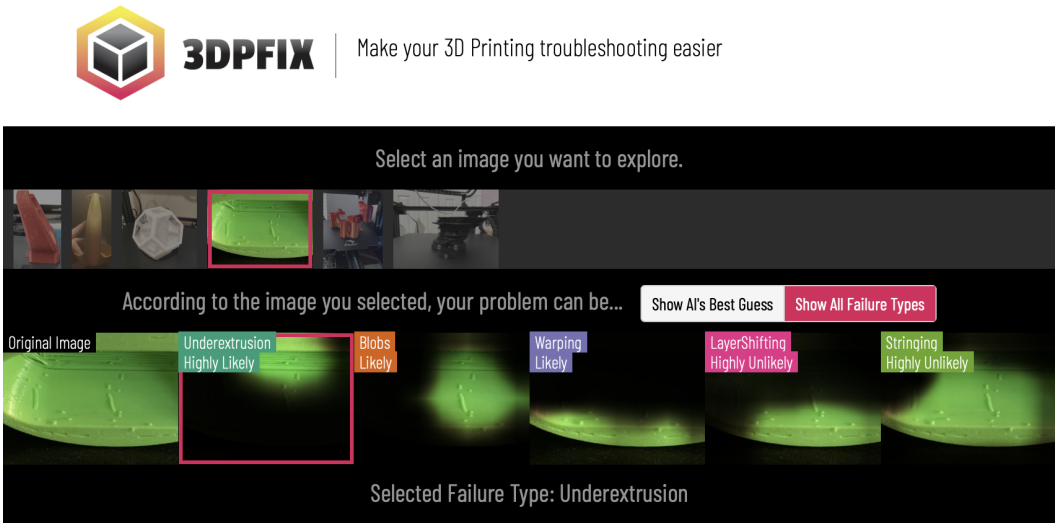


Fig. 2. This is the top section of the interface where users can select the 3D printing images they uploaded for diagnosis (upper subsection) and the corresponding failure type predictions with saliency maps generated by our models (lower subsection). Images are clickable tabs where red-colored borders indicate active selections. The system also displays the likelihood of each failure type prediction (Highly Likely: 75% - 100%, Likely: 50% - 75%, Unlikely: 25% - 50%, and Highly Unlikely: 0% - 25%). Users can click a specific failure prediction to explore further the solutions on the bottom section of the interface shown in Figure 3, 5, 4. Users can also use the toggle switch buttons to "Show All Failure Types" or "Show AI's Best Guess" which filters out the predictions below the 'Highly Likely' level.

it is called in the 3D printing community (jargon), what can cause it, and which visual characteristics commonly appear. Here users are able to learn about the type of failure, terminologies, visual characteristics, and the clue (what causes them), users move to the next step to learn about solutions.

#3. Hover-over functions for a detailed description of technical terms (DR3): In the 3D printing domain, there are many technical terms that refer to specific settings or various techniques to improve printing quality. Remote novices often face barriers due to unfamiliar technical terms during their troubleshooting processes, which makes them switch to searching for external sources. To minimize such intervention, as in Figure 4, 3DPFIX uses hover-over boxes, as many encyclopedic web services such as Wikipedia adopt, to further aid newcomers in learning technical terms used in the text when describing solutions. According to their knowledge levels, users can easily hover over the terms on demand and read a description along with the image.

#4. Curated solution searching flow (DR4): 3DPFIX curates two types of solutions, which leads users to follow a reasonable flow of troubleshooting to first look into the most general solution and then proceed to the custom or advanced solutions rooted in the visual information-seeking mantra [60]. Users can begin by exploring common solutions that are likely to solve many general issues by 'See Common Solutions' (Figure 5, (a)). They can also refer to the 'clue' to have a deeper understanding of what other phenomena could be there, which will help them self-diagnose issues. The clues about observations (e.g., popping noise while printing, the bed is wobbling) or settings (e.g., printing temperature was lower than 180 degrees when using PLA, installing new firmware) may lead users to the specific solution set. For example, if a user has an under-extrusion with the PLA as a material with a temperature set lower than 180 degrees, a low temperature of the print nozzle is likely to be a cause of not being able to completely melt and fuse PLA to safely 3D print the material [28]. The first-class trial must increase the nozzle temperature to around 210 degrees, which tends to work for many across various types of machines. If the user can spot any potential cause from the list of *clues*, the user can filter the relevant solutions by navigating through the given options. The 'Common Solution Did Not Work?' tab (Figure 5, (b)) unfolds solutions for more special cases for those who were not able to fix issues by the most common solutions to try these advanced solutions. They can also filter and sort solutions by difficulty level, where the color scheme of the solution cards reflects the difficulty level for intuitive comprehension. While novices would like to try the easiest options first, 'hard' solutions may alert them that they need more

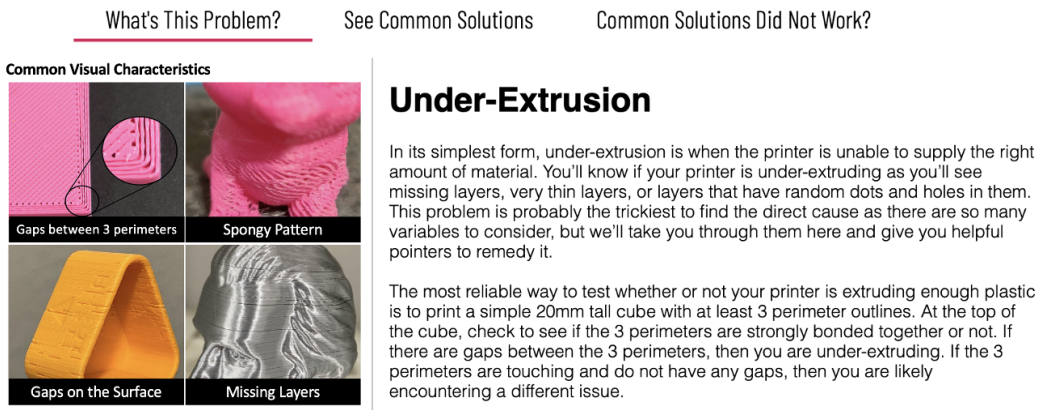


Fig. 3. The bottom section of the interface has 3 tabs to investigate solutions for the selected failure type by the module in Figure 2. The first tab, 'What's This Problem?', shows example photos about the common visual characteristics of a failure type, as well as an easy-understandable description on the right.

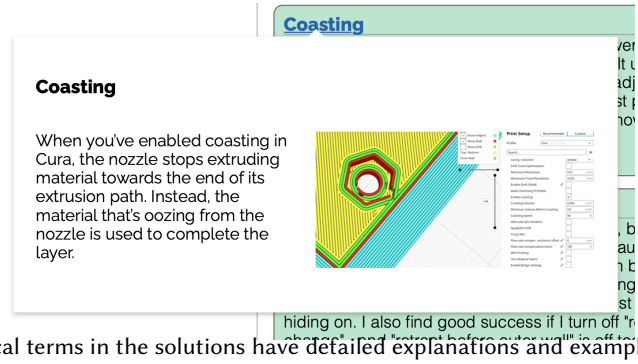
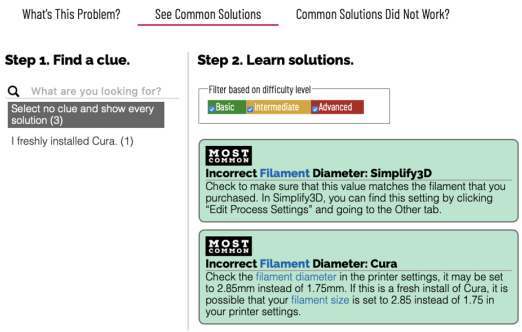


Fig. 4. Some technical terms in the solutions have detailed explanations and example photos when users hover over the terms (blue-colored with underlines). These underlined terms are also clickable on external websites with more comprehensive descriptions.

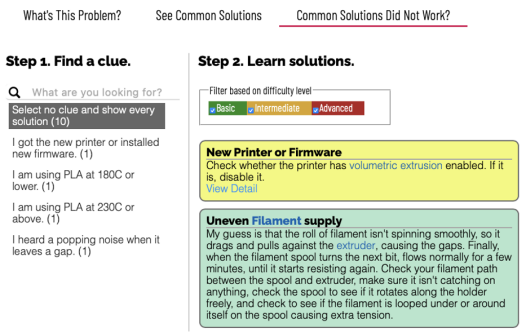
devotion. The solution card shows the title and a short description to convey the key information. If a userneeds more details, such as how to deal with the Cura, one of the most widely used slicing software, to increase the nozzle temperature, they can click on the 'View Detail' or 'View Video' in the card that will redirect them to the relevant online web document or a how-to video if needed.

4.2 Example User Scenario

Here, we present an example user scenario of 3DPFIX with a speculative user, Teddy. After auditing a Maker workshop from a local library and being introduced to many affordable 3D printers, Teddy recently bought a cheap 3D printer that costs less than $100 for his hobby projects at home. Being unable to meet with his workshop instructors due to many canceled in-person sessions due to the pandemic, he followed many how-to videos to install and set up the printer, then tried his first 3D printing using a 3D file he downloaded from Thingiverse [38]. Yet, he found that his 3D print has a lot of gaps on the surface, which makes the outer walls look separated from each other. While he is



(a) Tab for common solutions



(b) Tab for specific solutions

Fig. 5. The second and the third tabs of the bottom section are for investigating solutions. Users usually start with "See Common Solutions" where they can either search for existing clues in "Step 1. Find a clue." (on the left), or go straight to "Step 2. Learn solutions." (on the right) to browse the solution cards they think are relevant to their 3D printing failures. Solution cards are color-coded by their difficulty levels (Basic is green, Intermediate is in yellow, and Advanced is in red). A difficulty-level filter is also available. If users cannot find the solutions they need in the common solution tab, they can go to the last tab to see a more comprehensive batch of solutions provided by the system. All the navigation and filtering features are the same as in the previous tab.

eager to troubleshoot by himself, lots of general tutorials in how-to videos do not show how to address this specific issue. In fact, Teddy does not know what to call it nor how to describe the issue clearly, making the text search nearly impossible. Reminding of many supporting communities learned from the workshop he attended, he posted a photo of it to seek help from community users. Unfortunately, his post did not get any attention.

