Modelling and Classification of Optical Beam Profiles using Vision Transformer

Yu Dian Lim\textsuperscript{1} and Chuan Seng Tan\textsuperscript{1}

\textsuperscript{1}Affiliation not available

June 07, 2024
Modelling and Classification of Optical Beam Profiles using Vision Transformer

Yu Dian Lim, Member, IEEE, and Chuan Seng Tan, Fellow, IEEE

Abstract—A vision transformer (ViT) is developed to perform image classification on beam profiles coupled out from silicon photonics (SiPh) gratings. The image classification task is aimed to distinguish ‘converged’ and ‘diverged’ beam profiles, and the regions where the corresponding beam profiles are located above the SiPh gratings. Upon training with 1247 beam profile images, the ViT model is able to perform 6-category image classification task on 832 beam profile images with classification accuracy of 0.989. Since the training of ViT is probabilistic in nature, the ViT training is repeated for 100 times to test its robustness. Classification accuracy ranges from 0.83 to 0.99, where 82/100 runs with testing accuracy values of >0.95, are obtained.

Index Terms— Computer Vision, Machine Learning, Silicon Photonics, Vision Transformer

I. INTRODUCTION

R ecently, transformer neural networks have gained much attention among the machine learning community. Due to their self-attention mechanism, the transformer's ability to process long, parallel sequences of data exhibits superior performance compared to traditional recurrent neural networks (RNNs) in handling sequential data types [1].Attributing to this superior performance, transformer neural networks have been widely used in the field of natural language processing (NLP) [2–4]. Leveraging on its scaling success in NLP, Google Inc. attempted to apply transformer neural networks to image classification with minimal modifications, resulting in the vision transformer (ViT) model [5]. It has been found that the ViT model achieves excellent results compared to state-of-the-art convolutional neural networks (CNNs) which has been a conventional choice in image recognition. For instance, Maruricio et al. investigated that ViT models are more robust than CNNs for noisy and augmented images. Additionally, ViTs are able to learn information better with fewer images, as the images are divided into patches, allowing for the identification of diverse relationships among them [6]. Meanwhile, Sherwani et al. observed that vision transformers give better outputs compared to other deep learning algorithms when performing Alzheimer's disease detection via MRI scans. The ViT exhibits better performance than CNNs when trained on smaller datasets [7].

On the other hand, several works on the integration of machine learning-based image classification in the field of photonics have been reported. For instance, Xiong et al. demonstrated the usage of CNN classification techniques to recognize orbital angular momentum modes on vortex beams, achieving an average recognition accuracy exceeding 99% [8]. Thakur et al. reported the usage of a CNN model in spatio-temporal analysis and classification of biospeckle data, achieving classification accuracy ranging from 90 to 97% [9]. Smolina et al. reported the usage of a CNN model in the classification of topological phases in finite leaky photonic lattices using limited measurement data, with a classification accuracy of up to 90% [10].

In the realms of ViT in photonics, Jin et al. reported the application of ViT-empowered physics-driven deep neural network which can realize the generation of omnidirectional 3D holograms [11]. Meanwhile, Dong et al. propose a novel unsupervised autoencoder-based ViT to synthesize phase-only holograms. Both simulated and experimental results have demonstrated the effectiveness of the proposed method [12]. However, despite the advantages of ViTs over CNNs in image classifications, the application of ViT in photonics researches are limited. Utilizing the high parallelization, and requirement of less datasets to achieve high accuracy in image classification, more works on the application of ViT in photonics-related researches are called for.

II. BACKGROUND AND MOTIVATION

Over the past decade, silicon photonics (SiPh) has gained much attention due to its feasibility in realizing high speed, minimal loss data transmission applications [13]. attributed to the recent boom in artificial intelligence (AI) where mass amount of data to be generated, processed, and transmitted, the demand for data-related SiPh devices, such as transceivers in data centers, rises exponentially [14]. Besides its applications in data-related regimes, SiPh has also been widely used in free-space applications, including sensing [15], display [16], and Light Detection and Ranging (LiDAR) devices [17]. Apart from that, grating couplers in SiPh has been used in quantum computing related applications. One of the pioneering works was reported by Mehta et al., where gratings are used to perform optical addressing of trapped ion qubits via chip-to-free-space coupling of laser light [18,19]. The coupling of laser light from grating to trapped ion can be illustrated in Fig. 1(a). In optical addressing of trapped ions,
the application of gratings, in place of traditional optical fiber, greatly reduces the formfactor of ion trap quantum computing devices, and further enhances the manufacturability of quantum computing chips.

