The Annals of Applied Statistics
2025, Vol. 19, No. 2, 1292–1313
https://doi.org/10.1214/24-AOAS1996

CECNN: COPULA-ENHANCED CONVOLUTIONAL NEURAL NETWORKS IN JOINT PREDICTION OF REFRACTION ERROR AND AXIAL LENGTH BASED ON ULTRA-WIDEFIELD FUNDUS IMAGES

BY CHONG ZHONG[1,a], YANG LI[2,c], DANJUAN YANG[3,e], MEIYAN LI[3,f], XINGTAO ZHOU[3,g], BO FU[2,d], CATHERINE C. LIU[1,b] AND A. H. WELSH[4,h]

[1]*Department of Data Science and Artificial Intelligence, Hong Kong Polytechnic University,* [a]*chzhong@polyu.edu.hk,* [b]*macliu@polyu.edu.hk*

[2]*School of Data Science, Fudan University,* [c]*18110980006@fudan.edu.cn,* [d]*fu@fudan.edu.cn*

[3]*Eye Institute and Department of Ophthalmology, Eye & ENT Hospital, Fudan University,* [e]*docdanjuanyang@126.com,* [f]*limeiyan0406073@126.com,* [g]*doctzhouxingtao@126.com*

[4]*College of Business and Economics, Australian National University,* [h]*Alan.Welsh@anu.edu.au*

The ultra-widefield (UWF) fundus image is an attractive 3D biomarker in AI-aided myopia screening because it provides much richer myopia-related information. Though axial length (AL) has been acknowledged to be highly related to the two key targets of myopia screening, spherical equivalence (SE) measurement and high myopia diagnosis, its prediction based on the UWF fundus image is rarely considered. To save the high expense and time costs of measuring SE and AL, we propose the Copula-enhanced Convolutional Neural Network (CeCNN), a one-stop UWF-based ophthalmic AI framework to jointly predict SE, AL, and myopia status. The CeCNN formulates a multiresponse regression that relates multiple dependent discrete-continuous responses and the image covariate, where the nonlinearity of the association is modeled by a backbone CNN. To thoroughly describe the dependence structure among the responses, we model and incorporate the conditional dependence among responses in a CNN through a new copula-likelihood loss. We provide statistical interpretations of the conditional dependence among responses and reveal that such dependence is beyond the dependence explained by the image covariate. We heuristically justify that the proposed loss can enhance the estimation efficiency of the CNN weights. We apply the CeCNN to the UWF dataset collected by us and demonstrate that the CeCNN sharply enhances the predictive capability of various backbone CNNs. Our study supports the ophthalmology view that, besides SE, AL is an important measure of myopia.

1. Introduction. An important trend in ophthalmology research is to apply deep learning (DL) techniques to fundus images to aid the diagnosis and assessment of ophthalmological diseases (Cen et al. (2021), Kim et al. (2021), Li et al. (2021), among others). There are two main types of fundus images: traditional fundus images and advanced *ultra-widefield (UWF) fundus images* (Midena et al. (2022)). Compared to the former that measures a narrow visual range of 30°–75° (the orange dashed circle in Figure 1), the UWF fundus images offer a much broader 200° view of the fundus. These images are much more informative in *myopia screening* (e.g., the ellipsoids in Figure 1 reflect lesions associated with myopia that are outside the traditional image), although they require more advanced equipment.

In myopia screening the spherical equivalence (SE) acts as the gold standard for the degree of myopia; the larger the magnitude of SE, the higher the myopia status; the cut-off for high myopia is −8.0 dioptres (Kobayashi et al. (2005)). *High myopia status* is another important concern in myopia screening because high myopia can substantially increase the risk of

Received August 2024; revised November 2024.
Key words and phrases. Copula, convolutional neural network, multitask learning, myopia, ultra-widefield fundus image, 3D medical image object.

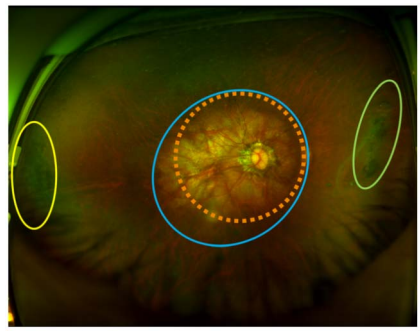

FIG. 1. *Advantages of UWF imaging in myopia-related pathology. The orange dashed circle represents the area covered by regular fundus images. The area within the dashed circle indicates lesions caused by peripheral laser spots, the region within the real circle shows extreme peripheral chorioretinal atrophy, and the area within the two ellipsoids contains pigmentary degeneration lesions.*

blindness (Iwase et al. (2006)). Ophthalmological practitioners have also recognised that the axial length (AL) may be meaningful to myopia screening, since AL is a crucial ocular component which combines information on anterior chamber depth, lens thickness, and vitreous chamber depth (Meng et al. (2011), Tideman et al. (2016)). Therefore, we are motivated to develop a one-stop scheme that jointly predicts SE and AL and diagnoses high myopia status based on UWF fundus images.

1.1. *Motivations*. There are two motivations for the present study: the ophthalmic need to integrate AL information into myopia prediction and our desire to model conditional dependence among responses in convolutional neural networks (CNN).

Motivation 1: *Integrating AL into myopia prediction*. The existing ophthalmology literature usually employs AL to predict SE or high myopia status (Mutti et al. (2007), Haarman et al. (2020), Zhang et al. (2024)), indicating that integrating information from AL should enhance myopia prediction. Nonetheless, precisely measuring AL in practice is costly and time-consuming (Oh et al. (2023)). This drives us to jointly predict SE and AL from the UWF fundus image biomarker. Specifically, we relate the bivariate responses to a tensor object (image) covariate through a CNN, the most widely used DL technique for multitask learning in computer vision.

Motivation 2: *Modeling conditional dependence among responses in a CNN*. A CNN naturally incorporates the dependence among responses that is explained by the common features from the image covariate. However, it is unclear whether a CNN learns the remaining unexplained *conditional dependence among responses given the image covariate*. Generally, a CNN is trained under an empirical loss that is the sum of mean squared error (MSE) losses or the sum of cross entropy losses for regression and classification tasks, respectively. Such empirical losses treat the responses as conditionally independent, given the image covariate; refer to our discussion in Section 4 for more details. Such a conditional independence assumption may be violated in practice. In our application there is a *strong correlation* between SE and AL (see Figure 2), and this strong correlation may not be completely explained by the UWF fundus image covariate. This drives us to model, interpret, and incorporate the conditional dependence among responses, given the image covariate into a CNN within the context of multitask learning for the purpose of enhancing the prediction of myopia.

1.2. *Related work*.

Multiresponse learning in statistics. In multiresponse models, multivariate classification and regression tree (CART) and its variants are widely used (De'Ath (2002), Loh and Zheng

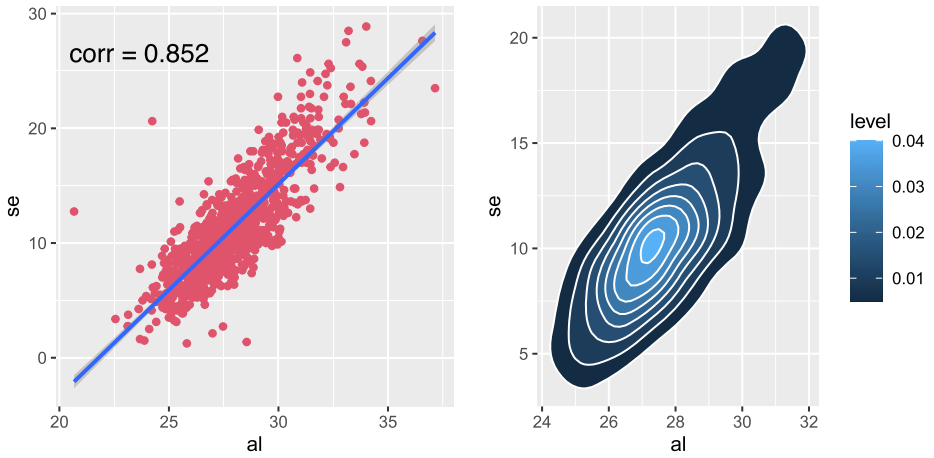

FIG. 2. *Left*: *Scatter plot*; *Right*:*Contour plot*; *x axis*: *AL*; *y axis*: *SE*.

(2013), Rahman, Otridge and Pal (2017), among others), although they do not take the dependence among responses into consideration. To model the dependence among responses, copula models (Sklar (1959)) have been used in various joint regression analyses (Song, Li and Yuan (2009), Panagiotelis, Czado and Joe (2012), Yang, Frees and Zhang (2020), among others). However, the methodologies in the above literature only study the influence of scalar covariates on multiple responses and may not be applicable to image (tensor) covariates. Recently, there has been some literature studying multiresponse tensor regression models (Raskutti, Yuan and Chen (2019), Chen et al. (2019), Zou, Ke and Zhang (2022), among others); such works tend to ignore one or more of the nonlinear associations, the dependence among responses, and the mixture of discrete-continuous responses.

Studies on dependence among responses in deep learning. In the field of computer vision, the existing DL literature rarely considers the conditional dependence among multiple responses, given the image covariate. In multitask learning, most of the literature either uses a simple empirical loss or uses a weighted sum of marginal MSE or cross entropy losses (Kendall, Gal and Cipolla (2018), Lin et al. (2019), among others). *Neither of these two practices incorporates the conditional dependence structure of the responses.* Meanwhile, those weights assigned to each marginal loss may not be well explained and may be difficult to learn from the data. In multiinstance learning the existing literature considers the spatial correlation between labels determined by small patches/instances of the image covariates (Song et al. (2018), Lai et al. (2024), among others). In multilabel learning the existing literature models the partial correlation among labels to guide the information propagation through graph convolutional networks (Chen et al. (2019), Sun et al. (2022), among others).

1.3. *CeCNN*: *Copula enhanced CNN framework*. Let $\boldsymbol{Y} = (y_1, \ldots, y_{p_1}, y_{p_1+1}, \ldots, y_{p_1+p_2})$ be a $(p_1 + p_2)$-dimensional response vector that allows both continuous and binary entries. We call such responses mixed-type responses. Without loss of generality, assume $y_j \in \mathbb{R}$ for $j \le p_1$ and $y_j \in \{0, 1\}$ for $p_1 < j \le p_1 + p_2$. That is, the pending ophthalmological multitask learning problem includes p_1 regression tasks and p_2 binary classification tasks simultaneously. Let $\mathcal{X} \in \mathbb{R}^{k_1 \times \cdots \times k_J}$ be a Jth-order tensor and $\mathcal{G} : \mathbb{R}^{k_1 \times \cdots \times k_J} \to \mathbb{R}^{p_1} \times \{0, 1\}^{p_2}$ be an unknown *nonlinear* function that maps a high order tensor to a $(p_1 + p_2)$-dimensional outcome. We formulate the following multiresponse mean regression that associates responses $\boldsymbol{Y}$ with a tensor object covariate $\mathcal{X}$:

$$E(\boldsymbol{Y}|\mathcal{X}) := \mathcal{G}(\mathcal{X}) = \{g_1(\mathcal{X}), \ldots, g_{p_1}(\mathcal{X}), \mathcal{S} \circ g_{p_1+1}(\mathcal{X}), \ldots, \mathcal{S} \circ g_{p_1+p_2}(\mathcal{X})\}, \tag{1}$$

where $g_j : \mathbb{R}^{k_1 \times \cdots \times k_J} \to \mathbb{R}$ are unknown single output nonlinear functions for $j = 1, \ldots, (p_1 + p_2)$, $\circ$ denotes composition of two functions, and $\mathcal{S}(z) = 1/(1 + e^{-z})$ is the sigmoid function that maps the real line to $(0, 1)$.

