Vector boson fusion versus gluon-gluon fusion Higgs boson production with full-event deep learning: Toward a decay-agnostic tagger

Cheng-Wei Chiang,[1,2,*] David Shih,[3,†] and Shang-Fu Wei[1,‡]

[1]*Department of Physics, National Taiwan University, Taipei, Taiwan 10617, Republic of China*
[2]*Physics Division, National Center for Theoretical Sciences, Taipei, Taiwan 10617, Republic of China*
[3]*NHETC, Department of Physics and Astronomy, Rutgers University, Piscataway, New Jersey 08854, USA*

(Received 7 November 2022; accepted 6 January 2023; published 25 January 2023)

We study the benefits of jet- and event-level deep learning methods in distinguishing vector boson fusion (VBF) from gluon-gluon fusion (GGF) Higgs production at the LHC. We show that a variety of classifiers (CNNs, attention-based networks) trained on the complete low-level inputs of the full event achieve significant performance gains over shallow machine learning methods (BDTs) trained on jet kinematics and jet shapes, and we elucidate the reasons for these performance gains. Finally, we take initial steps toward the possibility of a VBF vs GGF tagger that is agnostic to the Higgs decay mode, by demonstrating that the performance of our event-level CNN does not change when the Higgs decay products are removed. These results highlight the potentially powerful benefits of event-level deep learning at the LHC.

DOI: 10.1103/PhysRevD.107.016014

I. INTRODUCTION

The discovery [1,2] of the Higgs boson in 2012 was a monumental occasion, providing a capstone to decades of experimental and theoretical works in particle physics, and confirming the final missing piece of the Standard Model (SM).

Since the original discovery, much effort [3,4] has been devoted to measuring ever more precisely the couplings of the Higgs boson to other SM particles. Since the Higgs has numerous production modes and decay modes, measurements in many different final states are necessary to disentangle all the various effects and pin down the Higgs couplings to all the SM fields [5–11]. A key component of this program is distinguishing the vector boson fusion (VBF) production mode from other production modes, most predominantly gluon-gluon fusion (GGF). VBF is essential for measuring the Higgs couplings to the SM W/Z gauge bosons, thereby testing the most essential property of the Higgs, namely its role in electroweak symmetry breaking (EWSB).

Previous works [12,13] have studied the question of VBF vs GGF classification with machine learning methods (see also [14] for earlier work on high-level features for VBF vs GGF). The main thing that distinguishes VBF from GGF events is that VBF events come with two forward quark-initiated jets from the hard process, while GGF jets are going to be from initial-state radiation and will tend to be gluon-initiated. In Ref. [12], boosted decision trees (BDTs) trained on high-level physics variables such as invariant mass and rapidity difference of the leading jets, sum of transverse momenta of the Higgs decay products, and various jet shape variables were brought to bear on the question of VBF vs GGF classification, in the context of $H \to \gamma\gamma$ and $H \to WW^*$ final states specifically. Meanwhile, Ref. [13] studied the multiclass classification of multiple Higgs production modes (including VBF and GGF) in the boosted $H \to bb$ regime, considering BDTs trained on high-level features, as well as a specialized two-stream convolutional neural network (CNN), which was previously developed for boosted $H \to bb$ tagging [15], and was trained on event images made out of low-level inputs (the pixelated p_T's of all the particles in the event).

Experimental studies [5–10,16–19] have also used BDTs or dense neural networks (DNNs) on a variety of Higgs decay modes to discriminate VBF from GGF events, while other techniques such as recurrent neural networks (RNNs) were also found useful in practice [6]. The BDTs, DNNs, RNNs used by the experimental groups take the high-level features as input.