![Fig. 1. FDTD simulation model light coupled out from grating to free-space, Grating couplers with (b) R = 15 µm, (b) R = 60 µm](image)

In our previous works, we have attempted to improve the focusing of chip-to-free-space light by modifying the grating designs, such as varying the radius of curvature (R) of the pitch structure in gratings [20,21], or introducing mixed pitch gratings for enhanced focusing [22]. The R values of grating couplers are illustrated in Fig. 1(b) and (c). From our previous studies, we observed that changes in R values vary the shape of the corresponding beam profiles along y-axis, with limited impact on the beam profiles along x-axis (axis reference: Fig. 1(a)) [20,23]. As shown in Fig. 2, the beam waist along y-axis reduces as R reduced from 15 to 20 µm, after that, the beam waist increases gradually as R increases from 20 to 60 µm. The details of the variation of beam profiles with changing R values are discussed in our previous works [20,23].

![Fig. 2. FDTD-simulated beam profiles of 1092 nm light coupled out from gratings with various R values (Grating location: z = 5 µm, Beam profile location: z = 40 µm)](image)

Besides the grating design and the corresponding beam profiles, for the optical addressing of trapped ion with SiPh gratings, the beam propagation of light along z-axis (axis reference: Fig. 1(a)) is important. At various heights along z-axis, the beam profiles have different morphological appearance. For instance, the reactive field region ranges between the grating to $0.62 \sqrt{\frac{\Delta^2}{\lambda}}$ above the grating, where D is the linear dimension of the grating and $\lambda$ is the wavelength. At this region, the beam profile appears to be irregular, with visible fringes observed in the profile. In contrast, for beam profiles at the far-field region of $>\left(\frac{\Delta^2}{\lambda}\right)$, they possessed more Gaussian-like features, with less irregularities [24]. On the other hand, the height where the ions are trapped can be varied by adjusting the DC/RF signals fed into the ion trap, and the dimension of the ion trap [25–27]. Thus, to place the grating where it can accurately "aim" the coupled light towards the trapped ions, the corresponding beam profiles of light (from gratings) at various heights should be visualized and understood.

For the optical addressing of trapped ions, laser light ranging from visible to near-infrared (IR) wavelengths is often utilized, depending on the atomic energy levels in the trapped ions. In typical experimental setups, the beam profiles of light coupled out from gratings are often visualized using an IR camera. It is often difficult to determine the corresponding height position of a captured beam profile instance using the IR camera. Meanwhile, when using an IR camera for beam profiling, the grating couplers are often imperceptible to the IR camera. For instance, if we simultaneously observe the beam profiles from gratings with R = 15 µm and R = 60 µm under the IR camera, it is often challenging to ascertain which profile belongs to which grating, as the on-chip labeling of the grating is not visible under the IR camera. Hence, a real-time image recognition system capable of determining the corresponding grating and the height position of a beam profile instance is necessary.

To achieve this, a vision transformer (ViT) model can be applied to perform beam profile classification task on the light coupled out from gratings. As a preliminary attempt, finite-difference time-domain (FDTD) simulation technique is used to simulate beam profiles from R = 15 – 60 µm gratings. The simulated beam profile images are categorized into different categories, according to their beam focusing and their corresponding locations above the grating. Then, the categorized beam profiles are split into training and testing sets. For training sets, the beam profiles will be fed into the ViT model for parameters training. After the model training, the classification accuracy of the ViT model will be tested.

