Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 55-2 (2022) 390–395

Knowledge Distillation for Energy Consumption Prediction in Additive Manufacturing

Yixin Li*, Fu Hu*, Michael Ryan*, Ray Wang, Ying Liu***

**Department of Mechanical Engineering, School of Engineering, Cardiff University Cardiff, CF24 3AA, UK*

(e-mail: liy248@cardiff.ac.uk, huf4@cardiff.ac.uk, ryanm6@cardiff.ac.uk, LiuY81@cardiff.ac.uk).

***Unicmicro (Guangzhou) Co., Ltd, China*

(e-mail: ray.wang@unicmicro.com).

Abstract: Owing to the advances of data sensing and collecting technologies, more production data of additive manufacturing (AM) systems is available and advanced data analytics techniques are increasingly employed for improving energy management. Current supervised learning-based analytical methods, however, typically require extracting and learning valuable information from a significant amount of data during training. It is difficult to make a trade-off between latency and computing resources to implement the analytical models. As such, this paper developed a method utilizing the knowledge distillation (KD) technique for predicting AM energy consumption based on product geometry information to reduce computational burdens while simultaneously retaining model performance. Through a teacher-student architecture, layer-by-layer images of products and energy consumption datasets are used to train a teacher model from which the knowledge is extracted and used to build a student model to predict energy consumption. A case study was conducted to demonstrate the feasibility and effectiveness of the proposed approach using real-world data from a selective laser sintering (SLS) system. Comparisons between distilled and independently trained student models were made in terms of the root mean square error (RMSE) and training time. The distilled student model performed better (14.3947KWh/kg) and required a shorter training time (34s) than the complex teacher model.

Keywords: *Additive manufacturing; Knowledge distillation; Energy consumption; Machine learning*

1. INTRODUCTION

AM brings a new manufacturing paradigm that creates physical components by layering the layers of materials according to digitized 3D patterns produced by computer-aided design tools. This free-form manufacturing technology increases the design potential of design and improves manufacturability (Achillas *et al.*, 2015; Peng *et al.*, 2018). It can manufacture parts with customized and highly complex shapes on a single 3D printer, thus realizing free-form manufacturing and overcoming the shortcomings of traditional manufacturing techniques. Due to a reduction in material and resource utilization as well as other tooling requirements, AM has shown a great potential for energy-saving and clean environmental production in the past decades (Majeed *et al.*, 2021).

The advancement of data sensing and collection technologies has facilitated the development of more data regarding AM processes. Advanced data analytics techniques have also become more common for improving energy management strategies. The current AM machines are often embedded with different sensing devices where there is a massive amount of data generated during AM processes, while the most essential thing is to extract and learn valuable energy-relevant information and knowledge from this. Then the decision-making process is further performed from the extracted knowledge of data (Zhou and Yang, 2016). Considering the trade-off between model compression and performance when modelling with fixed architecture, such as neural networks, it is necessary to minimize performance loss while conserving computational resources. Knowledge distillation (Hinton, Vinyals and Dean, 2015) or KD plays a key role in the development of the lightweight computational method that trains a shallow model to mimic the generalization ability of a deep model that is achieved by teacher-student architecture (Cheng *et al.*, 2018).

This study presents a KD-based approach to predicting energy consumption for the SLS system. The remainder of the paper summarizes the literature on energy consumption analytics in AM systems in section 2.1 and discusses impact factors and challenges. In section 2.2, the literature of KD is described in terms of knowledge categories and distillation schemes. A pIn section 3, a proposed KD-based approach is illustrated, and a case study based on an SLS system is presented in section 4. The results and discussion are presented in section 5. Section 6 concludes the paper.

