ORCA – Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:https://orca.cardiff.ac.uk/id/eprint/155041/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Li, Yixin, Hu, Fu, Liu, Ying , Ryan, Michael and Wang, Ray 2023. A hybrid model compression approach via knowledge distillation for predicting energy consumption in additive manufacturing. International Journal of Production Research 61 (13) , pp. 4525-4547. 10.1080/00207543.2022.2160501

Publishers page: https://doi.org/10.1080/00207543.2022.2160501

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

A Hybrid Model Compression Approach via Knowledge Distillation for Predicting Energy Consumption in Additive Manufacturing

Yixin Li[a], Fu Hu[a], Ying Liu[a*], Michael Ryan[a], and Ray Wang[b]

[a]*Institute of Mechanical and Manufacturing Engineering, School of Engineering, Cardiff University, Cardiff, CF24 3AA, UK*

[b]*Unicmicro Co., Ltd, China*

[*]*Corresponding author. E-mail: LiuY81@Cardiff.ac.uk*

Abstract

Recently, additive manufacturing (AM) has received increased attention due to its high energy consumption. By extracting hidden information or highly representative features from energy-relevant data, knowledge distillation (KD) reduces predictive model complexity and computational load. By using almost predetermined and fixed models, the distillation process restricts students and teachers from learning from one model to another. To reduce computational costs while maintaining acceptable performance, a teacher assistant (TA) was added to the teacher-student architecture. Firstly, a teacher ensemble was combined with three baseline models to enhance accuracy. In the second step, a teacher ensemble (TA) was formed to bridge the capacity gap between the ensemble and the simplified model. As a result, the complexity of the student model was reduced. Using geometry-based features derived from layer-wise image data, a KD-based predictive model was developed to evaluate the feasibility and effectiveness of two independently trained student models. In comparison with independently trained student models, the performance of the proposed method has the lowest RMSE, MAE, and training time.

Keywords: Additive Manufacturing, Deep Learning, Neural Network Compression,

Knowledge Distillation, Energy Consumption Prediction

1. Introduction

Additive manufacturing (AM) is a typical example of smart manufacturing that exploits using new materials, changed the design and production paradigm, and led to the development of various applications (Kusiak 2018). In AM, products are fabricated by combining layers of materials according to 3D models (ISO/ASTM 2015). AM shows the merits of process flexibility and small-batch product customisation. The products are fabricated more efficiently than with traditional manufacturing (Watson and Taminger 2018; Machado, Winroth, and Ribeiro da Silva 2020). By bringing production closer to customers, 3D printing facilitates free-form product design and helps to make manufacturing more sustainable (Khorram Niaki and Nonino 2017). As a result of reducing tooling requirements and utilisation of materials and resources, AM has shown great promise for energy-saving and clean production over the past decades (Majeed et al. 2021).

Sustainable manufacturing boosts the long-term feasibility and production efficiency of a product cycle. Sustainability in manufacturing drives manufacturers to make better decisions in reducing material waste, and energy consumption and therefore improving their resource efficiency (Dunaway, Harstvedt, and Ma 2017). AM processes tend to have better material efficiency than conventional manufacturing while consuming more energy because of low productivity and other accessory requirements (Achillas et al. 2015). The challenge of tracking and optimising energy consumption in AM systems results from continuous data collection and sensing technology. Many current AM machines contain multiple sensing devices which generate an enormous amount of data during AM processes, and the main task is to extract and learn valuable energy-relevant information and knowledge from this data. Subsequent decision-making is based on the knowledge extracted from the data (Zhou and Yang 2016).

Therefore, conventional data analytics techniques might not be suitable in this complex scenario. In previous studies, process parameters have been extensively considered and analysed concerning AM energy consumption (Paul and Anand 2012a; Baumers et al. 2011; Peng et al. 2018; Z. Y. Liu et al. 2018; Baumers et al. 2013; Watson and Taminger 2018) while the impacts of part geometries are not considered in these studies. Previous studies have demonstrated that part geometries significantly influence the unit energy consumption of AM processes (Dunaway, Harstvedt, and Ma 2017; Qin, Liu, and Grosvenor 2018; Y. Yang, He, and Li 2020). It is common for CAD models to contain complex and varied geometry information. Some of these handcrafted geometries are difficult to describe. Methods used in previous studies for extracting geometric features for energy consumption prediction, such as statistically derived features (Verma and Rai 2013; Tian, Ma, and Alizadeh 2019; Y. Yang, He, and Li 2020) are capable of extracting general information about the geometry, while some detailed information regarding the internal structure is rarely considered.

Hand-crafted methods, however, make it difficult to extract representative features, especially for complex-shaped structures fabricated by AM processes. It has been shown that convolutional neural networks (CNNs) are capable of detecting hidden patterns in image data. As opposed to deep neural networks, CNNs can take image data directly as input, thereby eliminating the need for complex operations such as manual image pre-processing and feature extraction. Accordingly, this paper considered the nature of AM (layer by layer) and employed CNN to learn and extract representative features from highly complex geometries automatically. The KD-based approach was also used to optimise the energy consumption prediction, model. In addition to improving the accuracy of predictions, the inclusion of part geometries in unit energy consumption analysis provides valuable insights for AM designers to modify their designs and operators to determine the optimal part combinations for AM processes to save energy.

The use of deep learning-based approaches can provide predictive insights that are useful for identifying complex manufacturing patterns and allowing a system to make intelligent decisions throughout the production process. The complexity and parameters of deep neural networks limit their application to analytical models. Recently, many models have reached state-of-the-art performance levels. However, they are computationally inefficient and expensive due to their high latency and memory usage. A network layer tends to be deeper and more complex, requiring numerous parameters, and not suitable for real-time responses. The data collected from multiple sensors are heterogeneous with different types and dimensions. Therefore, a simple model will not accomplish tasks as effectively as a complex one. In addition, existing DL algorithms cannot handle the amount of storage across platforms. Hence, deep neural networks have limited storage, increasing their computational burden. Considering the trade-off between performance and model compression with fixed architectures, such as neural networks, is crucial when minimising performance loss. Teacher-student architectures (J. Cheng et al. 2018) can be used to optimise the design of neural networks utilising knowledge distillation (Hinton, Vinyals, and Dean 2015) or KD to accomplish the trade-off, which involves training a compact neural network with distilled knowledge of a large model. Several advantages can be gained from the methods of knowledge distillation, such as the ability to directly accelerate models without requiring special hardware or implementations. Developing KD-based approaches and exploring ways to improve their performance is still worthwhile.

Previous work (Li et al. 2022) exploited the KD under a typical teacher-student (T-S) scheme for energy consumption prediction in an AM system. Knowledge is extracted exclusively from the pre-trained teacher model, while the student model is not trained. Different model structures were not considered in the study, so the model capacity gap did not match. Based on historical AM energy-relevant data, this paper proposes a hybrid KD-based approach that exploits a redesigned T-S architecture to validate predictions of energy consumption in a specific AM system. In this paper, we present the following major contributions: (1) Combining traditional model compression techniques with KD in order to train an ensemble of teachers based on AM data. (2) To mitigate the information or feature loss from the AM image data input to the teacher ensemble, a new teacher assistant network is established at the intermediate stage of the T-S architecture based on a previously trained teacher ensemble. (3) In this paper, a KD-based model compression approach is introduced to compress the energy consumption predictive model (deep neural network) without reducing performance. As a result, the AM system can be managed and decisions made more efficiently.

The paper is divided into six sections, and the next section includes a description of current energy consumption analytics using computational methods in different AM systems, a description of traditional model compression techniques to reduce the computational burden at the algorithmic level, and a description of KD. The methodology for this study is described in section 3 using a hybrid KD-based technique, and section 4 describes the experiment setups and analysis. Section 5 presents the results and conclusions of the study. Section 6 summarises the paper and makes recommendations for future studies.

2. Related Work

As of now, research is being conducted to determine the relationship between part geometry and energy consumption, however, the focus of these studies has mainly been on part positioning and orientation. In this case, layer-wise geometry characteristics of the printed part have not been adequately considered in terms of how they may influence unit energy consumption (for example, the surface area of the printed part). In addition, these analytical approaches rely on manual feature extraction, which is not appropriate for fabricated components with complex geometries. Numerous studies have been conducted on monitoring and predicting energy consumption levels. The complexity and resource utilisation of prediction models at the algorithmic level, however, are not extensively discussed.

2.1. Energy Consumption Analytics in AM

The analysis of energy consumption is one of the most active areas of AM research. Analytical work can be performed at an early stage of the design and manufacturing process in order to produce eco-friendly products and processes. The paper (Peng et al. 2018) discussed the sustainability of AM from the life-cycle perspective, suggesting that research should focus on the AM process and system. Over the past few years, numerous studies have examined energy consumption analysis, focusing on the sources of energy consumption in specific AM systems. As a result of these changes, analysing and optimising energy consumption becomes more difficult. Several studies have demonstrated the correlation between energy consumption and operation characteristics, material attributes, the working environment, and design parameters. Over the past few years, extensive research has been conducted on the impact of various manufacturing processes on energy consumption in an additive manufacturing system. Due to the different control parameters required by different AM subprocesses, it has been difficult to analyse and optimise their power consumption. According to the study, there is a correlation between energy consumption and a variety of factors. Several factors contribute to this, including operation, design, material attributes, and work environment. As a result of evaluating the energy-relevant factors that influence energy consumption (summarised in Table 1), different analytical approaches were applied to the targeted AM system.

2.1.1. Review of impact factors on energy consumption in AM

According to Paul and Anand (Paul and Anand 2012b), the correlation between energy in selective laser sintering (SLS) and the total area of sintering is explained by two parameters, layer thickness and the orientation of the part. A slice's thickness was found to be inversely proportional to the energy required, while the component's orientation affected the input energy based on its geometry. A study conducted by Baumers et al. monitored energy consumption to provide a reliable classification of data and compared two different laser sintering platforms as part of the experiment. Energy consumption is influenced by both material and process properties, as well as geometry-related properties. In order to investigate the impact of impact factors on four different aspects of post-processing, this study examines four factors including the work environment, control parameters, part geometry, and machine settings that are associated with post-processing (Baumers et al. 2011). As part of the integrated process model, Tian et al (Tian, Ma, and Alizadeh 2019) considered the quality of manufactured parts in conjunction with the parameters that affect the power consumption of the fused filament fabrication (FFF) process, using linear regression analysis to determine those factors affecting the power consumption. As well as the mathematical method used by Yang et al. (Y. Yang et al. 2017a), other physical-based methods were developed by Lv et al. (Lv et al. 2021) for selective laser melting (SLM). Liu et al. reported that a better understanding of energy consumption has resulted in better performance and longer service lives for parts. An analysis of the energy consumption of EBM and SLM was conducted to identify the sources of energy consumption at the machine and process levels. It should be noted that AM subsystems such as laser scanning, building platforms, and gas circulation also consume a significant amount of energy at the machine level. At the process level, scanning speed, hatch spacing, and beam diameter can influence energy consumption significantly. Additionally, a material's microstructure may have an indirect impact on its energy consumption. In more studies, it has been confirmed that other process parameters such as scanning path and hatch space are closely correlated to energy consumption, while material powder absorption is also critical to AM energy consumption (Ma et al. 2018; Kellens et al. 2017). The recent advancements in machine learning (ML) algorithms have led to a growing number of researchers adopting data-driven methods for AM energy consumption modelling (Hu et al. 2021). Artificial neural networks (ANN) were used to model the unit energy consumed by the SLS system based on data and information collected from four sources, including material, geometry, working conditions, and process parameters (Qin, Liu, and Grosvenor 2017). The mask image projection SLA system's power consumption was estimated using multiple machine learning-based models based on the layer-wise geometry of the products (Y. Yang, He, and Li 2020).

Table 1 Energy consumption-related impact factors in AM systems in existing studies.

Existing Studies	AM System	Methodology	Impact Factors
Dunaway et al. (Dunaway, Harstvedt, and Ma 2017)	FDM	Statistical analysis	Part geometry.
Paul and Anand (Paul and Anand 2012b)	SLS	Mathematical analysis	Part orientation and slice thickness.
Hu et al.(Hu et al. 2021)	SLS	CNN-LSTM	Hatch speed, hatch space, hatch power, recoater speed, and the values of the dispense.
Liu et al. (Z. Y. Liu et al. 2018)	EBM/SLM	Mathematical analysis	Batch size, material types, laser scanning, building platform, gas circulation, scanning speed and hatch space.
Lv et al. (Lv et al. 2021)	SLM	Physical-based	Different machine subsystems, subprocesses, and working status.
Tian et al. (Tian, Ma, and Alizadeh 2019)	FDM	Linear regression	Process parameters (e.g., printing resolution, printing speed, nozzle temperature).
Qin et al. (Qin, Liu, and Grosvenor 2018)	SLS	ANN	material, geometry, working conditions, and process parameters.
Yang et al. (Y. Yang, He, and Li 2020)	SLA	Mathematical analysis	Part orientation, layer thickness, the curing time for stable layers, and curing time transition rate.
Ma et al. (Ma et al. 2018)	SLS	Mathematical and Genetic algorithm	Scanning path, and hatch space.