He instead decided to use the 3DPFIX. In a few seconds, he got his issue diagnosed, learning that it is a known issue called "under-extrusion" and might be because of the too-low temperature set for printing or potentially humid filaments, and many more. As the system shows grayscale saliency maps that highlight the area with the unique visual characteristics as shown in Figure 2, he is also able to learn about possible different failures that he may need to take another look for his future print trials. Although his issue seemed to be obviously under-extrusion, seeing the visual similarity between his and the sample image, Teddy also wants to check all other possible failure types to make sure that there are no other related issues. He taps on the 'Show All Failure Types' button and checks the sample images informing common visual characteristics under the 'What's this Problem?' as shown in Figure 2. Looking at all possible options, Teddy is confident that it is under-extrusion indeed, and learns how he can identify the same issue appearing in various forms in the future. Now, he has become more knowledgeable in that he needs to check whether there are any 'spongy patterns' on the surface.

Stepping to the solution stage, Teddy tried all the common solutions suggested for under-extrusion, but his issue remained unsolved. He then proceeded to the specific solutions tab, which covers most possible cases. By going through the *clues* that further hint about settings that cause the issue, he found that 'I heard a popping noise when it leaves a gap' matches his current settings. He expanded the card to see relevant solutions, which noted that it might be caused by wet filaments. He thoroughly read the linked web article, learning that it can even exacerbate his printing quality by other akin problems such as blobbing. One recommendation was to dry it using special equipment, while he chose to buy a new filament that was delivered in a sealed vinyl and cleared the issue.

4.3 Implementation

The back-end framework of the web-based user interface is built on Flask [18], a lightweight web application framework written in Python. Having a Python-based back end enables real-time model classification and visual explanation generation using our pretrained deep learning models, post-hoc explanation techniques (Grad-CAM), and data visualization libraries (PyTorch, Matplotlib, OpenCV). The front end is written in basic HTML, CSS, JavaScript, and additional libraries (D3, jQuery) for dynamic elements. The visual explanation is generated by Grad-CAM [59] PyTorch library, which is a popular post-hoc technique that computes the gradient saliency maps [61] to visualize the model prediction by highlighting 'important' pixels (i.e., changes in intensities of these pixels have the most impact on the prediction score). All maps are transformed into grayscale by applying a sigmoid (slope = 4) to the model-generated attention scores pixel-by-pixel. This allows segmenting the attention areas (transparent mapping) from non-attention areas (dark mapping). Users can assume that a model's judgment of a failure type was made by looking at the attention areas of this saliency map. The brighter the mapping, the higher the attention of the model. Then, each saliency map is passed to the front end as a clickable tab for each prediction. By choosing one, the interface will show the corresponding introduction, clues, and solutions relevant to the selected failure type at the bottom for the user's further investigation.

4.4 3DPFIX Dataset

At a high level, our system was implemented by extracting and organizing domain knowledge from ample Q&A records stacked in an online community for 3D printing troubleshooting discussions.

(1) To perform automatic 3D printing failure diagnosis:
   - We collected a raw post dataset that contains user-uploaded failure images and corresponding text discussions. We leveraged the rich text information that is associated with user-uploaded photos in Reddit posts, including body texts and discourses in threads, and developed a keyword-based automatic classification of the posts and, thus, the images.
   - Using this labeled image dataset, we trained individual binary classification models that manage one failure each. The resulting classifiers can be applied to detect 3D printing failures from the user-uploaded photo.

(2) To suggest solutions based on the diagnosed problems:
   - We used the classified posts to extract encyclopedic user suggestions addressing each 3D printing failure.
   - We then extracted high-quality comments by the user feedback score (upvotes) along with *clues* that indicate the specific cases from relevant threads as the most viable solutions.

4.4.1 Data Preparation. We collected a multi-modal dataset from the FixMyPrint subreddit with 28,030 images from raw 30,780 posts. This dataset covers all posts of this subreddit during its existence from 2014 to July 2021. The posts follow a common thread structure, with one post and subsequent comments. Often, multiple users discuss a specific 3D printing issue in varying depth in one thread.

4.4.2 Building Image Dataset by Text-based Classification. In order to automate the image-based diagnosis of failures through AI and to build a solution set that covers various cases accumulated over time in the community, we built an image dataset annotated with corresponding 3D printing failures. FixMyPrint's multi-modal (image and text) post dataset can work as a good source to determine the 3D printing failure type that each thread is trying to address. This grants us a unique opportunity in using the conversations to automatically label the images in the posts. Thus, our initial goal became developing a text-based automated classification approach to categorize the 3D printing failure images of each thread. First of all, as the discussions in this subreddit are highly domain-specific, they are different from everyday conversations or news articles that large language models are often trained on, such as BERT [16, 17, 74]. We decided to introduce heuristics into the classification and explore keywords-based approaches, as zero-shot classification techniques [74] seem not to be a viable solution owing to their low accuracy in the classification of posts. We used existing well-known troubleshooting archives as base documents to be compared with the posts, as those resources, such as Simplify3D, provide a pre-defined list of common 3D printing failures. For each printing failure type, they also provide an accompanying document with a detailed description of the phenomenon, why it may happen, and potential solutions for users. We specifically chose Simplify3D's print quality guide [62], which presents 27 types of failure, so we started with 27 base documents. We conducted an initial automatic classification of posts using cosine similarity [3] of the base documents on 26,232 posts that have at least one comment. Also, to increase the accuracy of keyword-based classification without any overlapping keywords shared in different document themes, we defined one representative word for each failure type, which we call 'failure-specific keyword', that can symbolize a unique failure type. The leading researcher manually weighted the failure-specific keywords for each type, (e.g., not sticking to the printing bed - bed, stringing or oozing - string, layer shifting - shift, under-extrusion - extrude, warping - warp, blobs - seam, poor surface above supports - support). We chose 5 of the most common types with distinct visual

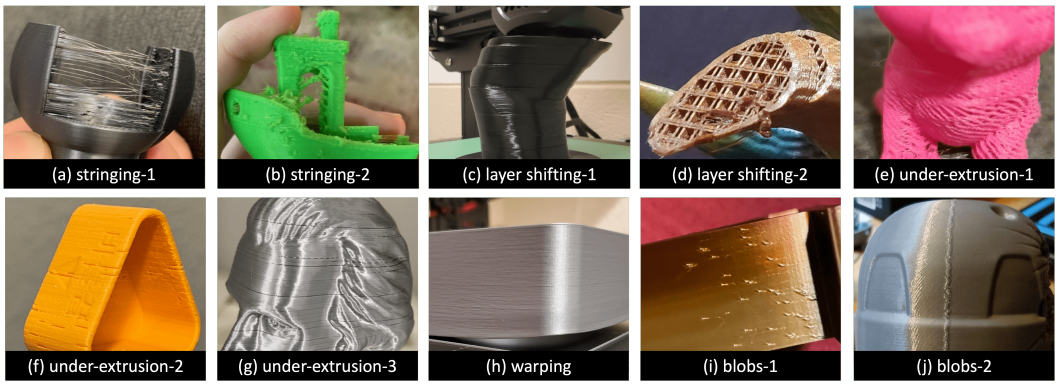


Fig. 6. Distinct visual features of five selected failure types: (a) stringing - fine strings, (b) stringing - branches, (c) layer shifting - shifting in a side view, (d) layer shifting - shifting in a top view, (e) under-extrusion - spongy surface, (f) under-extrusion - empty gaps on the surface, (g) under-extrusion - missing layers, (h) warping - bending at the bottom of the model, (i) blobs - sporadic zits, (j) blobs - z seam.

features as a starter to build an initial image dataset for a system demonstration. Since this work is not a deployment study, we chose to explore a small set of failure types first and then plan to expand the coverage by investigating more general ways to reduce manual labor. Increasing diversity of printer machines and advanced materials would alter the dimensions and complexity of un/known printing failures being discussed in the community in the future, we are to present a pipeline that further facilitates human-AI collaboration to quickly adapt to new trends in 3D printing. We detail this process in Section 6.3. Upon agreement between two researchers, including one 3D printing expert, we combined the similarity score results and validated the results with the high similarity using the failure-specific keywords (e.g., stringing - string, warping - warp). This approach achieved 73% accuracy from 100 randomly sampled posts from the whole dataset for the selected five types, indicating a decent performance of the keyword-based classification. .

4.4.3 Human Expert Validation on Image Data & Technical Evaluation of Classification Models. After the automatic classification using images, a human expert validated the classification results, defining the labels. In this process, we empirically found that there could be multiple visual characteristics in each failure type that we can categorize, as some examples in Figure 6 show. This finding was used to inform users of the common visual characteristics (as in Figure 3).