### III. FDTD Simulation and Beam Profiles Categorization

The beam profiles of light coupled out from gratings are simulated using finite-difference time-domain (FDTD) technique. Fig. 1(a) shows the simulation model. In a typical FDTD simulation, laser light with wavelength of 1,092 nm is input the grating via connecting waveguide. The wavelength of 1,092 nm is selected based on the energy levels in $^{88}$Sr$^+$ ions, where 1,092 nm corresponds to the ‘clear-out’ function in $^{88}$Sr$^+$ trapped ion quantum bits [21,28]. The input light will then propagate through the waveguide, and coupled out to free-space through the grating. The electric field (E-field) distribution of the output light will be recorded by the three-dimensional (3D) monitor. The gratings used in the simulation are 0.4 µm thick silicon nitride (SiN) etch-through grating with 0.5 duty cycle, 1.2 µm pitch or 1.1/1.2 µm mixed pitch, ~10 µm grating length along x-axis. The waveguide/grating
structure is sandwiched between 3 µm top oxide (TOX) layer and 3 µm bottom oxide (BOX) layer. After obtaining the E-field data in 3D data array, the beam profiles in xy-plane, along z-axis, are plotted and stored (axis reference: Fig. 1(a)).

To visualize the light coupling of light from grating to free-space, the E-field distribution of light in xz-plane is plotted, as shown in Fig. 3(a). The location of the grating is in z = 5 µm. Considering 3 µm TOX layer above the grating, only beam profiles from z = 10 µm onwards are taken into considerations. For the categorization of beam profiles by their location along z-axis, the beam profiles are segregated into Region A, B and C. Beam profiles in Region A falls between z = 10 to 20 µm, in Region B falls between z = 20 to 30 µm, and beam profiles in Region C falls between z = 30 to 40 µm.

Apart from categorizing the beam profiles according to their corresponding z-position, we attempted to perform categorization according to their beam morphologies too. From Fig. 2, it was observed that the beam profile from R = 20 µm grating is more focused in the y-axis, while the beam profiles from other investigated R values appeared to be relatively sparse, unfocused along y-axis. In terms of optical addressing of trapped ion qubits, focused beam profiles from R = 20 µm grating is undoubted the preferred option. However, it is also important for the vision transformer (ViT) model to identify unfocused beams from R = 15, 30, 40, 50, 60 µm gratings. Thus, we categorized beam profiles from R = 20 µm as ‘converged’ profiles and beam profiles from R = 15, 30, 40, 50, 60 µm gratings as ‘diverged’ profiles. However, such arrangement will result in imbalance distribution of beam profile images, as ‘diverge’ beam profiles will be five times more than ‘converged’ beam profile images. To have more ‘converged’ beam profiles, beam profiles from mixed pitch gratings with R = 20 µm are also included in the ‘converged’ category. From our prior study on mixed pitch gratings, the mixing of grating pitch only affects the beam waist along the x-axis, with minimal impact on the beam morphology along the y-axis [22]. Thus, all beam profiles from R = 20 µm mixed pitch gratings can be categorized as ‘converged’ beam profile images. The details of all categorizations used in this work is displayed in Table 1.

From the ‘Category’ column of Table 1, category 0, 1, 2, 3, 4 and 5 are labelled. This is to facilitate the ViT model, where the ViT uses the features (data used for classification: e.g. input beam profile images) to classify the label (data to be classified: category 0, 1, 2, 3, 4 or 5). The usage of these labelling will be mentioned in the next section. The typical beam profiles of each category are shown in Fig. 3(b) to (g). Across Region A, B, and C, it can be observed that beam profiles in Region A shows fringe-like morphologies. The fringe-like morphologies are visible in both converged and diverged profiles, as exhibited in Fig. 3(b) and (e). The fringe-like morphologies resemble the shape of the gratings shown in Fig. 1(b) and (c), as the beam profiles are relatively near to the gratings. In Region B, the fringes-like morphologies disappeared. However, a dual maxima E-field is observed across x-axis. This can be attributed to the two prominent beams (labeled as ‘1’ and ‘2’ in Fig. 3(a)) in the Region B, possibly due to the presence of multiple modes when coupling light from grating to free-space. The dual maxima phenomenon is observed in both converged and diverged profiles, as exhibited in Fig. 3(b) and (e). In Region C, only 1 prominent maximum along x-axis is observed, as the intensity of second beam (labelled as ‘2’) in Region B is reduced. On the other hand, comparing both converged and diverged categories, beam profiles in converged categories have smaller beam waist along y-axis, with lower y/x beam waists aspect ratio. In contrast, the beam profiles in diverged categories have larger beam waist along y-axis, with higher aspect ratio.