2. LITERATURE REVIEW

2.1 Energy Consumption Analytics in AM

(a) Impact factors

10.1016/j.ifacol.2022.04.225

The energy consumption of AM systems is related to various attributes in terms of material supplies, working principles, and parameter settings. Meanwhile, an AM system normally has several subsystems that involve different subprocesses, leading to difficulties when analyzing its power consumption. Existing studies have explored to identify energy consumption-related factors and build corresponding models for different AM processes based on different strategies or methods. Tian et al took process parameters into account to build a power consumption model using linear regression for the fused filament fabrication (FFF) system while simultaneously considering the quality of the produced parts (Tian, Ma and Alizadeh, 2019). Other researchers also developed energy consumption models by mathematical methods (Yang *et al.*, 2017) for the Stereolithography (SLA) system and physical-based methods (Lv *et al.*, 2021) for selective laser melting (SLM) systems. With the advances of ML algorithms, data-driven methods have been increasingly adopted by researchers for AM energy consumption modelling targeted to an SLS system (Qin, Liu and Grosvenor, 2018) and an SLA system (Yang, He and Li, 2020), respectively in recent years. Through existing studies, besides processing and material attributes, the geometry-related attributes were found significant influences on energy consumption.

(b) Challenges on energy consumption predictive modelling

In the training phase, a larger volume of computing resources and more complex models are often demanded to extract information from the redundant datasets (LeCun, 2019). A large-size model or an ensemble model is often likely to demonstrate a better effect on different tasks. For example, AM systems generate production data, where those considerable amounts of data will lead to the extra computational burden to modelling, which often is hard to deploy in the small devices due to some bottlenecks including (1) low inference speed and (2) high requirement on deployment resources such as storage. As for the deployment phase, latency and computing resources are demanding (Gou *et al.*, 2021). Targeting the different models to extract valuable energy-relevant information or effective structure, complex models is required. As a result, model compression (Bucila, Caruana and Niculescu-Mizil, 2006), which involves lowering parameters while maintaining excellent performance, has become a difficult challenge, and knowledge distillation is one of the approaches for model compression.

2.2 Knowledge Distillation

The KD method was firstly generalized by Hinton (Hinton, Vinyals and Dean, 2015). Knowledge distillation or KD is a model compression technique in which a small model is trained to mimic a large model (or ensemble of models) that has already been pre-trained. Many models have recently attained state-of-the-art performance. However, given both memory usages and high latency, their parameters make it computationally expensive and inefficient. Therefore, KD is required to obtain a small model from a large model (Wang and Yoon, 2021).

The distilled knowledge includes logit-based, feature-based and relation-based knowledge. To date, several studies have investigated logit-based knowledge and it refers to the neural response of the last output of the teacher model. Despite its simplicity, logit-based KD is effective. In student modelling, the probability distribution output by the teacher model is equivalent to similarity information between categories, which provides additional supervision signals that make learning easier. The aim is to directly mimic the final prediction of the teacher (Gou et al., 2021). A soft target or soft label is typically used in logit-based knowledge distillation, referring to the feature map that the larger network outputs after each convolutional layer (Ba and Caruana, 2014; Hinton, Vinyals and Dean, 2015). When the soft label or target is introduced, additional information can be included in classification tasks, showing that the teacher model is capable of generalization.

Distillation techniques include offline, online, and self-distillation. With this scheme, the entire training process occurs offline in two stages: 1) the large teacher model is trained on several training samples before distillation; and 2) the pre-trained teacher model extracts information, such as logits or intermediate features, to guide the student model during distillation (Gou et al., 2021). The teacher model will inevitably require a significant amount of training, however, the offline training of student models can be highly efficient when guided by the teacher model. Online distillation complemented offline distillation by further enhancing the performance of the student model. Teachers and students are both updated simultaneously, as well as the entire knowledge distillation architecture. Such model distillation exploited collaborative learning, where resemble a cluster of student networks to teach each other during the training process (Zhang et al., 2018). In self-distillation, the teacher and student model use the same network architecture (Zhang et al., 2019).

3. METHODOLOGY

3.1 Model Compression using Knowledge Distillation

This study employs the logit-based and offline KD technique to demonstrate the feasibility and algorithm performance of energy consumption prediction in a specific AM system. By minimizing a loss function, knowledge is distilled from the teacher model to the student, to match soft teacher logits and ground-truth labels (Ba and Caruana, 2014).

