Stellar spectral template library construction based on generative adversarial networks*

Jianghui Cai¹,², Zeyang Yan¹, Haifeng Yang¹,³,***, Xin Chen¹, Aiyu Zheng¹, Jing Hao¹, Xujun Zhao¹,³, and Yaling Xun¹,³

¹ School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China
² School of Computer Science and Technology, North University of China, Taiyuan 038507, China
³ Shanxi Laboratory of Big Data Analysis and Parallel Computing, Taiyuan University of Science and Technology, 030024, China

Received September 15, 1996; accepted March 16, 1997

ABSTRACT

Stellar spectral template libraries play an important role in the automated analysis of stellar spectra. Synthetic template libraries cover a very large parameter space but suffer from poor matching with observed spectra. In this study, we propose a synthetic-to-observed spectral translation (SOST) method based on generative adversarial networks. The SOST method is able to calibrate synthetic spectra by converting them to the corresponding observed spectra. We applied this method to Kurucz synthetic spectra and observed spectra data from the Large Sky Area Multi-Object Fiber Spectroscopic Telescope (LAMOST). After that, we constructed a stellar spectral library with uniform and broad parameter distributions using the SOST-corrected Kurucz synthetic spectra. Our stellar spectral template library contains 2431 spectra spanning a parameter space of 3500 to 8000 K for effective temperature (T eff), 0.0 to 5.0 dex for surface gravity (log g), and -2.0 to 0.5 dex for metallicity ([Fe/H]). The spectra in the library have a resolution of R ≈ 1800 and cover the wavelength range 3900-8700 Å. In order to verify the accuracy of this template library, we used the template library and the template-matching algorithm to derive the parameters of the PASTEL database. Compared to measurements using the original synthetic template library, the accuracies of the three parameters, T eff, log g, and [Fe/H], are improved, from 140 K, 0.31 dex, and 0.21 dex to 121 K, 0.26 dex, and 0.13 dex, respectively. In addition, we re-parameterised more than six million stellar spectra released by LAMOST DR8.

Key words. methods: data analysis – techniques: spectroscopic – stars: fundamental parameters – surveys

1. Introduction

In recent years, an increasing number of large spectroscopic surveys have been started as well as completed. These surveys have generated an unprecedented number of spectra. For instance, the ongoing LAMOST Galactic Spectroscopic Survey of the Galactic Anti-centre (LSS-GAC) (Luo et al. 2015) has already collected millions of low-resolution (R ~ 1800) spectra of stars. Faced with massive amounts of stellar spectroscopic data, researchers need automated analysis tools to efficiently extract useful astrophysical information about target stars. The stellar spectral template library is able to perform analytical tasks by matching observed spectra with template spectra, such as classification of stellar types and determination of stellar atmospheric parameters. These pieces of information are essential for understanding the formation, structure, and evolution of the Milky Way (Jofré et al. 2019). Several stellar parameter pipelines, such as SSP (Lee et al. 2008), LASP (Luo et al. 2015), and LSP3 (Xiang et al. 2015), have emerged to provide convenient solutions for spectral parameter determinations. These mainstream parametric measurement pipelines rely on synthetic template libraries or empirical template libraries; however, they all have corresponding shortcomings.

The synthetic template libraries comprise stellar spectra generated by theoretical models based on stellar properties. The synthetic template libraries offer noisless data covering a wide range of stellar parameter values. In a related work on template library construction based on LAMOST data, it has been pointed out that the influence of theoretical and instrumental factors could lead to inconsistencies between the synthetic model and observed spectra at the wavelengths of some features (Du et al. 2019). This mismatch may lead to unrealistic parameter measurements. On the theoretical side, the parameters that have been considered during spectral modelling are not comprehensive, as they lack factors such as microturbulence, rotation velocities, and convection. Assumptions regarding the physics of stellar models both in their interiors and atmospheres also impact the accuracy of any synthetic grid. These theoretical limitations have been mentioned in previous studies (Kurucz 2013; Prieto et al. 2018; Franchini et al. 2018). On the instrumental side, the modelling of instrumental factors and noise lacks perfection (Ballester et al. 2000; Martioli et al. 2012).

The empirical template libraries such as MILES (Sánchez-Blázquez et al. 2006; Falcón-Barroso et al. 2011) and ELODIE (Prugniel & Soubiran 2001; Prugniel et al. 2007) consist of high-quality, real observed spectra with stellar atmospheric parameters (Du et al. 2019; Royer et al. 2024; Prieto et al. 2004; Verro et al. 2022), having been created through the efforts of observers. However, a good empirical template library requires coverage of a wide range of parameters, and acquiring observed spectra that satisfy this requirement is a challenging task. Furthermore, the stellar labels determined using empirical template libraries are limited in accuracy. This is due to the errors present in the pa-

* The source code and pre-trained models are available at https://github.com/zeyangyan/SOST.
** Haifeng Yang email: hfyang@tyust.edu.cn

Article published by EDP Sciences to be cited as https://doi.org/10.1051/0004-6361/202349032
rameter labelling of the observed spectra, which are passed on to the spectra to be measured during the parameter measurement. Jofré et al. (2019) reviewed the latest efforts in this field (Ren et al. 2016; Jönsson et al. 2018) in order to assess the accuracy of stellar parameter measurement results and discussed the uncertainties of the measurement results from the perspectives of random uncertainties, systematic uncertainties, and biases.

Some works have been done by researchers to overcome the shortcomings of the current synthetic spectra. Works using traditional methods typically attempt to take into account a variety of complex physical effects and data properties when modelling spectra. Kovalev et al. (2019) incorporated non-local thermodynamic equilibrium (NLTE) effects into the model atmospheres when determining stellar parameters and abundances in medium-resolution spectra of FGK-type stars. Similarly, Amarsi (2016) investigated the departure from local thermodynamic equilibrium (LTE) of atomic oxygen and its impact on various oxygen lines in FGK-type stars, offering grids predicting 3D NLTE-based equivalent widths and abundance corrections. Additionally, Bialek et al. (2020) bridged the synthetic gap by augmenting the synthetic spectra with real observational features, such as by adding Gaussian noise, applying rotational and radial velocities, removing the continuum, and masking geodesic regions. However, these methods usually require a great deal of a priori knowledge and manual adjustments, and it is difficult for them to model the various effects in the spectra, which tend to overlook certain details or effects.