**Table I**

<table>
<thead>
<tr>
<th>Category</th>
<th>Details of All Categorizations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Region A – Converged (0)</td>
<td>Radius: 1.2 µm, mixed pitch (5:1/1.2) (mixing ratios: 2/8, 4/6, 5/5, 6/4, 8/2). Prior work: ref. [22]</td>
</tr>
<tr>
<td>Region B – Converged (1)</td>
<td>R = 20 µm, 1.1/1.2 µm, mixed pitch (mixing ratios: 2/8, 4/6, 5/5, 6/4, 8/2). Prior work: ref. [22]</td>
</tr>
<tr>
<td>Region C – Converged (2)</td>
<td>R = 20 µm, 1.2 µm, mixed pitch (mixing ratios: 2/8, 4/6, 5/5, 6/4, 8/2). Prior work: ref. [22]</td>
</tr>
<tr>
<td>Region A – Diverged (3)</td>
<td>R = 15, 30, 40, 50, 60 µm, Prior work: ref. [21]</td>
</tr>
<tr>
<td>Region B – Diverged (4)</td>
<td>R = 20 µm, 1.2 µm, mixed pitch (mixing ratios: 2/8, 4/6, 5/5, 6/4, 8/2). Prior work: ref. [22]</td>
</tr>
<tr>
<td>Region C – Diverged (5)</td>
<td>R = 20 µm, 1.2 µm, mixed pitch (mixing ratios: 2/8, 4/6, 5/5, 6/4, 8/2). Prior work: ref. [22]</td>
</tr>
</tbody>
</table>

**Fig. 3.** (a) E-field distribution of light in xz-plane, typical beam profile images of: (b) Region A – Converged (Category 0), (c) Region B – Converged (Category 1), (d) Region C – Converged (Category 2), (e) Region A – Diverged (Category 3), (f) Region B – Diverged (Category 4), (g) Region C – Diverged (Category 5)

**IV. ViT Training and Evaluation**

After simulating and categorizing the beam profile images, the images will be subject to vision transformer (ViT) model
training and evaluation. Fig. 4 illustrates the ViT model. As indicated in Table 1, 2079 beam profiles images from 6 categories are generated using finite-difference time-domain (FDTD) technique. The 2079 beam profiles are then split, where 60% (1247 images) will be used for ViT model training, and 40% will be used for ViT model evaluation (832 images). Within the ViT model, each beam profiles will first be split into 64 patches. The size of each patch is 8×8 pixels. For all 64 pixels in each patch, each pixel consists of 3 possible mixtures of red, green and blue (RGB) information. Thus, for each patch, the pixels and the RGB information will be flattened to a linear vector with 8×8×3 = 192 size. For a beam profile image that broke into 64 patches, the image will be resized to a two-dimensional (2D) data array with size of 64×192, where the 64 rows represent the number of patches, and 192 columns represent the patch size and the RGB information.

The 64×192 resized 2D data array will then be subjected to linear projection to dense neural network (DNN) layer with 192 nodes, followed by positional embedding where numerical values of 1 – 64 are added to each row of the 2D data array. After 1st layer of normalization, the 2D data array will be subjected to self-attention process, where the details of the self-attention process is described in ref. [1]. The output of the self-attention layer is then added with the output of positional embedding, followed by 2nd layer of normalization. Then, the 2D data array is subjected to multi-layer perceptron (MLP) with two layers of DNN with 192 nodes. The output of the MLP is added to the output of the 2nd normalization layer. When it comes to the end of ‘transformer encoder’ process, 2D data array with 64×192 size is flattened to 1D data array with 12288 size. The 1D data array is then subjected to MLP head where 2 layers of DNN with 128 nodes are used. After that, the corresponding category code (0, 1, 2, 3, 4 or 5) of the given image is then provided/predicted (‘provided’ when performing ViT training, ‘predicted’ when performing ViT evaluation). The activation function of all DNN layers is ‘gelu’ activation. The details of the fundamental concepts in ViT model are explained in ref. [5].