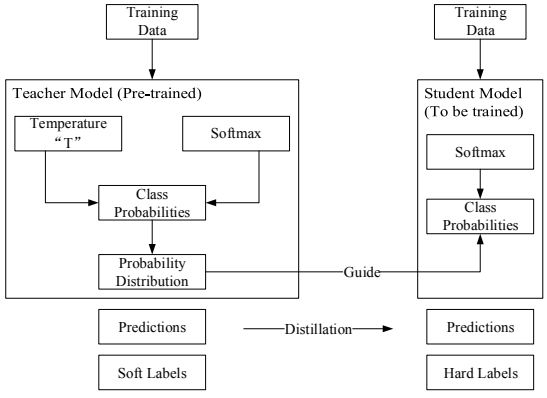

Figure 1 Knowledge distillation allows the soft labels to guide hard labels from teacher to student.

As shown in Figure 1, KD exploits the class possibilities of the large model as the soft target. Firstly, a model with excellent performance and generalization capacity, known as the teacher model, is trained. Second, all data is collected, and the predictions are generated using the pre-trained teacher model. The whole dataset, together with these predictions, produces knowledge, which is the soft labels. This process is referred to as knowledge distillation matching soft labels and hard labels. Finally, a smaller network is taught to serve as student models and to apply previously learned information.

The neural network uses a softmax layer to realize the conversion from logits to probabilities. Recalling the original softmax function in the classification model, and the logit generated in the last softmax layer ($\vec{z}$) in equation (1).

$$q_i = \frac{\exp(z_i)}{\sum_i \exp(z_j)} \quad where\ \vec{z} = \{z_1, z_2, \dots z_n\} \tag{1}$$

T is a hyperparameter called "temperature", and these values are called soft targets. However, utilizing the softmax layer's output value as a soft target has a drawback: when the entropy of the softmax output is relatively small, the values of negative labels are all near to zero, which means that their contribution to the loss function is minimal and negligible. According to Hinton (2015), a hyperparameter "temperature" T is useful and implemented in the softmax function, and these values After adding the variable T, the new softmax function is as follow:

$$q_i = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)} \quad q_i = \frac{1}{n}, where\ i \in \{1,2,3,\dots n\} \tag{2}$$

Equation (1) is a special form when T=1. When T rises, the output probability distribution tends to be smooth which will increase the information entropy of negative labels. If T becomes larger, the class probabilities will be softer, i.e. they become closer to each other, and in the extreme case, T tends to infinity. Therefore, to achieve this goal, the setup of T is greater than 1. This process is called distillation.

Simple models struggle to generalize successfully on a specific training dataset for the complicated problem, but the teacher model aims for prediction on all available data. The benefits are as follows: first, knowledge of the teacher model can teach a student how to generalize by predicting data outside of the training dataset. Second, soft targets provide more useful information than class labels, which indicates whether two classes are similar to each other.

3.2 Teacher-student Architecture by Regressing Logit with L2 Loss

The training environment for KD is frequently referred to as "teacher-student", with the large model serving as the teacher and the small model serving as the student. A simplified model is hard to solve the complex problem, not generalizing the training data. The knowledge of the teacher model can teach students how to generalize the model through available predictions other than training data. In addition, soft labels can provide more information than hard labels, which indicates the similarity between classes.

In the distillation, the objective function consists of distilling loss (corresponding to soft target) and student loss (corresponding to hard target). As shown in equations (3) and (4), a student similarity produces a softened class probability distribution P_j^T. The key of KD is the design of loss, which includes common cross-entropy L1 loss and L2 loss based on a soft target. By applying temperature T, L1 loss measures the gap between student model output and labels, using cross-entropy loss. L2 loss is the key distillation loss, which measures the difference between the output of the student model and the output of the trained teacher model after distilling (Ba and Caruana, 2014). α and T are hyperparameters, which are recommended to select 0.9 and the range of {3,4,5}, respectively (Hinton, Vinyals and Dean, 2015; Huang and Wang, 2017; Lan, Zhu and Gong, 2018; Cho and Hariharan, 2019).

$$Loss = \alpha L_1 + (1-\alpha)L_2 \tag{3}$$

$$where\ L_2 = -T^2 \sum_j^N P_j^T \log\left(q_j^T\right) \tag{4}$$

The training data is fed into both the teacher and student models, with the soft target being the softmax distribution created by the teacher model. The first part L1 of the loss function is the output of softmax of the student model and the cross-entropy of the soft target at the same temperature T. The second part of the function discusses how ground truth can reduce the risk of bias causing errors in the student model.