In this work we will revisit the question of VBF vs GGF event-level classification, exploring the benefits that machine learning methods (both shallow and deep) can bring to this problem. Our starting point will be a BDT trained on high-level features (HLFs) defined from the

*chengwei@phys.ntu.edu.tw
†shih@physics.rutgers.edu
‡b06202061@ntu.edu.tw

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI. Funded by SCOAP³.

leading two jets and the Higgs decay products; this baseline method is designed to characterize the previous state of the art from [12] and from the actual ATLAS and CMS analyses. To go beyond, we consider the following methods:

(i) Training a jet-level CNN to distinguish the leading two jets from VBF from their GGF counterparts, and adding the jet-CNN scores to the inputs of the HLF BDT.
(ii) Training an event-level CNN to distinguish full VBF events from full GGF events; we make full-event images out of the energy deposits of all the reconstructed particles in the event.
(iii) Training an event-level network based on the self-attention mechanism [20,21] as an interesting alternative to the event-level CNN. In such a self-attention model, we convert the input event into a sequence which directly records the detector-level information.

We will see that while augmenting the HLFs with the jet-CNN scores offers some gain in classification performance, a much bigger boost comes from the event-level classifiers trained on low-level inputs. We investigate the reasons for the performance gains of the event-level CNN and find it is due in part to additional hadronic activity beyond the leading two jets. Interestingly, this includes both additional jet activity, as well as unclustered hadronic activity in the event (i.e., hadronic activity that leads to softer jets below the jet p_T threshold). The pattern of soft radiation is different in VBF vs GGF events, again presumably due to differing quark vs gluon content in the initial states.

In this paper we will also highlight an added benefit of event-level classifiers trained on low-level inputs: they can be *Higgs decay mode agnostic.* Since the Higgs is a color singlet, the Higgs decay should be fairly well factorized from the VBF or GGF initial state jets, especially when it decays to electroweak states. Besides, the p_T-balance of the full event ensures that the kinematics of the Higgs can be well reconstructed from all the other final state objects. Using the diphoton mode as an explicit example, we will show that as long as our models take the whole event into account, adding information from the Higgs decay does not improve the performance of the classifier. This raises the possibility that a single VBF vs GGF classifier could be trained and deployed in a variety of Higgs analyses with different final states, with no loss in performance.

Much work in the literature has focused on boosted jet classification [15,22–88], but relatively less work has been done on event-level classification [13,15,40,56,66,89–93]. Our work illustrates the potential benefits of full event-level classification.

For simplicity, we will not consider SM backgrounds in this work; of course, these backgrounds are highly dependent on the Higgs final state. In certain decay modes such as $H \to ZZ^* \to 4\ell$ [6,11,94,95], the non-Higgs background is highly suppressed, so our work could directly apply there. For other decay modes where the SM background is less suppressed (e.g., $H \to \gamma\gamma$), we imagine the "universal" VBF vs GGF classifier could be combined with a Higgs decay classifier for full event classification including non-Higgs background rejection if necessary [5,7,19].

An outline of our paper is as follows. In Sec. II, we describe the simulation of our sample as well as the VBF preselection criteria and the numbers of training, validation, and testing sets for the classifier. In Sec. III, we describe the classifiers used in this study. We show the results in Sec. IV, which is comprised of a comparison of tagger performances, a discussion about what the event-level CNN has learned, and possible improvements of the BDT from adding information beyond the leading two jets. In Sec. V, we examine the p_T-balance of the full event and explore the possibility of the Higgs-decay-mode-agnostic classifier. Finally, we conclude in Sec. VI. Appendix A lists the structures of all the classifier models considered in this study. Appendix B examines an extension of CNN for our classification problem motivated by [13] and finds no further improvement. Appendix C studies and compares the impacts of using our parton showering setting in PYTHIA, the local dipole recoil scheme, with the default setting. The performance of each of our classifiers is found to improve in the former scheme.

II. SAMPLE PREPARATION

We use MadGraph5_aMC@NLO 2.7.3 [96] (MG5) with parton distribution functions (PDFs) of CT10 [97] to generate Higgs plus up to three jets events starting from pp collisions at $\sqrt{s} = 14$ TeV. The additional jets are matched using the MLM matching scheme [98] with parameters $xqcut = 30$ GeV and $qcut = 45$ GeV. For VBF we just use tree-level MG5, while for GGF we use a model generated by FeynRules 2.3.33 [99] following the effective vertex method.