2.1.2. Challenges of energy consumption modelling in AM

Energy consumption is one of the most critical factors in determining the sustainability of AM system. A variety of data sources have contributed to the understanding of the hidden knowledge and dependence of attributes and energy consumption of AM systems. The energy consumption rate of different processing technologies varies due to a variety of factors. These factors are identified throughout the AM process. A typical AM process consists of six steps (convert, locate and orient, add support structure, slice, build, and post-process). As part of this process, various attributes, including those that relate to energy consumption, are digitalised and connected. Due to the complexity of the data types and the data collection, traditional analytical methods are difficult to adapt to the prediction of AM energy consumption. Using deep learning (DL) techniques, complex manufacturing patterns can be distinguished through the use of predictive insights. It is possible to implement a system that supports intelligent decisions throughout the prediction process by using this method (Rai et al. 2021).

Energy and material flows have been examined in various AM processes in studies on environmental sustainability. AM production planning and optimising can benefit from the results of these studies, e.g., layer thickness (Paul and Anand 2012a), hatch space and scanning path (Ma et al. 2018; Kellens et al. 2017; Lv et al. 2021), and selecting appropriate process parameters. These studies have not investigated the influence of part geometry on energy consumption. Besides, energy consumption is found to be significantly affected by the geometry characteristics of CAD models. A geometry design adjustment may also contribute to improving energy efficiency in AM processes by reducing energy consumption. This will enable the process operators to further enrich our understanding of energy and AM processes. Consequently, some studies investigated the relationship between part geometry and energy consumption (Baumers et al. 2011; Paul and Anand 2012a; Y. Yang, He, and Li 2020; Balogun, Kirkwood, and Mativenga 2014).

Currently, research has been conducted to investigate the relationship between part geometry and energy consumption, but the focus of these studies, for instance, has primarily been on part positioning and orientation (Baumers et al. 2011; Y. Yang et al. 2017b) where the layer-wise geometry characteristics of the printed part are not well considered. For example, how the surface area of the printed part may affect the unit energy consumption. Furthermore, the analytical approaches in previous studies rely on manual feature extraction, which is not suitable for fabricated components with complex geometries. Several studies have been conducted on the subject of monitoring and predicting energy consumption levels using different approaches. However, there is a limited discussion of the complexity and resource utilisation of prediction models at the algorithmic level. In this paper, the unit energy consumption (Wh/g) is applied to assess the energy consumption of the AM system. A deep learning approach can be used to extract key features and hidden patterns in image data. Based on the features extracted from the geometry perspective, the KD model developed in this study enables the prediction of unit power consumption of different geometry designs before AM process begins, allowing AM designers to adjust or even optimise geometry designs accordingly.

In AM subsystems, statistical approaches to extracting information from redundant and heterogeneous data are no longer suitable to deal with numerous datasets. It is often necessary to use more sophisticated models in order to accomplish this task. Currently, supervised learning-based analytical models are usually trained by extracting relevant information from the data, but unlabelled data are not effectively utilized. The network layers tend to be deeper and the structure is more complex, requiring numerous parameters that cannot be altered in real-time. The deployment of ensemble models on tiny devices can be challenging due to bottlenecks such as (1) slow inference speeds and (2) high demand for deployment resources. During the deployment phase, latency and computational resources are required (Gou et al. 2021). There is insufficient storage available on the various platforms due to the volume of data required by DL algorithms. As a result of their limited storage capacity, deep neural networks have difficulty processing large amounts of data due to their computational burden. When considering the trade-off between model compression and performance for fixed architectures, such as neural networks, it is essential to minimise performance loss while conserving computing resources. Due to this, a faster and more compact model can approximate a function learned by a slower, larger, but higher performing model (Bucila, Caruana, and Niculescu-Mizil 2006), which entails lowering the number of parameters while retaining outstanding performance, has become a difficult task, with knowledge distillation (KD) being one of the techniques.

2.2. Knowledge Distillation

2.2.1. Background

KD is a technique for compressing large models by training a small model to imitate a large model (or ensemble of models) that has previously been pre-trained (Ba and Caruana 2014). State-of-art performance of current algorithms has lately been achieved by many models. The model structure makes it computationally expensive and inefficient due to excessive latency and memory usage. Therefore, KD is required to generate a tiny model from a large one (Wang and Yoon 2021). As shown in Figure 1, the teacher-student architecture follows a "distillation" process: the cumbersome model obtained in the training phase is viewed as a teacher, whereas the distilled model is viewed as a student. Teachers, who act as students' assistants, have a good learning capacity and pass a supervisory signal (knowledge) to students who have a lower learning capacity, increasing the student model's generalisation ability (Hinton, Vinyals, and Dean 2015). Due to the lack of interpretability of DL, the term "knowledge" can be understood as a mapping from input vectors to output vectors. Class possibility output from the teacher model is used as labels for the data and sent to the student model to be trained. This class possibility is called a soft target.

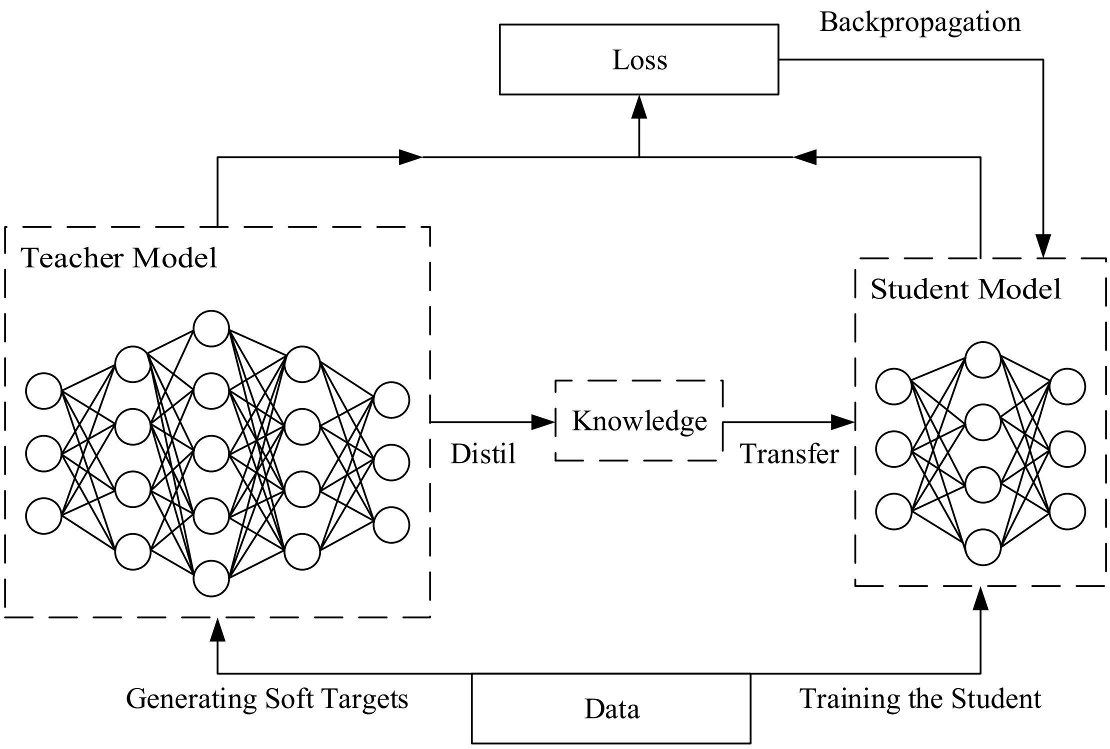

Figure 1. A classic teacher-student architecture with distillation.

Figure 1 Alt Text: There are three components in teacher-student architecture: a teacher network, a student network, and distillation. A knowledge distillation process involves transferring knowledge from a teacher to a student. The whole process is trained on a supervised data dataset.

As a transfer set with the soft target distribution is created using a complicated model with a high temperature in its softmax, a complicated model with a high temperature in its softmax is used to train a transfer set with the distilled model in its simplest form (Hinton, Vinyals, and Dean 2015).

$$q_i = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_i \exp\left(\frac{z_j}{T}\right)} \tag{1}$$

It differs from the original softmax by adding the hyperparameter T, matching the teacher and student model. When T of the student model is reset to 1, the loss function is the original softmax.

2.2.2. Logit-based knowledge distillation

Logit-based or response-based knowledge describes the neural response to the last output of the teacher model (shown in Figure 2). This method is primarily intended to mimic a teacher's final prediction (Gou et al. 2021). Due to its simplicity and efficacy, this kind of knowledge is widely used across different tasks and applications. A typical KD includes three parts: teacher net, student net, and knowledge transfer. The first detailed study of KD was to learn class distribution by using soft labels via a large pre-trained teacher and small student model. The model output of the teacher net is used as the input of the student net, and the latter approximates the former output (Hinton, Vinyals, and Dean 2015). As part of a traditional KD, knowledge is transferred from a complex model to a simplified model. Differently from the basis of the distilled model, Zhang et al. (Zhang et al. 2018) proposed mutual learning which replaced a classic teacher-student architecture with a group of students modelling learning and guiding each other. Besides, Kullback Leibler (KL) divergence was firstly introduced to determine the output probability from two students. The experiment results demonstrated a trend that the model performance increases as the number of students rise.

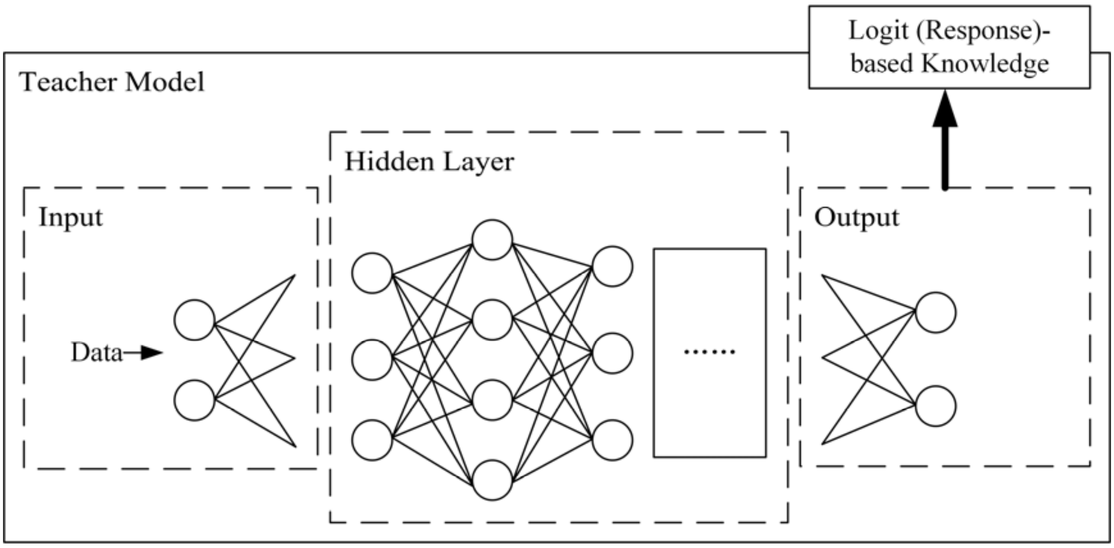

Figure 2. Categories of knowledge obtained from a teacher model.

Figure 2 Alt Text: Response-based or logic-based knowledge describes the neural response to the last output of the teacher model. Generally, this type of knowledge is used to mimic the final prediction of a teacher.

Classification tasks can be enhanced using soft labels, as they provide more information, reveal the degree of generalisation of the teacher model, and demonstrate its generalisation ability. The better teacher model, however, according to Cho and Hariharan (Cho and Hariharan 2019), will not lead to better student teaching with soft labels with regularization. Student models do not match the teacher because there is a capacity for mismatch samples, and they are biased towards primary losses. In this paper, a technique called early-stop distillation is presented, in which distillation is stopped before convergence occurs. (Xie et al. 2020) presented a training strategy using more noisy datasets to improve the student model performance by focusing on the data issue. (C. Yang et al. 2019) keeps using soft labels in the distillation process and applies constraints to optimise the model in successive generations. By combining the ground truth and secondary classes, the authors demonstrated that the second class could learn similarities effectively, preventing the student network from overfitting. A weakness with this teacher-student architecture, however, is that the different network structure between teacher and student networks affects the model performance. Phuong et al. proposed a new loss function and used a multi-exit architecture to conduct ensemble KD to ensure diversity distribution. By matching the output probabilities of early and later exits, the method allows early exits to mimic late, more accurate ones (Phuong and Lampert 2019). The logit-based method is relatively simple and effective. When a teacher model is used in student modelling, the probability distributions they output are equivalent to similarity information between categories and provide additional supervision signals, which makes it easier for students to learn (Gou et al. 2021). Although distillation efficiency strongly depends on softmax loss function, it cannot be achieved if there is a low-level vision problem (Romero et al. 2015).