Compared with traditional methods, deep learning methods have stronger non-linear modelling capabilities and are able to learn features and laws automatically without many manual adjustments. Deep learning technologies, including neural networks, convolutional neural networks (CNNs), and generative adversarial networks (GANs), are widely utilised in astronomy, particularly in spectral analysis. Recent studies have introduced innovative algorithms, such as ZETA-PAYNE by Strumit et al. (2022), that use CNNs to accurately analyse spectra from hot OBAB-type stars in SDSS-V data. Additionally, Gebran et al. (2022) have demonstrated the effectiveness of CNNs in determining stellar parameters from observed and synthetic spectra, offering practical guidance for their application in astronomy. Furthermore, deep learning techniques, such as LSTM neural networks, have been employed in studies, including that of Hu et al. (2022), to analyse spectral time series of Type Ia supernovae, enabling precise reconstruction of spectral sequences from a single observed spectrum. In response to the small number of O-type stars in the LAMOST data release, Zheng et al. (2020) proposed the SGAN for generating artificial O-type spectra. O’Briain et al. (2021) proposed Cycle-StarNet, a domain-adaptive method to narrow down the discrepancy between synthetic spectra and observed spectra. Gebran (2024) established a connection between stellar parameters and spectra by training a fully connected neural network, which can be used for library construction. Nevertheless, the reduction of the mismatch between the synthetic model and observed spectra is worth exploring.

In this paper, we propose a synthetic-to-observed spectral translation method called SOST. This method is based on a GAN and uses Kurucz synthetic spectra (Castelli 2005) (Kurucz model atmospheres with the radiative code SYNTHE) and LAMOST observed spectra as data sources. The well-trained SOST method can guide the calibration of synthetic spectra by providing observed spectra and correct errors that may arise during the synthetic spectrum modelling process. The SOST-corrected synthetic spectra can achieve a favourable match with the observed spectra. This addresses the past shortcomings of relying on either synthetic spectra or observed spectra data.

Additionally, we construct a stellar spectral template library using the SOST method. Our spectral library offers a set of stellar spectra with uniformly and extensively distributed parameters. The effective temperature \( T_{\text{eff}} \) ranges from 3500 to 8000 K, the surface gravity (log g) spans 0.0 to 5.0 dex, and the metallicity ([Fe/H]) varies from -2.0 to 0.5 dex. The wavelength coverage for the templates spans 3900 to 8700 Å at a resolution of approximately \( R = 1800 \). To assess the accuracy of our constructed library, this study employs a template-matching method based on \( \chi^2 \) minimisation (Jofré et al. 2010) to compare the parameters obtained using SOST with measurements from PASTEL (Soubiran et al. 2016) and the Large Sky Area Multi-Object Fibre Spectroscopic Telescope Data Release 8 (LAMOST DR8; Yan et al. 2022). The template-matching method determines the parameters by comparing the target spectrum with the template spectrum (Zwitter et al. 2008; Lee et al. 2008). Through comparison with external databases, the accuracy of this template was determined to be 1.21 K, 0.26 dex, and 0.13 dex for \( T_{\text{eff}}, \log g \), and [Fe/H], respectively. This work provides a powerful support for stellar studies and a powerful tool for a deeper understanding of stars in the Universe (Suda et al. 2008; Yoon et al. 2018).

This paper is organised as follows: The SOST and the stellar spectral template library construction method are introduced in Section 2. The spectral data used for the experiments as well as the designed experiments for stellar parameter determinations are described in detail in Section 3. Finally, we summarise this work in Section 4.

2. Method

In this section, the SOST, the SOST training process, and the stellar spectral template library construction method are described in detail. To facilitate the method construction, certain reasonable assumptions needed to be made. Synthetic and observed spectra lie in two related but distinct distribution domains, which we call the synthetic and observed spectral domains. Synthetic spectra and observed spectra both have some underlying physical characteristics (such as effective temperature, metallicity, and surface gravity), but they also contain distinctive information. This distinctiveness could arise from instrumental factors (such as line spread functions), errors introduced during data processing (such as pseudo-continuum normalisation), or variations not present in synthetic spectra (such as atmospheric absorption lines from Earth’s atmosphere). Learning the transformation between these two domains directly is complex and difficult. Therefore, we assumed that the spectrum can be decomposed into a content code and a style code. Specifically, content codes refer to the underlying physical characteristics of effective temperature, metallicity, and surface gravity, while style codes refer to the specific details of the spectrum, representing the unique properties of both synthetic and observed spectra. The unique properties include the features resulting from assumptions made during synthetic spectrum modelling and the instrumental factors of observed spectra.

Based on this assumption, we designed the SOST with an auto-encoder (Kingma & Welling 2013) and discriminator model in both the synthetic and observed spectra domains.

1 The instructions for using the Kurucz model can be found at the following URL: http://kurucz.harvard.edu/programs.html.
The auto-encoder model can realize the separation and reorganization of spectral content and style features. The content features extracted by the encoder have the same distribution, and the style features have different distributions. Cross-domain content features and style features can be combined by one of the generators in order to achieve cross-domain transfer of spectra. The discriminator plays the role of a teacher, evaluating the generated spectra, which in turn improves the quality of the generated domain migration spectra. The SOST allows us to construct a stellar spectral template library by transferring the desired synthetic spectra across domains to the corresponding observed spectra, enabling correction of the synthetic spectra.

