frontiers | Frontiers in Neurorobotics

TYPE Brief Research Report
PUBLISHED 11 January 2023
DOI 10.3389/fnbot.2022.1083835

OPEN ACCESS

EDITED BY
Xun Shen,
Tokyo Institute of Technology, Japan

REVIEWED BY
Xingguo Zhang,
Tokyo University of Agriculture and Technology, Japan
Haoyun Shi,
Great Wall Motor, China

*CORRESPONDENCE
Shacheng Liu
12012011@hnist.edu.cn

RECEIVED 29 October 2022
ACCEPTED 15 November 2022
PUBLISHED 11 January 2023

CITATION
Liu S (2023) Robust robot image classification toward cyber-physical system-based closed-loop package design evaluation.
Front. Neurorobot. 16:1083835.
doi: 10.3389/fnbot.2022.1083835

Robust robot image classification toward cyber-physical system-based closed-loop package design evaluation

Shacheng Liu*

Hunan Institute of Science and Technology, Yueyang, China

The package design can transmit the value of a product to consumers visually and can therefore influence the consumers' decisions. The traditional package design is an open-loop process in which a design can only be evaluated after the products are sent to the market. Thus, the designers cannot refine the design without any helpful advice. In this paper, a robust robot image classification is proposed to help the designers to evaluate their package design and improve their design in a closed-loop process, which is essentially the establishment of a cyber-physical system for the package design. The robust robot image classification adopts the total variation regularization, which ensures that the proposed robot image classification can give the right answers even if it is trained by noisy labels. The robustness against noisy labels is emphasized here since the historical data set of package design evaluations may have some false labels that can be equivalently regarded as disturbed labels from the true labels by noises. To validate the effectiveness of the proposed robot image classification method, experimental data-based validations have been implemented. The results show that the proposed method exhibits much better accuracy in classification compared to the traditional training method when noisy labels are used for the training process.

KEYWORDS

robust classification, robot image classification, noisy labels, total variation regularization, package design, cyber-physical systems

1. Introduction

1.1. Background and motivations

A product's value is first transmitted to the consumers by package design visually. A good design can significantly improve the sale performance of a product. Therefore, it is important for a designer to provide a design from the consumers' view of consumption. As shown in Figure 1, the traditional package design is an open-loop process. A designer must figure out the consumers' preferences according to the market survey.

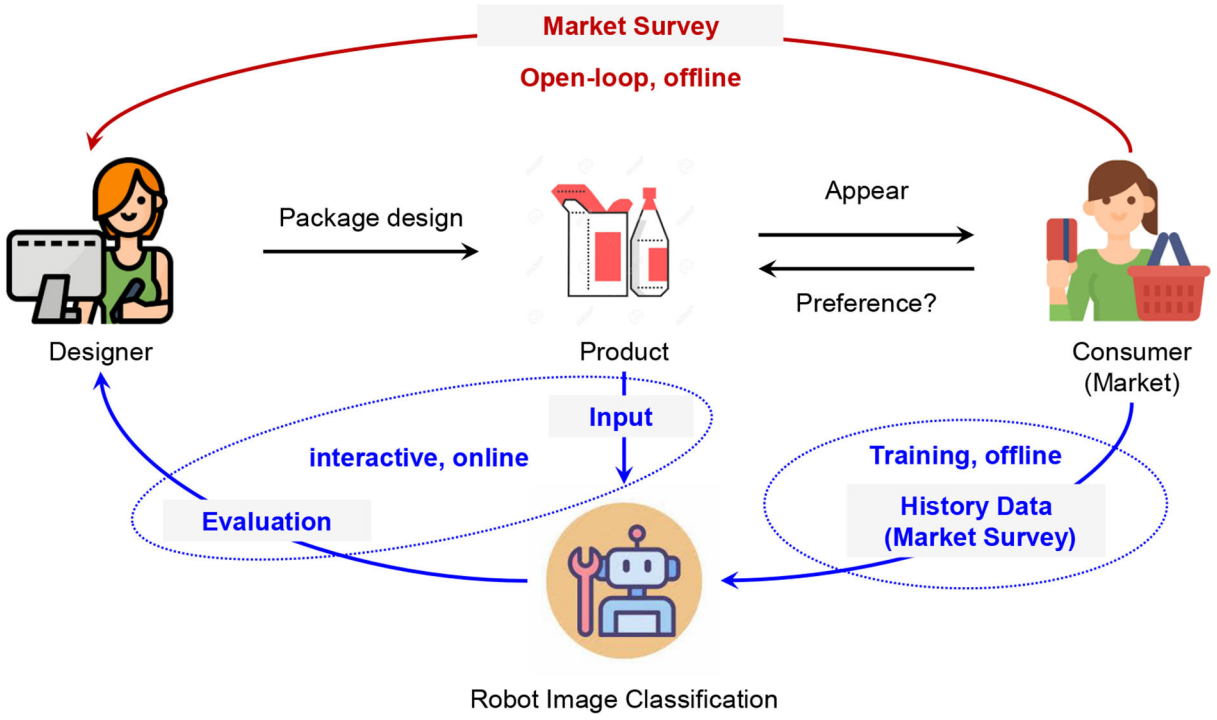

FIGURE 1
Comparison of the processes of traditional open-loop package design and robot-assisted interactive package design.

The evaluation of the package design for the products can only be implemented after sending the product to the market and obtaining the results of the market survey. The designer cannot refine the design with any helpful evaluation which reflects the preferences during the design process. This hurts the efficiency of the design and also brings risk to the sale performance. Robot image classification makes the interactive package design possible. In robot-assisted interactive package design, a robot image classifier can give an evaluation of the design. The evaluation can reflect the market preferences since the classification models are trained by the historical data of the market survey. Deep learning models have been applied to improve the package design evaluation performance (Shi, 2022). According to Zhao et al. (2018), a graphic design-based model is developed to predict the score of a design. According to Jolly et al. (2018), the conventional neural network has been used to classify a type of design. Definitely, the performance will be good with deep learning models if the labels are correctly given. However, there are lots of unclear or incorrect labels in the data set of market surveys (Xia et al., 2022), which can be regarded as noisy labels. Noisy labels can cause severe over-fitting issues in deep learning models. It is very important to improve the robustness of deep learning-based robot image classification for package design evaluation toward noisy labels.

1.2. Related research

Robot classification for package design aims to enable robots or computers to make esthetic decisions about images of the package design in a way of imitating human vision and esthetic thinking (Zhang et al., 2022). The image esthetic quality assessment methods can be categorized into two main streams. The traditional esthetic quality assessment is based on artificial design features. Nowadays, deep learning-based esthetic assessment methods are becoming popular.