2.2.3. Offline distillation

The offline distillation process includes 1) a complex teacher model is built based on data before distillation, and 2) a pre-trained teacher model provides extracted information, such as logits or intermediate features, to assist the student model with the distillation process (Gou et al. 2021; Hinton, Vinyals, and Dean 2015). Although training the student model offline is efficiently guided by the teacher model, training the teacher model is inevitable. Training takes a significant amount of time. Additionally, there is frequently a capacity gap between teachers and students, meaning that students rely on their teachers. Few studies examine the relationship between teacher and student models. By using the offline approach, knowledge transfer can be optimised in different areas of transferring knowledge including the type of knowledge (Hinton, Vinyals, and Dean 2015; Romero et al. 2015) and the design of loss functions for matching features or distribution matching (Huang and Wang 2017; Zagoruyko and Komodakis 2017; Mirzadeh et al. 2020).

2.2.4. Teacher-student architecture

A special focus will be placed on building T-S architecture. It may be difficult to learn from large deep neural networks due to gaps in modelling capabilities in comparison to student neural networks. A variety of research has been conducted to ensure that knowledge is effectively transferred to student networks. To alleviate the training gap between the teacher and student, a teacher assistant was applied to the T-S architecture (Mirzadeh et al. 2020). (Son et al. 2020) shows a similar mechanism, in which the researchers trained more teacher assistants iteratively, where the existing deeper teacher assistants were used to train new shallower ones. There have also been several recent methods that aim to minimize the difference between the structure of the student model and that of the teacher model. As part of the model compression technique, (Polino, Pascanu, and Alistarh 2018) synthesised the quantised teacher model as the student model, to improve the performance of the student model and to better fit the teacher model with behaviour. While conventional T-S architecture focuses on collecting knowledge from individual teachers, some studies have taken the approach of extracting knowledge from a group of teachers for more comprehensive guidance that will benefit student networks at different stages. Such as (You et al. 2017), different model compression techniques, such as voting strategies, were applied to unify dissimilarity information in intermediate representations between layers. Student networks with well-performed algorithms had superior accuracy to teacher networks with fewer parameters and faster inference speed. Another study exploited two approaches to leverage information from multiple teachers by updating student weights at the minibatch level and training students on multiple streams of information via data augmentation (Fukuda et al. 2017).

3. Methodology

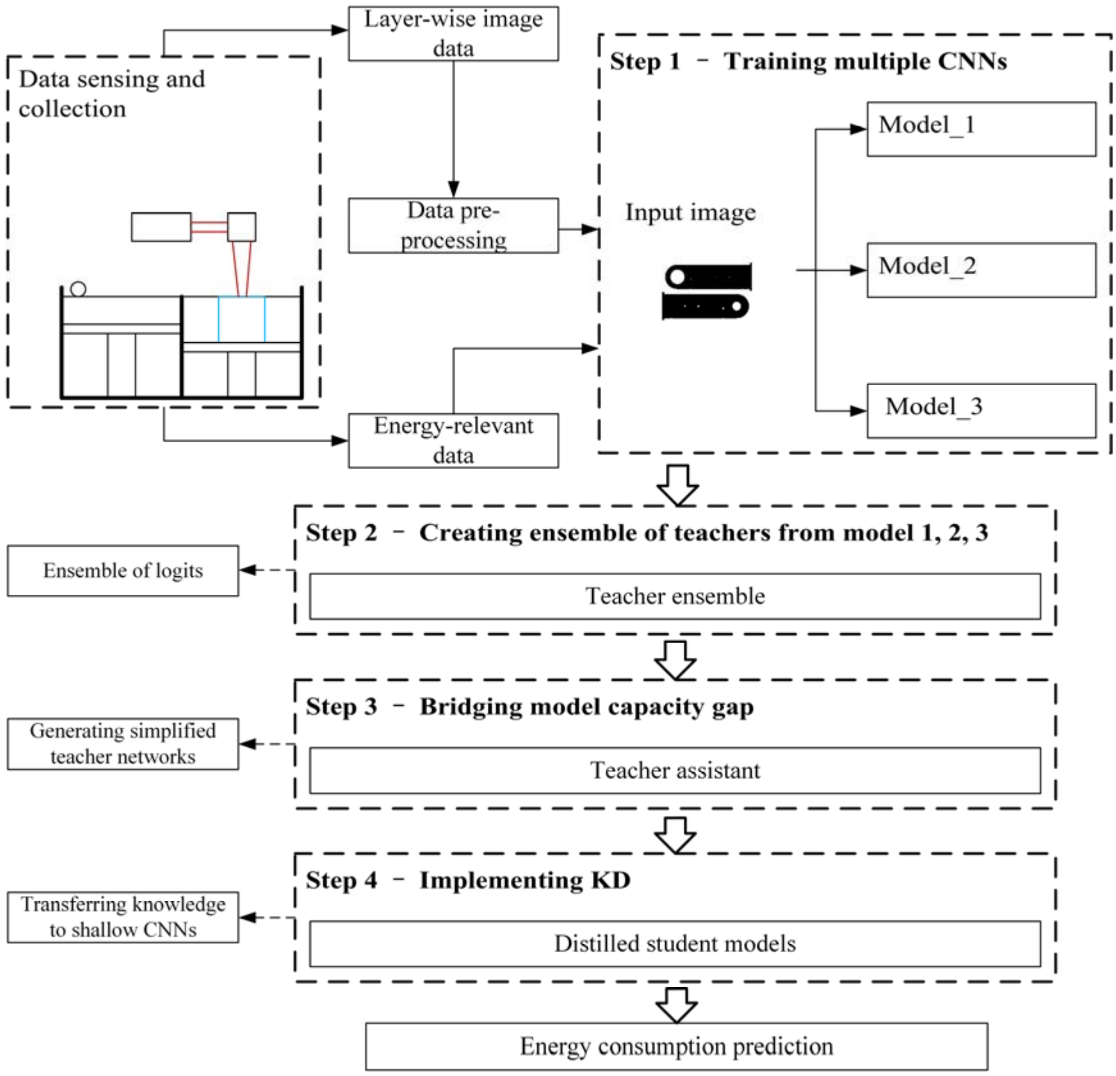

Figure 3. The workflow of the proposed approach.

Figure 3 Long Description: The framework consists of four major steps. Step one is to use layer-wise images and energy-relevant data to train multiple convolutional neural networks as the baseline model. The second step involves creating a teacher model to obtain an ensemble of logits. A TA network between teachers and students is built to alleviate the capacity gap between the models. Lastly, the KD process is used to transfer knowledge to shallow CNNs. Using these trained CNNs, energy consumption is predicted.

The feasibility and effectiveness of using layer-wise images of CAD models of printed objects to anticipate AM energy usage were established in a preliminary study (Li et al. 2022). In this study, by extending the teacher architecture and customising the student architecture, a hybrid model compression scheme for energy consumption prediction is achieved via both KD and other compressing techniques. The hybrid KD initialisation is divided into three components: teacher ensemble, teacher assistant, and distilled student by using KD, followed by the latter validation phase. To begin this process, a teacher ensemble is established to improve individual teachers' algorithm performance. The teacher ensemble is trained independently by combining three different convolutional neural networks (CNNs). To mitigate the difference in model complexity from the pre-trained teacher ensemble, a teacher assistant is applied. Besides, it will improve the generalisation ability of the student model. Through the distillation process, the other two low-precision student models that are not fine-tuned can learn from a teacher assistant. After this, a case study based on a specific AM system is carried out to show the feasibility and effectiveness of predicting energy consumed.

3.1. Creating Teacher Ensemble using Stacking Approach

3.1.1. Training multiple CNNs

It is possible to obtain better predictive performance using an ensemble of several learning algorithms than by utilising any one of the individual learning algorithms alone. The primary goal of using an ensemble is to find a hypothesis that is not necessarily present in the hypothesis space of the models on which it is built. Ensembles tend to produce better empirical findings when there is a large degree of variation across the models (J. Yang et al. 2013). To reduce the variance efficiently, an ensemble strategy for a deep neural network is presented. Three different learning algorithms hold relatively undesirable prediction performance through the experiments (experimental results are given in Table 3). A heterogeneous ensemble can achieve this by using different machine learning algorithms or different hyperparameters. In order to effectively enhance the overall performance of the teacher network while achieving diversity between different base learners, three different models participated and were integrated into the experiment (El-Rashidy et al. 2020). In addition, more weak learners have a negative impact on training and inference time. As a result of the stacking approach, it trains a model that is used to combine other models (models 1, 2, and 3 in this case). It is accomplished by splitting the data into two parts, training models 1, 2, and 3 testing these models, and then using the output as input to train the combined model.

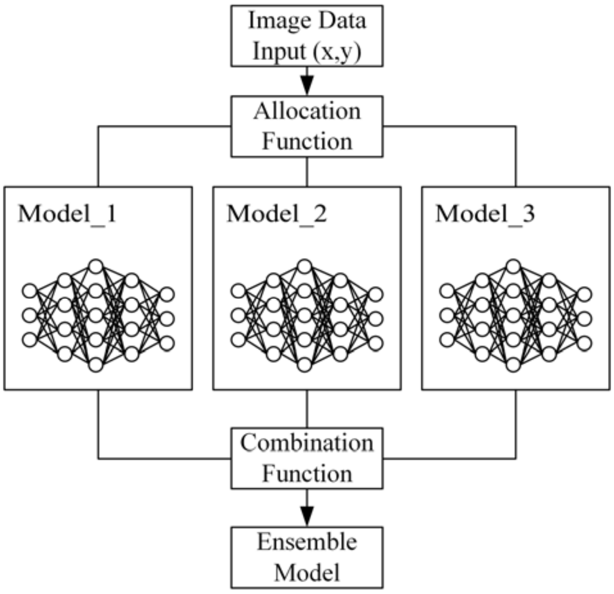

Figure 4. Neural network ensemble using an average of outputs of models.

Figure 4 Alt Text: The stacking approach combines the logits from different pre-trained CNNs. Models 1,2 and 3 are involved different models in terms of network structure.

3.1.2. Ensemble of predictions

There are many kinds of ensembles, which are theoretically capable of representing any other ensemble technique, including stacking. The stacking approach involves training a learning algorithm to combine predictions from different learning algorithms. This is accomplished by averaging the outputs of the models in the set. Ensemble involves stacking multiple models together, so input data for each model must be forward propagated. As a result, computations become more time-consuming, and evaluation (prediction) times increase. A unified soft label is derived from the output of multiple teacher networks and then used to guide the training of students (Fukuda et al. 2017). If $\boldsymbol{h}$ is the soft label of teacher, $\boldsymbol{T}$ is the total number of the models, and $\boldsymbol{w}$ is a set of weights, then the overall result can be shown as:

$$h(x) = \frac{1}{T}\sum_{i=1}^{T} w_i h_i(x)$$

$$where, w_i \geq 0, \sum_{i=1}^{T} w_i = 1 \tag{1}$$

The benefit of combining multiple models is that they tend to all make the same errors on the test set by combining errors from different models (Goodfellow, Bengio, and Courville Aaron 2016). From the published results, the bias from combining multiple models balances the variance. This results in predictions that are less sensitive to training data, training strategies, or the serendipity of one training run. Three different small CNNs are selected and trained separately on the AM slicing model image dataset. Their performance is then evaluated on a test set. Lastly, the ensemble model is evaluated by integrating the three models. The ensemble is expected to perform better than any model in it separately on the test set. As such, different teachers assign themselves the task of analysing the data and making predictions. Once the predictions are made, they are aggregated. With these data and soft labels, the model is trained to perform fairly in all teacher classes combined. An improved student model can be created by using labelled data for fine-tuning.