2.1. SOST

As shown in Fig. 1, the SOST consists of a content encoder, \( E_{\text{synth}} \) and \( E_{\text{obs}} \), a style encoder, \( E_{\text{ff}} \), a generator, \( G_{\text{synth}} \) and \( G_{\text{obs}} \), and a discriminator, \( D_{\text{synth}} \) and \( D_{\text{obs}} \), for the synthetic and observed spectra domains. The content encoder extracts the common features of the two domains, and the style encoder extracts the unique style features of the two domains. The common features of the two domains and the unique style features align with our assumption regarding ‘related but distinct domains’. The content features and the cross-domain style features are combined by the generator using the instance normalization (AdaIN) (Huang & Belongie 2017) to transform the observed spectra of LAMOST and the synthetic spectra of the Kurucz synthetic spectra transfer learning. The components of the SOST are briefly described in this subsection, and the functions of each component and the training process of the method are described in detail in Section 2.2.

2.2. Architecture and training

In order to achieve domain transfer between spectra while ensuring that the generated spectra do not lose their own information, we designed multiple loss functions to constrain the training process. We used spectral reconstruction loss (equation (3)) and cross-domain feature reconstruction loss (equation (4) and (5)) to train the auto-encoder for feature extraction and reduction. We used cycle consistency loss (equation (6)) to constrain the generated spectra and generative adversarial loss (equation (7)) to train the antagonism between the generated spectra and the real spectra so that the distribution of the cross-domain spectra matches the distribution of the target spectra. The guidance of the method during training was achieved through these
loss functions. In the following section, we provide a detailed
explanation of the components of equation (2). The overall loss
function consists of the following terms:

\[
\mathcal{L} = \lambda_{\text{GAN}} (\mathcal{L}_{\text{GAN}}^{\text{synth}} + \mathcal{L}_{\text{GAN}}^{\text{obs}}) + \lambda_{\text{id}} (\mathcal{L}_{\text{id}}^{\text{synth}} + \mathcal{L}_{\text{id}}^{\text{obs}}) + \lambda_{\text{cont}} (\mathcal{L}_{\text{cont}}^{\text{synth}} + \mathcal{L}_{\text{cont}}^{\text{obs}}) + \lambda_{\text{style}} (\mathcal{L}_{\text{style}}^{\text{synth}} + \mathcal{L}_{\text{style}}^{\text{obs}}),
\]

(2)

where the weights \(\lambda_{\text{GAN}}, \lambda_{\text{id}}, \lambda_{\text{cont}},\) and \(\lambda_{\text{style}}\) are hyperparameters that control the importance of each part of the loss function. The
individual loss functions are described in detail in the following
subsections. After conducting several empirical tests, we found
that setting \(\lambda_{\text{GAN}} = 1, \lambda_{\text{id}} = 10, \lambda_{\text{cont}} = 1,\) and \(\lambda_{\text{style}} = 1\) resulted in
satisfactory performance.

### 2.2.1. In-domain spectra reconstruction loss

The encoder-decoders should be able to map spectra to latent
representations and then back to original spectra within each do-
main. This process enables the auto-encoder model to learn both
the feature extraction and generative capabilities of spectra. As
seen in Eq. (2), \(\mathcal{L}_{\text{recon}}\) compresses the data from the high-
dimensional space to the potential space, and generator \(G\) does
the opposite; that is, it converts the potential space back to the
high-dimensional space. The encoder and generator pairs that
are inverse to each other are learned by minimising the difference
between the reconstructed spectrum and the input spectrum. In
the next step, the learned encoder and generator pairs can be used
for feature extraction and restoration of the spectral data. Equation
(3) demonstrates the detailed process of in-domain spectra
reconstruction, which involves feature extraction by the encoder
followed by reconstruction through the generator and comput-
ing the L1 loss with the original spectrum. The loss within the
synthetic spectra domain can be written as

\[
\mathcal{L}_{\text{recon}}^{\text{synth}} = \mathbb{E}_{X_{\text{synth}}} [\|G_{\text{synth}}(E_{\text{synth}}(X_{\text{synth}})) - X_{\text{synth}}\|_1],
\]

(3)

where \(X_{\text{synth}}\) is the synthetic spectrum sampled from the distribution of synthetic spectra \(p(X_{\text{synth}})\) and \(E_{\text{synth}}\) and \(G_{\text{synth}}\) are the content encoder, style encoder, and generator within the
synthetic spectra domain, respectively. The L1 loss (mean abso-
lute error) is utilised to compute the loss function.

### 2.2.2. Cross-domain feature reconstruction loss

For the content and style features obtained from sampling dur-
ing cross-domain translation, we wanted to be able to reconstruc-
t them after generating the cross-domain transfer spectra (gener-
ated by combining features across domains) by encoding them
again as shown in Fig. 3. Reconstruction of features ensures that
the encoder is able to extract the content features of the shared
latent space and the style features of the unique latent space,
which is consistent with the hypothesis presented at the begin-
ing of this section. Instead of using observed spectra for style
feature extraction during training, we directly adopted the stan-
dard Gaussian distribution as the prior distribution. This con-
straint ensures that the style potential space is regularised, which
helps generate diverse spectra when transferring across domains.