3.3 Energy Consumption Modelling based on Knowledge Distillation

Each overall energy usage in AM system is calculated. However, because the AM process is time-consuming, the energy consumption would inevitably rise as the manufacturing time increases. The overall energy usage is highly influenced by the duration of the procedure. Therefore, the unit energy consumption is utilised. In equation (5), the unit energy consumption E_U of each build, defines a target value for assessing the energy consumption level, where E_T is the total energy spent in the AM system and M_T is the total weight of manufactured products (Sreenivasan and Bourell, 2009; Qin, Liu and Grosvenor, 2017). Equation (6) details the computational method to assess energy consumption level, where e denotes subsystems consumed energy, and t is the time spent in each of subsystem.

$$E_U = \frac{E_T}{M_T} \tag{5}$$

$$E_T = \sum_e \left(\int_0^t E_e\right) \tag{6}$$

As shown in Figure 2, sensing devices are used to collect data from an SLS system including two types of data: layer-wise image data from each layer of the print, and energy output corresponding to each layer. After that, image data are

preprocessed since image size must be suitable for network input requirements. In the next stage, the predictive model using the KD approach is established, followed by the energy consumption prediction at the last stage.

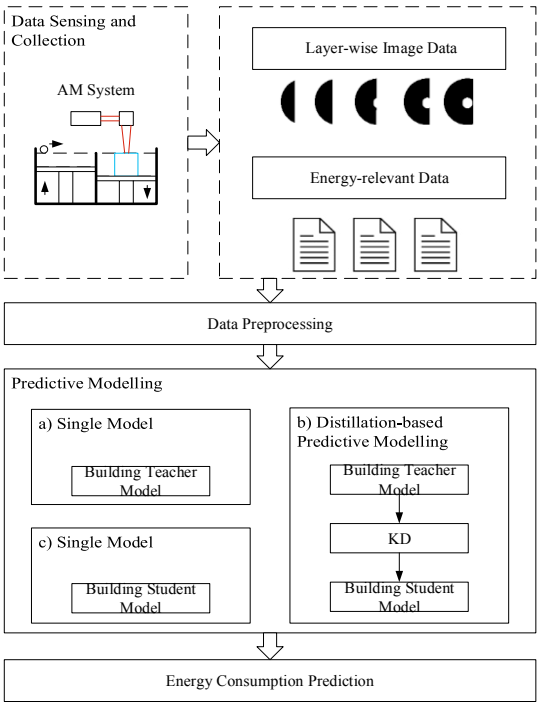

Figure 2 Demonstration of the proposed research framework for energy consumption predictive modelling via a) teacher model, b) distilled student model and, c) independently trained student model.

When the same batch of data is fed into teacher and student models at the same time, the predicted output of the teacher model is treated as a soft label and the real label as a hard label, and the student model's losses are calculated separately. Finally, the two losses are added in a weighted way to update the network parameters as the final loss (shown in Figure 3). Only the student model is utilized for predicting. This paper compares the student model trained via knowledge distillation with the student model trained independently (adding Softmax layer) to validate the results. The models are CNN architectures.

4. EXPERIMENTAL STUDY

4.1 Data Description

The data for the AM used in this case study consists of two types of data. One is related to the binary image data, which is collected from sliced files of 12 different builds layer by layer during manufacturing. Another is the unit energy consumption in the corresponding layer of the image. Each image is corresponding to a specific energy consumption value, and these values are reserved into a comma-separated value (CSV) file containing the unit energy consumption of each layer of the printed 3D model. The unit energy consumption is computed by equation (5) as shown in the previous section.