The human expert validation finalized the training dataset for stringing (N=439), layer shifting (N=631), under-extrusion (N=283), warping (N=199), and blobs (N=226), respectively. As real-world images in our dataset in Figure 7 shows, some prints might have visual characteristics of multiple failures at the same time. For example, a typical test 3D model for calibration in Figure 7 (a) has visual features of stringing (fine strings, branch-like structure) along with a spongy pattern on

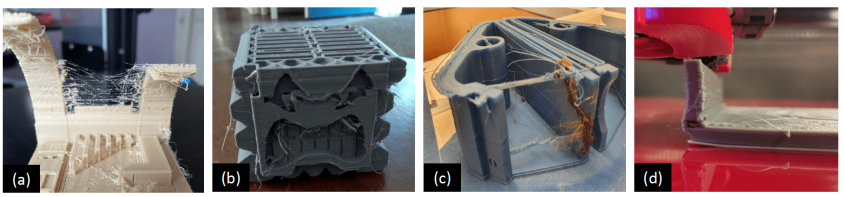


Fig. 7. Example images of prints that have multiple failures from the FixMyPrint dataset: (a) stringing and under-extrusion, (b) layer shifting and stringing, (c) layer shifting and stringing, and (d) layer shifting and warping

Failure type	Training set		Validation set		Test set	
	ACC	AUC	ACC	AUC	ACC	AUC
Stringing	97.15	99.66	88.51	92.59	81.89	96.13
Under-extrusion	95.55	99.18	82.14	87.89	85.88	93.55
Layer shifting	94.19	98.27	77.78	81.69	73.61	81.10
Warping	94.12	98.51	87.18	90.21	80.99	89.37
Blobs	95.57	99.26	91.11	95.17	87.50	92.03

Table 1. ACC (Accuracy) and AUC (Area under the ROC Curve) of binary classification models for 5 types of printing failure within 20 epochs with a batch size of 32. All models were trained using ResNet18, Adam optimizer, and 0.0001 as a learning rate. The models were initialized with the pretrained weights, and we fine-tuned entire layers.

the surface, which implies under-extrusion. On the other hand, Figure 7 (b) and (c) are showing stringing and layer shifting, but with different visual features of it. To capture these multiple classes, we trained a binary classification model (ResNet18 [29]) that one model can determine whether the input photo contains one type of failure. In this manner, one model is responsible for extracting the distinctive features of one failure type at a time. To avoid confusion during training, we also set the negative class samples for the training set that do not present any distinct visual features.

Since we carefully chose the training examples that present distinct visual cues for failures, as shown in Figure 6, our current dataset has a relatively small number of samples. We still chose a CNN-based approach to automate the classification over more classic algorithms (e.g., SIFT) or traditional ML approaches (e.g., XGBoost, random forest). This is because our dataset is scalable, as more cases can be discovered by communities. Also, even the same failure type can show different visual features. We believe that using CNN-based approaches can effectively address not only the large dataset but also automatically extract visual features without the hassle of defining manual features as the dataset grows. We chose ResNet18 [29] model for training and used Adam optimizer [39] with a 0.0001 learning rate. We initialized weights with the pretrained ResNet18 model and fine-tuned all layers during training. All five binary classification models converged within 20 epochs with a batch size of 32. Four models for stringing, under-extrusion, warping, and blobs showed an 80+% accuracy for the test set, while the layer shifting model showed slightly less test accuracy of 73.61% as summarized in Table 1. All five models showed a fairly good Area Under the Curve (AUC, [42]) with at least 80% AUC for the test set, implying that our models perform well in detecting one failure type overall.

4.4.4 Constructing the List of Suggestions by Extracting Key Conversations. By combining the automated text-based classification and the image-based human expert validation, we finally got the post-dataset annotated with the failure types. Aiming to extract various quality suggestions from labeled community posts per each failure type, this approach ensures the freedom to choose from multiple viable solutions. As there could be multiple causes and suggestions for a single problem, it is critical to cover a wide spectrum of failure cases accumulated in the community. To create a thorough list of suggestions, we first retrieved all comments that belong to each failure type and sorted comments by the number of upvotes (positive feedback), as it can reflect their quality and how helpful the contained suggestions are. One researcher went through the comments sorted in descending order of the upvote count and extracted suggestions. We extracted *clues* from the suggestions (See Figure 5), which can specify the causes of failures. We then used an online archive [62] to see if the suggestions were covered (*common solutions*). If not, we defined those suggestions as *specific* solutions. As described in Figure 5, this separation between common & specific solutions was used to guide users to the reasonable flow of troubleshooting, as common solutions can resolve the majority of issues.

5 S2: SUMMATIVE STUDY

We evaluate 3DPFIX through two steps. First, we conducted an experimental study that lets participants complete the predefined tasks in a controlled lab setting. We aimed to understand how several sub-components of 3DPFIX driven by our design requirements can individually work in improving remote novices' current practice in their troubleshooting process. To compensate for the artificial settings of the lab study and obtain more real-world feedback in natural and situated settings, we deployed 3DPFIX for 3 weeks and encouraged active 3D printing users to freely use 3DPFIX with the photos of their own failures that they have experienced so far.

5.1 Methodology

5.1.1 Study Structure. We first conducted an experimental study. Notable heuristic metrics (e.g., NASA TLX [27] or System Usability Scale [4]) have been developed and widely used to measure the general usability of an interactive system as a whole. However, complex interactive systems have many components (features) that support different objectives for each, thus, adopting such heuristic methods can hamper researchers from understanding how each sub-component helps users achieve a certain goal in detail [33]. Similarly, 3D printing troubleshooting can be done by collectively comprehending various elements, such as recognizing visual characteristics of a failed print, learning what can cause the specific problem, and finding the most applicable solutions to try out; 3DPFIX consists of several features to support users to achieve those elements. We aim to measure how 3DPFIX's components can improve users' experience in achieving the following objectives: (1) diagnosing/understanding failure types, (2) finding the applicable solutions, and (3) learning 3D printing knowledge and technical terms. We extracted these 3 main metrics by reviewing Besides, HCI researchers have gradually accommodated applying the experimental study as a method to evaluate an interactive system (e.g., [77]) under the two premises: (1) setting a baseline condition that can naturally capture a user's current practice, and (2) defining a task that is practical, specific, while having a clear "ground truths" to scientifically measure human users' task efficiency and task effectiveness. Thus, we chose to do an experimental study for evaluation and followed this comparative structure to better reveal *how each component of the system contributes to letting users fulfill the core sub-goals, which will lead to successful troubleshooting.* In our S2 setting, we define the baseline condition as participants' "best practice" using any online resources participants can use, following the most natural condition they can be under. To prevent any bias, we did not specify the resources they could use in the study, but we allowed them to use any methods, mentioning some examples such as Google search, online articles, and online communities as examples. In our dry-run, we found out that given a specific online resource, subjects' behavior was highly bound to the certain resource in completing the tasks. We saw this as artificial and leading to biased action. For the experimental condition, participants only used 3DPFIX to do the same task without any other online resources. We chose a within-subject study where participants were asked to perform troubleshooting using their familiar online resources (baseline condition: **C1**, hereinafter) and using 3DPFIX (experimental condition: **C2**, hereinafter) for a direct comparison [41].

5.1.2 Participants. For our experimental study, in order to ensure the reliability of the study, we first decided to recruit at least 12 participants, which is the most common sample size used by the CHI community [7]. For recruitment, we used convenience and snowball sampling strategies [12]. We first approached our acquaintances who were (1) interested in 3D printing or (2) working on a maker space in one of the two major research universities in the United States. We invited these contacts to participate in our study if they are interested in learning 3D printing but have no advanced knowledge about the domain. We also asked them to suggest potential participants

Fig. 8. The flow of S2 experimental study.

who might fit into our inclusion criteria. Participation was voluntary, and no compensation was provided. The IRB has been approved by the researchers' institutional boards at the moment we send an invitation. In total, 13 participants met our recruiting criteria (Male=5, Female=8). Their age range between 18 and 34. In the demographic survey, P8 and P10 self-evaluated that they have some extent of 3D printing knowledge and have experience in fixing troubles while the rest (N=11) assessed themselves as novices—just started learning or have limited experience in troubleshooting.

5.1.3 Study Procedure. Upon agreeing to participate, all participants were asked to finish the demographic survey. After the demographic survey, participants received a link to a remote, synchronous study link. The lead author (the facilitator, hereinafter) met each participant one-on-one on Zoom to run the sessions and communicate with the last author after finishing each session. At the beginning of the study, the facilitator asked for the e-signing of the consent form. After consenting, the study began with the study onboarding session containing an introduction, the purpose of the study, and the study steps. After finishing the study onboarding, we split participants into two groups for counterbalancing the ordering effect of two conditions. We chose this within-subject design to capture participants' perspectives about the direct comparison between C1 and C2. The first group of 7 participants used C1, baseline first (see the upper row in Figure 8) while the second group of 6 participants started from C2, experimental condition (see the lower row in Figure 8). In C1, participants were encouraged to use any preferred troubleshooting strategies, including reading online archives, watching tutorial videos, and searching posts or posting questions to online 3D printing communities. For C2, participants were only allowed to use 3DPFIX; other online resources were prohibited.

To run the within-subject study, we prepared two problem sets. Set #1 had examples of under-extrusion and layer shifting (PS1, hereinafter), and set #2 had blobs and stringing (PS2, hereinafter). To make their troubleshooting tasks realistic, we retrieved the four examples from the Ender3 [65] and 3DPrinting [64] subreddits with experts' consult in terms of (1) how likely the problem might occur and (2) the complexity of the problem to solve it to control the level of difficulties between PS1 and PS2[1]. In distributing PS1 and PS2 to each condition, participants always used P1 first (i.e., Phase 1 in Figure 8) and P2 next (i.e., Phase 2 in Figure 8), regardless of the order of conditions. Thus, we methodologically decoupled the problem sets and conditions.

Upon finishing each condition, participants completed the survey designed to evaluate their perceived and behavioral task efficiency, task effectiveness, and learning. Detailed questions are explained in the later subsection. In the first part of *task performance* survey, we collected their behavioral data and attitudinal perception about the condition they finished. In the later part of the *learnability performance* survey, we evaluated how each condition facilitated their learning of given failure types and helped them gain knowledge about 3D printing in general. For instance, we showed two different images and asked participants to identify the failure types. We note that the learnability survey was only asked in their first trial because participants can gain an understanding of four different types of failure in their first trial, already affecting their performance on the following condition, depending on how much they are engaged in using and learning. For

[1]The images in the problem sets are not included in our dataset used to train classification models.