Our goal is that synthetic spectra can be generated diversely by
different features provided by different observed spectra rather
than through a simple one-to-one correspondence between syn-
thetic and observed spectra. Equation (4) and (5) show the pro-
cess of the generator generating spectra in the cross-domain fol-
lowed by feature extraction by the encoder and computing the
L1 loss with the input features. The cross-domain feature recon-
struction can be achieved according to the following equation:

\[
\mathcal{L}_{\text{recon}}^{\text{obs}} = \mathbb{E}_{X_{\text{obs}}} [\|G_{\text{obs}}(E_{\text{obs}}(X_{\text{obs}})) - X_{\text{obs}}\|_1],
\]

(4)

\[
\mathcal{L}_{\text{recon}}^{\text{synth}} = \mathbb{E}_{X_{\text{synth}}} [\|G_{\text{synth}}(E_{\text{synth}}(X_{\text{synth}})) - X_{\text{synth}}\|_1],
\]

(5)

where \(c_{\text{synth}}\) and \(s_{\text{obs}}\) are the content code in the synthetic spectra
domain and the style code in the observed spectra domain,
respectively. The value of \(p(c_{\text{synth}})\) is given by \(E_{\text{synth}}(X_{\text{synth}})\), and \(q(s_{\text{obs}})\) is the prior \(N(0, 1)\).
2.2.3. Cycle consistency loss

Our goal is to ensure that spectra that have undergone cross-domain translation also retain their own information, rather than simply being mapped to the target domain (see Fig. 4). To achieve this goal, it was necessary to ensure that spectra can be transformed from one domain to another and can be reverse-transformed back to the original domain again. The principle of cycle consistency (Zhu et al. 2017) helped the generative model to achieve this goal, it was necessary to ensure that spectra can be transformed from one domain to another and can be reverse-transformed back to the original domain again. The principle of cycle consistency loss for synthetic spectra can be written as

\[
L_{\text{cy}} = \mathbb{E}_{X_{\text{obs}} \sim p(X_{\text{obs}}), X_{\text{synth-obs}} \sim p(X_{\text{synth-obs}})} \left[ \| G_{\text{obs}} \left( E_{\text{obs}}(X_{\text{obs}}), E_{\text{synth}}(X_{\text{synth}}) \right) - X_{\text{synth}} \|_1 \right],
\]

where \( X_{\text{synth-obs}} \) is the synthetic spectrum transformed across domains to the observed spectral domain, which is the main focus of this paper. The value of \( p(X_{\text{synth-obs}}) \) is given by \( G_{\text{obs}}(c_{\text{synth}}, x_{\text{obs}}) \).

2.2.4. Generative adversarial loss

The method, in addition to the auto-encoder model mentioned above, includes the discriminator model, as shown in Fig. 5, and the two parts are trained alternately to form the GAN together. The goal of the GAN is for the generator to produce spectra that are sufficient to deceive the discriminator. We wanted to generate the spectra transferred across domains to have the same data distribution as the spectra in the target domain, which requires matching the distribution of the synthetic spectra with the distribution of the observed spectra by the discriminator. In short, taking the generating adversarial process of synthetic spectra as an example, the task of the auto-encoder model is to generate synthetic spectra that are sufficiently similar to the observed spectra in order to fool the discriminator. The task of the discriminator is exactly the opposite, namely to distinguish the spectra generated by the generator from the real observed spectra, and the two models are trained alternately in order to generate the confrontation.

Taking the discriminator \( D_{\text{obs}} \) as an example, the loss function in equation (7) consists of two terms. One of the terms is related to the generator, and \( D_{\text{obs}}(X_{\text{synth-obs}}) \) is the probability that the discriminator will judge whether the generated spectra are observed spectra or not. We wanted the discriminator to be able to predict the generated spectra as negative samples, that is, the output of the discriminator tends to zero. For the generator, we wanted it to generate spectra that are sufficient to deceive the discriminator, that is, the result of \( D_{\text{obs}}(X_{\text{synth-obs}}) \) tends to one. In this way, the generator and the discriminator form an antagonistic relationship. For the second term, we required that the discriminator \( d \) is able to identify real observed spectra as positive samples, and the discriminator output tends to one. The same is true for the discriminator \( D_{\text{obs}} \). The discriminator \( D_{\text{synth}} \) was trained in the same way. The generator was trained alternately with the discriminator, forming a maximum-minimum loss training setup. In conclusion, the generative adversarial loss within the observed spectral domain is given as follows:

\[
L_{\text{GAN}} = \mathbb{E}_{X_{\text{obs}} \sim p(X_{\text{obs}}), X_{\text{synth-obs}} \sim p(X_{\text{synth-obs}})} \left[ \log \left( 1 - D_{\text{obs}}(X_{\text{synth-obs}}) \right) \right] + \mathbb{E}_{X_{\text{obs}} \sim p(X_{\text{obs}})} \left[ \log \left( D_{\text{obs}}(X_{\text{obs}}) \right) \right],
\]

where \( D_{\text{obs}} \) is a discriminator in the observed spectra domain. It attempts to distinguish between the synthetic spectra after cross-domain transfer \( X_{\text{synth-obs}} \) and the real observed spectra \( X_{\text{obs}} \).
Fig. 7. In-domain reconstruction and cycle reconstruction. This figure shows the residuals between the 5000 synthetic spectra in the test set and themselves after in-domain reconstruction and cycle reconstruction. These spectra cover most of the stellar parameter range. The blue part indicates the residuals after reconstruction of each spectrum in the wavelength range 4000 Å to 5500 Å. The colour depth is positively correlated with the number of spectra in this error. The red line in the figure indicates the average residuals, and the orange line indicates the average absolute residuals. The upper panel shows the in-domain reconstruction residuals as the relative residuals between $X_{\text{synth}} \rightarrow X_{\text{synth}}$ and the original spectra $X_{\text{synth}}$, and the lower panel shows the cycle reconstruction residuals as the relative residuals between the spectra $X_{\text{synth}} \rightarrow X_{\text{obs}} \rightarrow X_{\text{synth}}$ and the original spectra $X_{\text{synth}}$. The residuals for both the in-domain and cycle reconstructions are very small and can be neglected, as can be seen in the figure.

Fig. 8. Distribution of t-SNE visualisation domain. Spectra and features are represented visually using the t-SNE algorithm. The top-left and bottom-right panels show the visualised domain distribution of the synthetic spectra and observed spectra before and after cross-domain translation. The top-right and bottom-left show the corresponding feature distributions of the synthetic spectra and observed spectra extracted by the content encoder and the style encoder. The x-axis and y-axis coordinates represent the positions of data points in the new coordinate system after dimensionality reduction. The positions after dimensionality reduction can reflect the similarity between data points.