We model the unknown nonlinear regression functions $\mathcal{G}$ through a backbone CNN with $(p_1 + p_2)$ outputs. Although there are various types of backbone CNNs, such as LeNet (LeCun et al. (1998)), ResNet (He et al. (2016)), and DenseNet (Huang et al. (2017)), they share a similar architecture. Usually, the architecture of a CNN consists of many hidden layers, including convolution, pooling, and fully connected layers (LeCun, Bengio and Hinton (2015)). Fitting a CNN is then equivalent to optimizing a specific loss over the parameters contained in these hidden layers. In this paper we propose a new *copula-likelihood loss* to train backbone CNNs with mixed-type outputs. Specifically, to meet the emergent needs of ophthalmology practice, we focus on two sets of ophthalmic AI tasks arising from myopia screening, the regression-classification (R-C) task and the regression-regression (R-R) task, and derive the form of the proposed copula-likelihood loss in each task, respectively. The R-C task aims to jointly predict the AL and diagnose the high myopia. The proposed loss and the accompanying training procedure create a new AI framework called Copula-enhanced CNN (CeCNN). From a statistical perspective, we attempt to interpret the conditional dependence, modeled by CeCNN, and justify the enhancement in estimation brought by the proposed loss.

1.4. *Our contributions*. In this paper, motivated by incorporating AL into myopia screening to enhance myopia prediction, we present a nonlinear multiresponse regression where SE (or the Bernoulli variable of myopia status) and AL are associated with UMF fundus images (mode-3 tensors) through a backbone CNN. Specifically, we train the backbone CNN by optimizing a proposed *copula-likelihood loss* so as to accommodate the *conditional dependence* among responses that may not be captured by the backbone CNN itself. We now highlight our main contributions.

Our contributions are trifold. In ophthalmology we might be the first to jointly predict SE and AL in myopia screening based on UWF fundus images. The present study allows for one-stop measurement of SE, AL, and diagnosis of high myopia through one scan, saving manpower and time costs for the precise measurement of SE and AL. Numerical experiments demonstrate that, by incorporating the conditional correlation between SE and AL, our method enhances myopia prediction. In this sense our study might be seen as providing the first evidence of the ophthalmological view that, besides SE, AL is also an important measure of myopia.

In deep learning we contribute a new loss, which might be the first to model and use the conditional dependence among responses given the image predictor. We show that the traditional CNN with empirical loss naturally learns the dependence contributed by the common image predictor but ignores the conditional dependence among responses. In contrast, our proposed loss captures both dependencies; refer to equations (6) and (9) for illustration. Numerical results demonstrate that the proposed loss leads to better predictive performance compared to the empirical loss and the uncertainty loss for multitask learning (Kendall, Gal and Cipolla (2018)). It is anticipated that the proposed loss can be applied to other similar multitask applications in computer vision.

In statistics we might be the first to apply a CNN to relate multiple mixture-type responses to a tensor object covariate nonlinearly and also model the dependence among responses thoroughly. We show heuristically that optimizing the proposed copula-likelihood loss leads to lower estimation risk for the CNN weights in the asymptotic setting. Our study illustrates statistics harnessing AI through *extracting more information within data objects*.

The rest of the paper is organized as follows. Sections 2 and 3 introduce how the CeCNN works in regression-classification (R-C) and regression-regression (R-R) tasks, respectively.

Section 4 rethinks the copula-likelihood loss from the perspective of higher relative efficiency in estimation of CNN weights. Section 5 presents the performance of the CeCNN in myopia prediction on our UWF fundus image dataset. Section 6 carries out simulations on synthetic datasets for illustration. Section 7 concludes the paper with brief discussions. For reproducibility the computer code is available on GitHub https://github.com/Charley-HUANG/CeCNN.

2. Regression-classification task. We start with our R-C task in myopia screening. That is, under model (1), the response vector is $\boldsymbol{Y} = (y_1, y_2)$, where $y_1 \in \mathbb{R}_+$ and $y_2 \in \{0, 1\}$ denote the AL and the status of high myopia (1: high myopia or greater than eight diaoptres; zero, otherwise), respectively. The explanatory variable is $\mathcal{X} \in \mathbb{R}^{224\times224\times3}$, a UWF fundus image stored in red and green with 224×224 channelwise pixels. We construct the copula-likelihood loss for the R-C task in Section 2.1 and summarize the whole CeCNN procedure for the R-C task in Section 2.2. To characterize the joint distribution of the mixed-type responses (y_1, y_2), we adopt the commonly used copula model (Sklar (1959)), which models a joint distribution through a copula and the marginal distributions.

Without loss of generality, a two-dimensional (p-dim) copula $\mathbb{C}$ is a distribution function on $[0, 1]^2$, where each univariate marginal distribution is uniform on $[0, 1]$. One can always express a joint distribution F through a copula $\mathbb{C}$ and the marginal distributions as

$$F(y_1, y_2) = \mathbb{C}\big\{F_1(y_1), F_2(y_2)\big\},$$

where F_j denotes the jth marginal cumulative distribution function (CDF) of y_j for $j = 1, 2$. Let Φ be the CDF of the standard normal distribution $N(0, 1)$. The joint CDF under a *Gaussian copula* is

$$F(y_1, y_2) = \mathbb{C}(\boldsymbol{y}|\Gamma) = \Phi_2\big(\Phi^{-1}\big\{F_1(y_1)\big\}, \Phi^{-1}\big\{F_2(y_2)\big\}|\Gamma\big), \quad \Gamma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

where $\Phi_2(\cdot|\Gamma)$ denotes the CDF of the two-dimensional (p-dim) Gaussian distribution $\mathrm{MVN}(\mathbf{0}_2, \Gamma)$ with correlation matrix $\Gamma \in \mathbb{R}^{2\times2}$, and $\rho \in (-1, 1)$ characterizes the dependence between (y_1, y_2). In the presence of the covariate $\mathcal{X}$, the conditional joint CDF, given $\mathcal{X}$, is naturally written as

$$F(y_1, y_2|\mathcal{X}) = \mathbb{C}(\boldsymbol{y}|\Gamma, \mathcal{X}) = \Phi_2\big(\Phi^{-1}\big\{F_1(y_1|\mathcal{X})\big\}, \Phi^{-1}\big\{F_2(y_2|\mathcal{X})\big\}|\Gamma\big). \tag{2}$$

When both y_1 and y_2 are continuous, the closed form of the joint density function $f(y_1, y_2|\mathcal{X})$ is straightforward; refer to expressions (12) and (13) in Section 3. In this section we derive the closed form of the joint density when $y_1 \in \mathbb{R}$ and $y_2 \in \{0, 1\}$.

2.1. *Copula-likelihood loss.* We begin by modeling the marginal distributions $F_1(y_1|\mathcal{X})$ and $F_2(y_2|\mathcal{X})$. Note that the contour plot Figure 2(b) yields that y_1 is approximately normal on its margin. Meanwhile, y_2 is naturally Bernoulli. Therefore, under a CNN $\mathcal{G} = \{g_1, \mathcal{S} \circ g_2\}$, we model the marginal distributions of (y_1, y_2), given $\mathcal{X}$, as

$$y_1|\mathcal{X} \sim N\big(g_1(\mathcal{X}), \sigma^2\big), \qquad y_2|\mathcal{X} \sim \mathrm{Bernoulli}\big\{\mathcal{S} \circ g_2(\mathcal{X})\big\}. \tag{3}$$

Let n be the size of the training data. For $i = 1, \ldots, n$, let $\mu_{i1} = g_1(\mathcal{X}_i)$ be the marginal expectation of y_{i1}, given $\mathcal{X}_i$, $\mu_{i2} \equiv \Pr\{y_{i2} = 1|\mathcal{X}_i\} = \mathcal{S} \circ g_2(\mathcal{X}_i)$ be the marginal probability of $y_{i2} = 1$, given $\mathcal{X}_i$, and $z_{i1} = (y_{i1} - \mu_{i1})/\sigma$ be the standardized residual of y_{i1}. Let $u_{i1} = F_2(y_{i2}-)$ be the left-hand limit of the CDF F_2 at y_{i2} and $u_{i2} = F_2(y_{i2})$. Let $N(\cdot|\mu, \sigma^2)$ denote the density of $N(\mu, \sigma^2)$. The joint density for bivariate discrete-continuous variables (y_1, y_2), given $\mathcal{X}_i$, is

$$f(y_{i1}, y_{i2}|\mathcal{X}_i) = N\big(y_{i1}|\mu_{i1}, \sigma^2\big) \sum_{r=1}^{2} (-1)^r \mathbb{C}_1^2(z_{i1}, u_{ir}|\Gamma), \tag{4}$$

where $\mathbb{C}_1^2$ is the partial derivative of the Gaussian copula (2) with respect to the continuous coordinate y_{i1}. The joint density (4) is consistent with the bivariate version of Song, Li and Yuan ((2009), equation (9)).

Specifically, for $y_{i2} = 1$ in our case, up to a normalized constant c_0, we have

$$\begin{aligned}\mathbb{C}_1^2(z_{i1}, u_{i1}|\Gamma) &= c_0 \frac{1}{\sqrt{1-\rho^2}} \int_{-\infty}^{\Phi^{-1}(1-\mu_{i2})} \exp\left\{-\frac{1}{2}(z_{i1}, s)\Gamma^{-1}(z_{i1}, s)^T + \frac{1}{2}z_{i1}^2\right\} ds \\ &= c_0 \frac{1}{\sqrt{1-\rho^2}} \int_{-\infty}^{-\Phi^{-1}(\mu_{i2})} \exp\left\{-\frac{z_{i1}^2 - 2\rho z_{i1}s + s^2}{2(1-\rho^2)} + \frac{z_{i1}^2}{2}\right\} ds \\ &= \Phi\left(\frac{-\Phi^{-1}(\mu_{i2}) - \rho z_{i1}}{\sqrt{1-\rho^2}}\right) = 1 - \Phi\left(\frac{\Phi^{-1}(\mu_{i2}) + \rho z_{i1}}{\sqrt{1-\rho^2}}\right),\end{aligned}$$

and

$$\mathbb{C}_1^2(z_{i1}, u_{i2}|\Gamma) = c_0 \frac{1}{\sqrt{1-\rho^2}} \int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2}(z_{i1}, s)\Gamma^{-1}(z_{i1}, s)^T + \frac{1}{2}z_{i1}^2\right\} ds = 1.$$

Consequently, we have the conditional probability

$$\Pr\{y_{i2} = 1 | y_{i1} = \sigma z_{i1} + \mu_{i1}\} = \Phi\left(\frac{\Phi^{-1}(\mu_{i2}) + \rho z_{i1}}{\sqrt{1-\rho^2}}\right) \equiv C^*(\mu_{i2}, z_{i1}|\rho).$$

Hence, by taking the value of y_{i2} to be either 0 or 1, corresponding to (4), the closed form of the conditional joint density of (y_{i1}, y_{i2}), given $\mathcal{X}_i$, in our case is

$$l(y_{i1}, y_{i2}|\mathcal{X}_i) = N\big(y_{i1}; g_1(\mathcal{X}_i), \sigma^2\big) C^*(\mu_{i2}, z_{i1}|\rho)^{y_{i2}} \big(1 - C^*(\mu_{i2}, z_{i1}|\rho)\big)^{1-y_{i2}}.$$

Finally, we obtain the copula-likelihood loss, which is minus the log-likelihood for the training data

$$\begin{aligned}&\mathcal{L}_1\big(g_1, g_2|\{\boldsymbol{Y}_i\}_{i=1}^n, \{\mathcal{X}_i\}_{i=1}^n, \rho, \sigma\big) \\ &\quad= \frac{1}{2\sigma^2}\sum_{i=1}^n (y_{i1} - \mu_{i1})^2 \\ &\quad\quad - \left\{\sum_{i=1}^n \big[y_{i2}\log C^*(\mu_{i2}, z_{i1}|\rho) + (1 - y_{i2})\log\big\{1 - C^*(\mu_{i2}, z_{i1}|\rho)\big\}\big]\right\}.\end{aligned} \tag{5}$$

Note that on the right-hand side (RHS) of (5), the first summand is related to g_1 only, while the second summand associates μ_{i1} and μ_{i2} or, equivalently, g_1 and g_2 through a parameter ρ. Mathematically, when $\rho = 0$, $C^*(\mu_{i2}, z_{i1}|\rho)$ becomes $\mu_{i2} = \mathcal{S} \circ g_2(\mathcal{X}_i)$, implying that the second summand in the RHS of equation (5) reduces to a pure cross entropy loss of g_2 only. Thus, loss (5) is a more general form of the empirical loss. Both ρ and σ have explicit statistical interpretations. The scale parameter σ represents the standard deviation of $y_1|\mathcal{X}$, acting as the weight balancing the MSE loss and the cross-entropy-like loss. Let $X \overset{d}{=} Y$ denote the equality in distribution between random variables X and Y. We rely on the following theorem to interpret ρ in the R-C task.