4.2 Model Setup

As can be seen in Figure 3, a **teacher model** is well-executed and generalized by training a large model. All the data is obtained, and the predictions of the teacher model are then calculated. The total dataset with these predictions is called knowledge and the prediction themselves are usually called soft labels. The previously acquired knowledge is used to train a smaller network called the **student model**.

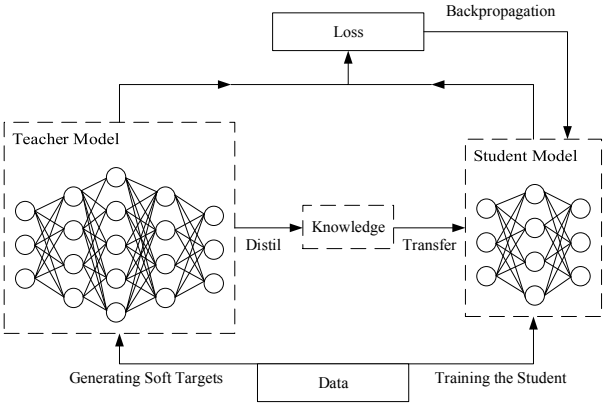

Figure 3 The architecture of the teacher and student model.

The complexity of a deep neural network comes from two dimensions: depth and width. It is often necessary to convert knowledge from deeper and wider neural networks to shallower and thinner ones. To achieve the distillation process of the computing model, the following parameters were used to construct the distiller, which includes a pre-trained teacher model, a student model to be trained, a student loss function, a distilling loss function (along with the hyperparameter "temperature" T) and α (=0.9) to weight the student and distilling loss. In the training step, the forward pass for both teacher and student models are carried out. According to equation (3), the loss is calculated by weighting the student loss and distillation loss with α and $(1-\alpha)$, respectively. The backpropagation is then performed. The image data and energy-relevant datasets are used to test on a student model. Presumably, the teacher model is trained and fixed before initializing the distiller, compiling the loss and hyperparameters, and distilling the teacher to the student model. In the training process, the model finally uses the softmax layer to calculate the loss value. After training, the soft label is generated when temperature T=3.

4.3 Performance Evaluation

Root mean squared error (RMSE) determines the accuracy of computing models, which is shown in equation (7), where a_t is the actual value and p_t is the predicted value.

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(p_t - a_t)^2} \tag{7}$$

5. RESULTS AND DISCUSSION

5.1 Prediction Accuracy Comparisons between KD-based Model and Prevailing Algorithms

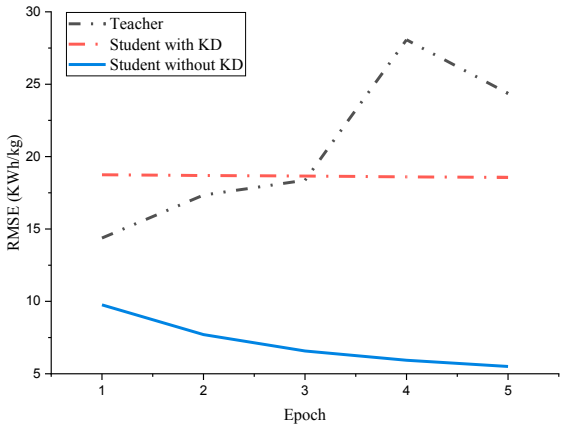

Figure 4 Comparison of RMSE between KD-based model and single models (at first 5 epoch).

The results obtained from three different model architectures can be compared in Figure 4. Distilled student model has the highest RMSE at the beginning and it shows a slight and constant decrease during training. In comparison with the distilled student model, the pre-trained teacher model (i.e. a complex and deep CNN architecture) occurred with successive increases in RMSE until epoch 4, where a sharp decrease is demonstrated. For the student model without the KD technique applied, i.e. a shallow and simplified CNN architecture, the RMSE declines at the first 5 epochs. Table 1 compares the experimental data on RMSE. The RMSE of the distilled model reported slightly less than others, with 14.3947 KWh/kg.

Table 1 The RMSE of Models at final epoch including teacher, distilled student, and independently trained student model.