As mentioned earlier, the ViT model was trained using 1247 beam profile images. The number of epochs is preliminary set at 100, where the 1247 beam profiles will be inserted into the ViT model for 100 times to optimize the internal parameters in the model. Meanwhile, validation split ratio of 0.5 is used, which means 50% of the training beam profiles will be used for validation. To minimize the possible underfitting and overfitting of ViT model, we used the value of ‘[training loss - validation loss]’ as the customized figure-of-merit when training the ViT model for best data-to-model fitting. Since the ViT model training can be time consuming, we included an early stopping mechanism: if the ‘[training loss - validation loss]’ values of each epoch did not reduce for 30 consecutive epochs, the training will stop, and the ViT model will be restored to the epoch where minimum ‘[training loss - validation loss]’ value is obtained. For our case, the training halted prematurely at the 56th epoch, and the model restored to the 26th epoch where minimum ‘[training loss - validation loss]’ value was obtained. The full Python code for the construction, training, and evaluation of ViT model used in this work, and the customized ‘[training loss - validation loss]’ figure-of-merit, is given in ref. [29].

Fig. 5(a) shows the training and validation losses. Several peaks can be observed, indicating that the losses are unusually high at several instances. As the model training can be probabilistic, we restored the ViT model to the epoch where the model best fits the training images. At the optimal 26th epoch, training loss - validation loss = -8.06. On the other hand, the training and validation accuracies are shown in Fig.
It can be observed that the training accuracy improved significantly between 1st to 5th epoch. Similar peaks occurred in Fig. 5(a) are also observed in Fig. 5(b). At the optimal 26th epoch, training accuracy - validation accuracy = 0.01. Having lower training loss than validation loss, and higher training accuracy that validation accuracy, suggested that the model is slightly overfitted [30]. However, the overfitting is insignificant, as reflected in the training and validation accuracy values in 26th epoch. The training and validation accuracies are 0.986 and 0.976, respectively, with slightly more than 1% deviations between the two. Thus, it can be deduced that the ViT model is currently well-fitted at the 26th epoch.

After the training of ViT model, the model is used to classify the beam profile images. Using the patched beam profile shown in Fig. 4, the beam profile image is obtained from R = 30 µm grating at z = 35 µm. Thus, it belongs to ‘Category 5: Region C, Diverged’ category (ref. Table 1). As referred to the previous explanation on the ViT model (Fig. 4), the beam profile image is first reshaped to 2D data array with 64×192 dimension, as shown in Fig. 6(a). The values in the data array ranges between 0 to 255, reflecting the 256-color code in the RGB regime. After the linear projection, positional embedding, and self-attention layers, the 64×192 data array transformed into Fig. 6(b), where most values in the data array is near zero, except several instances of high positive and high negative values to extract the key features in each patch.

As mentioned earlier, after the self-attention layer, the 64×192 data array will undergo MLP layer, and several addition/normalization steps in the transformer encoder block. Right after the transformer encoder block, the 64×192 data array will be flattened to 12288 1D data array, as shown in Fig. 7(a). After the 1st layer (Fig. 7(b)) and 2nd layer (Fig. 7(c)) of MLP head, the image classification layer in Fig. 7(d) shows that category 5 has higher value than other categories (from 0 to 4). This shows that the ViT model correctly classified the given beam profile image as ‘Category 5: Region C, Diverged’ category. The full Python for the layer-by-layer plots of the output of ViT model (shown in Fig. 6 and Fig. 7) in each layer can be found in ref. [31].