THEOREM 2.1. *Suppose the joint distribution of $\boldsymbol{Y} = (y_1, y_2)$ is given by the Gaussian copula (2) with correlation matrix Γ, where the marginal distributions of y_1 and y_2 are given by (3). Let $\mu_1 = g_1(\mathcal{X})$ and $\mu_2 = \mathcal{S} \circ g_2(\mathcal{X})$ be the conditional mean of y_1 and y_2,*

given image covariate $\mathcal{X}$, respectively. Let $z_1|\mathcal{X} = (y_1 - \mu_1)/\sigma$ be the standardized version of $y_1|\mathcal{X}$. Let $z_2|\mathcal{X} \sim N(\Phi^{-1}(\mu_2), 1)$ be a latent Gaussian score such that $(z_1, z_2)^T|\mathcal{X} \sim MVN\{[0, \Phi^{-1}(\mu_2)]^T, \Gamma\}$. Then we have $y_2 \overset{d}{=} I(z_2 > 0)$.

Theorem 2.1 tells that the discrete response y_2 is distributionally identical to the continuous latent Gaussian score z_2 under the Gaussian copula (2). It is validated by showing that $\Pr\{y_2 = 1|\mathcal{X}\} = \Pr\{z_2 > 0|\mathcal{X}\}$ and $\Pr\{y_2 = 1|\mathcal{X}, z_1\} = \Pr\{z_2 > 0|\mathcal{X}, z_1\}$ in Section A.1 of the Supplementary Material (Zhong et al. (2025)). As a result, the theorem indicates that the association matrix Γ of the Gaussian copula (2) becomes *the correlation matrix of the conditional joint distribution of z_1 (standardized y_1) and the latent Gaussian score z_2, given the image $\mathcal{X}$*. Our result is motivated by the Gaussian score correlation (Song (2007), Definition 6.3) that characterizes the association between two continuous variables. In the presence of a discrete Bernoulli variable y_2, we extend the concept of Gaussian score correlation from continuous-continuous cases to continuous-binary cases by the constructing a continuous latent variable z_2 as a replacement for y_2.

In summary, $\rho = \operatorname{corr}(y_1, z_2|\mathcal{X})$. From Theorem 2.1 and after some algebra, the full covariance structure between the two Gaussian variables (y_1, z_2) is

$$\operatorname{Cov}(y_1, z_2) = \rho\sigma + \operatorname{Cov}\big(g_1(\mathcal{X}), \Phi^{-1}\{\mathcal{S} \circ g_2(\mathcal{X})\}\big). \tag{6}$$

The latter summand in the RHS, $\operatorname{Cov}(g_1(\mathcal{X}), \Phi^{-1}\{\mathcal{S} \circ g_2(\mathcal{X})\})$, is determined by the image $\mathcal{X}$ and is captured by the CNN. However, the conditional correlation ρ between y_1 and z_2, given $\mathcal{X}$, is beyond the dependence/correlation explained by the image covariate.

The covariance structure (6) inspires natural estimators of ρ and σ. Once the marginal estimators of g_1 and g_2, denoted as $\hat{g}_1^0$ and $\hat{g}_2^0$, respectively, are obtained, one may: (i) estimate ρ as the Pearson correlation between y_1 and $\Phi^{-1}\{\mathcal{S} \circ \hat{g}_2^0(\mathcal{X})\}$, by treating $\Phi^{-1}\{\mathcal{S} \circ \hat{g}_2^0(\mathcal{X})\}$ as a realization of $z_2|\mathcal{X}$ and (ii) estimate the scale parameter σ as the standard deviation of the residuals from $\hat{g}_1^0(\mathcal{X})$. In summary, let $\boldsymbol{\mathcal{X}}$ denote all the images $\mathcal{X}$ in the training set. The estimators of the copula parameters (ρ, σ) are

$$\hat{\rho} = \operatorname{corr}\big(\boldsymbol{y}_1, \Phi^{-1}\{\mathcal{S} \circ \hat{g}_2^0(\boldsymbol{\mathcal{X}})\}\big), \qquad \hat{\sigma} = \operatorname{sd}\{\boldsymbol{y}_1 - \hat{g}_1^0(\boldsymbol{\mathcal{X}})\}. \tag{7}$$

Consequently, the proposed copula-likelihood loss (5) is specified by using the estimators $(\hat{\rho}, \hat{\sigma})$ in place of the unknown (ρ, σ).

2.2. *End-to-end CeCNN.* In Section 2.1 we formulated the proposed copula-likelihood loss accompanied by the estimators of the copula parameters. In this subsection we illustrate how statistics harnesses AI through the proposed CeCNN framework. The overall CeCNN framework has three modules, the warm-up CNN, copula estimation, and the C-CNN. The warm-up CNN module is basically a backbone CNN trained under the empirical loss, providing the marginal estimators needed for copula parameter estimation. The copula estimation module estimates the parameters (ρ, σ) based on the marginal estimators obtained by the warm-up CNN. The last C-CNN module is the core of the whole CeCNN framework where the backbone CNN is trained under the proposed copula-likelihood loss to incorporate the conditional dependence information. The three modules are summarized in Algorithm 1.

In Module 1, without loss of generality, we assume that the backbone CNN $\mathcal{G}$ has k_1 convolution (Conv) layers, k_2 pooling (Pool) layers, and one fully connected (F-C) layer (e.g., the LeNet and the ResNet backbone CNNs). The regression and classification tasks differ only in the F-C layer and share the Conv and Pool layers. All the numerous parameters included in the very deep hidden layers are updated by the Adam algorithm (Kingma and Ba (2014)) to optimize the empirical losses presented in lines 2 and 3, respectively, until convergence.

Algorithm 1 End-to-end CeCNN (regression-classification task)

Input: Training images $\boldsymbol{\mathcal{X}} = \{\mathcal{X}_i\}_{i=1}^n$ and training labels $\{\boldsymbol{Y}_i = (y_{i1}, y_{i2})\}_{i=1}^n$.

Output: CeCNN estimator $\hat{\mathcal{G}} = (\hat{g}_1, \mathcal{S} \circ \hat{g}_2)$.

Module 1: The warm-up CNN

1: Design a bivariate-output backbone CNN $\mathcal{G} = (g_1, \mathcal{S} \circ g_2)$, where $g_j = \text{F-C}_j \circ \text{Pool}(1, \ldots, k_2) \circ \text{Conv}(1, \ldots, k_1)$, $\text{F-C}_j(z) = \boldsymbol{w}_j^T z + b_j$, for $j = 1, 2$.

2: Obtain marginal estimator $\hat{g}_1^0 = \arg\min_{g_1} n^{-1} \sum_{i=1}^n (y_{i1} - g_1(\mathcal{X}_i))^2$.

3: Obtain marginal estimator $\hat{g}_2^0 = \arg\min_{g_2} -\sum_{i=1}^n [y_{i2} \log(\mathcal{S} \circ g_2(\mathcal{X}_i)) + (1 - y_i) \log(1 - \mathcal{S} \circ g_2(\mathcal{X}_i))]$.

Module 2: Copula estimation

4: Obtain estimators $(\hat{\rho}, \hat{\sigma})$ of the copula parameters by (7).

Module 3: The C-CNN

5: Determine the copula-likelihood loss $\mathcal{L}_1(g_1, g_2 | \mathcal{X}, \boldsymbol{Y}, \hat{\rho}, \hat{\sigma})$ in the form of (5).

6: Obtain $\hat{\mathcal{G}} = (\hat{g}_1, \hat{g}_2) = \arg\min_{g_1, g_2} \mathcal{L}_1$.

We view the outputs of the warm-up CNN as the marginal estimators $\hat{g}_1^0$ and $\hat{g}_2^0$ ($\hat{g}_2^0$ is the classification output before the sigmoid transformation). Then in Module 2, we obtain estimators $(\hat{\rho}, \hat{\sigma})$ following (7) and thus determine the copula-likelihood loss (5). In Module 3 we fix $(\hat{\rho}, \hat{\sigma})$ and only update the weights in the backbone CNN. This module fine-tunes the pretrained warm-up CNN. More detail about the training procedure in Module 3 is given in Section 7.

3. Regression-regression task. This section treats the specific regression-regression (R-R) task of predicting the clinically important, highly correlated responses SE and AL, using the proposed CeCNN.

Let $\boldsymbol{Y} \in \mathbb{R}^p$. We rewrite model (1) as the following equivalent multiresponse regression model:

$$\boldsymbol{Y} = \mathcal{G}(\mathcal{X}) + \boldsymbol{\epsilon}, \tag{8}$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)$ is a p-dimensional noise vector, $\boldsymbol{\epsilon} \perp \mathcal{X}$, $E(\epsilon_j) = 0$ for all $1 \le j \le p$. For any $p \ge 2$, (8) expresses multiresponse regression in a unified form. Expression (8) yields the following covariance structure among $(y_1, \ldots, y_p)$:

$$\text{Cov}(y_s, y_t) = \text{Cov}(\epsilon_s, \epsilon_t) + \text{Cov}\big\{g_s(\mathcal{X}), g_t(\mathcal{X})\big\}, \quad s, t = 1, \ldots, p. \tag{9}$$

Here $\text{Cov}(\epsilon_s, \epsilon_t)$ characterizes the *conditional dependence among* $\boldsymbol{Y}$*, given* $\mathcal{X}$, which originates from the model error $\boldsymbol{\epsilon}$ and is beyond the images; $\text{Cov}\{g_s(\mathcal{X}), g_t(\mathcal{X})\}$ characterizes the dependence of $\boldsymbol{Y}$, which is learned by the CNN and contributed by the image $\mathcal{X}$.

Under (8) the distribution of $\boldsymbol{Y} - \mathcal{G}(\mathcal{X}) | \mathcal{X}$ is the same as that of $\boldsymbol{\epsilon}$, and it is notationally convenient to express things in terms of $\boldsymbol{\epsilon}$. Thus, in Gaussian copula modeling, we first have to specify the marginal CDF, and the density of ϵ_j, for $j = 1, \ldots, p$, is simpler than stating we have to specify the conditional CDF and density of $Y_j - g_j(\mathcal{X}) | \mathcal{X}$. In this section we study two types of marginal densities.

Nonparametric error. We start from a general case where the marginal density of the error ϵ_j is unspecified. In this case we estimate the marginal CDF F_{ϵ_j} (and density f_{ϵ_j}) first for $j = 1, \ldots, p$, then estimate Γ, and finally we combine them to derive the copula-likelihood loss.

We begin with the warm-up CNN, equipped with the empirical MSE loss, and obtain the residuals $(e_{i1}, \ldots, e_{ip})$ on the training dataset, for $i = 1, \ldots, n$. To make sure $E(\epsilon_j) = 0$,

we use the mean-centered residuals $\tilde{e}_{i1}, \ldots, \tilde{e}_{ip}$ to estimate the marginal empirical CDF (and pdf) of ϵ_j by Gaussian kernel smoothing

$$\tilde{F}_{\epsilon_j}(t) = \frac{1}{n}\sum_{i=1}^{n} \Phi\{(t - \tilde{e}_{ij})/\psi_0\}, \qquad \tilde{f}_{\epsilon_j}(t) = \frac{1}{n\psi_0}\sum_{i=1}^{n} \phi\{(t - \tilde{e}_{ij})/\psi_0\}, \tag{10}$$

where $\psi_0 > 0$ is a tuning parameter acting as the bandwidth for both CDF and density estimation.

Once we obtain the estimates $\tilde{F}_{\epsilon_j}$, the correlation matrix Γ in the Gaussian copula (2) is estimated using

$$\hat{\gamma}_{sj} = \operatorname{corr}\big(\{\Phi^{-1}(\tilde{F}_{\epsilon_s}(\tilde{e}_{is}))\}_{i=1}^{n}, \{\Phi^{-1}(\tilde{F}_{\epsilon_j}(\tilde{e}_{ij}))\}_{i=1}^{n}\big), \quad s, j = 1, \ldots, p, \tag{11}$$

the Pearson correlation between the two Gaussian scores of the smoothed empirical CDFs of the centered residuals.