Model Type	RMSE (KWh/kg)
Teacher Model	14.9211
Student Model (Distilled)	14.3947
Student Model (Independently trained)	14.4394

The pre-trained teacher model produces the highest loss than both student models, which means the changes of output weight in each epoch shows more changes, while the distilled student model indicates less change and therefore a good model performance. The loss of students with KD has nearly half of the loss of the teacher model at 5.5079, while the individual student model has the lowest loss value at 4.2342.

Table 2 The loss of models including teacher, distilled student, and independently trained student model.

Model Type	Loss
Teacher Model	10.1005
Student Model (Distilled)	5.5079
Student Model (Independently trained)	4.2342

5.2 Computing Times Comparisons between KD-based Model and Prevailing Algorithms

Table 3 Training time of different computing models.

Model Type	Average Training Time / Epoch (s)
Teacher Model	108
Student Model (Distilled)	34
Student Model (Independently trained)	6

The student model is trained to consume less time than the teacher model. However, it still costs nearly 6 times more than that of an independently trained student model.

5.3 Discussion

The lower RMSE shows the model selection and fitting are better, and the data prediction is more accurate. The experimental result illustrates the feasibility and effectiveness of the KD-based model. As can be observed, the distilled student model reduces the training time and remain the moderate algorithm performance during the training process. A possible explanation for this might be that the loss function of the distilled model will effectively compute the model's execution effect by comparing the model's predicted value with the actual value that should be output. If the loss is large, the value will be spread through the network in the training process. The loss will be quite big if the predicted output is distant from the output. If the two numbers are almost identical, however, the loss will be quite low. Several factors could explain these observations. Firstly, the knowledge of the teacher model can be transferred to the student model by KD, and the student model produced has a better generalization impact than the identical model taught through direct training. Secondly, the parameters of a model effectively define the amount of knowledge that can be extracted from the data. if the accuracy of the present network is not satisfactory, then the teacher model can be trained to distil the knowledge with higher accuracy.

The training aims to model the relationship between input and output on the existing dataset because it is impossible to collect all data as training data and new data are always generated continuously. As the training data set is a sample of the real data distribution, the optimal solution on the training data set often deviates from the real optimal solution to some extent (the model capacity is not considered in this discussion). While in the KD technique, a teacher network with strong generalization ability is built. A student network is trained to learn the generalization ability of the teacher network as the distillation starts.

6. CONCLUSIONS

KD is a model compression technique that entails training a small (student) model to imitate a larger (teacher) model that has already been trained, dealing with the lightweight computational model, and maintaining excellent model

performance. To cope with AM energy modelling difficulties, a KD-based computational model was used in this study. With an RMSE of 14.3947 KWh/kg, the results show that the KD-based technique for predicting the energy consumption of the SLS system is practical and effective. Meanwhile, the distilled model decreases training time by 34 seconds for a training period when compared to the complex teacher model.

In the current stage, the results are acceptable, but they can be further improved and optimized on the algorithm and model structure in the later work. The purposed of KD is to compress the model to make it more suitable for the deployment environment. Therefore, in the future, there is abundant room for further progress in optimizing the proposed architecture which is intended for being embedded in edge devices allowing specialized hardware to be equipped with efficient deep networks without sacrificing performance significantly.

REFERENCES

Achillas, C. *et al.* (2015) 'A methodological framework for the inclusion of modern additive manufacturing into the production portfolio of a focused factory', *Journal of Manufacturing Systems*, 37, pp. 328–339. doi: 10.1016/j.jmsy.2014.07.014.

Ba, L. J. and Caruana, R. (2014) 'Do deep nets really need to be deep?', *Advances in Neural Information Processing Systems*, 3(January), pp. 2654–2662.

Bucila, C., Caruana, R. and Niculescu-Mizil, A. (2006) 'Model Compression', *Kdd*, 54(1), pp. 1–9.

Cheng, J. *et al.* (2018) 'Recent advances in efficient computation of deep convolutional neural networks', *Frontiers of Information Technology and Electronic Engineering*, 19(1), pp. 64–77. doi: 10.1631/FITEE.1700789.