In the traditional method of evaluating esthetic quality based on features of artificial designs, the image esthetics are assessed by the two-layer design by an expert. The lower layer has visual features and the higher layer has composition esthetic features. Images can be categorized into high and low esthetics by methods such as support vector machine, which has color matching, the contrast of images, and other features as inputs (Kumar et al., 2019; Wu et al., 2021).

In recent years, with the development of deep learning models, many researchers have started to apply convolutional neural networks to the tasks of image esthetic assessment. For example, according to Wang et al. (2019), convolutional neural networks have been used to specify some high-level abstract features from big data of the image. Furthermore, the structure

of the convolutional neural networks is adapted for image esthetic assessment by Darmawahyuni et al. (2019). However, the above methods assume that the used data for training are clean. However, according to the results of Xia et al. (2022), the classification based on esthetic assessment is different from the classification from the viewpoint of the consumers. The labels that reflect the viewpoint of the consumers can be obtained from the market survey. However, the labels obtained from the market survey may suffer the issue of having uncertainties, which makes the labels noisy. It is important to consider how to improve the accuracy in the establishment of robot image classification for package design when using noisy labels for training.

The classification with noisy labels was first addressed by Angluin and Laird (1988) and has just become a very hot topic in the machine learning community (Goodfellow et al., 2016). The first model for noisy labels, proposed by Angluin and Laird (1988) and Rooyen et al. (2015), is for binary classification, which is named the noise model of random classification. The random classification noise model has been extended to the noise model conditioned on classes in Natarajan et al. (2013), which is also known as the class-conditional noise model. The multi-class case has been developed in recent years (Goldberge and Ben-Reuven, 2017; Patrini et al., 2017). The research on classification with noisy labels has the following main streams:

- For class-conditional noise models, one popular method is to adopt robust loss functions to alleviate the issue caused by noisy labels (Ghosh et al., 2017; Feng et al., 2020; Lyu and Tsang, 2020; Ma et al., 2020). The method using robust loss functions works well under simple noises. When the noise is complex and with the high rate, the method performs very poorly.
- Another kind of method tries to improve the robust against label noise by sampling methods (Malach and Shalev-Shwartz, 2017; Han et al., 2018; Wei et al., 2020). In the sampling process, some samples are rejected to improve the robustness, which can be regarded as a semi-supervised learning process (Nguyen et al., 2020). The sampling methods suffer high computational cost and high model complexity.
- The third kind of method is based on estimating the noise transition matrix based on the assumption of the existence of anchor points (Liu and Tao, 2015; Patrini et al., 2017; Yu et al., 2018). The transition matrix is identifiable only if anchor points exist in all classes. When the transition matrix is obtained, the probability of the true labels can be recovered.

1.3. Contributions of this article

This paper extends the noise transition matrix estimation method into the robot image classification and proposes a robust robot image classification method for package design evaluation. The main contributions of this article are as follows:

- We adopt a problem formulation of the total variation regularization. The solution to the problem of total variation regularization is consistent with the real transition matrix and the probability distribution of the true labels.
- The stochastic gradient descent can be used to approach the solution that recovers the probability distribution of the true labels from the noisy labels.
- For the first time, a noisy label-aware package design evaluator can be established to help the designer improve their compositions in a closed-loop way.

The research of this article is a big step to establish a computer-vision cyber-physic system for package design evaluation since the issues raised by the noisy labels in the training process of a robot image classifier are resolved.

2. Problem description

Let $x \in \mathcal{X}$ be a package design composition. Let $y \in \mathcal{Y} = \{1, ..., C\}$ be the discrete score of the package design, which reflects the consumers' view toward the design. Note that $\mathcal{X}$ and $\mathcal{Y}$ represent the design space and score space, respectively. A high score of y means a high satisfaction with the design. The positive integer C represents the highest score. Note that the evaluation is stochastic which obeys a true probability distribution

$$p_{\text{true}}(y|x) := \Pr\{y|x\}, \tag{1}$$

for every $x \in \mathcal{X}$, different consumers have different opinions. Let $\mathbb{P}(\mathcal{Y})$ be the set of all possible probability measures defined on $\mathcal{Y}$. Let $\mathcal{E}_{\text{true}} : \mathcal{X} \to \mathbb{P}(\mathcal{Y})$ be the underlying function that outputs the probability distribution of true score by the consumer for a given design $x \in \mathcal{X}$, namely, $\mathcal{E}_{\text{true}} = p_{\text{true}}(y|x)$. In this article, we call $\mathcal{E}_{\text{true}}(\cdot)$ a real evaluator. Essentially, the evaluator $\mathcal{E}_{\text{true}}(\cdot)$ can be regarded as a classifier that gives a label of class $y \in \mathcal{Y}$ to each design. Although the basic function in mathematics is the same, we adopt the terminology "evaluator" here since there is an order that larger y means a better design.

Note that the real evaluator is unknown. It is necessary to construct an estimated evaluator from the available data set $\mathcal{D}_T = \{x(t), \tilde{y}(t)\}_{t=1,...,T}$. Here, we use $\tilde{y}(t)$ to denote the noisy score of the design $x(t)$. The noisy score $\tilde{y}(t)$ represents the randomness in the process of collecting data by market survey. For the extraction of the design $x(t)$, the following assumption holds throughout the rest of the paper.

Assumption 1. *The design $x(t)$ is extracted identically and independently from the design space.*

Although there is a true $y(t)$ for every $x(t)$, the observed score $\tilde{y}(t)$ is disturbed by noise. Here, we only consider the class-conditional noise. By adopting the class-conditional noise model, the following assumption on $\tilde{y}(t)$ holds throughout the paper.

Assumption 2. *The noisy score $\tilde{y}(t)$ does not relate to x and only depends on the true score $y(t)$. Namely,*

$$\Pr\{\tilde{y}(t)|y(t), x(t)\} = \Pr\{\tilde{y}(t)|y(t)\},\ \forall t = 1, ..., T. \tag{2}$$

Since the available data set $\mathcal{D}_T$ has noisy score $\tilde{y}(t)$, the estimated evaluator $\tilde{\mathcal{E}} : \mathcal{X} \to \mathbb{P}(\mathcal{Y})$ constructed by directly using $\mathcal{D}_T$ will give the noisy probability distribution $\tilde{p}(\tilde{y}|x)$ instead of true probability distribution $p(y|x)$, where the noisy probability distribution $\tilde{p}(\tilde{y}|x)$ is defined by

$$\tilde{p}(\tilde{y}|x) : = \Pr\{\tilde{y}|x\}. \tag{3}$$

It is necessary to improve the robustness of the estimated evaluator and guide the evaluator to give the probability distribution near the true probability distribution even it is trained by using noisy data set $\mathcal{D}_T$.