With the above estimates and under the Gaussian copula, based on Song, Li and Yuan ((2009), equation (7)), the copula-likelihood loss is given by

$$\mathcal{L}_2\big(\{g_j\}_{j=1}^{p} | \{\boldsymbol{Y}_i\}_{i=1}^{n}; \boldsymbol{\mathcal{X}}\big) = -\sum_{i=1}^{n} \frac{1}{2}\boldsymbol{q}_i^T(\boldsymbol{I}_p - \Gamma^{-1})\boldsymbol{q}_i - \sum_{i=1}^{n}\sum_{j=1}^{p} \log\{\tilde{f}_{\epsilon_j}(y_{ij} - g_j(\mathcal{X}_i))\}, \tag{12}$$

where $\boldsymbol{q}_i = (\Phi^{-1}\{\tilde{F}_{\epsilon_1}[y_{i1} - g_1(\mathcal{X}_i)]\}, \ldots, \Phi^{-1}\{\tilde{F}_{\epsilon_p}[y_{ip} - g_p(\mathcal{X}_i)]\})^T$. Obviously, if $\Gamma = \boldsymbol{I}_p$ (no correlation between the responses), loss (12) reduces to the sum of minus the log densities $\tilde{f}_{\epsilon_j}$.

Gaussian error. Since both SE and AL look normally distributed (based on the contour plot in Figure 2), we may simply adopt the Gaussian model error $\epsilon_j \sim N(0, \sigma_j^2)$, where σ_j is a scale parameter. With Gaussian error the likelihood contribution of $\boldsymbol{Y}_i | \mathcal{X}_i$ reduces to the multivariate Gaussian density directly. Thus, for training data $(\{\boldsymbol{Y}_i\}_{i=1}^{n}, \boldsymbol{\mathcal{X}})$, the copula-likelihood loss is

$$\mathcal{L}_3\big(\{g_j\}_{j=1}^{p} | \{\boldsymbol{Y}_i\}_{i=1}^{n}; \boldsymbol{\mathcal{X}}\big) = -\sum_{i=1}^{n} \log \mathrm{MVN}_p\{(y_{i1} - g_1[(\mathcal{X}_i)], \ldots, y_{ip} - g_p[(\mathcal{X}_i)]); \boldsymbol{0}_p, \Sigma\}, \tag{13}$$

where $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_p)\Gamma \operatorname{diag}(\sigma_1, \ldots, \sigma_p) \equiv (\sigma_{tj})_{p\times p}$ is the covariance matrix of $\boldsymbol{\epsilon}$. Specifically, in our bivariate application, if $\rho = 0$ (i.e., Γ is the identity), the copula-likelihood loss (13) simply reduces to the sum of the empirical MSE loss and some constant. Therefore, the copula-likelihood loss (13) is a generalization of the empirical MSE loss.

With Gaussian errors we only need to estimate the covariance matrix Σ or, equivalently, the correlation Γ and the marginal standard deviations $(\sigma_1, \ldots, \sigma_p)$. Intuitively, their estimates can be obtained from the empirical correlation matrix and marginal standard deviations of the residuals from warm-up CNNs. Let $(\hat{g}_{01}, \ldots, \hat{g}_{0p})$ be p-dimensional outputs of CNNs trained with the empirical MSE loss for $(g_1, \ldots, g_p)$ on the training dataset. Using the residuals $(e_{i1}, \ldots, e_{ip})$ for each observation, where $e_{ij} = y_{ij} - \hat{g}_{0j}(\boldsymbol{X}_i)$, we obtain estimates of the copula parameters Γ and σ_j as

$$\hat{\Gamma} \equiv (\hat{\gamma}_{tj})_{p\times p} = \operatorname{corr}\big(\{e_{it}\}_{i=1}^{n}, \{e_{ij}\}_{i=1}^{n}\big), \qquad \hat{\sigma}_j = \operatorname{sd}\big(\{e_{ij}\}_{i=1}^{n}\big). \tag{14}$$

Finally, we summarize the CeCNN for the regression-regression task in Algorithm 2. In this task the CeCNN again has the three-module structure with different copula parameters. For Gaussian error the copula parameters are the marginal SDs σ_j and the Pearson correlations γ_{tj}; for nonparametric error the copula parameters contain an infinite dimensional parameter f_{ϵ_j} and the transformed correlation γ_{tj}.

Algorithm 2 End-to-end CeCNN (regression-regression task)

Input: Training images $\boldsymbol{\mathcal{X}} = \{\mathcal{X}_i\}_{i=1}^n$ and training labels $\{\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{ip})\}_{i=1}^n$.

Output: CeCNN estimator $\hat{\mathcal{G}} = (\hat{g}_1, \ldots, \hat{g}_p)$.

Module 1: The warm-up CNN

Design a multi-output backbone CNN $\mathcal{G} = (g_1, \ldots, g_p)$, where $g_j = \text{F-C}_j \circ \text{Pool}(1, \ldots, k_2) \circ \text{Conv}(1, \ldots, k_1)$, $\text{F-C}_j(\boldsymbol{z}) = \boldsymbol{w}_j^T \boldsymbol{z} + b_j$, for $j = 1, \ldots, p$.

Obtain marginal estimators $\hat{g}_j^0 = \arg\min_{g_j} n^{-1} \sum_{i=1}^n (y_{ij} - g_j(\mathcal{X}_i))^2$, for $j = 1, \ldots, p$.

Module 2: Copula estimation

if Non-Gaussian **then**
    Estimate copula parameter Γ based on (11).
else if Gaussian **then**
    Estimate copula parameters (Γ, σ) based on (14).
end if

Module 3: The C-CNN

if Non-Gaussian **then**
    Define the loss $\mathcal{L}_2$ as (12). Obtain $\hat{\mathcal{G}} = (\hat{g}_1, \ldots, \hat{g}_p) = \arg\min_{g_1, \ldots, g_p} \mathcal{L}_2$.
else if Gaussian **then**
    Define the loss $\mathcal{L}_3$ as (13). Obtain $\hat{\mathcal{G}} = (\hat{g}_1, \ldots, \hat{g}_p) = \arg\min_{g_1, \ldots, g_p} \mathcal{L}_3$.
end if

4. Rethinking the copula-likelihood loss: Relative efficiency. An interesting question raised by the reviewers is what kind of "dependence" is learned by the copula-likelihood loss in addition to the "dependence" explained by the image covariates, which is learned automatically by the "baseline." Here the baseline refers to a reasonable backbone CNN equipped with the empirical loss. Our understanding is the proposed copula-likelihood loss *incorporates the conditional dependence among responses, given the image covariate*. Such conditional dependence cannot be learned by the backbone CNN equipped with the empirical loss. As a result, the estimators of the weights in the last fully-connected (F-C) layer of a CNN are asymptotically more efficient under the copula-likelihood loss than those under the empirical loss. In the following we confine our discussion to the bivariate-response learning task, considering our application to the UWF dataset.

In both R-C and R-R tasks, the backbone CNN with the empirical loss can only capture the correlation between y_1 and y_2, explained by the image covariate $\mathcal{X}$, and misses the conditional dependence structure given $\mathcal{X}$. This is evidenced by the fact that the CNN fitted under the empirical loss is the nonparametric maximum likelihood estimator on the space of CNNs under the conditional independence model assumption $y_1 \perp y_2 | \mathcal{X}$, that is, the special case $\rho = 0$ in our Gaussian copula modeling; referred to Propositions A.1 for R-R tasks and A.2 for R-C tasks in the Supplementary Material (Zhong et al. (2025)), respectively. Therefore, the baseline with the empirical loss may suffer from the risk of model misspecification if the true $\rho \neq 0$, incurring suboptimal prediction.

In contrast, the copula-likelihood loss delivers the conditional dependence information to the CNN through the correlation parameter ρ in the Gaussian copula. Specifically, in the R-R task, the loss incorporates the covariance structure of the model error as expressed in (9); in the R-C task, the loss incorporates $\rho = \text{corr}(z_1, z_2 | \mathcal{X})$, where z_1 is the standardized residual and z_2 is the latent Gaussian score defined in Theorem 2.1.

Suppose the depth of a backbone CNN is $L = k_1 + k_2 + 1$. Recall that, in Algorithms 1 and 2, the last F-C layer can be expressed as $H = (\text{F-C}_1, \text{F-C}_2)$; the stacking of $(1, \ldots, L-1)$ hidden layers can be expressed as $D = \text{Pool}(1, \ldots, k_2) \circ \text{Conv}(1, \ldots, k_1)$. Therefore, $\mathcal{G} = H \circ D$. By definition, $D : \mathbb{R}^{a \times b \times c} \to \mathbb{R}^K$ is a nonlinear function mapping a tensor $\mathcal{X}$ to K feature maps, where K is the width of the last F-C layer of the CNN. The width K varies

among different backbone CNNs (e.g., $K = 512$ in ResNet and $K = 4096$ in DenseNet). With K feature maps $D(\mathcal{X})$, the last F-C layer $H : \mathbb{R}^K \to \mathbb{R}^2$ is defined by

$$H \circ D(\mathcal{X}) = (\hat{y}_1, \hat{y}_2)^T, \qquad \hat{y}_j = \mathcal{A}_j\{\boldsymbol{w}_j^T D(\mathcal{X}) + b_j\}, \quad j = 1, 2, \tag{15}$$

where $\mathcal{A}_j$ are some specific activation functions and $\boldsymbol{w}_j \in \mathbb{R}^K$ and $b_j \in \mathbb{R}$ are the weights and bias of the jth output neuron, respectively. In the R-R task, the two activation functions $\mathcal{A}_j$ are both the identity projection; in the R-C task, the $\mathcal{A}_j$ are the identity projection and the sigmoid function, respectively.

Fitting a CNN is equivalent to optimizing some loss between $\hat{y}_j$ in (15) and the training responses y_j. In the R-R task, we denote by $\hat{\boldsymbol{w}}_j^{\text{emp}}$ the estimate of $\boldsymbol{w}_j$ under the empirical loss and denote by $\hat{\boldsymbol{w}}_j^{\text{cop}}$ the estimate of $\boldsymbol{w}_j$ under the copula-likelihood loss. We make the following assumption.

ASSUMPTION 1 (Uncovered feature maps). Let $\mathbf{1}_{\boldsymbol{w}_j} = \{k : w_{jk} \neq 0\}$ be the index set of nonzero entries in weights $\boldsymbol{w}_j$. Assume that $\mathbf{1}_{\boldsymbol{w}_1} \setminus \mathbf{1}_{\boldsymbol{w}_2} \neq \varnothing$ and $\mathbf{1}_{\boldsymbol{w}_2} \setminus \mathbf{1}_{\boldsymbol{w}_2} \neq \varnothing$.

Assumption 1 assumes that the two outputs y_1 and y_2 in the R-R tasks share different features of $\mathcal{X}$ extracted from the convolutional layers. This assumption naturally holds for those backbone CNNs that assign different F-C layers to different outputs (e.g., Liu et al. (2019), Lian et al. (2022)). For general backbone CNNs like ResNet and DenseNet, Assumption 1 may be examined by significance tests for black box learners (Dai, Shen and Pan (2024)) or by visual explanations for neural networks (Selvaraju et al. (2017)).

Let w_{jk} be the pth element of $\boldsymbol{w}_j$, for $k = 1, \ldots, K$. We are in a position to provide the following theorem on estimation of the weights in the last F-C layer of a CNN. The proof relies on transforming the linear model (15) into a seemingly unrelated regression model (Zellner (1962)); details are deferred to Section A.3 of the Supplementary Material (Zhong et al. (2025)). The theorem is confined to the R-R task; similar results may also hold on the R-C task, but the techniques are much more complicated.