3.2. Employing the Teacher Assistant at Intermediate Level

3.2.1. Pruning

A major challenge in pruning is to determine the less important parameters. Both reduced network complexity and decreased overfitting have been achieved by pruning networks. A larger neural network structure is pruned to obtain a smaller neural network. When pruning, a top-down approach is used, in which first a large network is constructed, and then the network structure is trained by removing or merging certain neurons or weights as necessary (Reed 1993). In the training process, the magnitude-based weight pruning (Han et al. 2015) achieves the sparsity of the model, by reducing the weight to zero. The method compares the absolute value of the weights with a set threshold and weights below the threshold are set to zero. This method entails first learning the connectivity via training. Contrary to conventional training, the connections are learned instead of learning the final value of weights. Followed by this, low-magnitude neurons are pruned out, which transfers a dense neural network to a sparse one, removing all connections with weights below the threshold. Lastly, after the network is retrained, a final weight is calculated for the remaining sparse connections to preserve accuracy. In general, pruning methods result in a reasonable compression rate without affecting accuracy. Therefore, the pruning approach is more appropriate for applications that require stable model performance. Pruning networks usually lead to a reduction in the size of the model but do not result in a significant increase in efficiency (e.g. more training or inference time) (Y. Cheng et al. 2017).

3.2.2. Teacher assistant

Teacher assistant (TA) was proposed by multiple phases of distillation based on recent work (Mirzadeh et al. 2020), according to which self-distillation is capable of improving the accuracy of the base model. When comparing a large deep neural network to a small neural network used by students, a model capacity gap can occur that degrades knowledge transfer. For effective knowledge transfer to student networks, adding a TA is proposed for reducing the complexity of the model in a controlled manner. The method involves minimising the structural differences between the student and teacher models, combined with network pruning and knowledge distillation, i.e., the student model is a simplified version of the teacher model. The model trains students by using TA, instead of a single large teacher. Teachers do not have the same level of expressiveness as students, so the TA is a smaller model of the same architecture. TA is considered to be able to translate teacher classifications that the student might not be able to express. This strategy (Figure 5) aims to introduce a TA that simplifies the knowledge of teachers to improve the generalisation ability of the student model. This model aims to approximate the previous model, while it ought to achieve better learning ability than the previous one. The soft label distribution of TA can be fitted more effectively by the student than that of the teacher ensemble. Furthermore, it allows softer targets to mitigate the weakening of the knowledge transfer directly from teacher to student.

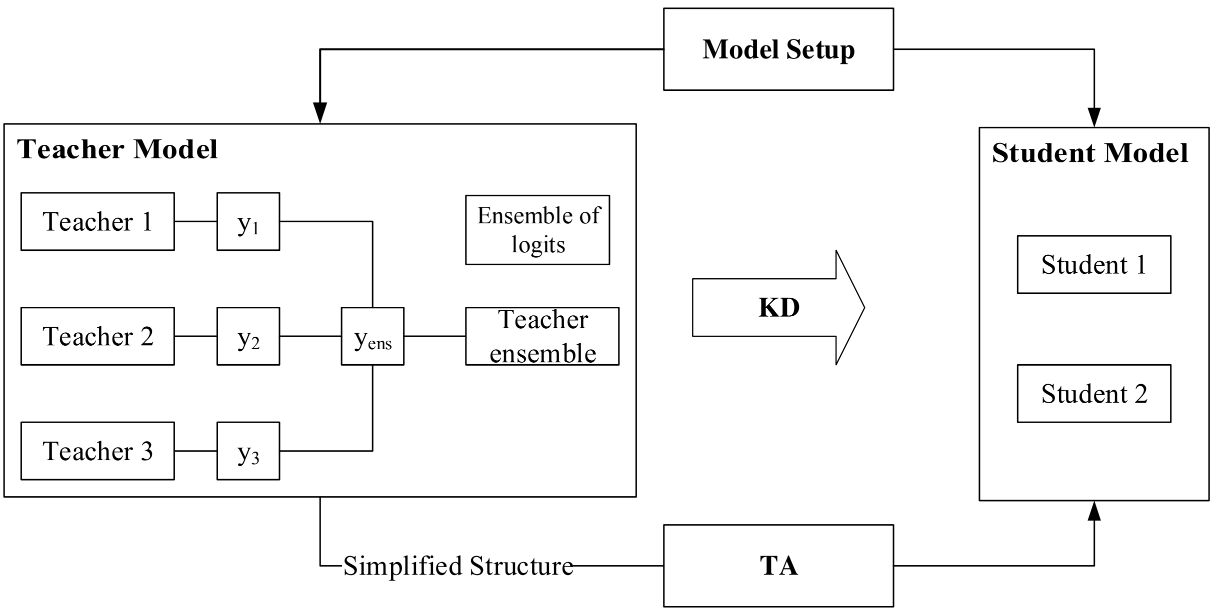

Figure 5. Teacher assistant in the proposed framework to increase the generalisation ability of the student model.

Figure 5 Long Description: A teacher assistant (TA) is built at the intermediate level of the teacher ensemble and the student. Compared to the previous teacher ensemble model, TA aims to approximate it while also providing better learning capabilities. It is important to focus on softer targets just to mitigate any weakening of the knowledge transfer directly from the teacher model to the student model.

3.3. KD-based Energy Consumption Predictive Modelling

3.3.1. Multi-stage KD with the hybrid optimisation approach

Using a group of teacher models can provide the student model with a variety of different types of information that can be more advantageous than learning from a single teacher. KD is the process of transferring probability information from a large model to a smaller one. The knowledge capacity of large models (such as very deep neural networks or ensembles of many models) is generally higher than that of small models, but this capability is sometimes under-utilised. To assign probability distributions to a large number of labels in deep neural networks for prediction tasks, a softmax layer is used, which outputs probabilities containing correlations between each class. Despite that, the model assigns a low probability to all incorrect labels in comparison to the correct ones, which results in one-hot codes (hard labels) that fail to take into account their prior probabilities (knowledge). KD defines the loss function as shown below:

$$L = -T^2 \sum_{i}^{N} q_i^T \log(p_i^T) \tag{2}$$

$$q_i = \sum_{j} w_i \, q_{ij} \tag{3}$$

where $\boldsymbol{q_i}$ is the averaging soft label generated from the teacher ensemble, $\boldsymbol{p_i}$ is the class probability distribution, and $\boldsymbol{T}$ is the temperature. A complex teacher model learns from both hard and soft labels, while a student model uses soft labels (part of total loss) to achieve knowledge transferring based on equations (4) and (5). The corresponding outputs of the teacher and student model are $\boldsymbol{L_{soft}}$ and $\boldsymbol{L_{hard}}$, respectively. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are used to control the trade-offs between two losses, which $\boldsymbol{\beta = (1 - \alpha)}$.

$$L_{student} = \alpha L_{soft} + \beta L_{hard} \tag{4}$$

$$L_{soft=} \sum_{i}^{N} q_i^T \log(p_i^T) \tag{5}$$

Students and teachers receive the training data, with the student model creating a soft target from the softmax distribution of the teacher model. This loss function is composed of the output of the softmax of the student model and the cross-entropy of the soft target at the same temperature. The second part of equation (4) discusses how ground truth can reduce the risks of bias leading to errors in student models.

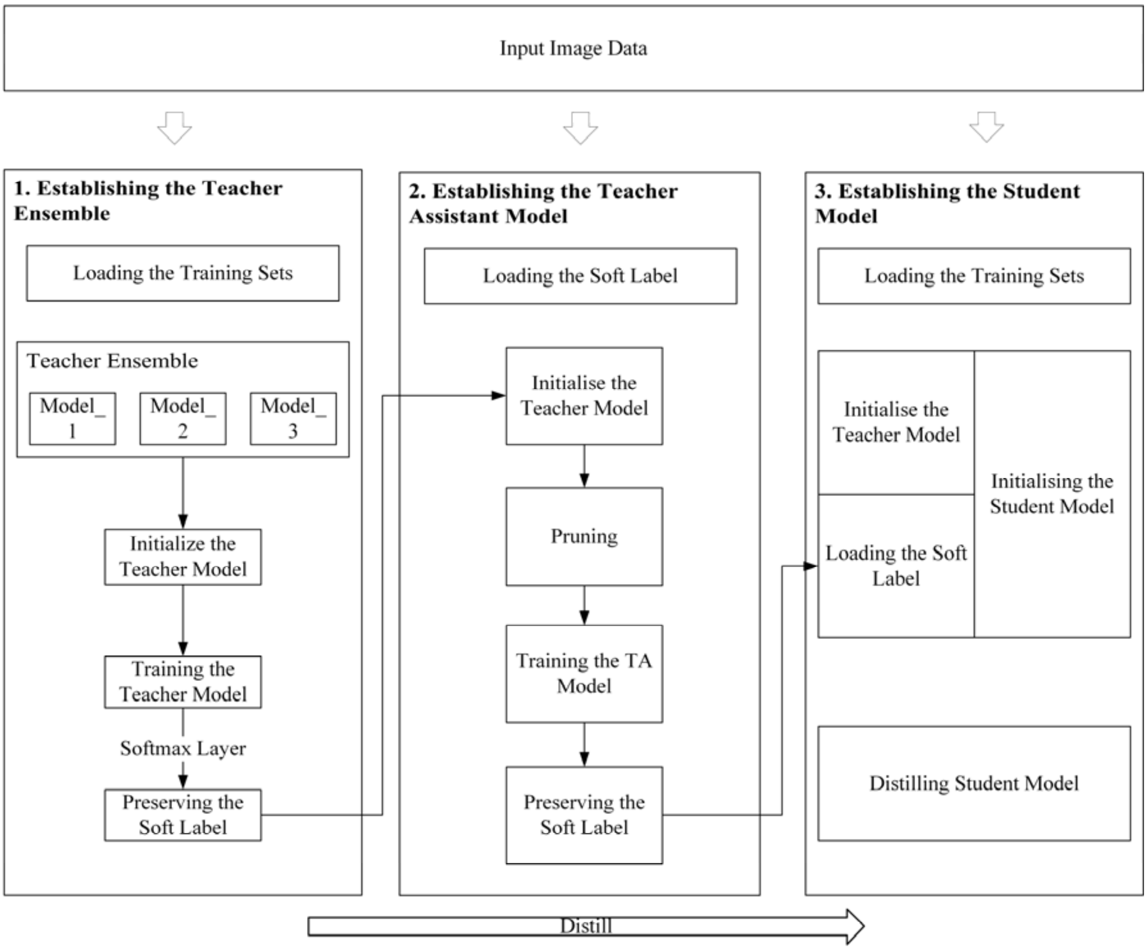

Figure 6. Multi-stage KD using a hybrid optimisation approach.

Figure 6 Alt Text: Three stages are included in the multi-stage KD: 1) establishing the ensemble of teachers, 2) establishing the assistant model, and 3) establishing the student model. Stage 1 involves training the teacher ensemble, with soft labels preserved at the end. Model capacity gaps are mitigated by TA in stage 2 because of model structural differences. Soft labels are transferred to students at the end.

As shown in Figure 6, the predictions of three teacher models (CNN) are firstly combined by using the stacking approach to establish the teacher ensemble. In this stage, a unified soft label is constructed by weighting the output of multiple teacher networks, which guides the training of the student networks. Secondly, this teacher ensemble with strong learning ability applies to a large and complex model trained by hard labels. The soft labels are generated at the softmax layer to calculate the primary loss, and then provide probability information (the soft labels for the next stage). These soft labels are finally preserved. Because of the nature of offline training, the teacher ensemble does not involve training in this stage, where temperature T does not change until the next two stages. Knowledge transfer can be degraded by a model capacity gap between deep neural networks and student neural networks. To ensure that knowledge is effectively transferred to student networks, the model needs to be simplified.

Stage two involves training a pruned model based on the ensemble of the teacher as a TA to minimise the difference between the ensemble and the student model, reduce redundant neurons, and reduce computational complexity. The training data and preserved soft labels are sent into this model to train. The magnitude-based weight pruning approach is adopted. This stage is used to transfer the last soft labels from the teacher ensemble.

In the third stage, a student model is built, with the outputs of the TA model using the same training set. By initialising the TA model, the student model can learn the probability information from TA. model is trained as the predictive model for energy consumption analysis.

3.3.2. Energy consumption analytical model

The overall energy consumption of AM is computed. Nevertheless, AM poses the challenge of consuming substantial time, leading to a rise in energy consumption. The overall energy utilisation strongly relies on the duration of the manufacturing process. Therefore, the unit energy consumption (Wh/g) is utilised to evaluate the energy consumption level. The unit energy consumption (E_U) determines the ratio between power rate and process productivity. The following equation can be expressed:

$$E_U = \frac{E_T}{M_T} \tag{6}$$

Denoted E_T and M_T are total energy usage and mass of a total part respectively, the unit energy consumption E_U can be computed by simply dividing.