Then, the problem we address in this article is summarized as follows.

Problem 1. *Suppose that data set $\mathcal{D}_T$ has been obtained from the market survey and Assumptions 1 and 2 hold. To find a robust evaluator $\mathcal{E}_\mathrm{r}$ by solving the following optimization problem.*

$$\min_{\mathcal{E}} \mathbb{D}(\mathcal{E}, \mathcal{E}_\mathrm{true}). \tag{4}$$

The challenging issue of solving Problem 1 is that the true probability distribution is unknown, which makes it difficult to find the direction of modifying the data set $\mathcal{D}_T$ for training.

3. Robust robot image classification

3.1. Noisy transition matrix

With Assumption 2, it is able to establish the relationship between noisy label posterior $\tilde{p}(\tilde{y}|x)$ and the true label posterior $p(y|x)$ as

$$\tilde{p}(\tilde{y}|x) = \sum_{y=1}^{C} p_y(\tilde{y}|y)p(y|x), \tag{5}$$

where $p_y(\tilde{y}|y) : = \Pr\{\tilde{y}|y\}$. The true label posterior $p(y|x)$ is essentially a vector-value function from $\mathcal{X}$ to $[0, 1]^C$, which is written as

$$p(y|x) = [\Pr\{y = 1|x\}, ..., \Pr\{y = C|x\}]^\mathsf{T}. \tag{6}$$

The noisy label posterior $\tilde{p}(\tilde{y}|x)$ is also a vector-valued function written as

$$\tilde{p}(\tilde{y}|x) = [\Pr\{\tilde{y} = 1|x\}, ..., \Pr\{\tilde{y} = C|x\}]^\mathsf{T}. \tag{7}$$

On the other hand, $p_y(\tilde{y}|y)$ is essentially a noisy transition matrix written as

$$p_y(\tilde{y}|y) = \begin{bmatrix} \Pr\{\tilde{y} = 1|y = 1\} & ... & \Pr\{\tilde{y} = C|y = 1\} \\ ... & ... & ... \\ \Pr\{\tilde{y} = 1|y = C\} & ... & \Pr\{\tilde{y} = C|y = C\} \end{bmatrix}. \tag{8}$$

Let $\boldsymbol{T}_\mathrm{n}$ be the notation of the noisy transition matrix instead of $p_y(\tilde{y}|y)$ and $\mathcal{T}_\mathrm{n} \subseteq [0, 1]^{C\times C}$ be the set of all possible $\boldsymbol{T}_\mathrm{n}$. Then, we can rewrite Equation (5) by

$$\tilde{p}(\tilde{y}|x) = \boldsymbol{T}_\mathrm{n}^\mathsf{T} p(y|x). \tag{9}$$

If the noisy transition matrix $\boldsymbol{T}_\mathrm{n}$ is available and $p(y|x)$ is identifiable with $\tilde{p}(\tilde{y}|x)$ (every $p(y|x)$ generate distinct $\tilde{p}(\tilde{y}|x)$), we can recover $p(y|x)$ from $\tilde{p}(\tilde{y}|x)$ by using $\boldsymbol{T}_\mathrm{n}$ according to Equation (9). It is reasonable to assume that the true label posterior $p(y|x)$ can be approximated by a parameterized model $\hat{p}(y|x, \theta)$ characterized by $\theta \in \Theta$. Namely, there exists one $\theta^* \in \Theta$ such that $\hat{p}(y|x, \theta^*) = p(y|x)$ for every $y \in \mathcal{Y}$ and $x \in \mathcal{X}$. Suppose that $\hat{p}(y|x, \theta)$ is differentiable with θ. Note that it is possible to represent $\hat{p}(y|x, \theta)$ by an expressive deep neural network characterized by θ.

For the learning objective, we adopt the expected Kullback-Leibler (KL) divergence, which is a standard objective. The expected KL divergence concerned here is constructed as follows:

$$L_{0,\mathrm{true}}(\theta) : = \mathbb{E}_{x\sim p_x(x)}\{D_\mathrm{KL}(\tilde{p}(\tilde{y}|x), \boldsymbol{T}_\mathrm{n}^\mathsf{T}\hat{p}(y|x, \theta))\}, \tag{10}$$

where $p_x(x)$ is the probability density function defined on $\mathcal{X}$. Note that $L_{0,\mathrm{true}}(\theta)$ has connections to the cross-entropy loss which is defined as

$$\begin{aligned} L_{\mathrm{ce,true}}(\theta) &: = \mathbb{E}_{(x,\tilde{y})\sim p(x,\tilde{y})}\{-\log(\boldsymbol{T}_\mathrm{n}^\mathsf{T}\hat{p}(y|x, \theta))\} \\ &= L_{0,\mathrm{true}}(\theta) + H(\tilde{y}|x), \end{aligned} \tag{11}$$

where $H(\tilde{y}|x)$ is the conditional entropy, namely, the entropy of $\tilde{y}$ under x. Note that $H(\tilde{y}|x)$ is a constant with respect to θ and $L_{\mathrm{ce,true}}(\theta)$ is minimized if and only if $L_{0,\mathrm{true}}(\theta)$. If $L_{\mathrm{ce,true}}(\theta)$ is optimized, we can say that $\boldsymbol{T}^\mathsf{T}\hat{p}(y|x) = \boldsymbol{T}^\mathsf{T}p(y|x)$ and $\hat{p}(y|x) = p(y|x)$. For practice, although it is only possible to empirically estimate $L_{\mathrm{ce,true}}(\theta)$, we can still prove the asymptomatic convergence. Namely, as the number of samples increases, with probability 1, we have $\boldsymbol{T}^\mathsf{T}\hat{p}(y|x) \to \boldsymbol{T}^\mathsf{T}p(y|x)$ and $\hat{p}(y|x) \to p(y|x)$.

3.2. Classification without noisy transition matrix

In our case, both $\boldsymbol{T}_\mathrm{n}$ and $p(y|x)$ are not available. Therefore, $\boldsymbol{T}_\mathrm{n}$ and $p(y|x)$ are partially identifiable which means that there might exist multiple pairs of $\boldsymbol{T}_\mathrm{n}$ and $p(y|x)$ with the same $\tilde{p}(\tilde{y}|x)$.