THEOREM 4.1 (Relative efficiency). *Let $\boldsymbol{\mathcal{X}}$ be the samples of the image covariates in the training set. In the R-R task, under Assumption 1 and other technical assumptions in Section A of the Supplementary Material, if $\mathbf{1}_{\boldsymbol{w}_1} \cap \mathbf{1}_{\boldsymbol{w}_2} = \varnothing$, as $n \to \infty$, for $k = 1, \ldots, K$, we have*

$$\Pr\{\operatorname{Var}(\hat{w}_{jk}^{\text{cop}} | \boldsymbol{\mathcal{X}}) \leq \operatorname{Var}(\hat{w}_{jk}^{\text{emp}} | \boldsymbol{\mathcal{X}})\} \to 1, \; j = 1, 2.$$

Note that both $\operatorname{Var}(\hat{w}_{jk}^{\text{cop}} | \boldsymbol{\mathcal{X}})$ and $\operatorname{Var}(\hat{w}_{jk}^{\text{emp}} | \boldsymbol{\mathcal{X}})$ are determined by the feature maps of the training image samples, and thus they are indeed stochastic. Theorem 4.1 shows that, in the asymptotic setting, the former is dominated by the latter. Since both $\hat{w}_{jk}^{\text{cop}}$ and $\hat{w}_{jk}^{\text{emp}}$ are unbiased for w_{jk}, Theorem 4.1 indicates that, with probability tending to 1, $\hat{w}_{jk}^{\text{cop}}$ is more efficient than $\hat{w}_{jk}^{\text{emp}}$. Therefore, compared with the empirical loss, in the R-R task, the copula-likelihood loss reduces $\mathrm{E}||\boldsymbol{w}_j - \hat{\boldsymbol{w}}_j||_2^2$, the estimation risk of the weights within the jth output neuron in the last F-C layer of a CNN.

5. Application to the UWF fundus image dataset. We apply the proposed CeCNN to myopia screening based on our UWF dataset. To evaluate the predictive capability of CeCNN, we conduct 10 rounds of five-fold cross-validation on the dataset.

5.1. *Data preparation.* The data collection process involved capturing 987 fundus images from the left eyes of 987 patients using the Optomap Daytona scanning laser ophthal-

moscope (Daytona, Optos, UK). All enrolled patients sought refractive surgery treatment and were exclusively myopia patients. To ensure homogeneity and accuracy, individuals with other ocular conditions, such as cataract, vitreoretinal diseases, or glaucoma as well as those with a history of trauma or previous ocular surgery, were excluded from the dataset. For image selection we required the fovea to be positioned at the center of the image and applied the criterion of gradability. Images were considered gradable if there was no blurring in the optic disc or foveal area and if less than 50% of the peripheral retinal area was obscured by eyelids or eyelashes. The UWF fundus images obtained during the study were exported in JPEG format and compressed to a resolution of 224×224 pixels to facilitate subsequent analysis.

Response variables. As stated before, the response variables considered include two continuous responses, SE and AL, and a binary myopia status response indicating whether the patient has high myopia or not. The ground truth values for AL and SE in our study are obtained through standard clinical procedures by professional ophthalmologists from the Eye & ENT Hospital of Fudan University, ensuring that our ground truth values are accurate and reliable. High myopia is defined according to SE value with a cut-off of -8.0 D. We will predict both AL value and myopia status (the R-C task) and predict SE and AL values (the R-R task).

5.2. *Structure of backbone CNNs*. To validate the generality of CeCNN, we select LeNet, ResNet-18, and DenseNet as our backbone CNNs. LeNet represents one of the simplest, original CNNs, while ResNet and DenseNet are acknowledged as two of the most effective and prominent CNN models in the field of computer vision.

Simplify backbone CNN to avoid overfitting. The conventional ResNet-18 and DenseNet models contain over ten million parameters, and the original LeNet model has over five million parameters. However, the size of our UWF dataset is quite limited, comprising a data size that is smaller than 1000. This discrepancy posed a significant challenge in the form of severe overfitting, which results in a large gap between the losses on the training set and the validation set (Goodfellow, Bengio and Courville (2016)). As a result, the predictive performances on our UWF test set are unsatisfactory, no matter what kind of loss is used to train the backbone CNN.

To mitigate the overfitting issue caused by the overparameterized backbone CNNs, we adopt the common strategy to reduce the number of the learning parameters by simplifying the neural network architecture (Han et al. (2015), Keshari et al. (2018), among others). Specifically, we removed the last two CNN blocks from ResNet18, removed the last dense block from DenseNet, and increased the filter sizes of the convolutional and pooling layers of LeNet. This substantial reduction in the parameter count effectively prevented overfitting. After simplification, for example, the parameter count of ResNet18 reduced from over ten million to around six hundred thousand. Figure 3 shows the architectures of the simplified versions of LeNet, ResNet18, and DenseNet. These simplified backbone CNNs are also set as the baselines for comparison.

Simplifying the backbone CNNs on our UWF dataset does NOT indicate that CeCNN has limitations in the choice of backbone CNN. In practical applications, if the dataset size is sufficient or if appropriate measures to address overfitting are implemented, it would not be necessary to use a simplified backbone CNN.

Architecture of the CeCNN. The complete architecture for applying the CeCNN model to the UWF dataset is illustrated in Figure 4. We begin with a warm-up CNN module where

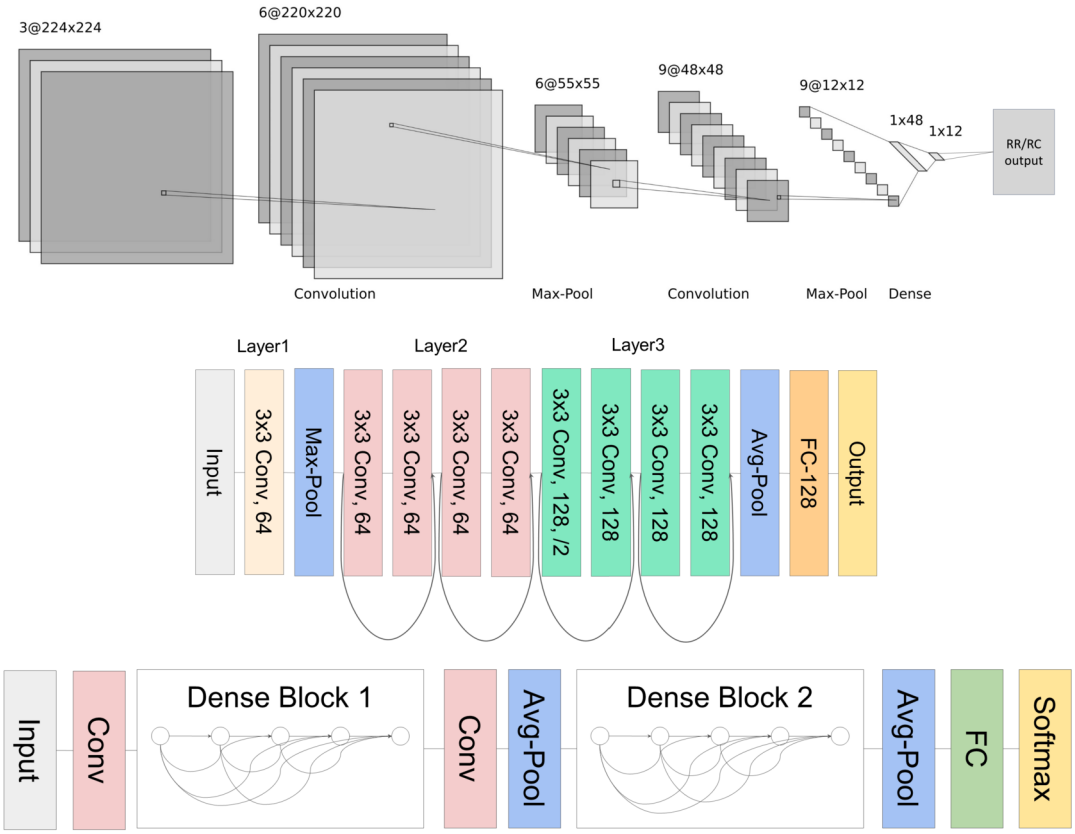

FIG. 3. *The architectures of backbone CNNs used. Top*: *The architecture of simplified LeNet*; *Middle*: *The architecture of simplified ResNet*; *Bottom*: *The architecture of simplified DenseNet*.

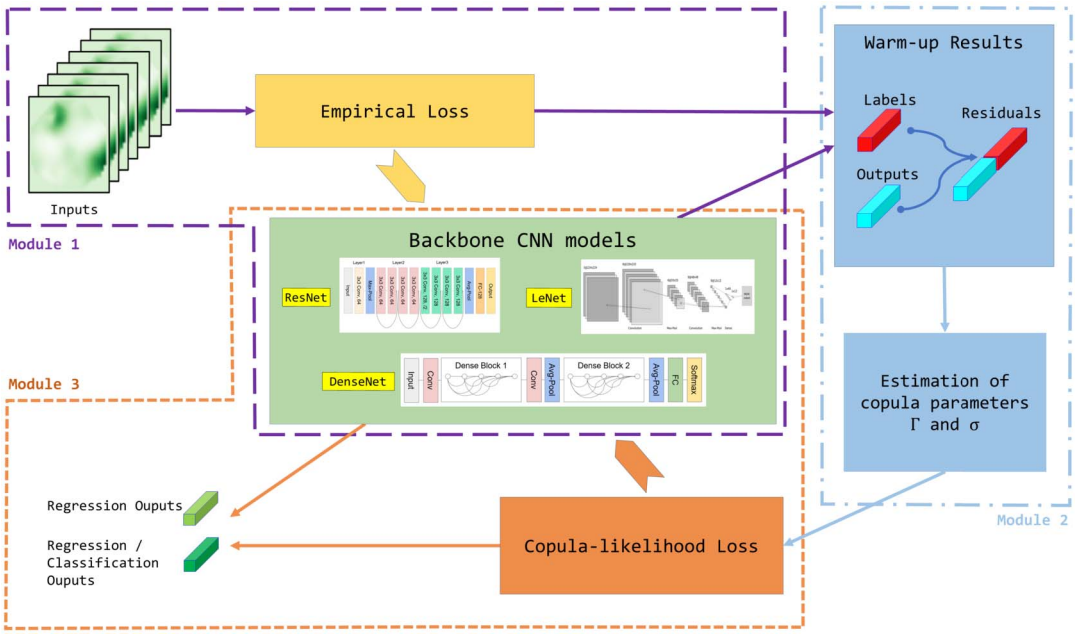

FIG. 4. *The whole architecture of the CeCNN framework*.

we train the backbone CNNs under empirical loss (Module 1). Then we estimate the copula parameter based on the residuals and Gaussian scores obtained in the warm-up CNN (Module 2). Finally, with the estimated copula parameter, we proceed to train our CeCNN models with the proposed copula-likelihood loss (Module 3).

5.3. *Results on the UWF dataset*. To evaluate the predictive capability of CeCNN, we consider two competitors for comparison. The first competitor is the *baseline*, where the backbone CNN is equipped with the empirical losses for the R-C and R-R tasks, respectively. The other competitor is the uncertainty loss (Kendall, Gal and Cipolla (2018)), a commonly used multitask learning loss in the DL community. Comparison with the uncertainty loss is deferred to Section B of the Supplementary Material (Zhong et al. (2025)). We do not compare with the Pareto weighted loss (Lin et al. (2019)) due to the high computational burden in tuning the weights.

We partition the full UWF dataset into the training data set, the validation set, and the testing set with a ratio of 6:2:2. In the R-C tasks, we evaluate predictive performance by classification accuracy and AUC for the classification task, and the root mean square error (RMSE) for the regression task; in the R-R tasks, we use the RMSE and the mean absolute error (MAE) as metrics.

Results on the R-C task. We present the boxplots of the evaluation metrics for the R-C tasks in Figure 5. Compared with the baseline, with a significance level of 0.05, the CeCNN significantly improves the AUC with both the DenseNet and the ResNet backbones. Furthermore, the CeCNN with the DenseNet backbone absolutely reduces the average regression RMSE by 2.887% and slightly improves the classification accuracy by 0.516%; the CeCNN with the ResNet backbone slightly reduces the RMSE by 0.120% and absolutely increases the classification accuracy by 0.994%; CeCNN with the LeNet backbone reduces the RMSE by 1.223% and increases the AUC by 1.227%, while slightly sacrificing classification accuracy by 0.145%. In summary, the CeCNN enhances the baseline in almost all tasks with various backbone CNNs; the enhancement on the DenseNet and the ResNet backbones is more significant than that on the LeNet backbone. We conjecture that this phenomenon may be caused by the different width of the last F-C layer of different CNNs, that is, the number K in Theorem 4.1. Recall that Assumption 1 in Section 4 requires the activated feature maps for the two tasks to be uncovered. When K is relatively small (e.g., $K = 12$ in the LeNet), this assumption may be violated, limiting the possible enhancement by the copula-likelihood loss. In contrast, for ResNet and DenseNet where K, the input dimension of the last F-C layer, is large, the enhancement from using the copula-likelihood loss is absolute.