	Task efficiency		Task effectiveness		Learnability
	Failure identification	Solution-seeking	Failure identification	Solution-seeking	
Behavioral	Q1	Q2	Q1	Q2	Q8
Attitudinal	Q3	Q5	Q4	Q6	Q7a–Q7d

Table 2. Type of measures in S2 and mapping of corresponding questions in our survey design.

those who already exhaustively browsed every possible failure type in the first condition, the knowledge might be used to answer the learnability survey in their second condition.

After finishing both conditions, all participants are invited to a semi-structured interview. The interviews went for 27 minutes on average. In the interviews, we focused on capturing the aspects of: (1) what are the notable benefits and drawbacks of using 3DPFIX in 3D printing troubleshooting and how the features in the system are related, (2) personal experience based on a direct comparison between 3DPFIX and their current practice, (3) possible improvements of 3DPFIX and how new design can improve people's 3D printing troubleshooting experience in general. 3DPFIX was deployed as a web application (Available at [Anonymized for the review]) that participants can access using any modern web browser. Participants were asked to share the screen during the study while the whole process was recorded for further analysis.

5.1.4 Measures. We measured participants' behavioral & attitudinal *task efficiency* (i.e., how using a system helped them cut down their effort), *task effectiveness* (i.e., how using a system helped them to yield a quality outcome), and *learnability* (i.e., how did using a system to help extend their internal representation of the 3D printing domain). In measuring the three directions, we designed the survey to capture a user's behavioral performance and attitudes towards C1 or C2. To measure task efficiency and effectiveness, we broke down the measure into (1) failure identification: detection of an accurate type of 3D printing failure, and (2) solution-seeking: the derivation of a reasonable action plan that can result in desirable outcomes. We did not ask participants to measure the time to complete each task, not to increase their workload. Instead, the first author reviewed the video recordings of the study and calculated the time taken to measure behavioral task efficiency for failure identification and solution-seeking. Table 2 shows how our measures are mapped with the specific questions we implemented through our survey.

The specific survey questions we implemented are as follows:

- **Q1**. What was the type of failure you identified? Please explain why.
- **Q2**. After your search, what would you do to fix the type of failure? Explain why you thought your plan can work.
- **Q3**. I was able to identify **the type of failure** with less effort and time.
- **Q4** I was able to identify **the type of failure** accurately using the tool(s).
- **Q5**. I was able to find **the solution** with less effort and time.
- **Q6**. I was able to identify **the solution** that would accurately work for this type of failure.
- **Q7a**. I was able to learn about **the type of failures** that the given images show in this condition.
- **Q7b**. I was able to learn the **3D printing-related knowledge** about this 3D printing failure.
- **Q7c**. I was able to learn about **useful suggestions and tips** that could resolve failures.
- **Q7d**. I was able to learn about **logical flow of troubleshooting** (problem identification - finding a clue/situation indicates the issue - solution-seeking) in 3D printing.
- **Q8**. [Attach an unseen image] What is the type of failure this image shows?
- **Q9**. Are there any other thoughts that you would like to share about your experiences?

In defining questionnaires above, we referred to standard usability-based questionnaires (e.g., NASA TLX [27], SUS [4]) and similar prior works (e.g., [8, 76]) that evaluated their interactive systems through an experimental study. Based on the prior work, we adapted the questions to

our three major metrics, mostly focusing on behavioral/attitudinal efficiency (time taken) and effectiveness (accuracy) for each metric. Additionally, for learnability, we added major obstacles identified (e.g., problem articulation, finding applicable solutions, difficult technical terms) to be components to measure. In implementing our survey, we made Q1, Q2, Q8, and Q9 open-ended short answers while the rest rated on a 1-7 Likert scale, ranging from strongly disagree to strongly agree. To evaluate the quality of participants' answers in Q1 and Q2, we recruited one 3D printing expert who has been using 3D printers for 4.5 years and has extensive knowledge in 3D printing and experience in troubleshooting. All participants' answers and condition information were de-identified when provided to the 3D printing expert for analysis. Given the participants' responses, including failure identification and the solution they chose to try, the expert rated participants' failure identification with True/False and the solution on a 1-7 Likert scale in terms of quality. In analyzing Q1 and Q2's behavioral effectiveness, we used Pearson's Chi-squared test for null hypothesis analysis by calculating the frequency of True and False labels under two conditions. For all ordinal responses in Q1 and Q2's behavioral efficiency and the 1-7 Likert Scale (Q3-7), we used the Kruskal-Wallis test. We used Python packages such as Scipy and statsmodels for all statistical analyses. Also, the facilitator used an iterative coding method [57] to analyze qualitative data. Upon completing each interview, we first built codes to extract unique or meaningful aspects based on responses in the transcript and wrote an analysis down to combine insights from different interviews and synthesize them. After finishing all the interviews, we finally did a consensus-based diagramming to structure our findings.

5.1.5 Deployment Procedure. The experimental study enabled us to understand how the several sub-components work together to help users troubleshoot the failures. To see how 3DPFIX can holistically work with users in more realistic, less-artificial settings, we deployed 3DPFIX for a short time to obtain feedback from active 3D printing users who had already experienced failures. We distributed a flyer through the online communities, including subreddits (e.g., FixMyPrint [66], 3DPrinting [64], Ender3v2 [65]), discord servers (e.g., Creality 3D [10]), and online forums (e.g., 3D Printing Space [19]), that have an active discussion about 3D printing matters, such as troubleshooting, materials, and printers. Users were encouraged to use the photos of the failed prints that they have experienced so far, without any verbal/functional restrictions of failure types, and submit voluntary feedback through Google Forms. Since it was fully voluntary, no compensation was provided to the users. In the survey, they were asked to answer four optional questions:

- How would you describe your overall experience using 3DPFIX? How do you think 3DPFIX could be helpful for novices?
- Do you find any difficulties or problems using 3DPFIX?
- Would you recommend 3DPFIX to other 3D printing novices? Why or why not?
- Do you have any additional comments to improve 3DPFIX?

From the deployment for 3 weeks, 3DPFIX obtained 30 upload attempts and written feedback from 6 3D printing users (U1-6).

5.2 Results

As shown in Figure 9 and Figure 10, 3DPFIX outperforms the baseline. As the distribution does not follow the normal distribution, we used a Kruskal-Wallis test to compare the two groups' perceived efficiency (Q3, 5), perceived effectiveness (Q4, 6), and behavioral learnability (Q8). The power analysis results show that our data collected from 13 participants contains enough power (> 80%) to correctly detect any significant difference between the baseline and experimental groups. The Cohen's d effect sizes are large enough so that a smaller sample size can still produce the desired

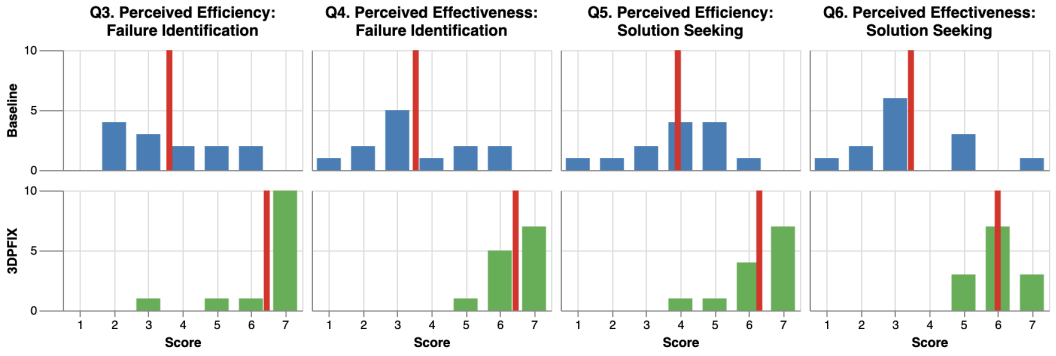


Fig. 9. Distribution of 13 participants' responses about their perception on a 7-point Likert scale, where 1 represents 'Strongly Disagree' and 7 is 'Strongly Agree', under two different conditions: baseline using online resources (top, blue-colored bar charts) and experimental using 3DPFIX (bottom, green-colored bar charts). A red line represents the mean value. For all 4 metrics on participants' perceived efficiency and effectiveness, using 3DPFIX outperforms the baseline.

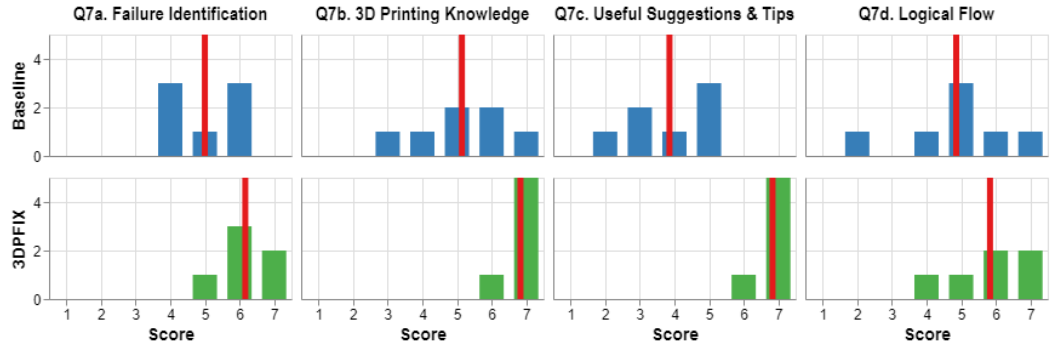


Fig. 10. Distribution of participants' perceived learnability responses in two conditions: baseline using online resources (top, blue-colored bar charts) and experimental using 3DPFIX (bottom, green-colored bar charts). When using 3DPFIX, all 4 metrics (on a 7-point Likert scale: 1 is 'Strongly Disagree' and 7 is 'Strongly Agree') showed increased performance compared to the baseline.