The samples are then showered and hadronized by PYTHIA 8.245 [100,101], and finally passed through the DELPHES 3.4.2 [102] fast detector simulation. Note that for the showering of VBF we toggle on the local dipole recoil option [103] in PYTHIA, which models the emission of additional jets in a way compatible with QCD [104–106]. This is achieved by specifying `SpaceShower:pTmaxMatch=2`, `SpaceShower:dipoleRecoil=on`, `TimeShower:QEDshowerByL=off`, and `PDF:lepton=off`. Also note that the detector configuration in DELPHES is based upon the default ATLAS card, while the inputs of the jet cluster module are EFlow objects instead of the default Tower objects. The jet clustering is done by FastJet 3.3.2 [107] using the anti-k_T [108] algorithm with $R = 0.4$. Jets are required to have $p_T > 25$ GeV in the simulation stage.

TABLE I. Numbers of training, validation, and testing sets for event-level classifiers.

	Training	Validation	Testing
VBF events	113 k	28 k	35 k
GGF events	89 k	22 k	28 k

TABLE II. Hyperparameters of the BDT.

Max depth	3
Learning rate	0.1
Objective	Binary logistic
Early stop	10 epochs
Evaluation metric	Binary logistic

In our sample preparation, we let the Higgs decay to two photons and use their invariant mass cut to select the required Higgs production samples. Although we generate samples in this particular Higgs decay mode, as discussed in the Introduction, we will demonstrate later that the full-event classifiers trained on low-level inputs are actually agnostic to the Higgs decay products, in that their performance does not suffer when those decay products are removed.

The samples used in the following analysis of this study are all extracted from the events passing the VBF preselection criteria, as inspired by experimental studies, that $N_\gamma \geq 2$, $120 \leq M_{\gamma\gamma} \leq 130$ GeV, $N_j \geq 2$, and $\Delta\eta_{jj} \geq 2$, with the jets in these criteria required to have $p_T > 30$ GeV. We have generated 500 k events each for the VBF and GGF samples and, after the VBF preselection, are left with 177 k events for VBF and 139 k for GGF.

Throughout this paper, we consider VBF as the signal and GGF as the background. For all of the event-level classifiers, the generated samples are split into training, validation, and testing sets as indicated in Table I. Since for any event we take the leading two jets as the samples of the jet-level classifier (i.e., jet-CNN), the numbers of the samples in different sets of the jet-level classifier are twice as those in Table I.

III. CLASSIFIER MODELS

A. BDT

We start by considering BDT models that are implemented in XGBoost 1.5.0 [109]. (The hyperparameters and the details of the BDT models are summarized in Table II.) We train three different BDTs based on the features summarized in Table III and Fig. 1. The first, "baseline," is based on six high level features from the study of VBF vs GGF classification in Ref. [12], which is inspired by ATLAS's setup [110]. This baseline BDT characterizes the discrimination power from the kinematics of the photons and the jets in the event.[1]

Based on the experimental setup, Ref. [12] further considers the jet shape variables [111] as additional input features, such as the girth summed over the two leading jets and the central/sided integrated jet shape. Including these jet shape variables leads to our second BDT, which we call "baseline + shape."

Finally, we consider the benefits of replacing the human-engineered jet shape variables of [12,111] with the output of a jet-level CNN classifier trained on VBF vs GGF jets. We call this the "baseline + jet-CNN" BDT. For more details on the jet-level CNN, see Sec. III B.

B. Jet-CNN

In this subsection, we introduce the VBF vs GGF jet-level CNN used in the "baseline + jet-CNN scores" BDT described in the previous subsection. The jet-level CNN is trained on jet images formed out of the leading two jets from the VBF and GGF events.[2] Our image preprocessing, which basically follows the procedure outlined in Ref. [45], contains image centralization, rotation, and flipping, followed by pixelation from the detector responses to the jet image. Finally, we pixelate the detector responses into images (10×10 pixels) for each of the following four channels: Tower E_T, Tower hits, Track E_T and Track hits. (Following the DELPHES particle flow algorithm: "Tower" means EFlowNeutralHadron or EFlowPhoton, and "Track" means EFlowTrack.)