To test the classification accuracy of the trained ViT model, we repeated the classification process (Fig. 6 and Fig. 7) on all 832 testing beam profile images. The outcome of the classification is shown in Fig. 8. It can be observed that the ViT model made no mistaken in classifying if a beam profile image is converged or diverged. However, some glitches occur when the ViT model tries to distinguish beam profiles occurring at various regions. For instance, in the ‘converged’ regime, the ViT model classified 1 beam profile to fall in Region B but it actually falls in Region A. Additionally, in the ‘diverged’ regime, the ViT model classified 3 beam profiles to fall in Region B but it actually falls in Region A. This can be due to the overlapping features of beam profiles in Region A and Region B, especially the profiles occurring near z = 20 µm height. Nevertheless, the trained ViT model correctly classified 823 beam profile images from 832 testing beam profile images, achieving outstanding classification accuracy of 0.989.
Fundamentally, the 1092 nm laser light coupled from gratings exhibits Gaussian propagation features as it propagates to free-space. The beam waist increases gradually as the beam propagates along z-axis (axis ref. Fig. 1) above Region A [22,32]. Meanwhile, it was shown in Fig. 2 that the beam waist (along y-axis) of the profile increases as the radius of curvature of the grating, R, increases from 20 to 60 µm. This means that, by varying the R values, it is possible to have two beam profiles, each falling in Region B and Region C, to have similar beam waist and beam shape. This was shown in Fig. 3(f) and (g), where the beam dimensions are almost identical. The difference between Fig. 3(f) and (g) is the electric field (E-field) distribution, where Fig. 3(f) that falls in Region B has a dual maxima E-field along x-axis, while Fig. 3(g) that falls in Region C has only one prominent peak along x-axis. It is often challenging for human eyes to distinguish such minor differences in beam features. Thus, a vision transformer (ViT) model can assist in performing the beam profile classification task.

However, as mentioned earlier, the training of ViT model can be probabilistic. Similarly, the splitting of beam profile images into training (60%) and testing (40%) beam profile images is also probabilistic. To test the robustness of the ViT model, we repeatedly-run the splitting-training-testing process for 100 times, with other factors, including ViT model architecture and customized figure-of-merit, to remain identical. Fig. 9 shows the outcome of the 100 runs. Generally, the testing accuracy ranges from 0.83 to 0.99, where 82/100 runs have testing accuracy values of >0.95. There are several instances where training accuracy is significantly lower than validation accuracy, this was also reflected in Fig. 9(b) where training accuracy - validation accuracy < 0. As the splitting-training-testing process is probabilistic, occasional outliers in accuracy values are unavoidable. Nevertheless, > 80 runs have ‘training accuracy – validation accuracy’ values falling between -0.05 to +0.05. This suggests that the ViT model has good level of consistency upon repeated splitting-training-testing runs.

**Fig. 7.** One-dimensional data array: (a) Right after the transformer encoder block, (b) at 1st layer of MLP head, (c) at 2nd layer of MLP head, (d) at image classification layer

**Fig. 8.** Confusion matrix on classifying 832 testing beam profile images

**V. DISCUSSION AND EXPLORATIVE INVESTIGATIONS**

Fundamentally, the 1092 nm laser light coupled from gratings exhibits Gaussian propagation features as it propagates to free-space. The beam waist increases gradually as the beam propagates along z-axis (axis ref. Fig. 1) above Region A [22,32]. Meanwhile, it was shown in Fig. 2 that the beam waist (along y-axis) of the profile increases as the radius of curvature of the grating, R, increases from 20 to 60 µm. This means that, by varying the R values, it is possible to have two beam profiles, each falling in Region B and Region C, to have similar beam waist and beam shape. This was shown in Fig. 3(f) and (g), where the beam dimensions are almost identical. The difference between Fig. 3(f) and (g) is the electric field (E-field) distribution, where Fig. 3(f) that falls in Region B has a dual maxima E-field along x-axis, while Fig. 3(g) that falls in Region C has only one prominent peak along x-axis. It is often challenging for human eyes to distinguish such minor differences in beam features. Thus, a vision transformer (ViT) model can assist in performing the beam profile classification task.

However, as mentioned earlier, the training of ViT model can be probabilistic. Similarly, the splitting of beam profile images into training (60%) and testing (40%) beam profile images is also probabilistic. To test the robustness of the ViT model, we repeatedly-run the splitting-training-testing process for 100 times, with other factors, including ViT model architecture and customized figure-of-merit, to remain identical. Fig. 9 shows the outcome of the 100 runs. Generally, the testing accuracy ranges from 0.83 to 0.99, where 82/100 runs have testing accuracy values of >0.95. There are several instances where training accuracy is significantly lower than validation accuracy, this was also reflected in Fig. 9(b) where training accuracy - validation accuracy < 0. As the splitting-training-testing process is probabilistic, occasional outliers in accuracy values are unavoidable. Nevertheless, > 80 runs have ‘training accuracy – validation accuracy’ values falling between -0.05 to +0.05. This suggests that the ViT model has good level of consistency upon repeated splitting-training-testing runs.