2.3. Stellar spectral template library construction method

The SOST can control the style of translation output by providing observed spectra. In order to generate better spectral data, a matching exercise between the synthetic spectra and the observed spectra is required. The matched observed spectra will be a better guide for the transfer of the synthetic spectra across domains. Here, the matching was done by minimising the $\chi^2$ distance. We defined $\chi^2$ as follows:

$$\chi^2 = \sum_{i} \frac{(S_i - O_i)^2}{\sigma_i^2},$$

where $S_i$ and $O_i$ are the flux densities of the synthetic spectra and the observed spectra of the i-th pixel, respectively. The term $N$ is the total number of pixels used to calculate the $\chi^2$ distance, and $\sigma_i$ is the flux density error of the observed spectra of the i-th pixel.

After obtaining the matched pairs of synthetic spectra and observed spectra, we input them into the corresponding encoder (as shown in Fig. 6) to get the corresponding content features and style features. The content features provided by the synthetic spectra were combined with the style features provided by the observed spectra in a generator for the observed spectra domain. This implemented a domain transfer from synthetic spectra to observed spectra, which we also refer to as calibration of the synthetic spectra. Next, we discuss how we used the calibrated synthetic spectra to construct synthetic template libraries with atmospheric parameters that are inherited from the original synthetic spectra.

3. Experiments and results

In this section, we first detail how the calibration of the synthetic spectra is achieved, including the construction of the experimental dataset and the pre-processing procedure. Then we
3.1. Calibration of synthetic spectra

In this section, we detail the experimental setup as well as the in-domain reconstruction, trans-domain transfer, and t-SNE visualisation of the domain distribution experiments. We show how SOST can compensate for the difference in the synthetic spectra with respect to the observed spectra, also known as the synthetic gap.

3.1.1. Experimental setup

The source of the observed spectra is the eighth data release of the LAMOST low-resolution spectral survey. To ensure the quality of the spectral data as well as a wider and uniform parameter distribution, we sampled the spectra according to the stratified sampling of the officially provided atmospheric parameters. We selected 50,000 spectra of the training set, 5,000 spectra of the validation set, and 5,000 spectra of the test set with an S/N greater than 30 in the g band. The source of the synthetic spectra is the Kurucz synthetic template library, and the parameter coverage of these synthetic spectra is 3500K $\leq T_{\text{eff}} \leq$ 8000K, in steps of 250 K; 0.0 dex $\leq$ log g $\leq$ 5.0 dex, in steps of 0.5 dex; and -2.5 dex $\leq$ [Fe/H] $\leq$ 0.5 dex, in steps of 0.25 dex. For generating the Kurucz synthetic templates used in our study, we utilised the Kurucz spectral synthesis code based on the ATLAS stellar atmosphere models provided by Bordone et al. (2004). The initial resolution of the synthetic data is 2000. To match the resolution of the LAMOST low-resolution spectra (1800), we smoothed the synthetic spectra using fast Fourier transform convolution with a Gaussian kernel (Zhang et al. 2020).

Although the Kurucz synthetic template library covers a wide range of parameters, their parameter distributions are relatively sparse, and generating synthetic spectra is relatively time-consuming. Ting et al. (2019) have demonstrated that the PAYNE model based on residual networks exhibits strong fitting capabilities, providing a good connection between physical parameters and synthetic spectra. To ensure consistency with the parameter space of observed spectra, we trained the PAYNE model using the Kurucz synthetic template library. The well-trained PAYNE model generated synthetic spectral datasets that match the parameter distribution of observed spectra. In this way, datasets of synthetic and observed spectra with consistent parameter distributions were obtained.

Next, we employed linear interpolation to interpolate the synthetic spectra and observed spectra onto the same wavelength grid, ensuring they are aligned in the same positions. The covered wavelength range extended from 3900 Å to 8700 Å, with a
Fig. 10. Some samples of SOST spectral templates. This figure shows 15 spectra of normal stellar types from the library from top to bottom based on temperature, from highest to lowest.

3.1.2. Reconstruction of synthetic spectra

As discussed in Section 2.2.1, SOST performs spectral reconstruction during training. We first show in this section that in-domain reconstruction is valid with cycle reconstruction, which is necessary for cross-domain translation.

The validity of the in-domain reconstruction and the cycle reconstruction is demonstrated in Fig. 7. The blue end of the spectrum contains most of the information needed to constrain stellar parameters, so only the blue end of the spectrum is sampled of 1 Å per pixel. For the fluxes, we performed pseudo-continuum spectral normalisation (Zhang et al. 2020, 2021; Cai et al. 2022, 2024). This removed the dispersion effect between the spectra and highlighted the chemical features, which helps when comparing and analysing the spectral features of the stars in order to study their physical properties and chemical composition. These steps play a key role in the pre-processing of spectral data.
Fig. 11. Comparisons of the results of parameter measurements using the Kurucz and SOST spectra libraries with the PASTEL database. The figure shows the comparison of the parameters $T_{\text{eff}}$, log $g$, and [Fe/H] from left to right. The samples are colour-coded according to their density. In the top row, the horizontal coordinates are the measurements provided by the PASTEL database, the vertical coordinates are the measurements from the Kurucz spectra library, and the diagonal dashed line is the reference line. The bias and scatter of the samples are labelled in the upper-left corner of the picture in the upper panel. In the bottom row, the vertical coordinates represent the results of the measurements using the SOST spectra library. As is shown, a lower bias and scatter are obtained when using the SOST spectra library compared to PASTEL database.

Fig. 12. Comparison of our results with the parameters of LAMOST DR8. A total of 6,542,528 spectral data were compared.

generally used in parameter measurements. Although the blue end of LAMOST covers wavelengths from 3700 Å $\sim$ 5900 Å, data below 4000 Å and above 5500 Å are excluded due to the low instrumental response at both ends, leaving this part of the spectrum of greater interest to us (Yang et al. 2022, 2023a,b).