Our jet-CNN model starts from a batch normalization layer [112], followed by several convolution layers and average pooling layers, which capture the features of the images. The sizes of the filters in convolution layers and pools in pooling layers are all 2×2. Due to the relatively small size of the images (10×10 pixels), the neural network (NN) does not need to be very deep. Since the image size shrinks as it passes through a pooling layer, the number of pooling layers is restricted. After the convolution and pooling layers, the images are then flattened and fully connected to three dense layers with 128 neurons respectively. The last dense layer with 2 neurons, activated by the SoftMax function, represents the final output score as probabilities. All the other dense layers and convolution layers use the ReLU activation function [113]. The model structure is plotted in Fig. 10.

[1]We have checked that a simple DNN trained on these high-level features does not outperform the BDTs, so we will focus on BDTs as our baseline.

[2]Another possible labeling scheme is to identify whether the jet is quark or gluon initiated, since VBF (GGF) events tend to contain more quark (gluon) jets. However, our trials show that both labeling schemes are equally useful when they are considered as features in the subsequent event-level BDT. We will focus exclusively on the process-labeling in the following study.

TABLE III. Summary of the features used in BDT. j_1 and j_2 mean respectively the p_T-leading and -subleading jets, while γ_1 and γ_2 mean respectively the p_T-leading and -subleading photons. In the jet shape variables, i represents the constituent of the jet and r is the distance between the constituent and the jet axis.

Baseline	1. m_{jj}, the invariant mass of j_1 and j_2
	2. $\Delta\eta_{jj}$, the absolute difference of the pseudorapidities of j_1 and j_2
	3. ϕ^*, defined by the ϕ-difference between the leading diphoton and dijet
	4. $p_{Tt}^{\gamma\gamma}$, defined by $\lvert(\mathbf{p}_T^{\gamma_1}+\mathbf{p}_T^{\gamma_2})\times\hat{t}\rvert$, where $\hat{t}=(\mathbf{p}_T^{\gamma_1}-\mathbf{p}_T^{\gamma_2})/\lvert\mathbf{p}_T^{\gamma_1}-\mathbf{p}_T^{\gamma_2}\rvert$
	5. $\Delta R_{\gamma j}^{\min}$, defined by the minimum $\eta-\phi$ separation between γ_1/γ_2 and j_1/j_2
	6. η^*, defined by $\lvert\eta_{\gamma_1\gamma_2}-(\eta_{j_1}+\eta_{j_2})/2\rvert$, where $\eta_{\gamma_1\gamma_2}$ is the pseudorapidity of the leading diphoton
Shape	7. the girth summed over the two leading jets $\sum_{j=1}^{2} g_j=\sum_{j=1}^{2}\sum_{i\in J^j} p_{T,i}^j r_i^j/p_T^j$
	8. the central integrated jet shape $\Psi_c=\sum_{j=1}^{2}\sum_{i\in J^j} p_{T,i}^j(0<r_i^j<0.1)/(2p_T^j)$
	9. the sided integrated jet shape $\Psi_s=\sum_{j=1}^{2}\sum_{i\in J^j} p_{T,i}^j(0.1<r_i^j<0.2)/(2p_T^j)$
Jet-CNN	10. the jet scores of the two leading jets, output by the jet-CNN, soon to be introduced in Sec. III B