Category	Measure	N	p-value	Effect Size (Cohen's d)	Power at N
Failure Identification	Task efficiency (Q3)	13	0.000107	-2.09471	1
	Task effectiveness (Q4)	13	0.000148	-2.01261	1
Solution-seeking	Task efficiency (Q5)	13	5.86E-05	-2.43937	1
	Task effectiveness (Q6)	13	0.000321	-2.0381	1
Learnability	Behavioral learnability (Q8)	7	0.012475	-1.63979	1
	Failure identification (Q7a)	7	0.048668	-1.30191	0.99
	Technical knowledge (Q7b)	7	0.095791	-0.78662	0.74
	Troubleshooting tips (Q7c)	7	0.001879	-3.17091	1
	Logical flow (Q8d)	7	0.240433	-0.69518	0.634

Table 3. Power analysis results of measures in S2 with the corresponding sample size. In evaluating learnability measures, only the first condition (N = 7) was used due to a possible learning effect in the second condition.

amount of power for statistical testing. Table 3 shows power analysis results of all measures with 13 participants, including p-value, Cohen's d, and effect size.

5.2.1 ***Effective Problem Identification #1: Finding the Right Representation.*** 3DPFIX significantly reduced participants' effort in identifying and learning the accurate naming of failure types by simplifying the process of exploring the right representation (keyword seeking). Using 3DPFIX, their perceived efficiency on failure identification significantly improved than the baseline. Compared to baseline (M_{base} = 3.62, SD_{base} = 1.44), participants thought that they were able to identify the failure type with less effort and time using 3DPFIX (M_{exp} = 6.46, SD_{exp} = 1.15, $p < 0.005$*). In terms of behavioral efficiency, there was no significant difference in the time taken to identify the problem ($p > 0.05$), but there was about 50% improvement on average completion time using 3DPFIX (M_{base} = 287, M_{exp} = 253 seconds). As a major difficulty of identifying the problem and finding solutions in the baseline, many participants pointed out finding the right representation for search (P6, "Since I do not know the exact term, I started searching with the keyword, 3D printing problem"). The difficulty of representation and the need for automated diagnosis were also recognized by users. *"The system would especially be helpful to novices because they may not know the relevant keywords to search for issues, and they may not even be aware of issues that are in their prints in the first place"* (U3). During C1, we observed that all participants engaged in a loop for finding the proper representation, such as reading articles or other online archives, and picked the keywords from resources for another round of search (shift representation). Most participants started from a general representation, such as '3D printing failures', '3D printer problems', and '3D printing common issues', since they do not know the exact technical terms to search. Some participants started with more descriptive representation by describing the visual characteristics in their own words. For example, P4 and P7 chose *'3D printing dots, bubbles'* and *'Bumps over the layer'* to describe blobs. P7 described the under-extrusion failure with *'Outer shell is not sticking to inner mesh*'. However, it was not always easy for people to articulate the failure, especially if the terms are more technology-oriented *"For me, it was easy to search for stringing and shifting as the terms are more related to daily life—easy to describe with everyday words—, but for under-extrusion and blobs, it was not easy"* (P11). *"For some images, it is easier to describe, but for some situations, I don't know how to describe. In this case, it is hard to search because I don't know the exact term [to begin with text-query for search]"* (P8). Finding the right representation was still challenging for participants who have some printing experience (P10). P10 noted that she relied on her previous knowledge and experience. However, some failures are also new, so she still needed an online search to investigate the right technical term.

5.2.2 ***Effective Problem Identification #2: Understanding Visual Characteristics of Printing Failure.*** 3DPFIX facilitated participants' effective problem identification by supporting them in understanding the common visual characteristics. Using 3DPFIX, their perceived effectiveness on failure identification significantly improved than the baseline. Participants felt that the accuracy of failure identification significantly increased (M_{base} = 3.92, SD_{base} = 1.33, M_{exp} = 6.31, SD_{exp} = 0.91, $p < 0.005$*).Also, from the expert review of their answers, the behavioral effectiveness (Q1) (accuracy of the answers) using 3DPFIX significantly improved than the baseline ($p = 0.006$* by Pearson's Chi-square test).

No matter how descriptive their initial representation (search keywords) is, all participants used an image-oriented search strategy, searching similar images under the 'Images' tab from Google search results or looking at representative images in the online articles. P6 explained her approach, starting with the general representation and comparing the images to locate similarities. If it failed, she went through the image tab. However, 8 participants (P1, 3-6, 8, 11-12) noted their struggles due to 3D printing failures' multi-appearance feature. They found it difficult to obtain images with the

same or similar visual characteristics. Also, such processes demand a lot of time and effort. *"Images may not look the same as the ones in the online resources. It requires some knowledge to diagnose"* (P2). This uncertainty eventually lowered the confidence in their decision. P12 pointed out that she could not even find diverse images to compare through in her current practice.

On the other hand, 3DPFIX helped participants understand the visual features of printing failures and thus was granted confidence in the failure type detection from two key features: (1) grayscale saliency maps (See Figure 3) and (2) common visual characteristics of the failure (See Figure 2). Twelve participants (P1, 3-13) emphasized their preference for the grayscale saliency maps (1), which support them in understanding which part is problematic. Highlighting the part to focus on, 3DPFIX improved participants' confidence in the failure type that they identified by double-checking AI's decision on all failure types that 3DPFIX supports. *"3DPFIX gives possible failure types, so I can take a look at all of those and pick the most relevant one with [confidence], I can make my own decision based on AI's prediction"* (P3). From the deployment, U1 mentioned: "I really like that it shows examples of other possible failures. It helps a lot to compare and figure out the actual issue before troubleshooting. A lot of the time, I find myself troubleshooting for random possible issues without knowing the exact problem. This tool just really helps to [narrow down the choices and gives a great, clear explanation of each problem]". In addition, several participants mentioned that they relied on a grayscale saliency map in assessing the AI's trustworthiness. They mentioned having saliency maps helped them trust the model's results (N = 5, P1, 4, 6, 8, 12), as focus given to unreasonable areas can help them to understand the AI made an unexpected mistake (P8). P1 remarked: *"If mask—highlighted area—is not highlighting the print, the AI's decision is probably wrong. It is the way to double-check AI's decision"*. However, one user (U3) from the deployment reported that the saliency maps were slightly hard to understand at first, which indicates that there should be additional descriptions about how to perceive the grayscale area of maps for laypeople.

Representative images that inform common visual characteristics of each failure also helped participants (P3, 4, 6-8, 12) identify the problem by comparing those representative images to the given images. *"That images with common visual characteristics really helped me find which problem is showing in the image. In the baseline, I wanted to see other images with blobs, but it took too much time to see if other people also have the problem that looks like the same as mine"* (P4). Similarly, U1 also mentioned that *"even if the AI's guess is wrong, there are multiple other examples for reference"*. P6 particularly appreciated learning the specific terms to describe the common visual characteristics of the system (e.g., spongy pattern, gaps on the surface, and missing layers for under-extrusion in Figure 3), noting that *"Only the experienced users can know [how to call it]. Since 3DPFIX can inform the specific keyword for each visual characteristic, I can use this for the further search"*.

5.2.3 ***Effective Solution-seeking: Easy Guidance to Solutions in Depth***. 3DPFIX improved participants' ability to find quality solutions and to set their troubleshooting plan more easily, helping them acquire 3D printing knowledge. Regarding the expert review of participants' solution quality, 3DPFIX significantly outperformed baseline (M_{exp} = 5.53, SD_{exp} = 1.93, M_{base} = 3.92, SD_{base} = 2.32, p = 0.006*). In deciding what to do to resolve the failure that they identified (solution), the participants perceived that they were able to get solutions with significantly less effort and time (p < 0.005*) using 3DPFIX (M_{exp} = 6.46, SD_{exp} = 0.63) than the baseline (M_{base} = 3.54, SD_{base} = 1.50). Even though correct answers were not given at the end, they felt more confident (p < 0.005*) on the solutions they came up with using 3DPFIX (M_{exp} = 6.0, SD_{exp} = 0.68) than the baseline (M_{base} = 3.46, SD_{base} = 1.55). Their confidence in their solutions under the baseline condition was significantly lower than that of using 3DPFIX. Interestingly, there was no significant difference in behavioral efficiency between the two conditions (p > 0.05), and even participants spent more time finding solutions using 3DPFIX on average (M_{exp} = 500, M_{base} = 376 seconds). Based on our observation,

most participants showed more engagement when using 3DPFIX. For example, they wished to explore every feature of the system and were more enthusiastic about learning by carefully going through the list of solutions and linked articles/videos. Even though they spent more time using 3DPFIX than baseline, they perceived that they used significantly less time and effort in deciding solutions to try, which can indicate increased engagement with the less mental load we assume.

Participants were able to effectively access a variety of solutions with depth using the 3DPFIX. In particular, they mentioned that the following features of 3DPFIX helped them to browse solutions: (1) providing common and specific solutions using different tabs, (2) 'View Detail/View Video' enables participants to review further relevant information, (3) visualizing difficulty levels & clue filtering, and (4) providing easy description when hovering over technical jargon (See Figure 5). In the baseline, many participants suffered from too many online resources available, whereas those were vague or too general. They do not know where to start and which suggestion is common and viable to start with. Novice participants (N = 11) were bewildered by the overflow of online resources as they found that many of them discussed similar suggestions in a very general manner. From the deployment, U5 raised the same concern by reflecting that searching on the internet can be overwhelming. P3 complained about the resources she was able to find, mentioning that *"It tells me to increase the printing temperature, [but how]?"*. P8 was concerned about missing the most critical and useful information from online supporting communities, by scrolling and skimming through general but not quite reasonable solutions that she can trust. On the contrary, participants (P4-5, 7-8, 12) appreciated the flow of 3DPFIX, which leads users to read the common solutions first and more specific solutions if needed, by separating solutions with two different tabs, 'Common Solutions' and 'Common Solutions Did Not Work?'. *"People really get stressed when there is too much information at once, it is better to go step by step. For common solutions, I feel confident those are the things I was looking for"* (P3). Providing direct links to relevant web resources for more detail using the 'View Detail' and 'View Video' buttons also supported participants' (P4-5, 7-9, 12) a better understanding of the solutions in-depth. While the short description contained in each solution card (as shown in Figure 5) enables quick decision-making, the linked video further assists in understanding what to do: *"For that suggestion about retraction distance, I could not understand what it does. That linked online resource did give me an answer to my question, 'increase the distance, BUT what would it do?'"* (P9) To novices, having the difficulty level labels and clue filtering function was helpful (P1-2, 4, 6-8), as it also gives them guidance on where to start *"It gives where to start, I would start with the basic one"* (P2). *"I think I will start with the basic solutions and if I think that would work, I will check the clue. I will find the best one considering my situation"* (P6). Participants also liked the easy navigation of 3DPFIX. All participants mentioned that 3DPFIX was easy to understand and how to use, and the simple interaction and layout were user-friendly.