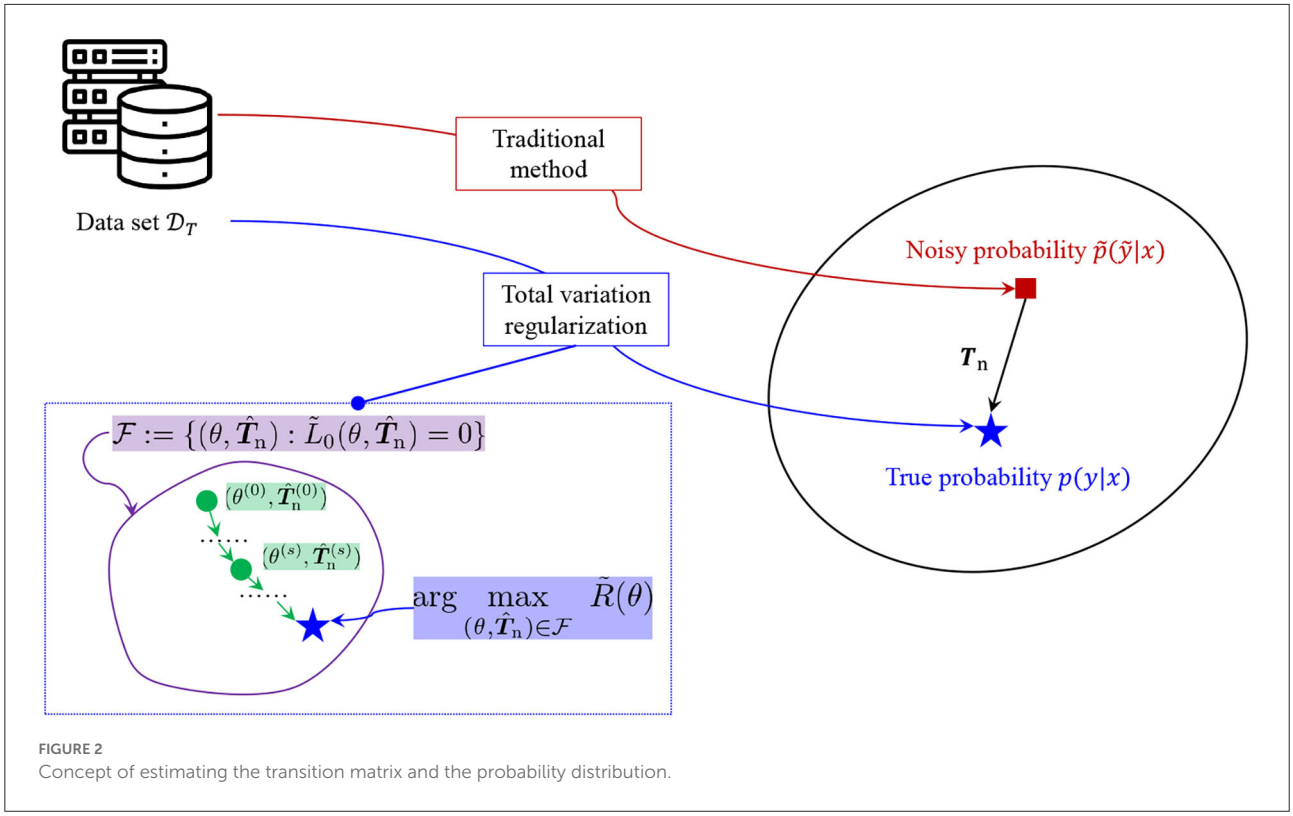

FIGURE 2
Concept of estimating the transition matrix and the probability distribution.

Then, it is impossible to identify the true T_n and $p(y|x)$ without any further assumptions.

Let $\hat{T}_n \in \mathcal{T}_n$ be an estimated noisy transition matrix. Then, the KL divergence is written as

$$L_0(\theta, \hat{T}_n) := \mathbb{E}_{x \sim p_x(x)}\{D_{KL}(\tilde{p}(\tilde{y}|x), \hat{T}_n^{\mathsf{T}} \hat{p}(y|x, \theta))\}. \quad (12)$$

The cross-entropy loss is then revised as

$$\begin{aligned} L_{ce}(\theta, \hat{T}_n) &:= \mathbb{E}_{(x,\tilde{y}) \sim p(x,\tilde{y})}\{-\log(\hat{T}_n^{\mathsf{T}} \hat{p}(y|x, \theta))\} \\ &= L_0(\theta, \hat{T}_n) + H(\tilde{y}|x), \end{aligned} \quad (13)$$

In the case without noisy transition matrix, we should find θ and $\hat{T}_n$ to minimize $L_{ce}(\theta, \hat{T}_n)$ or equivalently make $L_0(\theta, \hat{T}_n) = 0$. As the same with the case with noisy transition matrix, we can empirically obtain the estimation of $L_{ce}(\theta, \hat{T}_n)$ based on samples of $x, \tilde{y}$-pairs and also optimize it by adjusting θ and $\hat{T}_n$. Although it is possible to ensure the convergence of $\hat{T}_n \hat{p}(y|x)$ to $T_n p(y|x)$ with infinite sample size, the convergence of $\hat{p}(y|x)$ to $p(y|x)$ may not be guaranteed if $p(y|x)$ is not identifiable with $\tilde{p}(\tilde{y}|x)$ (Patrini et al., 2017).

Most of the existing methods adopt a two-step method. In the first step, the noisy transition matrix is estimated. Then, the estimated noisy transition matrix is used for neural network training. The estimation of the noisy transition matrix is based on the anchor points (Yu et al., 2018; Xia et al., 2022), which are defined as follows.

Definition 1. *An point x is called an anchor point for class $i = 1, ..., C$ if* $\Pr\{y = i|x\} = 1$.

For an anchor point x for class $i = 1, ..., C$, Equation (9) can be transformed to

$$\tilde{p}(\tilde{y}|x) = T_n p(y|x) = T_{n,i}. \quad (14)$$

Then, it is possible to estimate T_n based on the estimation of $\tilde{p}(\tilde{y}|x)$ from the data with noisy labels. Note that the two-step method cannot be applied if anchor points can not be obtained from data clearly. However, estimating the noisy label posteriors has a dramatically worse overfitting issue than estimating the true label posteriors. Therefore, the estimation of the transition matrix suffers from the estimated noisy label distributions, which are not accurate and the performance deteriorates sharply.

Transition matrix estimation also suffers from poorly estimated noisy label posteriors and the performance deteriorates sharply.