4. Experimental Study

An SLS process involves the application of a carbon dioxide laser beam to a powder to sinter or fuse it. A temperature of almost the melting point of the material is maintained in the chamber. For each layer specified by the design, the powder was fused using a laser at a specific location. When a layer is completed, a piston lowers the layer thickness by the same amount every time a layer is completed (Wong and Hernandez 2012; Gao et al. 2015). In this case study, an SLS machine (EOS P700) was used to fabricate the product. With a building envelope measuring 740 mm by 400 mm by 590 mm (x, y, and z), two CO2 lasers can sinter nylon materials (PA2200 and PA3200GF) on the EOS P700. The original polyamide-12 is PA2200, and PA3200GF contains 40% glass beads to increase stiffness (Qin, Liu, and Grosvenor 2018). The material used in this experimental study is PA2200.

4.1. Data Description

In the target SLS system, the image data are processed and trained by CNN using convolution operations. In this type of neural network, 2D images are recognized and categorized and features are extracted from sliced models based on layer-wise slices. In the case study, layer-wise images represent the product design data obtained from sliced CAD models, which illustrate the geometric features of each layer. In the dataset, 12 builds of product CAD models are included which contain more than 10000 layer-wise images in totally (Hu et al. 2021). Data in a comma-separated value (CSV) file is collected by a power meter, which shows the actual energy consumption in this SLS system. Each layer in the dataset has an individual unit energy consumption ranging from 4 to 200 (Wh/g). Each CSV file with unit energy consumption is corresponding to the layer-wise images of the sliced 3D model.

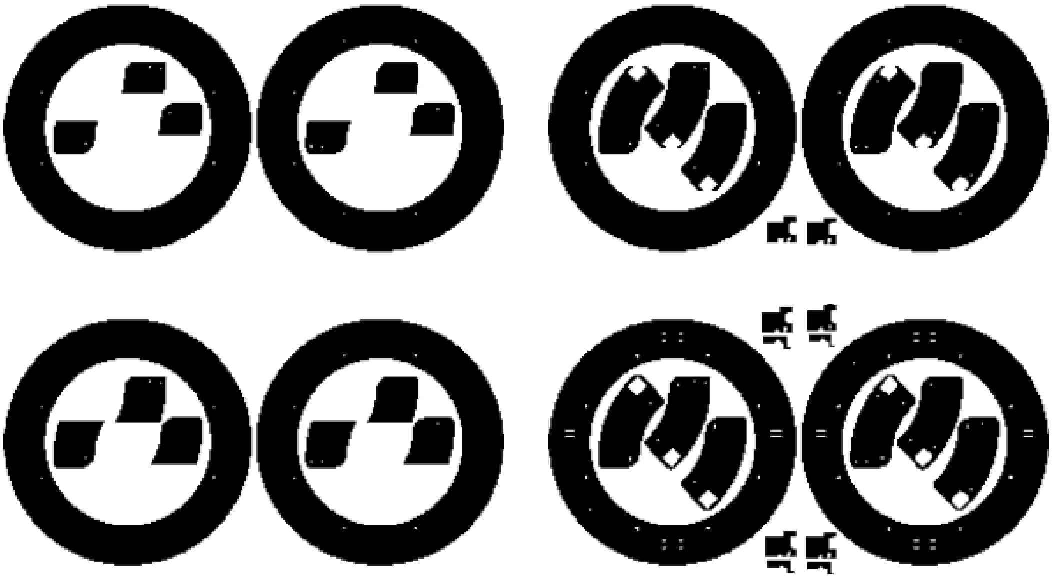

Figure 7 Layer-wise image data.

Figure 7 Alt Text: Some examples of the grayscale images used as the input data from the sliced CAD model.

As can be seen in Figure 7, the image data have been sliced from the 3D model data to form the image data. It is a complex build that has been created using an SLS system throughout the entire process. A number of geometric features are included in the design of the part. In some cases, they can take the form of a standard shape or form, whereas, in others, they can be a one-off, tailored to your specifications. A layer-by-layer geometry feature extraction can be done by using the different features within Images associated with certain layers at a given time in order to extract the layer-by-layer geometry feature.

Table 2 Statistic information of the unit energy consumption of each build.

	Minimum	Maximum	Mean	Standard deviation	Median
Unit energy consumption of a build (Wh/g)	221.07	769.81	382.53	142.24	345.45
Layer-wise unit energy consumption (Wh/g)	3.37	81.28	12.62	9.18	9.58

Unit energy consumption values are presented in Table 1 including minimums, maximums, means, standard deviations, and medians. A build consumes 142.24 Wh/g of energy on average, with a standard deviation of 142.24 Wh/g. The median value is lower than the mean value, which indicates that the majority of energy consumption falls between the minimum value and the mean value. In the table, layer-by-layer information regarding unit energy consumption is also provided.

4.2. Experiment Setups

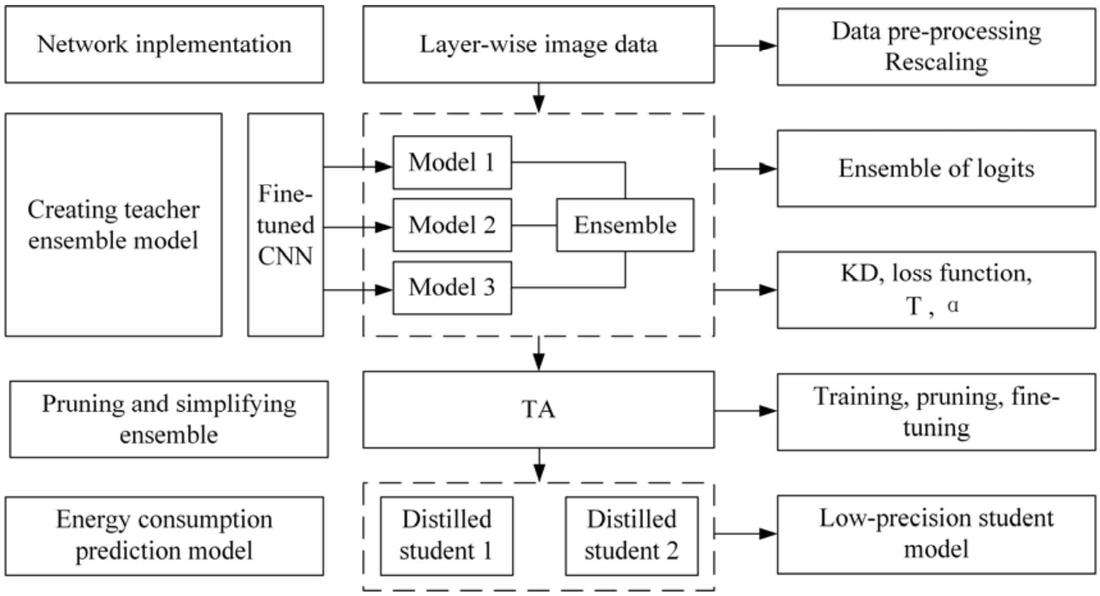

Figure 8. The model training process and implementation flow of the proposed approach

Figure 8 Alt Text: The experimental setups for the proposed methodology. The left-hand side indicates the training process of the methodology, while the right-hand side is the output of each stage.

Data processing is applied before the experiment. All images are collected and resized to 128×128. For ease of processing, all images are changed to binary. The rescaled and binary images are used as the dataset of which 80% are used as the training set and 20% are the testing set. The predictive model is established based on T-S architecture including a fine-precision teacher ensemble model, a pruned TA, and a low-precision student model. The teacher ensemble model is trained by three different CNN models to obtain the best performance of all. All models were used to train the energy consumption predictive model based on the same dataset. To show prove the effectiveness and feasibility of the proposed KD-based approach. There are four groups of the experiment: 1) training each CNN separately by image datasets, 2) training teacher ensemble, 3) training under the T-S scheme without a TA network involved and 4) training with the proposed approach. For experiments 3 and 4, the other two CNNs are built to serve as student models for comparative usage.

Firstly, three separate models are directly trained by the AM dataset to set as the baseline model. The major use of CNNs is for image-driven pattern recognition, which is greatly influenced by variables such as the number of layers and the depth of each layer. CNN is defined by adding convolutional layers, pooling layers, and full-connected layers. Energy consumption according to each layer can be more precisely computed when the model is denser and wider, leading to higher computational costs. As a result, the model complexity must be balanced against the computational cost. The parameters used for compiling and training each CNN are the same. The batch size is set at 32 with 20 epochs, which allows the model to reach the local minima. Validation data comes from 20% of the training data. Model 1 is in its simplest form with approximately 4.43M parameters. It is usually composed of several convolutional layers followed by a pooling layer, so it can be used in place of a fully connected layer. After adding a high-level convolution layer and neurons, model 2 and model 3 contains more than 8.76M and 10.68M parameters, respectively. The three models are used to train the ensemble model to yield better results, however, this will lead to a computational burden while training.

Secondly, the ensemble model uses the same input layer that was applied to all pre-trained models. Averaging is performed to merge the three models' outputs in the top layer of the ensemble. The merged model has more parameters ($\approx$ 6.7M), which is much denser than previous models. Using hold-out validation data, these weights can be optimised based on the predictions from each model so that the ensemble can be improved slightly. An ensemble blending method is often referred to as a weighted average ensemble. In the case of teacher ensembles, a student may not be able to distil complete knowledge from many teachers, thus resulting in a decline in model accuracy.

Thirdly, to reduce the size of the teacher ensemble while minimising accuracy losses, neural network pruning is used as a model compression technique. It aims to remove more of the less relevant parameters. Weights are essentially set to zero by pruning based on magnitude. By removing weights that are already close to zero, or low in magnitude, the effect on the network is minimised. A normal constructing process includes training, pruning and fine-tuning. Therefore, the model ensemble starts with 50% sparsity (50% zeros in weight) and ends with 80% sparsity to fine-tune the pre-trained teacher ensemble from the previous step. This experiment aims to justify whether the model complexity affects the final predictions of the student. If so, an early-stop mechanism is adopted to alleviate this phenomenon.

To bridge the model performance gap between teacher and student, the fourth step applies a TA, in which the TA is a pruned version of the teacher ensemble. The key to distillation is the loss function. The output of the student model is a conventional softmax function to match the output of hard labels from the teacher model, until the teacher model finishes training, denoted as $\boldsymbol{L_{soft}}$. The temperature parameter is added to the softmax classifier of the completed teacher network, being a fitting target for the student network with the same temperature parameter $\boldsymbol{T}$, denoted as $\boldsymbol{L_{hard}}$. Hyperparameter $\boldsymbol{\alpha}$ is introduced to determine the total loss function followed by the final training. A distiller will be constructed using the following parameters, which include a pre-trained teacher ensemble model, a student model to be trained, and a student loss function ($\boldsymbol{\alpha}$ and $\boldsymbol{T}$ are hyperparameters, which are set to select 0.9 and 3 respectively).

4.3. Evaluation Metrics

In the evaluation of algorithm performance, the mean absolute error (MAE) and root mean squared error (RMSE) are used. Equation (8) and (9) shows MAE and RMSE respectively.

$$MAE = \frac{\sum_{t=1}^{N} |p_t - a_t|}{N} \tag{8}$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^{N} (p_t - a_t)^2} \tag{9}$$

In the above equations, a_t refers to the actual value, p_t represents the predicted value, and N is the number of data points.

5. Results and Discussion

5.1. Benchmark Training

An illustration of RMSE and MAE of three different CNNs over 20 epochs can be seen in Figures 9 and 10. The RMSE and MAE of teacher ensembles directly trained from image data in this figure were significantly lower than others. At the beginning of the training, the teacher ensemble had the second-lowest RMSE and MAE. These values decrease as the training continues. As a result of the ensemble approach, the prediction for energy consumption is better and errors are reduced as expected. TA is then pruned based on this teacher ensemble in the next stage. Based on the results shown in Table 3, the ensemble model has shown a trend of decreasing errors with the lowest values being 9.94 Wh/g and 6.72 Wh/g in RMSE and MAE, respectively. High performance, however, increases the computational cost of the model. Using the proposed methodology, the trade-offs between performance and size are realised. KD-based approaches aim to reduce the model size and maintain significant performance. The results of the experiments are shown in the following section.