5.2.4 ***Improved Learnability: 3D Printing Knowledge & Technical Terms.*** As Figure 10 shows, 3DPFIX significantly improved participants' perceived learning abilities in three learnability metrics. Participants perceived that they were able to learn about the failure types given as the task materials were significantly better ($p < 0.05$). than the baseline (M_{base} = 5.0, SD_{base} = 0.93) when they use 3DPFIX (M_{exp} = 6.17, SD_{exp} = 0.69). Results proved that 3DPFIX can improve participants' experience in getting 3D printing knowledge about specific failures (M_{base} = 5.14, SD_{base} = 1.25, M_{exp} = 6.8, SD_{exp} = 0.37, $p < 0.05$). Perceived learnability about troubleshooting tips was notably improved using 3DPFIX (M_{exp} = 6.83, SD_{exp} = 0.37) than baseline (M_{base} = 3.86, SD_{base} = 1.12), which reflects that 3DPFIX can provide more diverse & exhaustive suggestions than the baseline ($p < 0.005^{*}$). In understanding the logical troubleshooting flow, the two groups' responses did not significantly vary ($p = 0.24$). Still, we see the mean value (M_{base} = 4.86, SD_{base} = 1.46) was slightly higher by 18% when using 3DPFIX (M_{exp} = 5.8, SD_{exp} = 1.07). In acquiring technical terms, 3DPFIX's

hover-over function for the technical terms (as shown in Figure 4) improved novice participants' ability to learn technical terms. *"If I am not using 3DPFIX, I need to do another search for that term and get back to the sources that I was looking at. It takes twice the effort. The hover over really saved my time and effort"* (P4). U1 also reported that the highlighted definition of common 3D printing terminology is specifically good for newcomers.

5.2.5 ***User Feedback & Limitations.*** Our deployment flyer informed that 3DPFIX is an ongoing project with plans for future expansion of its scope. Neither the flyer nor the system itself imposed any verbal or functional restrictions on users regarding the types of failure cases they could upload, ensuring a fair and natural setting. Despite the current scope of 3DPFIX, the majority of users (N = 5) from the deployment favored the current ability and potential of 3DPFIX for novices. They expressed their willingness to recommend 3DPFIX to novices as a quick and efficient means of diagnosing problems (U1-4), obtaining immediate feedback (U2, 5), and learning about common failures (U1, 5) through an easy-to-follow workflow (U4). U6, while more neutral, highlighted the need for 3DPFIX to expand its failure case knowledge base, which is one of the main focuses of our future work for a longitudinal deployment study. As potential areas for improvement, two participants (U2, 5) shared their experiences with false positive issues. They mentioned that 3DPFIX correctly identified the failure but also erroneously labeled a different type as 'Highly likely' failure. U2 suspected that the photo orientation might be a factor, while U5 attributed it to a cluttered background. Despite maintaining anonymity by not collecting personal identity information, we observed that at least one participant attempted multiple uploads. U5, for instance, reported conducting two separate trials using photos of the same print, each taken against a different background. In their initial attempt with a cluttered background, the system incorrectly identified the background as an 'under-extrusion' issue. However, in the second attempt with a clear background, U5 noted that the system accurately diagnosed the real issue as 'blobs' without any false positives in the background. Both U2 and U5 suggested including clear instructions for novices on how to capture photos to achieve optimal AI performance. We believe that enhancing the classification model's performance, along with providing clear instructions, will be a primary focus of our future work to make 3DPFIX more effective in real-world scenarios. Additionally, U3 pointed out that the solution tab is text-heavy and recommended incorporating visual aids to help users understand how specific methods can influence print results. Lastly, within the first three weeks of deployment, we received 6 written feedback responses out of 30 upload attempts, recognizing the possibility that some users may have made multiple attempts. We encouraged users to voluntarily submit separate Google Forms following their interaction with the system. We acknowledge that the process of providing individual written feedback might have been daunting for some users, potentially resulting in a lower number of feedback responses compared to the total number of upload attempts. In our longitudinal deployment study, we plan to implement a built-in feedback system that enables users to submit their feedback directly within the 3DPFIX interface, rather than redirecting them to Google Forms. Additionally, we intend to expand our reach to potential user pools by distributing the flyer to makerspaces, printing centers, and 3D printing workshops. This broader outreach will allow us to gather feedback from users with diverse levels of expertise, enhancing our understanding of 3DPFIX's performance and usability.

6 DISCUSSION & IMPLICATIONS FOR DESIGN

We offer high-level insights we learned through S1 and S2 in terms of the factors that system designers can consider when leveraging social annotation in building new applications for troubleshooting and a broader scope of applications for exploratory tasks.

6.1 Leveraging Social Annotation in Designing Troubleshooting Interfaces

In applying social annotation in the domain of 3D printing troubleshooting, our main considerations were (1) considering a user-perceived complexity of sub-tasks, i.e., among several small and large interactions a user makes in troubleshooting, which sub-task typically blocks them from making progress, and (2) Social annotation-AI transfer feasibility; i.e., would using social annotation can result in building an AI model that performs well. Literature in human-AI collaboration has shown that users would have a higher valuation of the AI-driven features if the task automated by AI is perceived cognitively challenging [20, 76]. However, the way to realize AI-driven design and its performance would be bounded by the quantity and quality of the data. When defining the scope of the AI-driven features in supporting troubleshooting, designers may consider the following four segments:

- **S1. Demanding sub-task, High-performing AI**: The segment of *Promise*. The AI-driven feature can help novices through cognitive offloading, thus the feature in this segment would likely be perceived as useful. Designers may put the highest priority on building the AI that belongs to this segment.
- **S2. Demanding sub-task, Low-performing AI**: The segment of *Challenge*. A designer can expect a highly positive user-side effect by implementing this feature. To realize this feature, designers may focus on diagnosing the factors that negatively affect performance.
- **S3. Easy sub-task, High-performing AI**: The segment of *Potential*. While it is possible to build a viable AI-driven feature that works in a user's task, the way to offer the feature should be carefully considered and designed. Even if the task itself is not taxing, this feature can be useful if novices conduct this type of task repetitively throughout. For instance, for users who frequently "translate" unknown terms in a complex article, having a high-performing domain-specific lexical simplification model can be useful.
- **S4. Easy sub-task, Low-performing AI**: The segment of *Unwanted*. There is no strong reason to implement the feature that falls into this segment, as user-side merit is unclear while the cost of developing the high-performing AI is expensive.

Considering the two factors could help a designer scope the AI's role in troubleshooting and reduce the risk of building an AI that would be perceived as not useful or not feasible. In our case, we found formulating a lexical query can be a learning block that hampers novices from building up their knowledge. Replacing the lexical query with an image-based approach was feasible due to the social annotation collected in online resources we targeted to use. In general, we think it's worth considering ranking the possible sub-tasks in terms of task complexity and feasibility considering available social annotation.

6.2 Improving the Way to Communicate in Future Troubleshooting Designs

Among the many learning blocks, the most evident block for the remote novices we identified in S1 was formulating a query using accurate domain terminology. Ultimately, the ultimate role of AIs in troubleshooting is to connect a novice's inquiry with relevant social annotation accumulated in online resources. In that sense, it is crucial to further develop how the AI can work better to elaborate on a user's situation and connect to the relevant solutions. A mode for representing one's representation to the system must be carefully designed based on the deep consideration of the nature of the target task. We introduce two new directions.

- **Vision-based Human-AI Loop**: Rapid progress in computer vision has enabled vision-based AI models to outperform humans in many tasks. However, when it comes to the way we interact with vision-based models, the way we interact is one-time use rather than iterative. We are at the early stage of designing a feedback loop between humans and vision-based models [22, 25, 68].

We expect that devising a new form of communication loop that can realize the elaboration and adjustment of a vision-based question can result in opening an intriguing design space in troubleshooting. As of early work, Attention Branch Network [21] and Convolutional Human-in-the-Loop [44] have been introduced in the field of computer vision. A recent study in CSCW started to apply such a feedback loop in HCI and CSCW domains [25].

- **Multi-modal Interaction between Human and AI**: Recent progress in Visual Question Answering (VQA) has developed powerful model architectures that combine images and texts [79]. While we expect such multi-modal-based communication designs can open a new way to interact with systems for a troubleshooting task and beyond, we identified relatively little research related to this direction in HCI.