3.3. Construction of equivalence class and partial order

It is necessary to deeply investigate the generating process of the class-conditional label noise and establish equivalence class

and partial order for noisy transition matrix $\boldsymbol{T}_{\mathrm{n}}$. According to Zhang et al. (2021), partial order of the transition matrices led by the contraction property of the stochastic is shown to be able to find the true class posterior. Here, we summarize the idea of constructing equivalence class and partial order for noisy transition matrix, which is the theoretical basis of our proposed robust robot image classification.

The definition of transition matrix equivalence is as follows.

Definition 2. *Transition matrix equivalence is essentially an equivalence relation of a pair of matrices with an ordered product. We say* $(\boldsymbol{U}, \boldsymbol{V}) \sim (\boldsymbol{U}', \boldsymbol{V}')$ *if* $\boldsymbol{U}\boldsymbol{V} = \boldsymbol{U}'\boldsymbol{V}'$. *Besides, for a given matrix* $\boldsymbol{W}$, *the equivalence class associated with* $\boldsymbol{W}$ *is defined by*

$$[W] := \{(\boldsymbol{U}, \boldsymbol{V}) : \boldsymbol{U}\boldsymbol{V} = \boldsymbol{W}\}. \tag{15}$$

For noisy transition matrix $\boldsymbol{T}_{\mathrm{n}}$, there is also an equivalence class $[T]$. For a pair $(\boldsymbol{U}, \boldsymbol{V}) \in [\boldsymbol{T}]$, we can form an optimal solution that minimizes (Equations 12, 13) by setting

$$\hat{\boldsymbol{T}}_{\mathrm{n}} = \boldsymbol{V}, \tag{16}$$

and

$$\hat{p}(y|x, \theta) = \boldsymbol{U}^{\mathsf{T}} p(y|x). \tag{17}$$

Note that the potential optimal solutions are infinite and only $(\boldsymbol{I}, \boldsymbol{T})$ is the true pair for our interest. Thus, it is important to investigate other conditions to direct us to the pair $(\boldsymbol{I}, \boldsymbol{T})$ among infinite optimal solutions.

For any given optimal solution $\hat{\boldsymbol{T}}_{\mathrm{n}}$, $\hat{p}(y|x, \theta)$ of Equations (12) and (13), there exists a matrix $\boldsymbol{U}$ that satisfies $\hat{p}(y|x) = \boldsymbol{U}^{\mathsf{T}} p(y|x)$ if anchor points exists in data set $\mathcal{D}_T$ for each class i (Zhang et al., 2021). Thus, we have the following assumption throughout the paper.

Assumption 3. *The obtained data set* $\mathcal{D}_T$ *has at least one anchor point for each class* $i = 1, ..., C$.

With the absolute existence of anchor points as stated in Assumption 3, it is able to find a condition to break the transition matrix equivalence and obtain the desired pair $(\boldsymbol{I}, \boldsymbol{T})$.

Let $\boldsymbol{v}$ and $\boldsymbol{w}$ be two categorical probabilities. For any $\boldsymbol{v}$ and $\boldsymbol{w}$, the total variation distance is defined by

Definition 3. *The total variation distance between two categorical probabilities is*

$$\mathbb{D}_{\mathrm{TV}}(\boldsymbol{v}, \boldsymbol{w}) := \frac{\|\boldsymbol{v} - \boldsymbol{w}\|_1}{2}, \tag{18}$$

where $\|\cdot\|_1$ *is the* ℓ_1 *norm.*

Based on the theory of Markov chains (Moral et al., 2003), we have that $\boldsymbol{v} \mapsto \boldsymbol{U}^{\mathsf{T}}\boldsymbol{v}$ is a contraction mapping with the total variation distance, namely,

$$\mathbb{D}_{\mathrm{TV}}(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{v}, \boldsymbol{U}^{\mathsf{T}}\boldsymbol{w}) \leq \mathbb{D}_{\mathrm{TV}}(\boldsymbol{v}, \boldsymbol{w}), \forall \boldsymbol{v}, \boldsymbol{w}, \forall \boldsymbol{U} \in \mathcal{T}_{\mathrm{n}}. \tag{19}$$

Inputs: data set $\mathcal{D}_T$; step size α; exponential decay rates β_1, β_2; update rate ϵ;
Initialization: initial parameter vector $\hat{\theta}^{(0)}$; initial noise transition matrix $\hat{\boldsymbol{T}}_{\mathrm{n}}^{(0)}$; initial first moment vectors $m_\theta^{(0)} = 0, m_{\boldsymbol{T}_{\mathrm{n}}}^{(0)} = 0$; initial second moment vectors $v_\theta^{(0)} = 0, v_{\boldsymbol{T}_{\mathrm{n}}}^{(0)} = 0$;

for $s = 1, 2, 3, ..., S$ **do**

randomly extract $\mathcal{D}_s \subset \mathcal{D}_T$ (update $\mathcal{D}_s$ in every iteration)

get gradients $g_\theta^{(s)} = \nabla_\theta \mathcal{L}(\hat{\theta}^{(s-1)}, \hat{\boldsymbol{T}}_{\mathrm{n}}^{(s-1)})$ and $g_{\boldsymbol{T}_{\mathrm{n}}}^{(s)} = \nabla_{\boldsymbol{T}_{\mathrm{n}}} \mathcal{L}(\hat{\theta}^{(s-1)}, \hat{\boldsymbol{T}}_{\mathrm{n}}^{(s-1)})$ based on $\mathcal{D}_s$

calculate prior first moment estimate $m_\theta^{(s)} = \beta_1 m_\theta^{(s-1)} + (1 - \beta_1) g_\theta^{(s)}$

calculate prior first moment estimate $m_{\boldsymbol{T}_{\mathrm{n}}}^{(s)} = \beta_1 m_{\boldsymbol{T}_{\mathrm{n}}}^{(s-1)} + (1 - \beta_1) g_{\boldsymbol{T}_{\mathrm{n}}}^{(s)}$

calculate prior second raw moment estimate $v_\theta^{(s)} = \beta_2 v_\theta^{(s-1)} + (1 - \beta_2)(g_\theta^{(s)})^2$

calculate prior second raw moment estimate $v_{\boldsymbol{T}_{\mathrm{n}}}^{(s)} = \beta_2 v_{\boldsymbol{T}_{\mathrm{n}}}^{(s-1)} + (1 - \beta_2)(g_{\boldsymbol{T}_{\mathrm{n}}}^{(s)})^2$

calculate corrected first moment estimate $\hat{m}_\theta^{(s)} = m_\theta^{(s)}/(1 - \beta_1^s)$, $\hat{m}_{\boldsymbol{T}_{\mathrm{n}}}^{(s)} = m_{\boldsymbol{T}_{\mathrm{n}}}^{(s)}/(1 - \beta_1^s)$