Table 3. The model performance of predicting energy consumption between baseline models and an ensemble model

Model No.	RMSE (Wh/g)	MAE (Wh/g)
1	11.61	9.01
2	10.05	6.88
3	12.72	12.03
Ensemble	9.94	6.72

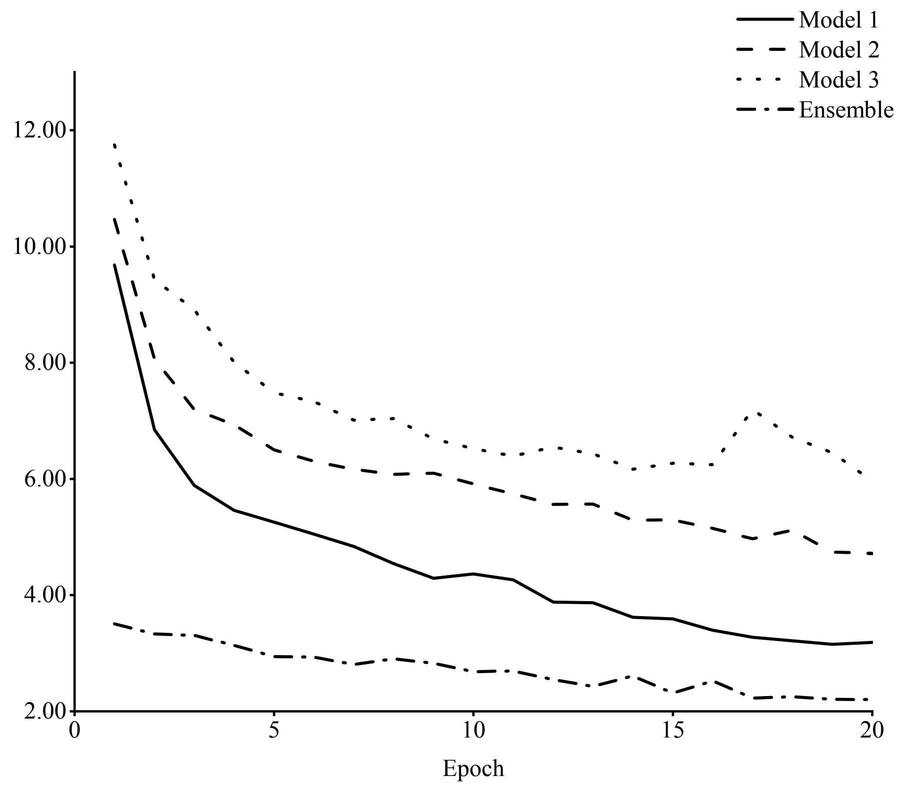

Figure 9. Comparisons of RMSE of trained models in training. Models 1 to 3 are the CNNs, and the model ensemble combines predictions from all.

Figure 9 Alt Text: A line graph shows the root mean square error (RMSE) of three baseline models and the teacher ensemble. The teacher ensemble has significantly fewer errors. As a result of the ensemble approach, the prediction for energy consumption is better and errors are reduced as expected

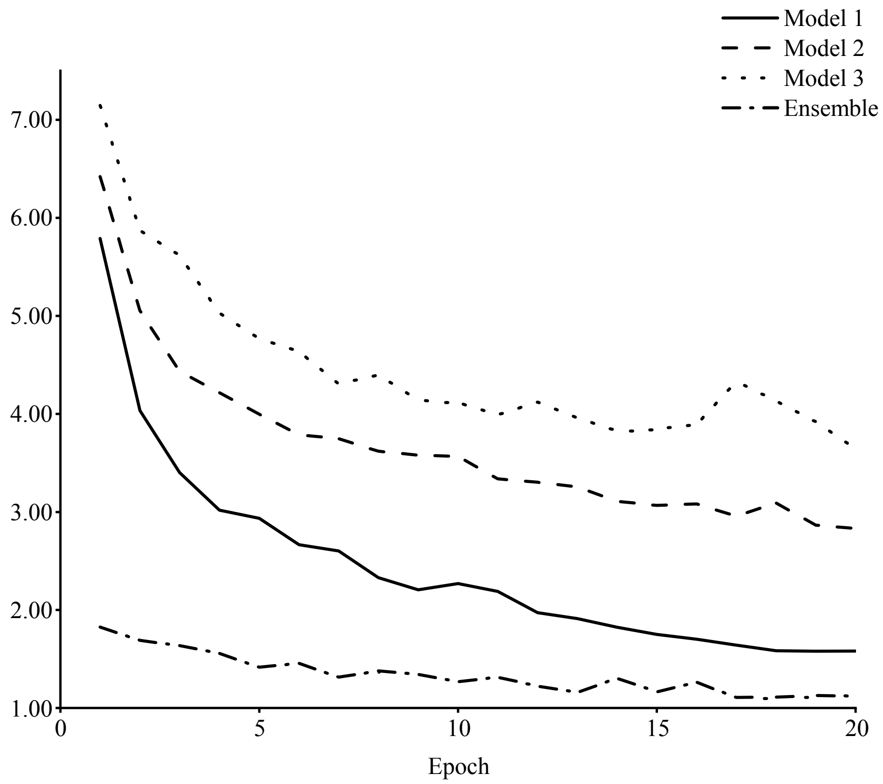

Figure 10. Comparisons of MAE of trained models in training. Models 1 to 3 are the CNNs, and the model ensemble combines predictions from all.

Figure 10 Alt Text: A line graph plotting the mean absolute error (MAE) of three baseline models and teacher ensemble. Significantly fewer errors occur in the teacher ensemble. As a result of the ensemble approach, the prediction for energy consumption is better and errors are reduced as expected.

5.2. Experimental Results

The performance of the model is expressed in Table 3 as the number of parameters, model size, RMSE, and MAE. Using the same AM image datasets, three different CNNs are trained separately. As the number of layers of CNN increases, the RMSE and MAE of models 1 to 3 vary. In addition, they require a considerable amount of time to train using AM data. Based on epochs, model 3 takes 179 seconds, whereas model 1 takes 101 seconds. The ensemble model shows a high level of computational complexity due to the large number of parameters involved, as evidenced by its size of approximately 13.01M, which requires a longer training time of 259 seconds per epoch. In the second experiment, predictions are combined using three CNNs. The RMSE and MAE, on the other hand, significantly reduce to 9.94 Wh/g and 6.72 Wh/g, respectively. In order to facilitate the generalization capacity of students, a teacher assistant is expected to simplify information and knowledge from a teacher ensemble. As compared to the ensemble model, the TA has a slightly smaller model size and a lower error rate. This network was able to achieve faster setup times and a lower error level between the student and teacher ensembles as a result of the pruning process.

Table 4 Illustration of model performance in terms of baseline, model ensemble, and TA models.

Experiment No.	Model	#Params	Model Size (MB)	Time (s)	RMSE (Wh/g)	MAE (Wh/g)
Experiment 1	1	~4.42M	~12.20	101	11.61	9.01
	2	~8.77M	~49.12	184	10.05	6.88
	3	~10.68M	~87.74	179	12.72	12.03
Experiment 2	Ensemble	~13.01M	~149.06	259	9.94	6.72
	TA	/	/	90	18.14	11.08

According to Table 5, experiment 3 did not involve a TA as an intermediary stage between the teacher ensemble and the students. As the parameters are increased, students A and B (not fine-tuned) begin to outperform the ensemble teacher network but remain faster. Generally, the size of the model is greater than what is expected from independent training. It has been shown that distilled student models have advantages over independently developed ensemble models in terms of training time and model size. In experiment 4, TA is used at an intermediate stage between the teacher ensemble and the students.

Table 5 Illustration of model performance in terms of distilled student models. a: training without TA b: training with TA.

Experiment No.	Model	#Params	Model Size (MB)	Time (s)	RMSE (Wh/g)	MAE (Wh/g)
Experiment 3[a]	Distilled student A	~1.19M	~1.94	95	9.14	5.48
	Distilled student B	~2.20M	~3.72	110	9.56	6.16
Experiment 4[b]	Distilled student A	~1.19M	~1.94	**98**	**9.03**	**5.30**
	Distilled student B	~2.20M	~3.72	**106**	**7.39**	**3.77**

Figures 11 and 12 illustrate a downward trend in the overall pattern. The RMSE of model A is 9.29 Wh/g and the MAE of model B is 10.06 Wh/g. After training the first two models, TA and direct training using KD are applied. When TA was applied to Model A, its RMSE decreased by 2.8%. There is a similar pattern in model B, which has a significant decrease in RMSE by 26.5%. The value of model A decreased from 5.727 to 5.3, while the value of model B decreased by 45.4%. It, therefore, appears that the knowledge of a teacher model can be transferred to a student model, resulting in a model that is more general than the same model trained directly. A solution to the capacity gap between the teacher and student models is to introduce a TA in the middle, whose effect is ideally at the level of both models at the same time.

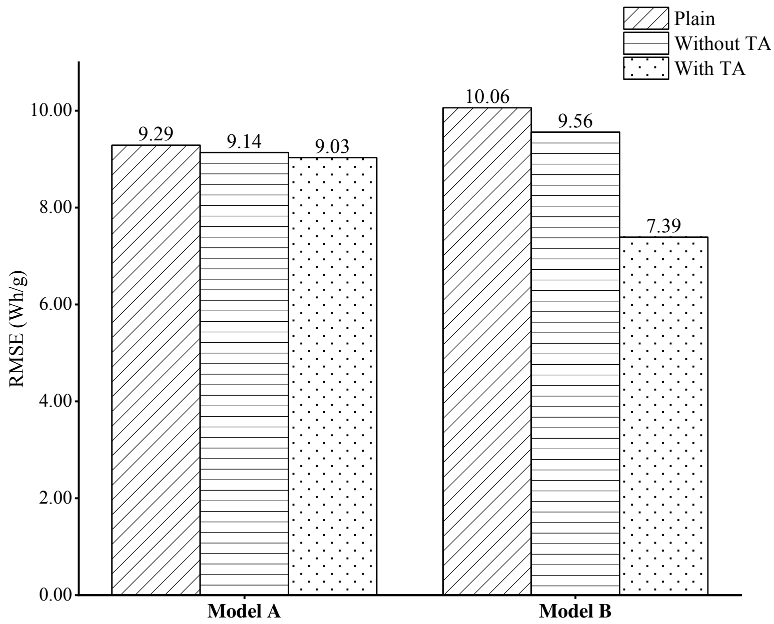

Figure 11. RMSE of models. Distilled students A and B are compared to models A and B.

Figure 11 Alt Text: The bar chart indicates the model performance of distilled student models A and B in terms of RMSE. Three groups of results are compared including plain model, employing TA, and not employing TA. Among all, distillation with TA has reached the best algorithm performance.

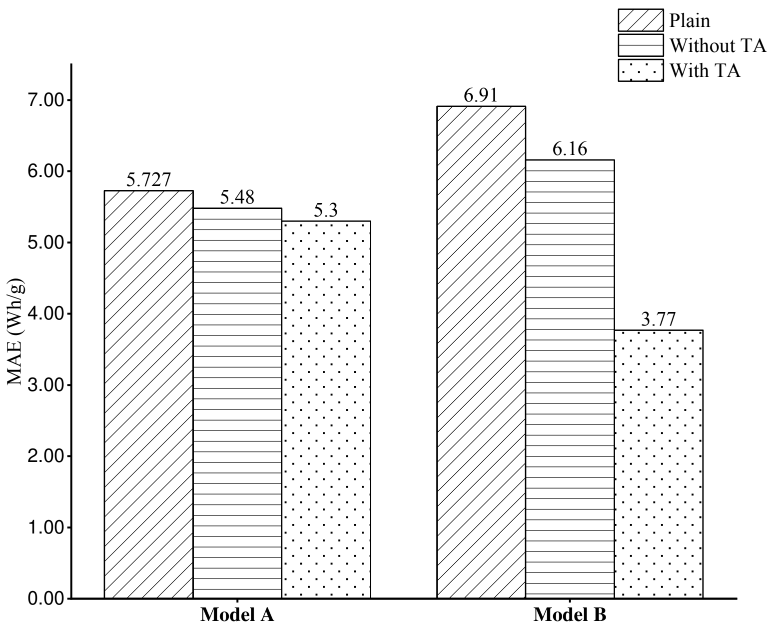

Figure 12. MAE of models. Distilled students A and B are compared to models A and B.

Figure 12 Alt Text: The bar chart indicates the model performance of distilled student models A and B in terms of MAE. Three groups of results are compared including plain model, employing TA, and not employing TA. Among all, distillation with TA has reached the best algorithm performance.

Table 6 The comparison of experimental results using geometry features and process parameters.

	RMSE (Wh/g)		**MAE (Wh/g)**	
	Model A	Model B	Model A	Model B
Part geometry	9.03	7.39	5.30	3.77
Part geometry + process parameters	9.08	7.64	5.41	2.85

An additional experiment is added which considers the parameters of the process in the model. As part of the predictive model, these process parameters are integrated with part geometry attributes, including the maximum and minimum measured values of dispenser speed, recoater speed, hatch power, hatch speed, and hatch width. Table 6 provides a summary of the experimental results. The results indicate that the difference between the predictions (between the part geometry alone and the part geometry combined with the process parameters) is marginal. The reasons for this may be due to the slight differences in the process parameters of each build and the relatively low impact of process parameters on unit energy consumption.