6.3 Social Annotation for Scalable 3DPFIX Dataset

We initially built our 3DPFIX dataset for classifying the cases based on the 27 base documents defined by Simplify3D Print Quality Guide [62]. This raises an important question; are the failure types that the Simplify3D guide presented comprehensive enough to capture contemporary trends of printing failures? As observed in the actual practice, there likely emerges new failure types, terminologies, and visual features as printing technology evolves with new printers and advanced slicing algorithms resolve known major printing issues (e.g., Ultimaker Arachne). Therefore, renowned online archives may fail to provide inclusive support cases timely. To verify this assumption, we first examined the recent subset of community discourses by retrieving *1,000 recent images* from the FixMyPrint dataset, and manually annotated the images by checking visual cues and corresponding comments. The annotation was done by the first author who has extensive knowledge of 3D printing failures and verified by another author who is a 3D printing expert with over 10 years of experience. We discovered the 20 most recent common failure types reported by community members:

under-extrusion, stringing, blobs, [z-seam], layer shifting, warping,
overheating, vibration, bed leveling, too close/far print nozzle, over-extrusion, layer separation & [delamination], gaps between infill and outer wall, visible lines on the top, bridging, weak infill, [spaghetti, z-binding, pillowing, elephant foot]

Failure types covered by the initial 3DPFIX dataset are highlighted in blue. The initial dataset covers 6 types (having z-seam as a subclass of blobs) of types above, about 30% of failure types recently reported by the community. Aligning with our assumption, we found 4 new cases and 6 new terminologies (highlighted in red and [brackets], respectively), which are not a part of the Simplify3D guide. This gap between static guides and actual cases with the growing need for remote novices motivated us to build a pipeline, scalable and generalizable that extracts domain knowledge accumulated in online communities from social annotations and constructs a large dataset of images for automated detection and solution pool accumulated by community users. More importantly, the pipeline approach eventually reduces manual human labor needed for labeling data, making our system sustainable with new cases powered by human-AI collaboration in the future. 3DPFIX dataset pipeline (see Figure 11 in Appendix for a high-level view) consists of three steps:

(1) Visual feature dictionary: from the subset of online posts database, defining a dictionary of failure types and visual features using comments (i.e., social annotation)
(2) Failure image dataset: expanding the dataset by annotating posts by referencing the visual feature dictionary & comments
(3) Solution pool: extracting solutions from annotated posts

In addition to what is already shown in Figure 6, we present a failure type dictionary with distinct visual features, including all 20 types of printing failure spotted in Figure 12 in Appendix. We

envision that this can be used as a useful reference to expand the 3DPFIX dataset for new emerging 3D printing issues, for example, through human annotations on a minimal set of images along with state-of-art ML techniques, such as few-shot learning approaches (e.g., [36]) that effectively trains the model with a large unannotated dataset and a few annotated examples.

6.4 Longitudinal Deployment Study

In the S2 setting, we provided a specific error type to participants and asked them to solve the problem. To compensate for artificiality in the lab settings, we deployed the system for a short period of time and received comments from real users with their own failures. This enabled us to obtain feedback under less-artificial settings. As a future work, we plan to conduct a deployment study to improve remote novices' troubleshooting experience in the long run, to better understand 3DPFIX's ability to provide reliable & scalable failure diagnosis and solution suggestions powered by 3DPFIX pipeline, in realistic situations.

6.5 Alternative Design Choices & Multi-modal dataset

In this work, we were able to collect invaluable comments and suggestions to improve the design of 3DPFIX through S2. One user from S2's mini-deployment mentioned that having more visual aids in presenting solutions would help users better understand the concepts. Similarly, in future work, we will actively collect feedback from real users and keep improving the design of 3DPFIX. Also, we did not disclose the numerical representation (e.g., 90% probability) but only showed categorical information (e.g., highly likely) in providing AI's decision rationale, according to previous work's observation [75] that numerical values can confuse novices. Considering that AIs have become more prevalent and that numerical values can be beneficial for educational purposes, we will consider having the option to switch to numerical representations for our future deployment study. Currently, 3DPFIX considers visual information (image) and text (solution). For example, instead of having sound data in the dataset, the current design encourages users to consider a 'clue'. For example, if they heard a popping noise during printing, the filament likely needs to be dry. It is not only limited to visual and textual information, and we believe that building a multi-modal dataset, such as audio data, along with visual/text, is important. Collecting videos instead of images would be an effective way to collect sound data as well, which can provide additional evidence in diagnosing failures.

7 CONCLUSION

In this paper, we designed 3DPFIX to support remote novices' troubleshooting experience based on the design requirements identified through S1. In implementing 3DPFIX, we leveraged social annotation accumulated in online resources, hoping that such an approach can improve the way remove novices resolve issues more efficiently and effectively. Our S2 showed that participants' task efficiency, task effectiveness, and learnability-related performance are significantly higher when using the 3DPFIX than when relying on their current practice. We hope this work can motivate engendering a full-fledged system that can improve people's realistic troubleshooting practice in the 3D printing domain and beyond.

ACKNOWLEDGMENTS

This work is partially funded by National Science Foundation, IIS-2213842 and Future of Work Grant No.2026513.

REFERENCES

[1] 2009. Stack Overflow. https://stackoverflow.com

[2] Celena Alcock, Nathaniel Hudson, and Parmit K Chilana. 2016. Barriers to using, customizing, and Printing 3D designs on thingiverse. In *Proceedings of the 19th international conference on supporting group work*. 195–199.

[3] Mohammad Alodadi and Vandana P Janeja. 2015. Similarity in patient support forums using tf-idf and cosine similarity metrics. In *2015 International Conference on Healthcare Informatics*. IEEE, 521–522.

[4] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.

[5] Felix Baumann and Dieter Roller. 2016. Vision based error detection for 3D printing processes. In *MATEC web of conferences*, Vol. 59. EDP Sciences, 06003.

[6] Alexander Berman, Francis Quek, Robert Woodward, Osazuwa Okundaye, and Jeeeun Kim. 2020. "Anyone Can Print": Supporting Collaborations with 3D Printing Services to Empower Broader Participation in Personal Fabrication. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. 1–13.

[7] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 981–992.

[8] Minsuk Choi, Cheonbok Park, Soyoung Yang, Yonggyu Kim, Jaegul Choo, and Sungsoo Ray Hong. 2019. Aila: Attentive interactive labeling assistant for document classification through attention-based deep neural networks. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–12.

[9] John Joon Young Chung, Jean Y Song, Sindhu Kutty, Sungsoo Hong, Juho Kim, and Walter S Lasecki. 2019. Efficient elicitation approaches to estimate collective crowd answers. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[10] Shenzhen Creality 3D Technology Co. 2020. Creality 3D Printers Official Discord Server. https://discord.com/invite/Ay3sBqXAG7. Accessed: 2022-01-15.

[11] Print Everything Discord Community. 2020. Print Everything Discord Server. https://discord.com/invite/DzZmRms8j7. Accessed: 2022-01-15.

[12] John W Creswell and Cheryl N Poth. 2016. *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

[13] Ugandhar Delli and Shing Chang. 2018. Automated process monitoring in 3D printing using supervised machine learning. *Procedia Manufacturing* 26 (2018), 865–870.

[14] PFDM Dudek. 2013. FDM 3D printing technology in manufacturing composite elements. *Archives of metallurgy and materials* 58, 4 (2013), 1415–1418.

[15] Zeynep Erden, Georg Von Krogh, and Seonwoo Kim. 2012. Knowledge sharing in an online community of volunteers: the role of community munificence. *European Management Review* 9, 4 (2012), 213–227.

[16] Hugging Face. 2016. Hugging Face. https://huggingface.co/. Accessed: 2021-09-05.

[17] Hugging Face. 2020. Hugging Face Documentation: Pretrained models. https://huggingface.co/transformers/pretrained_models.html. Accessed: 2021-09-06.

[18] Flask. 2022. Web development, one drop at a time. https://flask.palletsprojects.com/en/2.0.x/. Accessed: 2022-01-16.

[19] 3D Printing Space Forum. 2023. https://3dprintingspace.com/. Accessed: 2023-07-13.

[20] Jonas Frich, Lindsay MacDonald Vermeulen, Christian Remy, Michael Mose Biskjaer, and Peter Dalsgaard. 2019. Mapping the landscape of creativity support tools in HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–18.

[21] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. 2019. Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10705–10714.

[22] Yuyang Gao, Siyi Gu, Junji Jiang, Sungsoo Ray Hong, Dazhou Yu, and Liang Zhao. 2022. Going Beyond XAI: A Systematic Survey for Explanation-Guided Learning. *arXiv preprint arXiv:2212.03954* (2022).

[23] Yuyang Gao, Tong Sun, Rishab Bhatt, Dazhou Yu, Sungsoo Hong, and Liang Zhao. 2021. Gnes: Learning to explain graph neural networks. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 131–140.

[24] Yuyang Gao, Tong Steven Sun, Guangji Bai, Siyi Gu, Sungsoo Ray Hong, and Zhao Liang. 2022. Res: A robust framework for guiding visual explanation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 432–442.

[25] Yuyang Gao, Tong Steven Sun, Liang Zhao, and Sungsoo Ray Hong. 2022. Aligning eyes between humans and deep neural network through interactive attention alignment. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–28.

[26] Betty Gray. 2005. Informal learning in an online community of practice. *International Journal of E-Learning & Distance Education/Revue internationale du e-learning et la formation à distance* 19, 1 (2005).

[27] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[28] Zachary Hay. 2021. Best 3D Printing Temperatures for PLA, TPU, ABS, & More. https://all3dp.com/2/the-best-printing-temperature-for-different-filaments/

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[30] Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1844–1850.

[31] Sungsoo Hong, Minhyang Suh, Nathalie Henry Riche, Jooyoung Lee, Juho Kim, and Mark Zachry. 2018. Collaborative dynamic queries: Supporting distributed small group decision-making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[32] Sungsoo Hong, Minhyang Suh, Tae Soo Kim, Irina Smoke, Sangwha Sien, Janet Ng, Mark Zachry, and Juho Kim. 2019. Design for Collaborative Information-Seeking: Understanding User Challenges and Deploying Collaborative Dynamic Queries. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[33] Sungsoo Ray Hong, Sonia Castelo, Vito D'Orazio, Christopher Benthune, Aecio Santos, Scott Langevin, David Jonker, Enrico Bertini, and Juliana Freire. 2020. Towards Evaluating Exploratory Model Building Process with AutoML Systems. *arXiv preprint arXiv:2009.00449* (2020).

[34] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–26.