calculate corrected second moment estimate $\hat{v}_\theta^{(s)} = v_\theta^{(s)}/(1 - \beta_2^s)$, $\hat{v}_{\boldsymbol{T}_{\mathrm{n}}}^{(s)} = v_{\boldsymbol{T}_{\mathrm{n}}}^{(s)}/(1 - \beta_2^s)$

update parameter $\hat{\theta}^{(s)} = \hat{\theta}^{(s-1)} - \alpha \hat{m}_\theta^{(s)}/\sqrt{\hat{v}_\theta^{(s)} + \epsilon}$

update parameter $\hat{\boldsymbol{T}}_{\mathrm{n}}^{(s)} = \hat{\boldsymbol{T}}_{\mathrm{n}}^{(s-1)} - \alpha \hat{m}_{\boldsymbol{T}_{\mathrm{n}}}^{(s)}/\sqrt{\hat{v}_{\boldsymbol{T}_{\mathrm{n}}}^{(s)} + \epsilon}$

end for

Output: solution $\hat{\theta}^{(S)}$ and $\hat{\boldsymbol{T}}_{\mathrm{n}}^{(S)}$

Algorithm 1. One-step algorithm for estimating $\hat{T}_{\mathrm{n}}$ and $\tilde{p}(y|x, \theta)$ based on stochastic gradient descent.

With the above discussions, we can define partial order in the equivalence class $[\boldsymbol{T}_{\mathrm{n}}]$ as follows.

Definition 4. *The transition matrix partial order by the total variation distance is expressed as*

$$(\boldsymbol{U}, \boldsymbol{V}) \preceq (\boldsymbol{U}', \boldsymbol{v}') \Leftrightarrow \mathbb{D}_{TV}(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{v}, \boldsymbol{U}^{\mathsf{T}}\boldsymbol{w}) \leq \mathbb{D}_{TV}(\boldsymbol{U}'^{\mathsf{T}}\boldsymbol{v}, \boldsymbol{U}'^{\mathsf{T}}\boldsymbol{w}), \forall \boldsymbol{v}, \boldsymbol{w}. \tag{20}$$

Note that $(\boldsymbol{I}, \boldsymbol{T})$ is the unique element for the greatest total variation (Zhang et al., 2021). Therefore, it is able to find $(\boldsymbol{I}, \boldsymbol{T})$ by gradually increasing the total variation.

3.4. Total variation regularization

First, we define the expected total variation distance by

$$R(\theta) := \mathbb{E}_{x_1 \sim p(x)} \mathbb{E}_{x_2 \sim p(x)} \{\mathbb{D}_{\mathrm{TV}}(\hat{p}_1, \hat{p}_2)\}, \tag{21}$$

where $\hat{p}_i := \hat{p}(y|x = x_i, \theta), i = 1, 2$. We summarize Theorem 2 in Zhang et al. (2021) here.

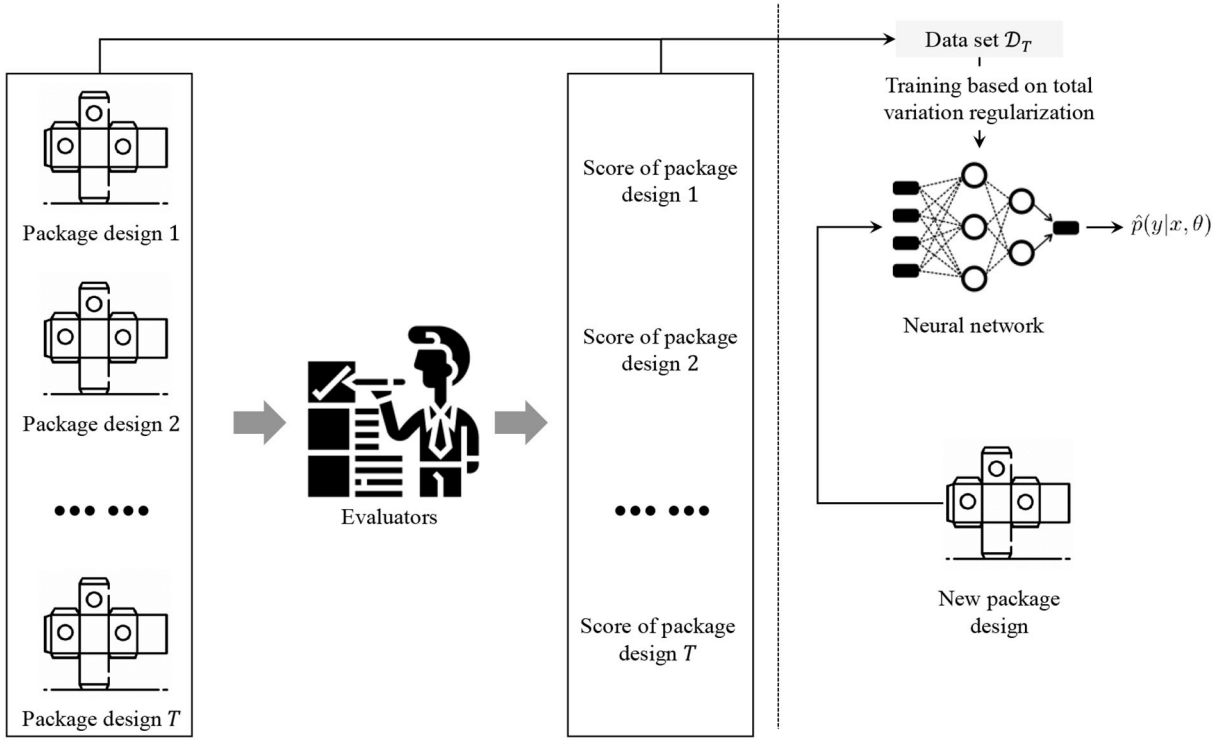

FIGURE 3
Process of implementation of the proposed robust robot classification for evaluating package design.