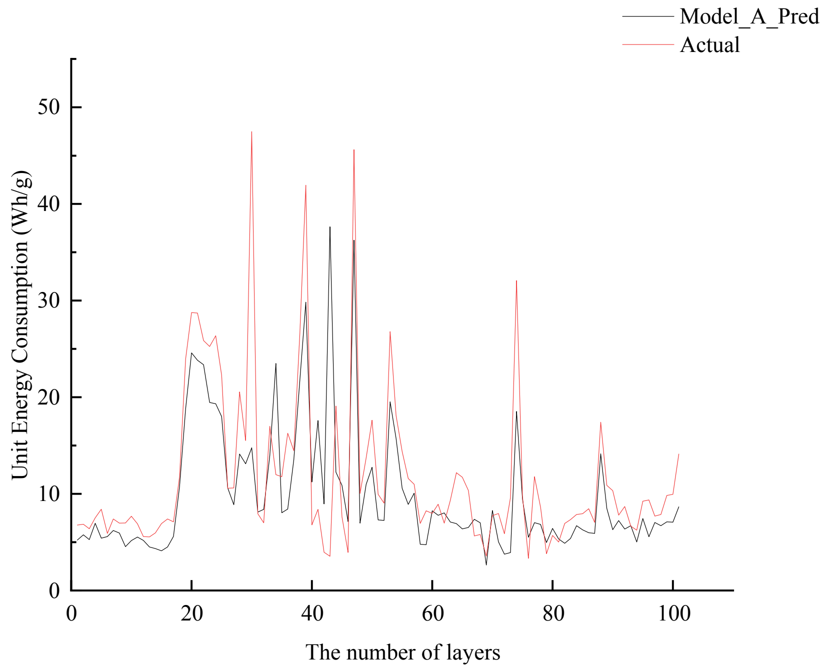

Figure 13. A comparison of predicted and actual unit energy consumption for model A.

Figure 13 Alt Text: The comparison of predicted and actual unit energy consumption for student model A with employing TA.

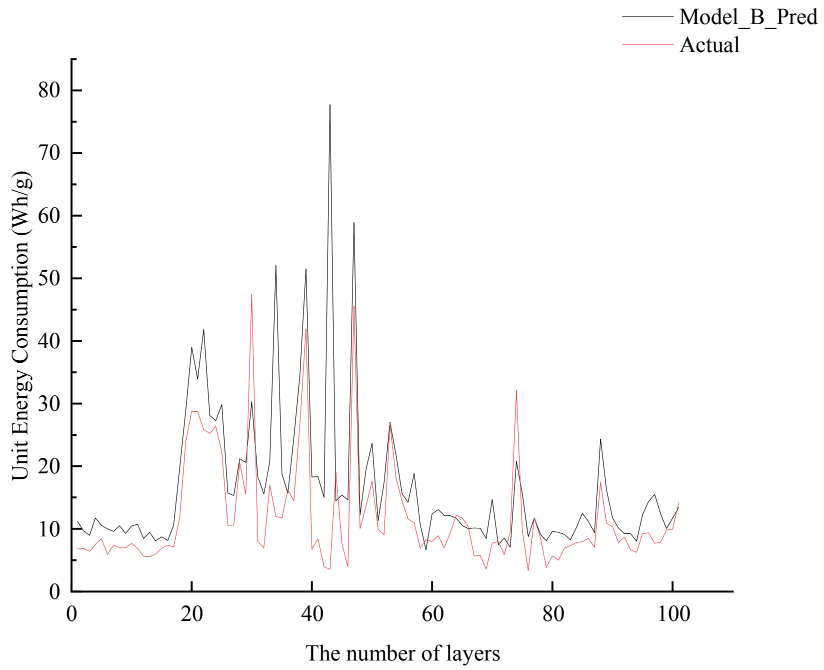

Figure 14. A comparison of predicted and actual unit energy consumption for model B.

Figure 14 Alt Text: The comparison of predicted and actual unit energy consumption for student model B with employing TA.

These two results from Figure 13 and 14 are based on the two student models after KD (Model A and Model B, respectively). There is a small error between actual and predicted unit energy consumption in Model A, ranging from approximately 12.5 Wh/g to 27.5 Wh/g, while there is a slight increase in Model B. This provided information regarding the predicted unit energy consumption in a particular layer. As a result, designers have access to relatively accurate estimates of the energy consumption levels of their design structures and part combinations during the AM production process.

5.3. Discussion

5.3.1. The impact of building teacher ensemble

Three CNNs utilised in experiment 1 vary differently in the network architecture. As the number of layers increases, a deeply layered architecture exhibits the benefit of reducing errors. The ensemble model normally yields better results than any single model in the ensemble separately, because it combines weak learners to reduce the variance finally optimising the model output. Three baseline models with diversity are trained to improve performance. As shown in Table 2, when adding the outperformed ensemble members to the ensemble, the result can be mitigated and show merit in optimising error rates. As for the knowledge obtained from different data in the AM dataset, multiple teachers provide better "guidance" than a single teacher does. The purpose of using an ensemble is to find a better hypothesis not covered by the models on which the ensemble is built by stacking different models describing different hypotheses about the data. The teacher network after an ensemble of predictions helps to integrate and synthesise a variety of illustrations of knowledge representations drawn from individual teachers' networks. According to the experimental results from Table 3, when a single model was used, the error rate was much higher than when an ensemble was used. More complex CNN reaches a high requirement of predictions of image data. It should be emphasised that every model has its weaknesses. Increasing the model size and the number of parameters leads to better performance in the ensemble model, as shown in Table 1. Ensembles appear to be more effective but require more computation, which results in long prediction times when deployed. Based on the performance of the integrated model, the number of models will be determined (normally, an odd number of models will be considered) (B. Liu, Long, and Chou 2016; Troussas et al. 2020). In addition, the training time increases with an increase in the number of base models. As a result, three base models were considered. The experimental results indicate that three different learning algorithms have relatively poor prediction performance (Table 3). Different machine learning algorithms or different hyperparameters can be used to achieve this in a heterogeneous ensemble. To enhance the overall performance of the teacher network while achieving diversity among the base learners, three different models were integrated (El-Rashidy et al. 2020).

5.3.2. The roles of TA in bridging the model capacity gap

The increase in layers of the final CNN might also limit the use of the ensemble model in deployment. Using the same algorithms as the model, model complexity and parameters are positively correlated with the ability of the model to represent data. Therefore, a pruning approach is carried out to balance the trade-off between model compression and training time. This pruning strategy has negative impacts on RMSE and MAE which may affect the final performance of distilled models, while it is set as a simplified version of a pre-trained teacher ensemble. As part of the T-S scheme, this pruned neural network is utilized as a TA, as a crucial stage between the teacher ensemble and distilled student. As shown in the experimental result in Table 3, the model has significantly improved in computational time and becomes more accurate than the ensemble.

A teacher ensemble's performance increases, which allows it to provide better supervision for the student by acting as a better predictor. The teacher ensemble, however, continues to gain complexity so that the student cannot fully match the predictions from the teacher ensemble, which makes KD process less effective. Therefore, a TA in the middle stage is used to mitigate this model capacity, connecting the teacher ensemble and the student model, whose effect is preferably above the average of the effect of the teacher and student models. By using soft labels generated with the teacher ensemble, the TA plays a critical role in mitigating errors according to comparative experiments 3 and 4. The second observation in Figures 11 and 12 is that adding a TA helps improve the results compared to a distilled student model without a TA. When data becomes more certain to teachers, it makes their logits (soft labels) less soft. Through matching soft labels, knowledge transfer is weakened.

Despite the disadvantages of fewer variables in the student model, soft labels provided by teachers are integral to alleviating these problems. Using mimicked student models on training sets allows for relatively high performance. Differently from other compression techniques, the proposed KD-based approach introduces a multi-step model of transferring important probability information from a deep teacher model to a shallow student model through a TA network while considering model performance and model size. The TA is distilled from the teacher, while the students are distilled only from the TA. As part of such a T-S architecture, the probability information extracted from the teacher is transferred to the student model while minimising error. In general, a large model brings more accuracy, while a student might not perform well. The distilled student model shows degraded performance due to the large structural difference between the teacher ensemble network and the student network.

6. Conclusions

AM is a new manufacturing paradigm of small-batch customised production adopted in various industries. This illustrates the advantages of fabricating products layer by layer to meet complicated geometrical requirements. Due to the growing concerns about sustainability, the current development of additive manufacturing processes requires energy management and optimization. In the context of AM processes, energy consumption is an important consideration. Various impact factors in AM systems produce heterogeneous data, which increases the complexity of analysing AM energy consumption. In this study, the unit energy consumption (Wh/g) is used to assess the energy consumption of the AM system. The extraction of key features and hidden patterns from image data can be accomplished using a deep learning approach. Based on the features extracted from the geometry perspective, this study examines KD and conducts a prior prediction of unit energy consumption. It is expected that layer-wise geometry-related features will provide valuable information that will greatly facilitate AM designers in considering potential energy consumption based on assessed unit energy consumption levels before generating and optimising geometry designs. As a result of deep learning models developed in this study, AM designers can predict the power consumption of different geometry designs before starting the design process, allowing them to adjust or even optimise their geometry designs accordingly.

The proposed approach uses a KD-based model to reduce computational costs while maintaining acceptable model performance. For the first step in the training process, a teacher ensemble was constructed by stacking three separately trained CNNs to improve the performance of the teacher model. To minimize the loss from outputs, a TA was generated from the pre-trained teacher ensemble. In this manner, the gap in model capacity between the teacher ensemble and the student can be bridged. In the third step of the experiment, KD is utilized to transfer knowledge from a complex teacher model to a simplified student model. Compared to independently trained students, distilled students exhibit better results, particularly in terms of training time and computation complexity. Instead of training a model directly, distilled models are used to reduce the size of the model, making it suitable for deployment.

The teacher-student architecture has great potential to compress the model in terms of model size and computing time. Even though the shallow student network may have more parameters in some cases, it is still less computationally intensive than the deep model, making it easier to deploy to smaller edge devices. To further improve deep neural networks (especially CNNs) performance and accelerate the processing speed, the reduction of convolutional computation, the most time-consuming part of CNNs, will be considered and explored. For the limited resources, the performance bottleneck mainly lies in the large number of multiplication operations required for convolutional computation, and the large number of weight and model parameters involved in the computation will bring more computational burden.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author, Y. LIU. The data are not publicly available due to their containing information that could compromise the research interests of the OEM manufacturers of AM machinery.

References

Achillas, Ch, D. Aidonis, E. Iakovou, M. Thymianidis, and D. Tzetzis. 2015. "A Methodological Framework for the Inclusion of Modern Additive Manufacturing into the Production Portfolio of a Focused Factory." *Journal of Manufacturing Systems* 37. The Society of Manufacturing Engineers: 328–39. doi:10.1016/j.jmsy.2014.07.014.

Ba, Lei Jimmy, and Rich Caruana. 2014. "Do Deep Nets Really Need to Be Deep?" *Advances in Neural Information Processing Systems* 3 (January): 2654–62.

Balogun, Vincent A., Neil D. Kirkwood, and Paul T. Mativenga. 2014. "Direct Electrical Energy Demand in Fused Deposition Modelling." In *Procedia CIRP*. Vol. 15. doi:10.1016/j.procir.2014.06.029.

Baumers, Martin, Chris Tuck, Ricky Wildman, Ian Ashcroft, Emma Rosamond, and Richard Hague. 2013. "Transparency Built-In." *Journal of Industrial Ecology* 17 (3): 418–31. doi:https://doi.org/10.1111/j.1530-9290.2012.00512.x.

Baumers, Martin, Christopher Tuck, David Bourell, R Sreenivasan, and Richard Hague. 2011. "Sustainability of Additive Manufacturing: Measuring the Energy Consumption of the Laser Sintering Process." *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture* 225: 2228–39. doi:10.1177/0954405411406044.

Bucila, Cristian, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. "Model Compression." *Kdd* 54 (1): 1–9. http://users.iems.northwestern.edu/~nocedal/PDFfiles/dss.pdf%0Ahttp://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.67.7952&rep=rep1&type=pdf%255Cnhttp://books.nips.cc/papers/files/nips15/AA03.pdf%250Ahttps://ieeexplore.ieee.org/document/85783.

Cheng, Jian, Pei song Wang, Gang Li, Qing hao Hu, and Han qing Lu. 2018. "Recent Advances in Efficient Computation of Deep Convolutional Neural Networks." *Frontiers of Information Technology and Electronic Engineering* 19 (1): 64–77. doi:10.1631/FITEE.1700789.

Cheng, Yu, Duo Wang, Pan Zhou, and Tao Zhang. 2017. "A Survey of Model Compression and Acceleration for Deep Neural Networks," 1–10. http://arxiv.org/abs/1710.09282.