The residual values corresponding to the spectral data of the test set and the histogram of the residual distribution are shown in Fig. 7. The red line indicates the average residual corresponding to each angstrom in the spectral data, and the orange line indicates the average absolute residual. The calculation results show that the average absolute residual value is less than 1% and is basically negligible, and this proves that the SOST can extract potential features from spectra and reconstruct spectra based on these features.

The phenomenon of having more residual values (exceeding one standard deviation) at certain positions of emission lines and absorption lines, such as the Balmer Hbeta line, G-line, and MgI line, does exist. The reasons for these variations remain unclear, and the complex variations of emission lines and absorption lines in the spectra may be one factor.

3.1.3. Visualisation of domain distribution using t-SNE

In this section, we use the method introduced in Section 2.3 to pair the synthetic and observed spectra in the test set. The matched spectral pairs are fed into the encoder in the SOST for feature extraction. The synthetic spectra provide content features, the observed spectra provide style features, and the two features are combined to achieve domain transfer from the Kurucz synthetic spectra to the LAMOST observed spectra. The t-SNE (Van der Maaten & Hinton 2008) is a non-linear dimensionality reduction and data visualisation algorithm that maps high-dimensional data onto a low-dimensional space in order to be able to better present the similarities and differences between the data. We used t-SNE to visualise the cross-domain translation process.

As shown in Fig. 8, we compressed the synthetic and observed spectra pairs, the content features of the synthetic and observed spectra, the stylistic features, and the synthetic and observed spectra pairs after the cross-domain translation to two dimensions using t-SNE, respectively. In the upper-left panel, one can see that there is a more significant difference between the synthetic spectra and the observed spectra. This is the synthetic
gap mentioned earlier, which is the gap we aim to overcome. The upper-right panel shows the common content features extracted by the content encoder. The lower-left panel shows that the style encoder extracts the style features of each of the two domains, which matches our hypothesis. The lower-right panel shows that the data distributions of the two domains are mixed together after the cross-domain translation, which proves that SOST overcomes the differences between the two domains better.

3.2. Estimation of stellar spectral parameters

In previous experiments, the successful transfer of the synthetic spectra across domains was demonstrated by t-SNE visualisation. In this section, we explore the effect of the corrected synthetic spectra on parameter estimation. In this experiment, template spectra were generated using SOST and parameter estimation was performed using a method based on $\chi^2$ minimization for template-matching. The estimated parameters were compared with the stellar atmospheric parameters published by the PASTEL database and LAMOST DR8. Different spectroscopic surveys usually employ different spectral analysis methods to obtain the stellar atmospheric parameters and chemical compositions and to compare them with different spectroscopic surveys in order to assess the reliability of the generated spectra (Kassounian et al. 2019).

3.2.1. Stellar spectral template library construction

We used the low-resolution A, F, G, and K type stellar parameter catalogues published by LAMOST as a reference. Grids were constructed with a range of steps of 150 K, 0.25 dex, and 0.15 dex between the $T_{\text{eff}}$ of 3500-8000 K, log $g$ of 0.0-5.0 dex, and [Fe/H] of -2.0-0.5 dex, respectively. In order to minimise the effect of LASP measurement errors, we expelled grids with fewer than five stars. A stellar parameter was randomly selected as the parameter of the template spectra within the grid, and a uniform and wide-coverage parameter distribution was ultimately constructed. The metallicity considered in our study is [Fe/H], which is consistent with the official data from LAMOST. We generated synthetic spectra corresponding to this parameter distribution using the PAYNE model. The pairs of synthetic spectra and observed spectra were then matched by the $\chi^2$ distance. Our test set included a wide range of atmospheric parameters, and to reduce the computational effort, we used the data from the test set directly for matching. The synthetic spectra were then transferred across domains to observed spectra using the SOST, where the spectral atmospheric parameter values after the cross-domain translation were inherited from the synthetic spectra. The corrected synthetic spectra form a stellar spectral template library with a broad and uniform distribution of stellar parameters. The stellar spectral template library contains 2413 spectra, and the spectra in the library have a resolution of $R \sim 1800$ covering the wavelength range 3900-8700 Å.

Our method has several advantages over previous methods of constructing synthetic template libraries and empirical template libraries. Firstly, the introduction of the PAYNE model made it easy for us to generate synthetic spectra data with any parameter. Second, the SOST better enables correction of the synthetic spectra, overcomes the synthetic gap, and does not require the a priori knowledge of the atmospheric parameters of the observed spectra. Finally, the advantages based on the first two points allowed us to easily construct a stellar spectral template library with a wider parameter coverage and more uniform parameter distributions.

The atmospheric parameter distributions of the ELODIE (Wu et al. 2011) library, which LASP derives stellar atmospheric parameters, and the SOST spectral library are shown in Fig. 9. As can be seen from the figure, our constructed stellar spectral template library has a wider parameter coverage and more uniform parameter distribution. The influence of clusters and vacancies on the parameter distribution of the traditional empirical template libraries of the measurement results has been over-

Fig. 13. SOST and LAMOST DR8 parameter differences as a function of the spectral S/N. The black dashed line in the figure is the reference line where the difference is zero. We binned the S/N ([6, 50], [50, 100], [100, 150], [150, 200], and [200, 250]). The mean deviation is shown in orange, and the scatter of the residuals is indicated by the red line in the figure.
come. Figure 10 shows some samples of SOST spectral templates. All spectra in the SOST spectral templates are normalised by pseudo-continuum normalisation. For spectra with \([\text{Fe/H}]\) greater than 0.3, the absorption lines appear deeper. Next, we use the SOST spectral library for parameter estimation in order to further demonstrate the validity of this stellar spectral template library.