Cho, J. H. and Hariharan, B. (2019) 'On the Efficacy of Knowledge Distillation', *Microelectronics Reliability*, 13(6), p. 444. doi: 10.1016/0026-2714(74)90354-0.

Gou, J. *et al.* (2021) 'Knowledge Distillation: A Survey', *International Journal of Computer Vision*, 129(6), pp. 1789–1819. doi: 10.1007/s11263-021-01453-z.

Hinton, G., Vinyals, O. and Dean, J. (2015) 'Distilling the Knowledge in a Neural Network', pp. 1–9. Available at: http://arxiv.org/abs/1503.02531.

Huang, Z. and Wang, N. (2017) 'Like What You Like: Knowledge Distill via Neuron Selectivity Transfer'. Available at: http://arxiv.org/abs/1707.01219.

Lan, X., Zhu, X. and Gong, S. (2018) 'Knowledge distillation by on-the-fly native ensemble', *Advances in Neural Information Processing Systems*, 2018-Decem(Nips), pp. 7517–7527.

LeCun, Y. (2019) '1.1 Deep Learning Hardware: Past, Present, and Future', in *2019 IEEE International Solid-State Circuits Conference - (ISSCC)*. IEEE, pp. 12–19. doi: 10.1109/ISSCC.2019.8662396.

Lv, J. *et al.* (2021) 'A novel method to forecast energy consumption of selective laser melting processes', *International Journal of Production Research*, 59(8), pp. 2375–2391. doi: 10.1080/00207543.2020.1733126.

Majeed, A. *et al.* (2021) 'A big data-driven framework for sustainable and smart additive manufacturing', *Robotics and Computer-Integrated Manufacturing*, 67(March 2019), p. 102026. doi: 10.1016/j.rcim.2020.102026.

Peng, T. *et al.* (2018) 'Sustainability of additive manufacturing: An overview on its energy demand and environmental impact', *Additive Manufacturing*, 21(April), pp. 694–704. doi: 10.1016/j.addma.2018.04.022.

Qin, J., Liu, Y. and Grosvenor, R. (2017) 'A Framework of Energy Consumption Modelling for Additive Manufacturing Using Internet of Things', *Procedia CIRP*, 63, pp. 307–312. doi: https://doi.org/10.1016/j.procir.2017.02.036.

Qin, J., Liu, Y. and Grosvenor, R. (2018) 'Multi-source data analytics for AM energy consumption prediction', *Advanced Engineering Informatics*, 38, pp. 840–850. doi: 10.1016/j.aei.2018.10.008.

Sreenivasan, R. and Bourell, D. L. (2009) 'Sustainability study in Selective Laser Sintering - An energy perspective', *20th Annual International Solid Freeform Fabrication Symposium, SFF 2009*, pp. 257–265.

Tian, W., Ma, J. and Alizadeh, M. (2019) 'Energy consumption optimization with geometric accuracy consideration for fused filament fabrication processes', *International Journal of Advanced Manufacturing Technology*, 103(5–8), pp. 3223–3233. doi: 10.1007/s00170-019-03683-5.

Wang, L. and Yoon, K. J. (2021) 'Knowledge Distillation and Student-Teacher Learning for Visual Intelligence: A Review and New Outlooks', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(8), pp. 1–40. doi: 10.1109/TPAMI.2021.3055564.

Yang, Y. *et al.* (2017) 'Energy Consumption Modeling of Stereolithography-Based Additive Manufacturing Toward Environmental Sustainability', *Journal of Industrial Ecology*, 21(S1), pp. S168–S178. doi: https://doi.org/10.1111/jiec.12589.

Yang, Y., He, M. and Li, L. (2020) 'Power consumption estimation for mask image projection stereolithography additive manufacturing using machine learning based approach', *Journal of Cleaner Production*, 251, p. 119710. doi: 10.1016/j.jclepro.2019.119710.

Zhou, K. and Yang, S. (2016) 'Understanding household energy consumption behavior: The contribution of energy big data analytics', *Renewable and Sustainable Energy Reviews*, 56, pp. 810–819. doi: 10.1016/j.rser.2015.12.001.