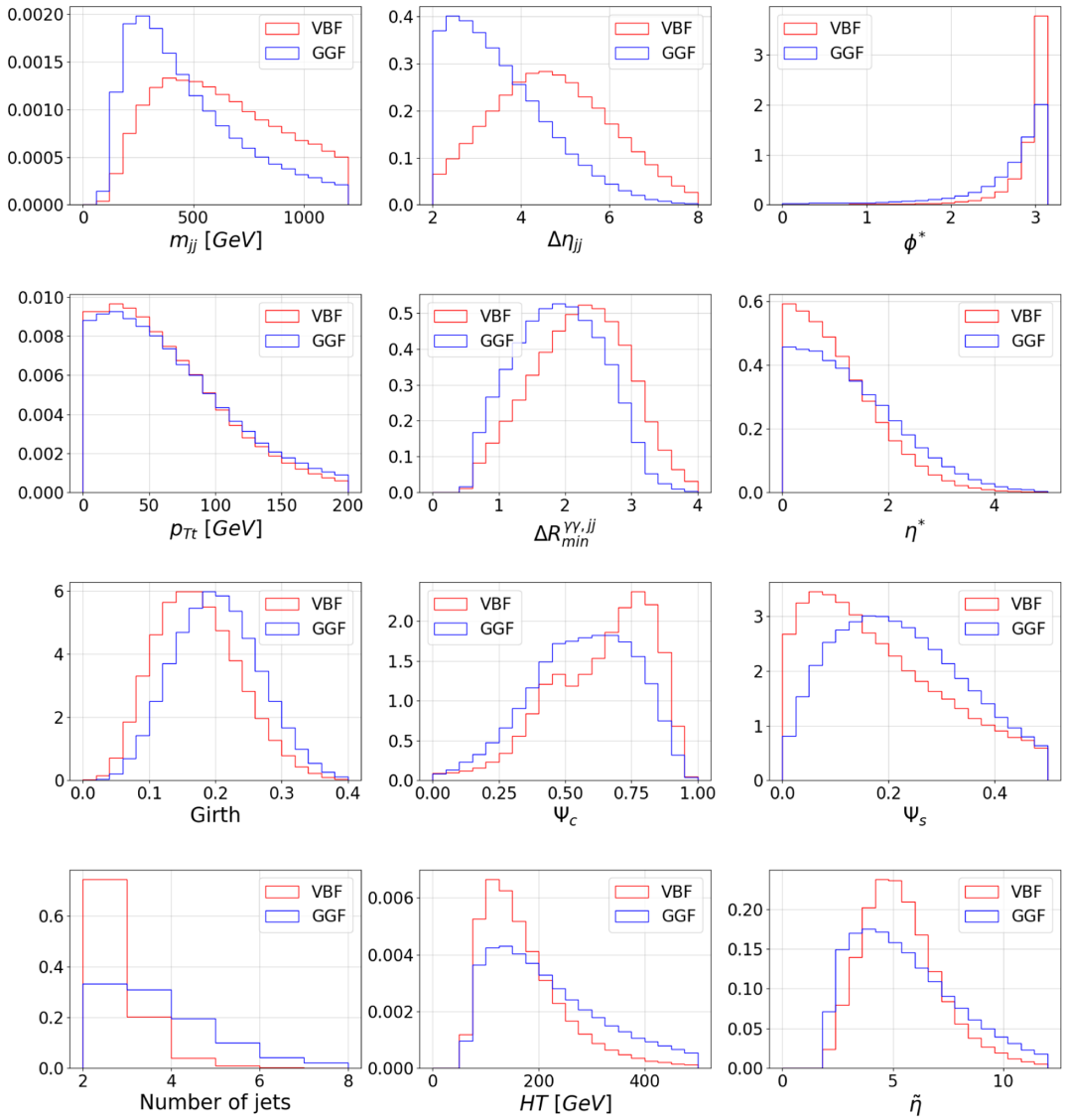

FIG. 1. Distributions of BDT input variables. All histograms are normalized so that the area under each curve is one.

TABLE IV. Hyperparameters for the jet-CNN tagger.

Optimizer	Adam
Loss function	Categorical cross entropy
Early stopping	20 epochs
Batch size	1024

The CNNs in this study are all implemented in TensorFlow 2.0.0 [114] with KERAS [115] as its high-level API. We use Adam [116] as our optimizer during the training stage with the categorical cross entropy loss function in all of our NN models. By monitoring the loss of the validation set, early stopping is implemented to prevent overfitting in all of the NN and BDT models. The hyperparameters of the model are summarized in Table IV.