### 3.2.2. Accuracy assessment by the PASTEL database

PASTEL is a database of stellar atmospheric parameters brought together from multiple sources. The parameters it provides are obtained from detailed analyses of high-resolution high-S/N spectra, thus providing more accurate measurements, and are often used to assess the accuracy of parameter measurements. In this study, the first step to assessing our obtained parameters was to remove the null values of the parameters in the PASTEL database. Then, a cross-match with the low-resolution spectral data from the eighth LAMOST release was performed with a matching radius of 3 arcsec. Stars with \(T_{\text{eff}}\) between 3500 K and 8000 K were selected. Finally, for \(T_{\text{eff}}\), \(\log g\), and \([\text{Fe/H}]\), 2190, 1872, and 1837 samples that can be compared were obtained.

In this subsection, we use the template spectral before and after the SOST correction to parameterise the observed spectra in the PASTEL database. Typically, one assesses the quality of the predictions using the deviation and scatter between the predictions and the labels of the stars. Good predictions usually have a low bias and scatter. Importantly, the bias quantifies the accuracy, whereas the scatter quantifies the precision of the method.

As can be seen in Fig. 11, after SOST correction, the predictions of the stellar atmospheric parameters \(T_{\text{eff}}\), \(\log g\), and \([\text{Fe/H}]\) show a significant improvement. The bias decreases from 144 K, -0.08 dex, and -0.33 dex before correction to -20 K, -0.01 dex, and 0 dex, respectively, while the dispersion decreases from 140 K, 0.31 dex, and 0.21 dex to 121 K, 0.26 dex, and 0.13 dex, respectively, before correction. The significant improvement of these metrics suggests that the SOST-corrected spectra are closer to the observed spectra, and therefore more accurate measurements of stellar parameters can be obtained. This further demonstrates the effectiveness of the SOST construction template library.

### 3.2.3. Comparison with the results of the LAMOST DR8 data survey

LAMOST DR8 is the eighth data release of LAMOST, a large sky area multi-object fibre-optic spectroscopic telescope developed by China for large-scale spectroscopic surveys. LAMOST uses LASP to estimate the \(T_{\text{eff}}\), \(\log g\), and \([\text{Fe/H}]\) of stars. In our study, we selected a total of 6,542,528 spectral data in the \(T_{\text{eff}}\) range of 3500 to 8000 K for parameter measurements and compared the results with those released by LAMOST DR8. Subsequently, we utilised a template-matching method based on \(\chi^2\) minimisation, using the SOST spectral library, to derive the stellar parameters of the LAMOST DR8 data. This method selects two templates with the minimum \(\chi^2\) for the target spectrum, assigns weights based on their \(\chi^2\) distances, and then calculates the weighted average as the measurement result. From Fig. 12, it can be seen that compared with the parameter...
ric results published by LAMOST DR8, our results have a mean- 
measurement deviation of -40 ± 121 K for T eff, -0.05 ± 0.22 dex for 
log g, and -0.01 ± 0.16 dex for [Fe/H]. Overall, our results are in 
good agreement with the parameter results published by LAM- 
OST DR8, indicating that the template library we constructed 
achieves satisfactory results in stellar parameter measurements.

The relationship between the parameter difference and S/N 
of the measurements using the SOST and LAMOST DR8 is 
shown in Fig. 13. We chose spectra with S/N between 6 and 250. 
This choice was due to the fact that the small number of spectra 
with S/N greater than 250 lose statistical significance. From the 
figure, it can be seen that with the increase in S/N, the devia-
tion and scatter of our predictions from the LASP residuals do 
not change significantly, and they still agree more closely with 
the LASP results. When the S/N is low (less than ten), the accu-
0.6, -0.2]; [-0.2, 0.2]; and [0.2, 1.0]. In each grid, the di-
gress may be due to the small number 
of stars with log g less than two in ELODIE and the template li-
brary used by LASP, which leads to the deviation of the results 
(Chen et al. 2022). For stars with [Fe/H] in the range [-3.0, -1.5], our SOST method generally gives higher values, and the differ-
ences are more pronounced. This may be due to limitations in 
the parameter coverage of spectral data currently released using 
LAMOST DR8. Our template library lacks adequate coverage 
of low-metallicity stars, causing the results to be pulled towards 
more metal-abundant quantities during template matching.

Taken together, our measurements differ in some cases from 
the parameters of LAMOST DR8, and these differences may be 
influenced by the coverage of the template library and stellar 
properties. It is important to note that despite the differences, 
our estimates still show some degree of agreement with the at-
mospheric parameters published by LAMOST.

4. Conclusions

In traditional stellar spectral template library construction ef-
forts, the construction is usually based on observed spectra or 
synthetic spectra. We have proposed a template library construc-
tion method, called SOST, that is different from previous meth-
ods. This method is based on a GAN and uses an auto-encoder
to separate and recombine the content and style features of spec-
tra. The content features of synthetic spectra are cross-domain 
combined with the style features of observed spectra, achieving 
a cross-domain transfer of synthetic spectra. This process ac-
complishes the correction of synthetic spectra, and the corrected 
synthetic spectra can have a good match with the observed spec-
tra that are to be measured. The distinctive feature of this method 
lies in its elimination of the need for spectral expertise and man-
ual feature extraction.

With this approach, we calibrated the Kurucz synthetic spec-
tra using LAMOST observed spectra, which in turn led to the 
construction of a widely distributed and uniform template library 
of stellar spectra. The template library includes a total of 2413 
spectra. It covers a parameter space ranging from 3500 to 8000 K 
for effective temperature, 0.0 to 5.0 dex for surface gravity, and 
-2.0 to 0.5 dex for metallicity, while T eff, log g, and [Fe/H] were 
constructed in grids with step ranges of 150 K, 0.25 dex, and 
0.15 dex, respectively. We used the library to perform param-
eter measurements of the PASTEL database, and after method 
correction, the accuracy of the predictions of the atmospheric 
parameters T eff, log g, and [Fe/H] was significantly improved, 
decreasing from 140 K, 0.31 dex, and 0.21 dex to 121 K, 0.26 
dex, and 0.13 dex. Finally, we also used our stellar spectral tem-
plate library to estimate the atmospheric parameters of more than 
6.5 million spectral data released by LAMOST DR8.