Results on the R-R task. We present boxplots of the evaluation metrics for the R-C tasks in Figure 6. On average, when using the Gaussian error, CeCNN reduces the RMSE of SE by 0.717%, 2.232%, and 3.827% and reduces the RMSE of AL by 10.378%, 3.683%, and 5.262% on the LeNet, the ResNet, and the DenseNet backbones, respectively. With a significance level of 0.05, the improvements in the AL prediction with all of the LeNet, the ResNet, and the DenseNet backbones are significant, and the improvement in the SE prediction with the DenseNet backbone is also significant. When using the nonparametric error, the predictive performances of the CeCNNs are similar to those using Gaussian error. It is not surprising that the CeCNN enhances predicting AL more than it enhances predicting SE since the marginal variance of AL is smaller than that of SE. In the joint Gaussian log-likelihood (13), a smaller variance puts a larger weight on the corresponding loss, and hence the optimizer tends to optimize more on AL than SE.

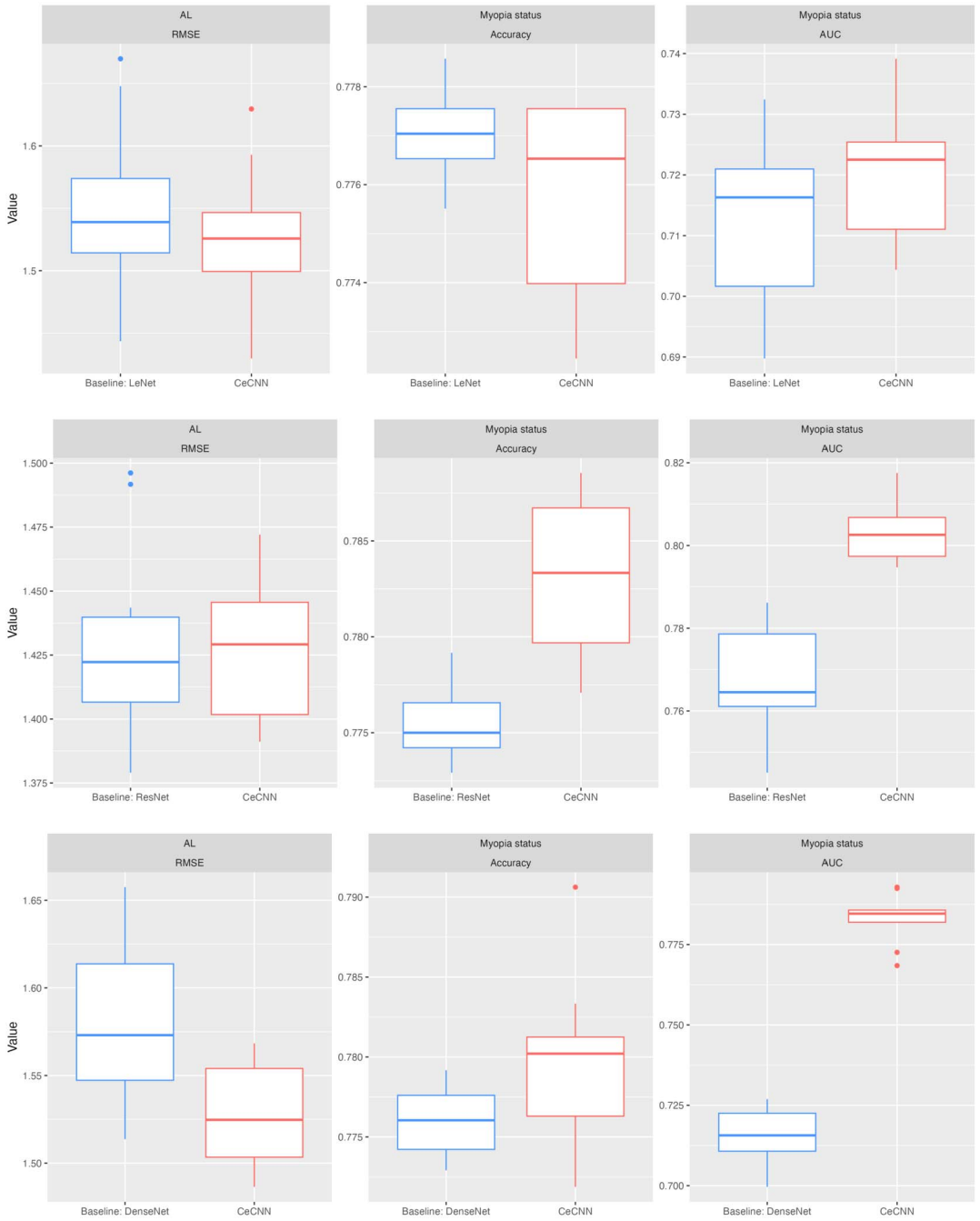

FIG. 5. *Boxplots of RMSE, classficiation accuracy and AUC in* 10 *rounds of five-fold validation of the R-C tasks. Top*: *With LeNet backbone*; *Middle*: *With ResNet backbone*; *Bottom*: *With DenseNet backbone*.

6. Simulations on synthetic data. We carry out simulations on synthetic datasets for both R-C and R-R tasks to illustrate the superiority of CeCNN compared with the baseline. To match our application scenarios, in both R-C and R-R tasks, we consider correlated bivariate responses and a single image covariate. We present the simulation details for the regression-classification and regression-regression tasks in Sections 6.1 and 6.2, respectively.

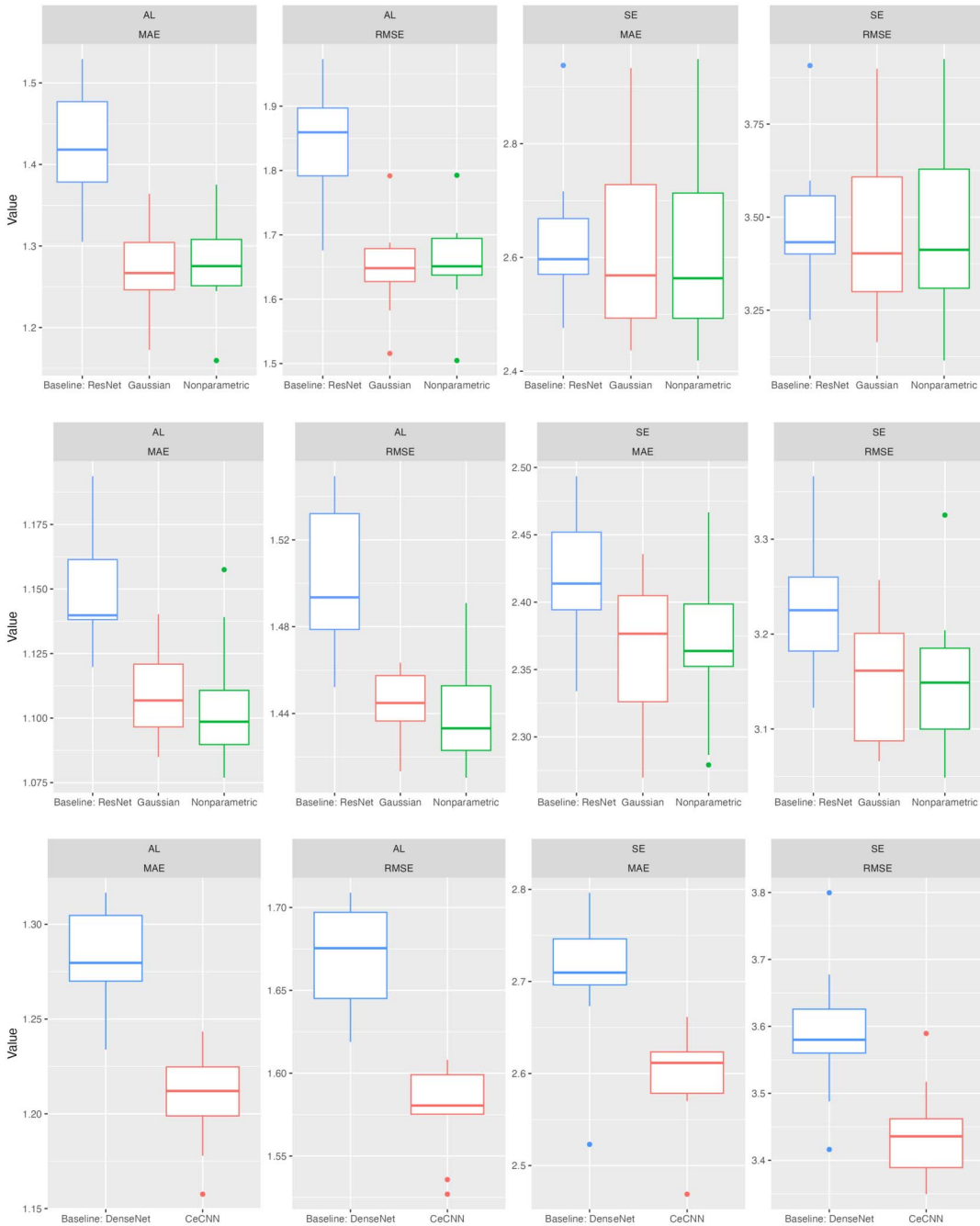

FIG. 6. *Boxplots of RMSE and MAE in* 10 *rounds of five-fold validation of the R-R tasks. Top*: *With the LeNet backbone*; *Middle*: *With the ResNet backbone*; *Bottom*: *With the DenseNet backbone*.

6.1. *Simulated regression-classification task.* On this task the synthetic image covariate is generated as a grey matrix $X_i \in \mathbb{R}^{9\times 9}$ that can be divided into nine independent blocks of 3×3 square block matrices $X_{t,s} \in \mathbb{R}^{3\times 3}$,

$$X = \begin{pmatrix} X_{1,1} & X_{1,2} & X_{1,3} \\ X_{2,1} & X_{2,2} & X_{2,3} \\ X_{3,1} & X_{3,2} & X_{3,3} \end{pmatrix}, \qquad X_{t,s} = \begin{pmatrix} X_{t,s}^{(1,1)} & X_{t,s}^{(1,2)} & X_{t,s}^{(1,3)} \\ X_{t,s}^{(2,1)} & X_{t,s}^{(2,2)} & X_{t,s}^{(2,3)} \\ X_{t,s}^{(3,1)} & X_{t,s}^{(3,2)} & X_{t,s}^{(3,3)} \end{pmatrix}, \quad t, s = 1, 2, 3.$$

For $k, l = 1, 2, 3$, we set $X_{3,3}^{(k,l)} \sim N(1, 0.5^2)$ as elements of the dark block and $X_{t,s}^{(k,l)} \sim N(0, 1^2)$ as elements of the bright blocks, for $(t, s) \neq (3, 3)$. Based on this covariate, we generated the synthetic continuous response $y_1 \in \mathbb{R}$ and the synthetic binary response $y_2 \in \{0, 1\}$ using model (1).

For each block $X_{t,s}$, we define the block operators $S_{t,s}, S_{t,s}^* : \mathbb{R}^{3\times 3} \to \mathbb{R}$ for responses y_1 and y_2, respectively. Then the two responses were generated as

$$y_1 \sim N\Big\{ \sum_{1\le t,s\le 3} S_{t,s}(X_{t,s}), 1^2 \Big\}, \qquad y_2 \sim \text{Bernoulli}\Big\{ \mathcal{S}\Big[\sum_{1\le t,s\le 3} S_{t,s}^*(X_{t,s}) \Big] \Big\}. \tag{16}$$

For the block operators $S_{t,s}$, we set

$$S_{1,1} = \sum_{1\le l,k\le 3} \tanh(X_{1,1}^{(k,l)}), \qquad S_{2,2} = \sum_{1\le l,k\le 3} X_{2,2}^{(k,l)}, \qquad S_{3,3} = \tanh\Big(\sum_{1\le l,k\le 3} X_{3,3}^{(k,l)} \Big),$$

and $S_{t,s} = 0$ for other blocks. Under this setting the function $g_1 = \sum_{1\le t,2,\le 3} S_{t,s}$ is a nonlinear function that can be modeled by a CNN. For the block operators $S_{t,s}^*$, we set

$$S_{2,2}^* = S_{2,2} = \sum_{1\le l,k\le 3} X_{2,2}^{(k,l)}$$

and $S_{t,s}^* = 0$ for other blocks so that $E(y_2) = 1/2$. The two responses are dependent in this case since they share the same block operator on block $X_{2,2}$.