Our jet-CNN takes a jet image as its input and outputs a score ranging from 0 (GGF-jet) to 1 (VBF-jet). The scores of leading and subleading jets can thus be useful features for subsequent event-by-event classification. The distributions of the jet-CNN scores and the receiver operating characteristic (ROC) curve for the jet-CNN are shown in Fig. 2, where the horizontal (vertical) axis label TPR (FPR) refers to the true positive rate (false positive rate). The area under the ROC curve (AUC) of the jet-CNN is 0.734, which is less than an efficient classifier. However, we will show that the jet-CNN scores are indeed useful information in the subsequent event-level classification. Instead of training and testing separate taggers for the leading and subleading jets respectively, we utilize one tagger which is trained on mixed samples including the leading and subleading jets. Our trial explicitly shows that such training does not sacrifice the performance of the classifier.

C. Event-CNN

A potentially more powerful way to perform event-level classification is to leverage the capabilities of deep learning to predict the VBF vs GGF label directly from the lowest-level features of each event (in our case, the 4-vectors of all the particles in the event). In this paper we consider two approaches to this, a CNN trained on whole-event images, to be described in this subsection, and a self-attention model trained on sequences of the particle 4-vectors, to be described in the next subsection.

Our whole-event images are preprocessed similarly to the jet images of the previous subsection. However, unlike jets, the whole event is not a localized object, nor is there an approximate boost or rotation invariance. So the preprocessing consists of just the following steps: we first move the ϕ coordinate of the weighted center to the origin, and flip the image vertically or horizontally to make the upper-right quadrant more energetic than all the other quadrants. Finally, the detector responses are pixelated into images with 40×40 pixels for each of the six channels, which includes the same four channels used in the jet-CNN and two additional ones recording the hits and E_T of the isolated photons.

An example of single event images is shown in Fig. 3. The left plot shows the isolated photon E_T and Tower E_T combined with Track p_T of an event before the preprocessing, while the right plot is after the preprocessing.

We employ a toy ResNet model [117] in our event-CNN. Two convolution layers form a residual block in ResNet. There are shortcuts connecting the residual blocks, enabling us to deepen our model without suffering from the degradation problem. The sizes of filters in the convolution layers and pools in the pooling layers are all 3×3. The detailed model structure of the event-CNN is shown in Fig. 11. The hyperparameters are the same as those in Table IV.

In order to extract information from both the local jet-level and global event-level features, Ref. [13] adopts a two-stream CNN architecture, where one stream processes an image of the highest p_T non-Higgs jet in the event, and the other stream processes the full-event image. Motivated by this, we further study the performance of an extension of our full-event CNN in Appendix B, using a similar

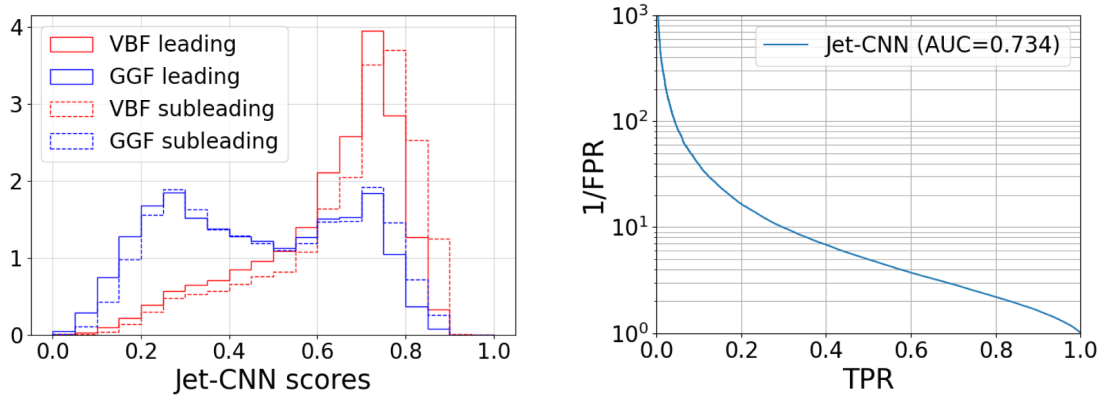

FIG. 2. Distributions of the jet-CNN scores (left) and the ROC curve of the jet-CNN (right). All histograms on the left are normalized so that each area under the curve is one.