Theorem 1. *Suppose Assumption 3 holds. Let* $\tilde{L}_0(\theta, \hat{\boldsymbol{T}}_\mathrm{n})$ *and* $\tilde{R}(\theta)$ *be the empirical estimates of* $L_0(\theta, \hat{\boldsymbol{T}}_\mathrm{n})$ *and* $R(\theta)$ *by using data set* $\mathcal{D}_T$*, respectively. Suppose* Θ *is compact. Let* $\tilde{\theta}, \tilde{\boldsymbol{T}}_\mathrm{n}$ *be an optimal solution of the following optimization problem:*

$$\max_{\theta\in\Theta,\, \hat{\boldsymbol{T}}_\mathrm{n}\in\mathcal{T}_\mathrm{n}} \tilde{R}(\theta) \text{ s.t. } \tilde{L}_0(\theta, \hat{\boldsymbol{T}}_\mathrm{n}) = 0. \tag{22}$$

Then, $\tilde{\boldsymbol{T}}_\mathrm{n}$ *is a consistent estimator of* $\boldsymbol{T}_\mathrm{n}$*, and* $\hat{p}(y|x, \tilde{\theta}) \to p(y|x)$ *with probability 1 as* $T \to \infty$.

Theorem 1 claims that we can obtain a consistent estimate of noisy transition matrix by solving (Equation 22). The true probability distribution can also be obtained. Note that the constrained problem (Equation 22) can be solved by introducing Lagrangian

$$\mathcal{L}(\theta, \hat{\boldsymbol{T}}_\mathrm{n}) := \tilde{L}_0(\theta, \hat{\boldsymbol{T}}_\mathrm{n}) - \lambda\tilde{R}(\theta), \tag{23}$$

where $\lambda \in \mathbb{R}^+$ is a positive number that controls the importance of the regularization term. Therefore, the unconstrained problem (Equation 23) is called the optimization problem with total variation regularization.

3.5. Proposed algorithm

Then, we present the algorithms for estimating the transition matrix $\boldsymbol{T}_\mathrm{n}$ and also the probability distribution simultaneously. The concept of the simultaneous estimation is illustrated in Figure 2.

Note that $\mathcal{L}(\theta, \hat{\boldsymbol{T}}_\mathrm{n})$ is differentiable with respect to $\hat{\boldsymbol{T}}_\mathrm{n}$. Therefore, it is possible to use gradient-based optimization to find a local minimum for $\hat{\boldsymbol{T}}_\mathrm{n}$ and $\tilde{p}(y|x, \theta)$. To make sure that $\hat{\boldsymbol{T}}_\mathrm{n} \in \mathcal{T}_\mathrm{n}$, we can use softmax to an unconstrained matrix and optimize $\mathcal{L}(\theta, \hat{\boldsymbol{T}}_\mathrm{n})$ by stochastic gradient descent. The proposed algorithm is adapted from Adam algorithm (Kingma and Ba, 2015) and is summarized as follows.

Remark 1. *Algorithm 1 solves Problem 1 under Assumptions 1, 2, and 3 by making* $\mathbb{D}(\mathcal{E}, \mathcal{E}_{true}) \to 0$ *with probability 1 if* $S, T \to \infty$.

The proposed method can be regarded as a perfect version of the clustering-based method. In the clustering algorithm-based method, some of the data with wrong labels are forbidden. However, there is no existing clustering algorithm-based method that can make sure that all the data with wrong labels are forbidden and all the data with correct labels remain, which can be ensured by the proposed method if Assumptions 1, 2, and 3 hold.

FIGURE 4
Examples of experimental data set: **(A)** one example in class 1 (with score 1); **(B)** one example in class 2 (with score 2); **(C)** one example in class 3 (with score 3); **(D)** one example in class 4 (with score 4); **(E)** one example in class 5 (with score 5).

FIGURE 5
Illustrative examples of the results: **(A)** one design with a score of 5 in the test set; **(B)** one design with a score of 3 in the test set.

3.6. Implementation of the proposed robust robot classification

The process of implementing the proposed robust robot classification for evaluating package design is illustrated in Figure 3. First, we construct the data set $\mathcal{D}_T$ to train the neural networks. A total of T package designs are collected with different levels. Then, several evaluators are asked to give labels to each package design. The evaluators make the labels from the viewpoint of a customer. Then, total variation regularization is applied to data set $\mathcal{D}_T$ to train neural networks that output the estimated probability distribution. For any given package design, with the outputted probability distribution by the trained neural networks, the score or the classification will be determined by

$$i(x) = \arg\max_i \hat{p}(y = i|x, \theta). \tag{24}$$

4. Validations

In this section, the results of experimental data are presented to validate the effectiveness of the total variation regularization-based robust robot classification method. The results show that the proposed method can improve the accuracy of the classification even using noisy labels compared to the traditional method.

4.1. Experimental data set and methods

A total of 5,000 different package designs have been collected. All package designs are categorized into six classes, namely, $C = 5$. In every class, there are 1,000 samples. The categorization has been implemented by some experienced evaluators. In this experiment, we regard the label given by experienced evaluators as real ones and the noisy labels are generated by using the following kinds of label noises:

- (Pair.) defines the pair flipping noise for labeling, which is introduced by Han et al. (2018);
- (Symm.) represents the symmetric noise for labeling, which is introduced by Patrini et al. (2017);
- (Rand.) is random noise generated by Dirichlet distribution mixing with the identity matrix.

Figure 4 shows five examples of experimental data sets from five different classes. It is reasonable to use the labels given by experienced evaluators since experienced evaluators can give relatively precise labels according to their rich experience in the market and package design. In addition, 70% of the data set has been used for training and the rest is for testing. The data for training have noisy labels and the data for testing are with true labels.

TABLE 1 Accuracy (%) on test set.

Noise rate	Methods	Symm.	Pair	Rand
15%	Traditional	94.11	87.24	43.71
	Proposed	99.42	99.37	98.02
	Baseline	99.67	99.62	99.77
25%	Traditional	91.98	83.57	36.22
	Proposed	99.38	99.27	97.66
	Baseline	99.67	99.62	99.77
35%	Traditional	90.65	81.73	32.98
	Proposed	99.27	99.15	95.32
	Baseline	99.67	99.62	99.77
45%	Traditional	88.91	78.31	29.047
	Proposed	99.26	99.12	94.29
	Baseline	99.67	99.62	99.77
55%	Traditional	87.29	77.12	26.97
	Proposed	99.21	99.05	93.91
	Baseline	99.67	99.62	99.77

In this validation, sequential convolutional neural network (CNN) has been used for classifier models.

4.2. Results and discussions

Figure 5 gives two illustrative examples of the validation results with random noise whose noise rate is 35%. (Prop.) is short for the proposed method. (Trad.) is short for the traditional method in which noisy labels are used for training as true labels. The proposed method gives results that are consistent with the true labels despite using noisy labels. On the other hand, the traditional method gives results different from the true labels. Since true labels and noisy labels both exist, the traditional method gets confused and sometimes gives results that are not consistent with both.