**Fig. 9.** (a) Testing, training, and validation accuracies, (b) Training accuracy – validation accuracy, of 100 repeated runs
To further explore the possibilities of improving the classification accuracy, we attempted to scale-up the ViT model. The ViT model was scaled-up to 2, 3, 4 transformer encoders, with 256, 512, and 1024 nodes in the MLP head layer (MLP head size), as shown in Fig. 10. Generally, increasing the number of transformer encoders do not yield significant improvement on the classification accuracies. For instance, when MLP head size = 128, the classification accuracies reduced to 0.948 and 0.956 when more transformer encoders are used. Similarly, when number of transformer encoder = 1, the classification accuracies reduced to 0.982, 0.926, and 0.983 when MLP head size increases to 256, 512, and 1024. The scaling up of ViT model will result in the increase of model parameters. For instance, scaling up the ViT model from MLP head size = 128, to number of transformer encoders = 1; to MLP head size = 1024, to number of transformer encoders = 4, increases the model parameters from 1,862,726 to 14,581,446. As transformer models are fundamentally data-hungry models [33], higher number of model parameters may result in high level of uncertainties when training it. Since only 1247 beam profile images are available in this study, a smaller scale ViT model of MLP head size = 128 with only 1 transformer encoder is preferred.

![Classification accuracies of ViT model with various architectural combinations](image)

As mentioned earlier, transformer model is fundamentally data-hungry. For its application in natural language processing (NLP), the parameters to be trained often range from 70 million to over 16 billion [4,34]. Thus, a large amount of training data is required. In the context of ViT model for image recognition, the ViT model can be flexibly-scaled to accommodate various data size. For instance, Dosovitskiy et al. from Google Inc. reported usage of ViT model on image dataset with 10 million to 30 million images [5]. For smaller datasets, Pandya et al. reported training of ViT model using less than a thousand image for facemask detection [35]. Meanwhile, Sherwani et al. reported using 6400 imaging scans for Alzheimer’s disease detection [7]. The number of images used in these reported works benchmarked favorably with the 1247 beam profile images used in this work.

For possible expansions of this work, the number of patches, and the size of each patch can be scaled-up to accommodate larger ViT model with more trainable parameters. Having more patches, and more pixels in each patch can capture finer features of beam profiles in each category. However, this will result in exponential increase in model training time (for this work, the model training time ranges between 60 – 300s). We have provided the full Python code on the ViT model construction, training, evaluation, and customized figure-of-merit in ref. [29]. Meanwhile, the full Python code for layer-by-layer plots of the ViT model can be found in ref. [31] to provide deeper understanding on the fundamentals of ViT model. By providing these Python codes, we hope to expand the scope of this work to various beam profile categorizations, or other photonics regimes.

VI. CONCLUSION

In this study, a vision transformer (ViT) model is applied to perform a classification task on beam profiles coupled out from silicon photonics (SiPh) gratings to free space. Beam profile images are first simulated using finite-difference time-domain technique. The beam profiles are split into ‘converged’ and ‘diverged’ groups, where ‘converged’ group are beam profiles coupled out from mixed pitch SiPh gratings with radius of curvature (R) of 20 µm; while ‘diverged’ group are beam profiles coupled out from SiPh gratings with R = 15, 30, 40, 50, and 60 µm. Within each group, the beam profiles are further categorized into Region A, B, and C, according to their corresponding heights above the SiPh gratings. The beam profile images are then used to train ViT model with customized figure-of-merit, where minimum difference between training loss and validation loss is desired. By using the trained ViT model to classify 832 beam profile images, classification accuracy of 0.989 is achieved. Upon repeating the ViT training process for 100 times, classification accuracy ranges from 0.83 to 0.99, where 82/100 runs have testing accuracy values of >0.95 is obtained. By providing the full Python codes of the ViT model in ref. [29,31], readers can further expand the scope ViT model classification on a wider scope of optical-related applications.

ACKNOWLEDGMENT

This work was supported by the Ministry of Education of Singapore AcRF Tier 2 (T2EP50121-0002 (MOE-000180-01)) and AcRF Tier 1 (RG135/23, RT3/23). The preparation of Python codes in ref. [29,31] is partly assisted by the generative AI tool, ChatGPT.

REFERENCES