Cho, Jang Hyun, and Bharath Hariharan. 2019. "On the Efficacy of Knowledge Distillation." *Microelectronics Reliability* 13 (6): 444. doi:10.1016/0026-2714(74)90354-0.

Dunaway, Daniel, James Dillon Harstvedt, and Junfeng Ma. 2017. "A Preliminary Experimental Study of Additive Manufacturing Energy Consumption." *Proceedings of the ASME Design Engineering Technical Conference* 4: 1–8. doi:10.1115/DETC2017-67864.

El-Rashidy, Nora, Shaker El-Sappagh, Tamer Abuhmed, Samir Abdelrazek, and Hazem M El-Bakry. 2020. "Intensive Care Unit Mortality Prediction: An Improved Patient-Specific Stacking Ensemble Model." *IEEE Access* 8: 133541–64. doi:10.1109/ACCESS.2020.3010556.

Fukuda, Takashi, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. 2017. "Efficient Knowledge Distillation from an Ensemble of Teachers." *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* 2017-Augus: 3697–3701. doi:10.21437/Interspeech.2017-614.

Gao, Wei, Yunbo Zhang, Devarajan Ramanujan, Karthik Ramani, Yong Chen, Christopher B. Williams, Charlie C.L. Wang, Yung C. Shin, Song Zhang, and Pablo D. Zavattieri. 2015. "The Status, Challenges, and Future of Additive Manufacturing in Engineering." *CAD Computer Aided Design* 69. doi:10.1016/j.cad.2015.04.001.

Goodfellow, Ian, Yoshua Bengio, and Courville Aaron. 2016. *Deep Learning*. MIT Press.

Gou, Jianping, Baosheng Yu, Stephen J. Maybank, and Dacheng Tao. 2021. "Knowledge Distillation: A Survey." *International Journal of Computer Vision* 129 (6). Springer US: 1789–1819. doi:10.1007/s11263-021-01453-z.

Han, Song, Jeff Pool, John Tran, and William J. Dally. 2015. “Learning Both Weights and Connections for Efficient Neural Networks.” *Advances in Neural Information Processing Systems* 2015-Janua: 1135–43.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2015. “Distilling the Knowledge in a Neural Network,” 1–9. http://arxiv.org/abs/1503.02531.

Hu, Fu, Jian Qin, Yixin Li, Ying Liu, and Xianfang Sun. 2021. “Deep Fusion for Energy Consumption Prediction in Additive Manufacturing.” *Procedia CIRP* 104: 1878–83. doi:https://doi.org/10.1016/j.procir.2021.11.317.

Huang, Zehao, and Naiyan Wang. 2017. “Like What You Like: Knowledge Distill via Neuron Selectivity Transfer.” http://arxiv.org/abs/1707.01219.

ISO/ASTM. 2015. “Standard Terminology for Additive Manufacturing – General Principles – Terminology (ISO/ASTM52900).” *International Organization for Standardization: Geneva, Switzerland* i: 1–9.

Kellens, Karel, Raya Mertens, Dimos Paraskevas, Wim Dewulf, and Joost R Duflou. 2017. “Environmental Impact of Additive Manufacturing Processes: Does AM Contribute to a More Sustainable Way of Part Manufacturing?” *Procedia CIRP* 61: 582–87. doi:https://doi.org/10.1016/j.procir.2016.11.153.

Khorram Niaki, Mojtaba, and Fabio Nonino. 2017. “Additive Manufacturing Management: A Review and Future Research Agenda.” *International Journal of Production Research* 55 (5). Taylor & Francis: 1419–39. doi:10.1080/00207543.2016.1229064.

Kusiak, Andrew. 2018. “Smart Manufacturing.” *International Journal of Production Research* 56 (1–2). Taylor & Francis: 508–17. doi:10.1080/00207543.2017.1351644.

Li, Yixin, Fu Hu, Michael Ryan, Ray Wang, and Ying Liu. 2022. “Knowledge Distillation for Energy Consumption Prediction in Additive Manufacturing.” *IFAC-PapersOnLine* 55 (2): 390–95. doi:10.1016/j.ifacol.2022.04.225.

Liu, Bin, Ren Long, and Kuo Chen Chou. 2016. “IDHS-EL: Identifying DNase i Hypersensitive Sites by Fusing Three Different Modes of Pseudo Nucleotide Composition into an Ensemble Learning Framework.” *Bioinformatics* 32 (16): 2411–18. doi:10.1093/bioinformatics/btw186.

Liu, Z Y, C Li, X Y Fang, and Y B Guo. 2018. “Energy Consumption in Additive Manufacturing of Metal Parts.” *Procedia Manufacturing* 26: 834–45. doi:https://doi.org/10.1016/j.promfg.2018.07.104.

Lv, Jingxiang, Tao Peng, Yingfeng Zhang, and Yuchang Wang. 2021. “A Novel Method to Forecast Energy Consumption of Selective Laser Melting Processes.” *International Journal of Production Research* 59 (8). Taylor & Francis: 2375–91. doi:10.1080/00207543.2020.1733126.

Ma, Feng, Hua Zhang, K K B Hon, and Qingshan Gong. 2018. “An Optimization Approach of Selective Laser Sintering Considering Energy Consumption and Material Cost.” *Journal of Cleaner Production* 199: 529–37. doi:https://doi.org/10.1016/j.jclepro.2018.07.185.

Machado, Carla Gonçalves, Mats Peter Winroth, and Elias Hans Dener Ribeiro da Silva. 2020. “Sustainable Manufacturing in Industry 4.0: An Emerging Research Agenda.” *International Journal of Production Research* 58 (5). Taylor & Francis: 1462–84.

doi:10.1080/00207543.2019.1652777.

Majeed, Arfan, Yingfeng Zhang, Shan Ren, Jingxiang Lv, Tao Peng, Saad Waqar, and Enhuai Yin. 2021. “A Big Data-Driven Framework for Sustainable and Smart Additive Manufacturing.” *Robotics and Computer-Integrated Manufacturing* 67 (March 2019). Elsevier Ltd: 102026. doi:10.1016/j.rcim.2020.102026.

Mirzadeh, Seyed Iman, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. “Improved Knowledge Distillation via Teacher Assistant.” *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, 5191–98. doi:10.1609/aaai.v34i04.5963.

Paul, Ratnadeep, and Sam Anand. 2012a. “Process Energy Analysis and Optimization in Selective Laser Sintering.” *Journal of Manufacturing Systems* 31 (4): 429–37. doi:https://doi.org/10.1016/j.jmsy.2012.07.004.

———. 2012b. “Process Energy Analysis and Optimization in Selective Laser Sintering.” *Journal of Manufacturing Systems* 31 (4). The Society of Manufacturing Engineers: 429–37. doi:10.1016/j.jmsy.2012.07.004.

Peng, Tao, Karel Kellens, Renzhong Tang, Chao Chen, and Gang Chen. 2018. “Sustainability of Additive Manufacturing: An Overview on Its Energy Demand and Environmental Impact.” *Additive Manufacturing* 21 (April). Elsevier: 694–704. doi:10.1016/j.addma.2018.04.022.

Phuong, Mary, and Christoph Lampert. 2019. “Distillation-Based Training for Multi-Exit Architectures.” *Proceedings of the IEEE International Conference on Computer Vision* 2019-Octob (308036): 1355–64. doi:10.1109/ICCV.2019.00144.

Polino, Antonio, Razvan Pascanu, and Dan Alistarh. 2018. “Model Compression via Distillation and Quantization.” *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, no. 2015: 1–21.

Qin, Jian, Ying Liu, and Roger Grosvenor. 2017. “A Framework of Energy Consumption Modelling for Additive Manufacturing Using Internet of Things.” *Procedia CIRP* 63: 307–12. doi:https://doi.org/10.1016/j.procir.2017.02.036.

———. 2018. “Multi-Source Data Analytics for AM Energy Consumption Prediction.” *Advanced Engineering Informatics* 38: 840–50. doi:10.1016/j.aei.2018.10.008.

Rai, Rahul, Manoj Kumar Tiwari, Dmitry Ivanov, and Alexandre Dolgui. 2021. “Machine Learning in Manufacturing and Industry 4.0 Applications.” *International Journal of Production Research* 59 (16). Taylor & Francis: 4773–78. doi:10.1080/00207543.2021.1956675.

Reed, R. 1993. “Pruning Algorithms-a Survey.” *IEEE Transactions on Neural Networks* 4 (5): 740–47. doi:10.1109/72.248452.

Romero, Adriana, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. “FitNets: Hints for Thin Deep Nets.” *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–13.

Son, Wonchul, Jaemin Na, Junyong Choi, and Wonjun Hwang. 2020. “Densely Guided Knowledge Distillation Using Multiple Teacher Assistants,” 9395–9404.

http://arxiv.org/abs/2009.08825.

Tian, Wenmeng, Junfeng Ma, and Morteza Alizadeh. 2019. “Energy Consumption Optimization with Geometric Accuracy Consideration for Fused Filament Fabrication Processes.” *International Journal of Advanced Manufacturing Technology* 103 (5–8). The International Journal of Advanced Manufacturing Technology: 3223–33. doi:10.1007/s00170-019-03683-5.

Troussas, Christos, Akrivi Krouska, Cleo Sgouropoulou, and Ioannis Voyiatzis. 2020. “Ensemble Learning Using Fuzzy Weights to Improve Learning Style Identification for Adapted Instructional Routines.” *Entropy* 22 (7). doi:10.3390/e22070735.

Verma, Anoop, and Rahul Rai. 2013. “Energy Efficient Modeling and Optimization of Additive Manufacturing Processes.” *24th International SFF Symposium - An Additive Manufacturing Conference, SFF 2013*, no. 1: 231–41.

Wang, Lin, and Kuk Jin Yoon. 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): 1–40. doi:10.1109/TPAMI.2021.3055564.

Watson, J K, and K M B Taminger. 2018. “A Decision-Support Model for Selecting Additive Manufacturing versus Subtractive Manufacturing Based on Energy Consumption.” *Journal of Cleaner Production* 176: 1316–22. doi:https://doi.org/10.1016/j.jclepro.2015.12.009.

Wong, Kaufui V., and Aldo Hernandez. 2012. “A Review of Additive Manufacturing.” *ISRN Mechanical Engineering* 2012: 1–10. doi:10.5402/2012/208760.

Xie, Qizhe, Minh Thang Luong, Eduard Hovy, and Quoc V. Le. 2020. “Self-Training with Noisy Student Improves Imagenet Classification.” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 10684–95. doi:10.1109/CVPR42600.2020.01070.

Yang, Chenglin, Lingxi Xie, Siyuan Qiao, and Alan L. Yuille. 2019. “Training Deep Neural Networks in Generations: A More Tolerant Teacher Educates Better Students.” *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 5628–35. doi:10.1609/aaai.v33i01.33015628.

Yang, Jing, Xiaoqin Zeng, Shuiming Zhong, and Shengli Wu. 2013. “Effective Neural Network Ensemble Approach for Improving Generalization Performance.” *IEEE Transactions on Neural Networks and Learning Systems* 24 (6): 878–87. doi:10.1109/TNNLS.2013.2246578.

Yang, Yiran, Miao He, and Lin Li. 2020. “Power Consumption Estimation for Mask Image Projection Stereolithography Additive Manufacturing Using Machine Learning Based Approach.” *Journal of Cleaner Production* 251. Elsevier Ltd: 119710. doi:10.1016/j.jclepro.2019.119710.

Yang, Yiran, Lin Li, Yayue Pan, and Zeyi Sun. 2017a. “Energy Consumption Modeling of Stereolithography-Based Additive Manufacturing Toward Environmental Sustainability.” *Journal of Industrial Ecology* 21 (S1): S168–78. doi:https://doi.org/10.1111/jiec.12589.

———. 2017b. “Energy Consumption Modeling of Stereolithography-Based Additive Manufacturing Toward Environmental Sustainability.” *Journal of Industrial Ecology* 21: S168–78. doi:10.1111/jiec.12589.

You, Shan, Chang Xu, Chao Xu, and Dacheng Tao. 2017. “Learning from Multiple Teacher Networks.” *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* Part F1296: 1285–94. doi:10.1145/3097983.3098135.

Zagoruyko, Sergey, and Nikos Komodakis. 2017. “Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer.” *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 1–13.

Zhang, Ying, Tao Xiang, Timothy M. Hospedales, and Huchuan Lu. 2018. “Deep Mutual Learning.” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 4320–28. doi:10.1109/CVPR.2018.00454.

Zhou, Kaile, and Shanlin Yang. 2016. “Understanding Household Energy Consumption Behavior: The Contribution of Energy Big Data Analytics.” *Renewable and Sustainable Energy Reviews* 56. Elsevier Ltd: 810–19. doi:10.1016/j.rser.2015.12.001.