In Table 1, the test accuracy is reported. The traditional method uses noisy labels for training. The baseline is obtained by using the true labels for training. It is obvious that the proposed method outperforms the traditional method in terms of accuracy and shows very close accuracy with the baseline. As the percentage of noise rate increases, the accuracy of the traditional method decreases dramatically while the proposed method only has a very slight deterioration in accuracy. For the types of noise, even with random noise, the proposed method still gives very accurate results. The above results show the effectiveness of the proposed method.

5. Conclusion

In this paper, a novel robust robot image classification method for package design evaluation has been introduced. The proposed method can give high accuracy in classification even with noisy labels in the training process. In the proposed method, the loss function for training is total variation regularization whose optimal solution is consistent with the true probability distribution of the labels. The proposed method has been validated by experimental data and it exhibits outperformed accuracy compared to the traditional method. With the proposed robot image classification, it is possible to establish a close-loop package design process, in which the designers can use the robot to help them improve their design.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Angluin, D., and Laird, P. (1988). Learning from noisy examples. *Mach. Learn.* 2, 343–370. doi: 10.1007/BF00116829

Darmawahyuni, A., Nurmaini, S., Sukemi, Caesarendra, W., Bhayyu, V., Naufal Rachmatullah, M., et al. (2019). Deep learning with a recurrent network structure in the sequence modeling of imbalanced data for ECG-rhythm classfier. *Algorithms* 12, 1–12. doi: 10.3390/a12060118

Feng, L., Shu, S., Lin, Z., Lv, F., Li, L., and An, B. (2020). "Can cross entropy loss be robust to label noise," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence* (Yokohama), 2206–2212.

Ghosh, A., Kumar, H., and Sastry, P. (2017). "Robust loss functions under label noise for deep neural networks," in *Proceedings of the Theirty-FIrst AAAI Conference on Artificial Intelligence* (San Francisco, CA), 1919–1925.

Goldberge, J., and Ben-Reuven, E. (2017). "Training deep neural networks using a noise adaptation layer," in *Proceedings of International Conference of Machine Learning* (Sydney).

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep Learning*. Cambridge: MIT Press.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., et al. (2018). "Co-teaching: robust training of deep neural networks with extremely noisy labels," in *Proceedings of Advances in Neural Information Processing Systems* (Montréal, QC), 8527–8537.

Jolly, S., Iwana, B., Kuroki, R., and Uchida, S. (2018). "How do convolutional neural networks learn design," in *Proceedings of 24th International Conference on Pattern Recognition* (Beijing), 1085–1090.

Kingma, D., and Ba, J. (2015). "Adam: a method for stochastic optimization," in *Proceedings of 3rd International Conference on Learning Representations* (San Diego, CA).

Kumar, B., Naren, J., Vithya, G., and Prahathish, K. (2019). A novel architecture based on deep learning for scene image recognition. *Int. J. Psychosoc. Rehabil.* 23, 400–404. doi: 10.37200/IJPR/V23I1/PR190251

Liu, T., and Tao, D. (2015). Classification with noisy labels by importance reweighting. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 447–461. doi: 10.1109/TPAMI.2015.2456899

Lyu, Y., and Tsang, I. (2020). "Curriculum loss: robust learning and generalization against label corruption," in *Proceedings of the International Conference on Learning Representations* (Addis Ababa).

Ma, X., Huang, H., Wang, Y., Romano, S., Erfani, S., and Bailey, J. (2020). "Normalized loss functions for deep learning with noisy labels," in *Proceedings of International Conference of Machine Learning* (Vienna).

Malach, E., and Shalev-Shwartz, S. (2017). "Decoupling "when to update" from "how to update,"" in *Proceedings of Advances in Neural Information Processing Systems* (Long Beach, CA), 960–970.

Moral, P. D., Ledoux, M., and Miclo, L. (2003). On contraction properties of markov kernels. *Probabil. Theory Related Fields* 126, 395–420. doi: 10.1007/s00440-003-0270-6

Natarajan, N., Dhillon, I., Ravikumar, P., and Tewari, A. (2013). "Learning with noisy labels," in *Proceedings of Advances in Neural Information Processing Systems* (Lake Tahoe), 1196–1204.

Nguyen, D., Mummadi, C., Ngo, T., Nguyen, T., Beggel, L., and Brox, T. (2020). "Self: Learning to filter noisy labels with self-ensembling," in *Proceedings of the International Conference on Learning Representations* (Addis Ababa).

Patrini, G., Rozza, A., Menon, A. K., Nock, R., and Qu, L. (2017). "Making deep neural networks robust to label noise: a loss correction approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 1944–1952.

Rooyen, B. V., Menon, A., and Williamson, R. (2015). "Learning with symmetric label noise: the importance of being unhinged," in *Proceedings of Advances in Neural Information Processing Systems* (Montréal, QC), 10–18.

Shi, L. (2022). Design of packaging design evaluation architecture based o deep learning. *Sci. Program.* 2022, 4469495. doi: 10.1155/2022/4469495

Wang, Q., Du, P., Yang, J., Wang, G., Lei, J., and Hou, C. (2019). Transferred deep learning based waveform recognition for cognitive passive radar. *Signal Process.* 155, 259–267. doi: 10.1016/j.sigpro.2018.09.038

Wei, H., Feng, L., Chen, X., and An, B. (2020). "Combating noisy labels by agreement: a joint training method with co-regularization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 13726–13735.

Wu, D., Zhang, C., Ji, L., Ran, R., Wu, H., and Xu, Y. (2021). Forest fire recognition based on feature extraction from multi-view images. *Traitement du Signal* 38, 775–783. doi: 10.18280/ts.380324

Xia, B., Sakamoto, H., Wang, X., and Yamasaki, T. (2022). Packaging design analysis by predicting user preference and semantic attribute. *ITE Trans. Media Technol. Appl.* 10, 120. doi: 10.3169/mta.10.120

Yu, X., Liu, T., Gong, M., and Tao, D. (2018). "Learning with biased complementary labels," in *Proceedings of the European Conference on Computer Vision* (Munich), 68–83.

Zhang, J., Yu, X., Lei, X., and Wu, C. (2022). A novel capsnet neural network based on mobilenetv2 structure for robot image classification. *Front. Neurorobot.* 16, 1007939. doi: 10.3389/fnbot.2022.1007939

Zhang, Y., Niu, G., and Sugiyama, M. (2021). "Learning noise transition matrix from only noisy labels via total variation regularization," in *Proceedings of International Conference of Machine Learning*.

Zhao, N., Cao, Y., and Lau, R. (2018). What characterizes personalities of graphic designs. *ACM Trans. Graph.* 37, 1–15. doi: 10.1145/3197517.3201355