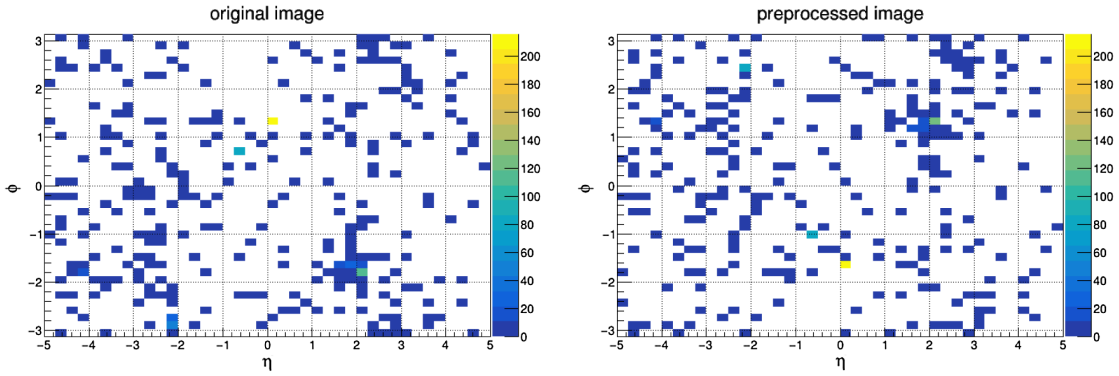

FIG. 3. The isolated photon E_T and Tower E_T combined with Track p_T of an event without preprocessing (left) and after preprocessing (right). The color of each pixel indicates the energy in units of GeV.

structure containing three streams of CNN, dealing with event images and leading two jet images respectively. However, we find no improvement from our original single-stream event-CNN. This does not contradict the works of Ref. [13] since they did not compare the performance of their two-stream CNN against a single-stream CNN consisting of just the full-event classifier.

D. Self-attention

For comparison, we also consider another whole-event low-level-feature classifier based on the technique of self-attention [20], which is used in the famous transformer model [21] dealing with sequence-to-sequence tasks. The original motivation of this model is to use the multihead attention layers to capture the correlation among elements in the input sequence. Inspired by this idea, instead of representing an event as an image, we view the event as a sequence, where the elements of the sequence are the p_T, η, ϕ, and electric charge of the 100 highest-p_T reconstructed particles in the event (with zero padding for events with fewer than 100 particles). In principle, the self-attention network could be advantageous over event-level images, because it is not subject to the information loss induced by pixelation. Also, a nice property of the self-attention mechanism is that it preserves permutation invariance of the inputs (as does a CNN).

The implementation of the self-attention model is based on TensorFlow 2.5.0 and KERAS. The model structure of the self-attention model is shown in Fig. 12. There are three five-head attention layers at the beginning, followed by a global average pooling (GAP) layer, which converts the sequence of detector responses into a single vector by taking the element-wise average. Dense layers are not implemented before the GAP layer to keep permutation invariance of the input sequence. Then the model is passed into seven dense layers. The hyperparameters are listed in Table V.

TABLE V. Hyperparameters of the self-attention model.

Optimizer	Adam
Loss function	Categorical crossentropy
Early stopping	50 epochs
Batch size	1024

IV. RESULTS

A. Comparison of methods

The performance of the event-level classifiers defined in the previous section is shown in Fig. 4. As an explicit example, Table VI lists both FPRs and AUCs at the working point where the TPR is fixed at 0.3. From Fig. 4 and Table VI, we can easily compare the different event-level classifiers.