[35] Sungsoo Ray Hong, Jorge Piazentin Ono, Juliana Freire, and Enrico Bertini. 2019. Disseminating Machine Learning to domain experts: Understanding challenges and opportunities in supporting a model building process. In *CHI 2019 Workshop, Emerging Perspectives in Human-Centered Machine Learning. ACM.*

[36] Shell Xu Hu, Da Li, Jan Stühmer, Minyoung Kim, and Timothy M Hospedales. 2022. Pushing the limits of simple pipelines for few-shot learning: External data and fine-tuning make a difference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9068–9077.

[37] Nathaniel Hudson, Celena Alcock, and Parmit K Chilana. 2016. Understanding newcomers to 3D printing: Motivations, workflows, and barriers of casual makers. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 384–396.

[38] MakerBot Industries. 2008. Thingiverse. https://www.thingiverse.com/. Accessed: 2022-01-15.

[39] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[40] Aniket Kittur, Andrew M Peters, Abdigani Diriye, and Michael Bove. 2014. Standing on the schemas of giants: socially augmented information foraging. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 999–1010.

[41] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction*. Morgan Kaufmann.

[42] Charles X Ling, Jin Huang, Harry Zhang, et al. 2003. AUC: a statistically consistent and more discriminating measure than accuracy. In *Ijcai*, Vol. 3. 519–524.

[43] MatterHackers. 2016. 3D Printer Troubleshooting Guide. https://www.matterhackers.com/articles/3d-printer-troubleshooting-guide. Accessed: 2021-08-26.

[44] Masahiro Mitsuhara, Hiroshi Fukui, Yusuke Sakashita, Takanori Ogata, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. 2019. Embedding human knowledge in deep neural network via attention map. *arXiv preprint arXiv:1905.03540* 5 (2019).

[45] Meredith Ringel Morris, Jarrod Lombardo, and Daniel Wigdor. 2010. WeSearch: supporting collaborative search and sensemaking on a tabletop display. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 401–410.

[46] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. 2010. What do people ask their social networks, and why? A survey study of status message Q&A behavior. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1739–1748.

[47] Lora Oehlberg, Wesley Willett, and Wendy E Mackay. 2015. Patterns of physical design remixing in online maker communities. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 639–648.

[48] Konstantinos Paraskevoudis, Panagiotis Karayannis, and Elias P Koumoulos. 2020. Real-time 3D printing remote defect detection (stringing) with computer vision and artificial intelligence. *Processes* 8, 11 (2020), 1464.

[49] Peter Pirolli. 2009. An elementary social information foraging model. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 605–614.

[50] Peter Pirolli and Stuart Card. 1995. Information foraging in information access environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 51–58.

[51] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA,

USA, 2–4.
[52] Alun Preece. 2018. Asking 'Why'in AI: Explainability of intelligent systems–perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management* 25, 2 (2018), 63–72.
[53] Thierry Rayna, Ludmila Striukova, and John Darlington. 2015. Co-creation and user innovation: The role of online 3D printing platforms. *Journal of Engineering and Technology Management* 37 (2015), 90–102.
[54] RepRap. 2021. Print Troubleshooting Pictorial Guide. https://reprap.org/wiki/Print_Troubleshooting_Pictorial_Guide. Accessed: 2021-08-26.
[55] Prusa Research. 2018. Prusa3D Discord Server. https://discord.com/invite/cjk3FuJ. Accessed: 2022-01-15.
[56] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 269–276.
[57] Johnny Saldaña. 2015. *The coding manual for qualitative researchers*. Sage.
[58] Aécio Santos, Sonia Castelo, Cristian Felix, Jorge Piazentin Ono, Bowen Yu, Sungsoo Ray Hong, Cláudio T Silva, Enrico Bertini, and Juliana Freire. 2019. Visus: An interactive system for automatic machine learning model building and curation. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*. 1–7.
[59] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.
[60] Ben Shneiderman. 2003. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*. Elsevier, 364–371.
[61] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).
[62] Simplify3D. 2021. Print Quality Troubleshooting Guide. https://www.simplify3d.com/support/print-quality-troubleshooting/. Accessed: 2021-08-26.
[63] 3D SOURCED. 2022. The Ultimate 3D Print Quality Troubleshooting Guide 2022. https://rigid.ink/pages/ultimate-troubleshooting-guide. Accessed: 2021-08-26.
[64] 3DPrinting Subreddit. 2010. https://www.reddit.com/r/3Dprinting/. Accessed: 2022-01-15.
[65] Ender3 Subreddit. 2018. https://www.reddit.com/r/ender3/. Accessed: 2022-01-15.
[66] FixMyPrint Subreddit. 2014. https://www.reddit.com/r/FixMyPrint/. Accessed: 2022-01-15.
[67] Prusa3D Subreddit. 2016. https://www.reddit.com/r/prusa3d/. Accessed: 2022-01-15.
[68] Tong Steven Sun, Yuyang Gao, Shubham Khaladkar, Sijia Liu, Liang Zhao, Young-Ho Kim, and Sungsoo Ray Hong. 2023. Designing a Direct Feedback Loop between Humans and Convolutional Neural Networks through Local Explanations. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–32.
[69] Anthony Tang, Melanie Tory, Barry Po, Petra Neumann, and Sheelagh Carpendale. 2006. Collaborative coupling over tabletop displays. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 1181–1190.
[70] James J Thomas. 2005. *Illuminating the path:[the research and development agenda for visual analytics]*. IEEE Computer Society.
[71] Tiffany Tseng and Mitchel Resnick. 2014. Product versus process: representing and appropriating DIY projects online. In *Proceedings of the 2014 conference on Designing interactive systems*. 425–428.
[72] Carlota V. 2019. Ai Build implements AI to detect and correct 3D printing errors in real time. https://www.3dnatives.com/en/ai-build-3d-printing-errors-110620195/ Accessed: 2021-09-06).
[73] Russell Wade, Nigel P Garland, and Gary Underwood. 2017. Challenges of 3d printing for home users. (2017).
[74] Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 38–45.
[75] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. 2020. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[76] Chuan Yan, John Joon Young Chung, Kiheon Yoon, Yotam Gingold, Eytan Adar, and Sungsoo Ray Hong. 2022. FlatMagic: Improving Flat Colorization through AI-driven Design for Digital Comic Professionals. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.
[77] Chuan Yan, John Joon Young Chung, Kiheon Yoon, Yotam Gingold, Eytan Adar, and Sungsoo Ray Hong. 2023. FlatMagic: Improving Flat Colorization through AI-driven Design for Digital Comic Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.
[78] Mark Zastrow. 2020. 3D printing gets bigger, faster and stronger. https://www.nature.com/articles/d41586-020-00271-6 Accessed: 2021-01-08.
[79] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6720–6731.

A APPENDIX

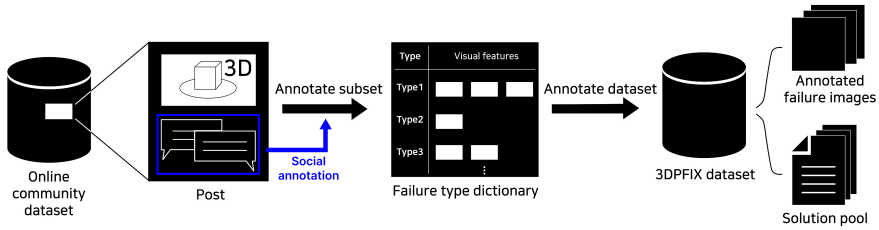


Fig. 11. 3DPFIX dataset building pipeline that leverages social annotation (comments).

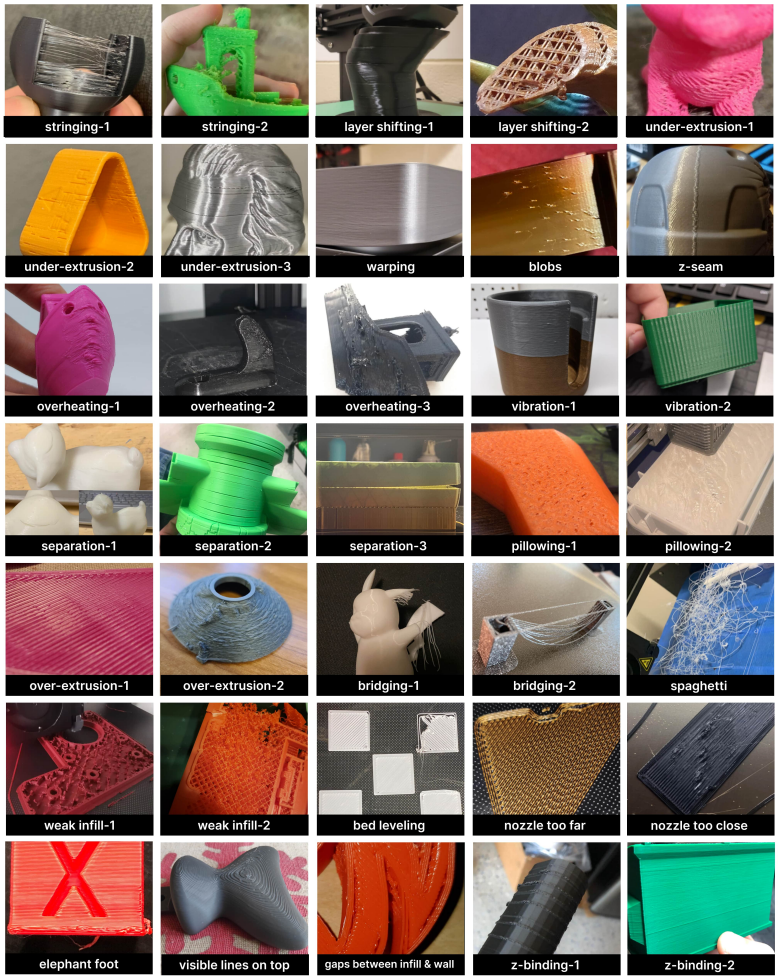


Fig. 12. Failure type dictionary for 20 failure types spotted from recent posts in the online community.

Received January 2023; revised July 2023; accepted November 2023