6.2. *Simulated regression-regression task.* For the regression-regression task, we generated the synthetic image covariate in a similar way to that of the regression-classification task. For this task we set five dark blocks $X_{t,s}^{(k,l)} \sim N(1, 0.5^2)$ for $k, l = 1, 2, 3$ and $(t, s) \in \{(1, 1), (2, 2), (3, 3), (1, 3), (3, 1)\}$. The remaining blocks were set to be the aforementioned bright blocks $N(0, 1^2)$. The responses were generated from model (8), where the model error $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \epsilon_{i2}) \sim \text{MVN}(\mathbf{0}_2, \Sigma)$ with

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

This covariance setting yields a strong correlation and two different levels of marginal variation (variances of 1 and 4), corresponding to the AL and SE, respectively. We set the nonlinear regression functions g_1 and g_2 as the sum of block operators so that

$$g_1(X) = \sum_{1\le t,s\le 3} S_{t,s}(X_{t,s}), \qquad g_2(X) = \sum_{1\le t,s\le 3} S_{t,s}^*(X_{t,s}).$$

For y_1 the block operators $S_{t,s}$ are given by

$$S_{1,1} = \sum_{1\le l,k\le 3} \tanh(X_{1,1}^{(k,l)}), \qquad S_{2,2} = \sum_{1\le l,k\le 3} X_{2,2}^{(k,l)}, \qquad S_{3,3} = \tanh\Big(\sum_{1\le l,k\le 3} X_{3,3}^{(k,l)} \Big),$$

and $S_{t,s} = 0$ for other pairs of (t, s). For y_2 the the block operators $S_{t,s}^*$ are given by

$$S_{1,3}^* = \sum_{1\le l,k\le 3} \tanh(X_{1,3}^{(k,l)}), \qquad S_{2,2}^* = \sum_{1\le l,k\le 3} X_{2,2}^{(k,l)}, \qquad S_{3,3}^* = \tanh\Big(\sum_{1\le l,k\le 3} X_{3,3}^{(k,l)} \Big),$$

and $S_{t,s}^* = 0$ for other pairs of (t, s). Under this setting, marginally, we have $E(y_1) = E(y_2)$. Note that the block operators $S_{t,x}$ and $S_{t,s}^*$ can be viewed as the input feature maps of the last F-C layer of the CNN. Then our setting guarantees that Assumption 1 holds.

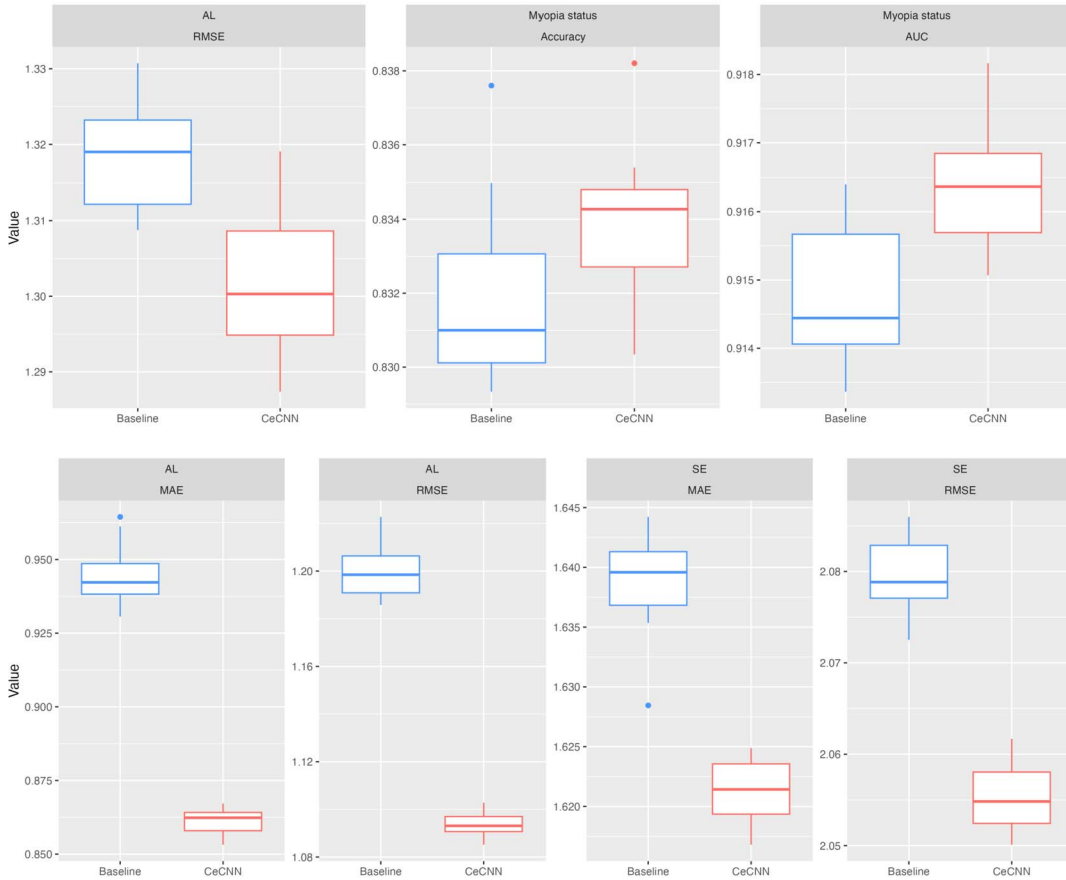

FIG. 7. *Top*: *Boxplots of RMSE, classification accuracy and AUC in* 10 *rounds of five-fold validation of the R-C tasks on the synthetic dataset*; *Bottom*: *Boxplots of RMSE and MAE for two responses in* 10 *rounds of five-fold validation of the R-R tasks on the synthetic dataset*.

6.3. *Simulation results*. In total, we generated $n = 10{,}000$ synthetic images and the corresponding pairs of responses in our simulations. To save computation time, we adopt a simple backbone CNN in both the baseline and the CeCNN. The backbone model used in the simulations consists of two basic convolutional layers and two fully connected layers. With such a simple backbone CNN and the relatively large data size, the overfitting issue is not severe in our simulations. The baseline is then defined as the same backbone CNN equipped with the empirical loss. For the R-R task, we used MSE, AUC, and classification accuracy as the evaluation metrics; for the R-R tasks, we use RMSE as the evaluation metric.

From Figure 7 we find that CeCNN improves the prediction performance of the baseline in both R-C and R-R tasks. Recall that Assumption 1 holds under our R-C simulation setting. The simulation result on the R-R task illustrates the improvement brought by CeCNN that we established theoretically in Theorem 4.1.

7. Discussion. In this paper we propose the CeCNN, a new one-stop ophthalmic AI framework for myopia screening based on the UWF fundus image. The CeCNN models and incorporates the conditional dependence among mixed-type responses in a multiresponse nonlinear regression through a new copula-likelihood loss to train the backbone CNN. We provide explicit statistical interpretations of the conditional dependence and justify the enhancement in estimation brought about by the proposed loss through a heuristic inferential procedure. In our UWF dataset, the CeCNN succeeds in enhancing the predictive capability

of various deep learning models in measuring SE, AL, and diagnosing high myopia simultaneously.

Why not iteratively update the copula parameters. One referee questioned why we do not iteratively update the estimates of the copula parameters Γ and σ in module 2, during each epoch updating of the backbone CNN in module 3 of the CeCNN. Indeed, we estimate the copula parameter Γ from the outputs of the backbone CNN trained under the empirical loss. It is called the warm-up CNN since we only train this CNN once to estimate Γ and σ. We attempt to answer the reviewer's question from the point of view of estimating $\boldsymbol{w}_j$, the weights in the last F-C layer of the backbone CNN.

In the CeCNN, by replacing the true Γ with its estimate based on the residuals and Gaussian scores obtained from the warm-up CNN, the induced estimator $\hat{\boldsymbol{w}}_j^{\text{cop}}$ is the feasible generalized least squares (FGLS) estimator (Avery (1977)). Since the FGLS estimator is asymptotically equivalent to the GLS estimator using the true copula parameters (Prucha (1984)), using the estimated copula parameters in the copula-likelihood loss will not lose estimation efficiency asymptotically. This justifies our CeCNN architecture.

On the other hand, we conduct five runs of five-fold cross-validation to evaluate the idea of iteratively updating the copula parameters; refer to Section C in the Supplementary Material (Zhong et al. (2025)). Iteratively updating the copula parameters Γ and σ is closely related to the alternating minimization (AM) algorithm for seemingly unrelated regression (Jain and Tewari (2015)). The AM algorithm requires the optimizer to converge in each iteration of updating the copula parameters. In deep learning, convergence of the optimizer is sensitive to the choice of initial values. In our numerical studies, we find that: (i) iteratively updating the copula parameters can converge very slowly when using the ImageNet pretrained initials, (ii) too frequently updating copula parameters may result in poorly trained epochs and lead in wrong directions in training, and (iii) the weights trained by the warm-up CNN module are good initial values to ensure convergence, while iteratively updating the copula parameters does not enhance prediction in the R-C task and leads to mere progress in the R-R task. It would be an interesting future direction to explore the performance of iteratively updating the copula parameters on larger datasets.

Why only reduce the model size to avoid overfitting. Another referee questioned why we do not consider other techniques to overcome overfitting, such as hyperparameter tuning, regularization, and data augmentation.

The CeCNN framework relies on a backbone CNN. In both the warm-up CNN and the C-CNN modules, we maintain the default settings for most of the hyperparameters in the backbone CNN, allowing ophthalmic practitioners to use the CeCNN conveniently without further tuning. We carefully tune the learning rate on the warm-up CNN such that the backbone CNN converges under the training loss and no longer improves in the validation loss. Then in the C-CNN module, we reduce the learning rate of the warm-up module by multiplying by 0.1 to seek a more precise direction for optimization. We do not consider dropout rate tuning of the backbone CNN since, unfortunately, we found that dropout leads to poor prediction performance in both the R-C and the R-R tasks. We conjecture the reason is that dropout cannot preserve the variance in each hidden layer after the nonlinear activation and thus may not be suitable for regression tasks (Özgür and Nar (2020)).

Adding regularization to the CNN weights is generally thought to be a good way to overcome overfitting. However, in practice, it is challenging to specify the tuning parameter λ for the regularization. In our practice we try several candidate values for the tuning parameters and present two examples of the loss traces in Section D of the Supplementary Material (Zhong et al. (2025)). We find that, on our UWF dataset, with a small λ, the overfitting issue is still serious. Otherwise, with a large λ, the predictive performance on the test set is

worse than that with no regularization. Therefore, we do not consider regularization in our application to our UWF dataset.

Data augmentation is effective in improving DL models with limited data size. There are a wealth of approaches to data augmentation in computer vision including image manipulation, image eraser, and image mix, among others. We believe that suitable data augmentation can lead to better predictive performance of both the baseline and the CeCNN. Nonetheless, it is challenging to find "suitable" data augmentation methods since generalizing qualified synthetic images based on the very limited data size is difficult (Yang et al. (2022)). To avoid possible error induced by poorly generated images, we do not consider data augmentation procedures in our application. Seeking an appropriate data augmentation approach for our UWF dataset is beyond the scope of the present study, but it deserves treating in separate work.

Future work. The present paper has shown that the proposed copula-likelihood loss can reduce the asymptotic estimation risk in the R-R task, while theoretically proving the enhancement in the R-C task is much more difficult, pending another separate work to resolve. From the application perspective, this paper only considers bivariate myopia screening tasks on a single eye. The proposed loss was successfully applied to a quadravariate case where the images of left and right eyes are included to predict SE and AL for both eyes (Li et al. (2024)). It is anticipated that, besides myopia screening in ophthalmology, CeCNN can also be applied to other multitask learning scenarios. Furthermore, an interesting future direction is to associate the CeCNN framework with large models such as Vision Transformers (Dosovitskiy et al. (2020)). This can be accomplished by replacing the backbone CNNs by transformers if the overfitting issues are well resolved.