The automatic calibration principle of SOST is versatile and 
can be applied to reduce discrepancies among various astronom-
ical projects. Moreover, we plan to explore its applicability to 
various astronomical projects. The work in constructing the 
template library contributes to stellar parameter measurements, 
establishing a solid foundation for future studies on the stellar pop-
ulations, kinematics, and chemistry of the Galactic disc, along 
with its evolutionary history. Our results further demonstrate the 
significance of deep learning in advancing astronomical tech-
niques.

Acknowledgements. We thank the anonymous referee for valuable sugges-
tions. The work is supported by the National Natural Science Foundation of China (Grant Nos. U1931209, 62272336, 62306204), the Projects of Sci-
technology Cooperation and Exchange of Shanxi Province (Grant 
Nos. 2022004041101037, 2022004041101033), the Fundamental Research Pro-
gram of Shanxi Province (Grant Nos. 20210302122323, 202103021224275, 
202103021223267), the Taiyuan University of Science and Technology Scientific 
Research Initial Funding (No. 20210205), the Reward for Outstanding 
Doctoral Work in Shanxi (No. 20222107).

Guoshoujing Telescope (the Large Sky Area Multi-Object Fiber Spectroscopic 
Telescope, LAMOST) is a National Major Scientific Project built by the Chinese 
Academy of Sciences. Funding for the project has been provided by the National 
Development and Reform Commission. LAMOST is operated and managed by 
the National Astronomical Observatories, Chinese Academy of Sciences.

References

Bialek, S., Fabbro, S., Venn, K. A., et al. 2020, Monthly Notices of the Royal 
Astronomical Society, 498, 3817
Cai, J., Yang, Y., Yang, H., Zhao, X., & Hao, J. 2022, ACM Transactions on 
Knowledge Discovery from Data (TKDD), 16, 1
Cai, J., Zhang, M., Yang, H., et al. 2024, Expert Systems with Applications, 
123587
Castelli, F. 2005, Memoria della Societ`a Astronomica Italiana Supplement, v. 8, 
p. 25 (2005), 8, 25
Chen, Y. H., Li, G. W., & Shu, H. 2022, Research in Astronomy and Astro-
physics, 22, 050008
Series, 240, 10
& Astrophysics, 532, A95
Appendix A: SOST architecture

Here, we summarise the detailed architecture of each sub-module of SOST in Table A.1. Each row represents the operations applied by each network layer, including convolution, normalisation, and activation functions. The CONV layer denotes a standard 1D convolutional layer; the DCONV layer represents a transposed convolutional layer; FC stands for a fully connected layer; IN stands for InstanceNorm1d; ReLU and LeakyReLU are used as activation functions; InstanceNorm1d and AdaIN are used for normalisation; AdaptiveAvgPool1d and AvgPool1d layer denote a pooling layer. For each layer, N, K, S, and P represent the number of filters, the size of the filter kernel, the stride of the convolution operation, and the size of the padding for the convolution operation, respectively. The input spectrum has 4800 pixels. After dimensionality reduction by the encoder, the content code dimension was $256 \times 298$, while the style code dimension was $8 \times 1$.

Table A.1. Detailed architecture of the SOST sub-module.

<table>
<thead>
<tr>
<th>Layer</th>
<th>Content_Encoder_Down − sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CONV − (N64, K7, S4), IN(N64), ReLU</td>
</tr>
<tr>
<td>2</td>
<td>CONV − (N128, K4, S2), IN(N128), ReLU</td>
</tr>
<tr>
<td>3</td>
<td>CONV − (N256, K4, S2), IN(N256), ReLU</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layer</th>
<th>ResidualBlock</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CONV − (N256, K7, S1, P3), IN(N256), ReLU</td>
</tr>
<tr>
<td>2</td>
<td>CONV − (N256, K7, S1, P3), IN(N256)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layer</th>
<th>Style_Encoder_Down − sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CONV − (N64, K7, S4), IN(N64), ReLU</td>
</tr>
<tr>
<td>2</td>
<td>CONV − (N128, K7, S2), IN(N128), ReLU</td>
</tr>
<tr>
<td>3</td>
<td>CONV − (N256, K7, S2), IN(N256), ReLU</td>
</tr>
<tr>
<td>4</td>
<td>AdaptiveAvgPool1d(output_size = 1)</td>
</tr>
<tr>
<td>5</td>
<td>CONV − (N8, K1, S1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layer</th>
<th>Decoder_MLP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FC(IN = 8, OUT = 256), ReLU</td>
</tr>
<tr>
<td>2</td>
<td>FC(IN = 256, OUT = 256), ReLU</td>
</tr>
<tr>
<td>3</td>
<td>FC(IN = 256, OUT = 3072)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layer</th>
<th>Decoder_Up − sampling</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DCONV − (N128, K4, S2), AdaIN(N128), ReLU</td>
</tr>
<tr>
<td>2</td>
<td>DCONV − (N64, K4, S2), AdaIN(N64), ReLU</td>
</tr>
<tr>
<td>3</td>
<td>DCONV − (N1, K7, S4)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Layer</th>
<th>Discriminator</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DCONV − (N64, K4, S2), LeakyReLU</td>
</tr>
<tr>
<td>2</td>
<td>DCONV − (N128, K4, S2), IN(N128), LeakyReLU</td>
</tr>
<tr>
<td>3</td>
<td>DCONV − (N256, K4, S2), IN(N128), LeakyReLU</td>
</tr>
<tr>
<td>4</td>
<td>DCONV − (N512, K4, S2), IN(N128), LeakyReLU</td>
</tr>
<tr>
<td>5</td>
<td>DCONV − (N512, K3, S1)</td>
</tr>
<tr>
<td>6</td>
<td>AvgPool1d(K1, S2)</td>
</tr>
</tbody>
</table>