Acknowledgments. The authors are grateful for the instructive guidance from the Chief Editor, the Associate Editor, and the two anonymous referees. The authors also thank Mr. Qiuyi Huang for his help in programming.

Chong Zhong and Yang Li are the co-first authors.

Catherine C. Liu and Bo Fu are the co-corresponding authors.

Funding. Chong Zhong is partially supported by Postdoc Fellowship of CAS AMSS-PolyU Joint Laboratory of Applied Mathematics and ZZPC, PolyU.

Bo Fu and Yang Li are partially supported by the National Natural Science Foundation of China (12071089, 71991471).

Danjuan Yang and Meiyan Li are partially supported by the Shanghai Rising-Star Program (21QA1401500). Meiyan Li is also supported by the National Natural Science Foundation of China (82371091).

Catherine C. Liu is partially supported by a grant (GRF15301123) from the Research Grants Council of the Hong Kong SAR, and ZZQ2, PolyU.

A. H. Welsh is partially supported by the Australian Research Council Discovery Project DP230101908.

SUPPLEMENTARY MATERIAL

Technical proofs and additional numerical studies (DOI: 10.1214/24-AOAS1996SUPPA; .pdf). This file contains technical proofs and tables and figures that provide additional numerical studies.

Code for simulations (DOI: 10.1214/24-AOAS1996SUPPB; .zip). This file contains Python code to reproduce the simulations on synthetic data.

REFERENCES

AVERY, R. B. (1977). Error components and seemingly unrelated regressions. *Econometrica* **45** 199–209. MR0451543 https://doi.org/10.2307/1913296

CEN, L.-P., JI, J., LIN, J.-W., JU, S.-T., LIN, H.-J., LI, T.-P., WANG, Y., YANG, J.-F., LIU, Y.-F. et al. (2021). Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nat. Commun.* **12** 4828.

CHEN, Z.-M., WEI, X.-S., WANG, P. and GUO, Y. (2019). Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 5177–5186.

DAI, B., SHEN, X. and PAN, W. (2024). Significance tests of feature relevance for a black-box learner. *IEEE Trans. Neural Netw. Learn. Syst.* **35** 1898–1911. MR4710268 https://doi.org/10.1109/tnnls.2022.3185742

DE'ATH, G. (2002). Multivariate regression trees: A new technique for modeling species–environment relationships. *Ecology* **83** 1105–1117.

DOSOVITSKIY, A., BEYER, L., KOLESNIKOV, A., WEISSENBORN, D., ZHAI, X., UNTERTHINER, T., DEHGHANI, M., MINDERER, M., HEIGOLD, G. et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

GOODFELLOW, I., BENGIO, Y. and COURVILLE, A. (2016). *Deep Learning. Adaptive Computation and Machine Learning*. MIT Press, Cambridge, MA. MR3617773

HAARMAN, A. E., ENTHOVEN, C. A., TIDEMAN, J. W. L., TEDJA, M. S., VERHOEVEN, V. J. and KLAVER, C. C. (2020). The complications of myopia: A review and meta-analysis. *Investig. Ophthalmol. Vis. Sci.* **61** 49–49.

HAN, S., POOL, J., TRAN, J. and DALLY, W. (2015). Learning both weights and connections for efficient neural network. *Adv. Neural Inf. Process. Syst.* **28**.

HE, K., ZHANG, X., REN, S. and SUN, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 770–778.

HUANG, G., LIU, Z., VAN DER MAATEN, L. and WEINBERGER, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4700–4708.

IWASE, A., ARAIE, M., TOMIDOKORO, A., YAMAMOTO, T., SHIMIZU, H., KITAZAWA, Y., GROUP, T. S. et al. (2006). Prevalence and causes of low vision and blindness in a Japanese adult population: The Tajimi study. *Ophthalmology* **113** 1354–1362.

JAIN, P. and TEWARI, A. (2015). Alternating minimization for regression problems with vector-valued outputs. *Adv. Neural Inf. Process. Syst.* **28**.

KENDALL, A., GAL, Y. and CIPOLLA, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 7482–7491.

KESHARI, R., VATSA, M., SINGH, R. and NOORE, A. (2018). Learning structure and strength of cnn filters for small sample size training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 9349–9358.

KIM, K. M., HEO, T.-Y., KIM, A., KIM, J., HAN, K. J., YUN, J. and MIN, J. K. (2021). Development of a fundus image-based deep learning diagnostic tool for various retinal diseases. *J. Personalized Med.* **11** 321.

KINGMA, D. P. and BA, J. (2014). Adam: A method for stochastic optimization. Preprint. Available at arXiv:1412.6980.

KOBAYASHI, K., OHNO-MATSUI, K., KOJIMA, A., SHIMADA, N., YASUZUMI, K., YOSHIDA, T., FUTAGAMI, S., TOKORO, T. and MOCHIZUKI, M. (2005). Fundus characteristics of high myopia in children. *Jpn. J. Ophthalmol.* **49** 306–311.

LAI, Q., ZHOU, J., GAN, Y., VONG, C.-M. and CHEN, C. P. (2024). Single-stage broad multi-instance multi-label learning (BMIML) with diverse inter-correlations and its application to medical image classification. *IEEE Trans. Signal Process.* **8** 828–839.

LECUN, Y., BENGIO, Y. and HINTON, G. (2015). Deep learning. *Nature* **521** 436–444.

LECUN, Y., BOTTOU, L., BENGIO, Y. and HAFFNER, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* **86** 2278–2324.

LI, Y., HUANG, Q., ZHONG, C., YANG, D., LI, M., WELSH, A., LIU, A., FU, B., LIU, C. C. et al. (2024). OUCopula: Bi-channel multi-label copula-enhanced adapter-based CNN for myopia screening based on OU-UWF images. In *IJCAI*.

LI, Z., GUO, C., LIN, D., NIE, D., ZHU, Y., CHEN, C., ZHAO, L., WANG, J., ZHANG, X. et al. (2021). Deep learning for automated glaucomatous optic neuropathy detection from ultra-widefield fundus images. *Br. J. Ophthalmol.* **105** 1548–1554.

LIAN, C., LIU, M., WANG, L. and SHEN, D. (2022). Multi-task weakly-supervised attention network for dementia status estimation with structural MRI. *IEEE Trans. Neural Netw. Learn. Syst.* **33** 4056–4068.

LIN, X., ZHEN, H.-L., LI, Z., ZHANG, Q.-F. and KWONG, S. (2019). Pareto multi-task learning. *Adv. Neural Inf. Process. Syst.* **32**.

LIU, M., ZHANG, J., ADELI, E. and SHEN, D. (2019). Joint classification and regression via deep multi-task multi-channel learning for Alzheimer's disease diagnosis. *IEEE Trans. Biomed. Eng.* **66** 1195–1206.

LOH, W.-Y. and ZHENG, W. (2013). Regression trees for longitudinal and multiresponse data. *Ann. Appl. Stat.* **7** 495–522. MR3086428 https://doi.org/10.1214/12-AOAS596

MENG, W., BUTTERWORTH, J., MALECAZE, F. and CALVAS, P. (2011). Axial length of myopia: A review of current research. *Ophthalmologica* **225** 127–134.

MIDENA, E., MARCHIONE, G., DI GIORGIO, S., ROTONDI, G., LONGHIN, E., FRIZZIERO, L., PILOTTO, E., PARROZZANI, R. and MIDENA, G. (2022). Ultra-wide-field fundus photography compared to ophthalmoscopy in diagnosing and classifying major retinal diseases. *Sci. Rep.* **12** 19287.

MUTTI, D. O., HAYES, J. R., MITCHELL, G. L., JONES, L. A., MOESCHBERGER, M. L., COTTER, S. A., KLEINSTEIN, R. N., MANNY, R. E., TWELKER, J. D. et al. (2007). Refractive error, axial length, and relative peripheral refractive error before and after the onset of myopia. *Investig. Ophthalmol. Vis. Sci.* **48** 2510–2519.

OH, R., LEE, E. K., BAE, K., PARK, U. C., YU, H. G. and YOON, C. K. (2023). Deep learning-based prediction of axial length using ultra-widefield fundus photography. Korean journal of ophthalmology. *Korean J. Ophthalmol.* **37** 95–104.

ÖZGÜR, A. and NAR, F. (2020). Effect of dropout layer on classical regression problems. In 2020 28*th Signal Processing and Communications Applications Conference* (*SIU*) 1–4. IEEE.

PANAGIOTELIS, A., CZADO, C. and JOE, H. (2012). Pair copula constructions for multivariate discrete data. *J. Amer. Statist. Assoc.* **107** 1063–1072. MR3010894 https://doi.org/10.1080/01621459.2012.682850

PRUCHA, I. R. (1984). On the asymptotic efficiency of feasible Aitken estimators for seemingly unrelated regression models with error components. *Econometrica* **52** 203–207. MR0729216 https://doi.org/10.2307/1911468

RAHMAN, R., OTRIDGE, J. and PAL, R. (2017). Integratedmrf: Random forest-based framework for integrating prediction from different data types. *Bioinformatics* **33** 1407–1410.

RASKUTTI, G., YUAN, M. and CHEN, H. (2019). Convex regularization for high-dimensional multiresponse tensor regression. *Ann. Statist.* **47** 1554–1584. MR3911122 https://doi.org/10.1214/18-AOS1725

SELVARAJU, R. R., COGSWELL, M., DAS, A., VEDANTAM, R., PARIKH, D. and BATRA, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* 618–626.

SKLAR, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris* **8** 229–231. MR0125600

SONG, L., LIU, J., QIAN, B., SUN, M., YANG, K., SUN, M. and ABBAS, S. (2018). A deep multi-modal CNN for multi-instance multi-label image classification. *IEEE Trans. Image Process.* **27** 6025–6038. MR3852320 https://doi.org/10.1109/TIP.2018.2864920

SONG, P. X.-K. (2007). *Correlated Data Analysis*: *Modeling*, *Analytics*, *and Applications*. *Springer Series in Statistics*. Springer, New York. MR2377853

SONG, P. X.-K., LI, M. and YUAN, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics* **65** 60–68. MR2665846 https://doi.org/10.1111/j.1541-0420.2008.01058.x

SUN, K., HE, M., XU, Y., WU, Q., HE, Z., LI, W., LIU, H. and PI, X. (2022). Multi-label classification of fundus images with graph convolutional network and lightgbm. *Comput. Biol. Med.* **149** 105909.

TIDEMAN, J. W. L., SNABEL, M. C., TEDJA, M. S., VAN RIJN, G. A., WONG, K. T., KUIJPERS, R. W., VINGERLING, J. R., HOFMAN, A., BUITENDIJK, G. H. et al. (2016). Association of axial length with risk of uncorrectable visual impairment for europeans with myopia. *JAMA Ophthalmol.* **134** 1355–1363.

YANG, L., FREES, E. W. and ZHANG, Z. (2020). Nonparametric estimation of copula regression models with discrete outcomes. *J. Amer. Statist. Assoc.* **115** 707–720. MR4107674 https://doi.org/10.1080/01621459.2018.1546586

YANG, S., XIAO, W., ZHANG, M., GUO, S. ZHAO, J. and SHEN, F. (2022). Image data augmentation for deep learning: A survey. Preprint. Available at arXiv:2204.08610.

ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *J. Amer. Statist. Assoc.* **57** 348–368. MR0139235

ZHANG, S., CHEN, Y., LI, Z., WANG, W., XUAN, M., ZHANG, J., HU, Y., CHEN, Y., XIAO, O. et al. (2024). Axial elongation trajectories in Chinese children and adults with high myopia. *JAMA Ophthalmol.* **142** 87–94.

ZHONG, C., LI, Y., YANG, D., LI, M., ZHOU, X., FU, B., LIU, C. C. and WELSH, A. H. (2025). Supplement to "CeCNN: Copula-enhanced convolutional neural networks in joint prediction of refraction error and axial length based on ultra-widefield fundus images." https://doi.org/10.1214/24-AOAS1996SUPPA, https://doi.org/10.1214/24-AOAS1996SUPPB

ZOU, C., KE, Y. and ZHANG, W. (2022). Estimation of low rank high-dimensional multivariate linear models for multi-response data. *J. Amer. Statist. Assoc.* **117** 693–703. MR4436306 https://doi.org/10.1080/01621459.2020.1799813