First of all, "BDT: baseline" has the lowest AUC since it only considers the high-level kinematic features in an event. Indeed, including additional information on the jet shape variables can improve a little, but not as much as using the jet-CNN score as an input. Notably, our jet-CNN scores serve as a better feature than the jet shape variables,

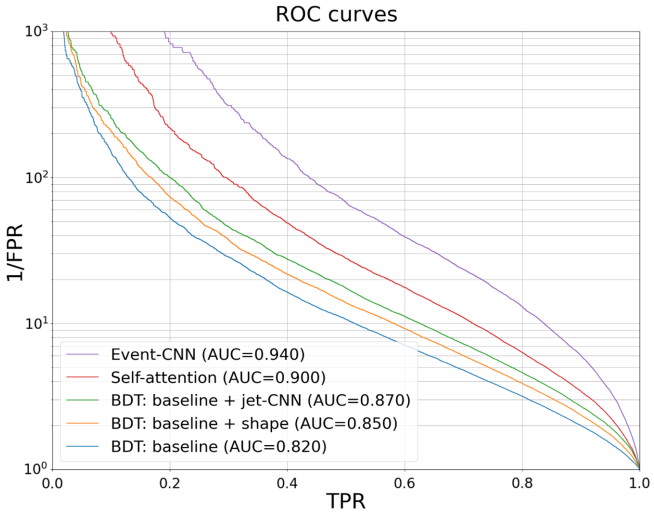

FIG. 4. ROC curves of several event-level classifiers.

TABLE VI. Performance comparison at TPR = 0.3.

	FPR	AUC
BDT: baseline	0.035	0.820
BDT: baseline + shape	0.027	0.850
BDT: baseline + jet-CNN	0.022	0.870
Self-attention	0.010	0.900
Event-CNN	0.003	0.940

with the former reducing the FPR from the baseline by a factor of 1.6 while the latter only by a factor of 1.3. Therefore, despite a low AUC of the jet-CNN as shown in Fig. 2, its score still provides valuable information. We have also checked that combining jet shape variables and jet-CNN scores in the input features together did not provide extra improvement in the AUC, indicating that the jet-CNN has learned all the information contained in the human-engineered jet shape variables.

Second, we see that our self-attention model and event-CNN both perform better than the BDTs. This is understandable because the BDTs only take into account high-level variables or features of the two leading jets and photons only, while the self-attention and event-CNN taggers take in the entire event and catch more features therein.

Finally, the event-CNN is the most powerful classifier among all considered taggers. Its inverse FPR is roughly a factor of 2.5 better than the self-attention model for most of the TPR. Its AUC reaches 0.940 and the FPR is reduced by a factor of 11.7 from the baseline at the assumed working point in Table VI.

B. Saliency maps of event-CNN

To further investigate what the event-CNN has learned, we examine its saliency maps [118]. Let the input pixel x be identified as $x_{c,h,w}$, where c is the channel index, h is the height index, and w is the width index. The saliency is defined by the gradient of the i-th class score P^i with respect to the input pixel $x_{c,h,w}$,

$$w^i_{c,h,w} \equiv \frac{\partial P^i}{\partial x_{c,h,w}}, \tag{1}$$

where the gradient is calculated by back-propagation. In our case, we only deal with binary classifiers, so it suffices to only consider the VBF class score P. Putting $w_{c,h,w}$ together according to the indices, one can obtain the saliency maps. However, what we are actually interested in is the saliency according to the standardized pixels $y_{c,h,w}$ which have no scale difference across channels,

$$x_{c,h,w} \to y_{c,h,w} = \frac{x_{c,h,w} - \mu_c}{\sigma_c}, \tag{2}$$

where ${\sigma_c}^2$ and μ_c are the variance and mean of the channel c in the whole sample, including the training, validation, and testing sets. Hence, we will consider the following gradient,

$$\tilde{w}_{c,h,w} \equiv \frac{\partial P}{\partial y_{c,h,w}} = w_{c,h,w} \times \sigma_c. \tag{3}$$

Finally, we arrange $\tilde{w}$ according the c, h, w indices and then plot its absolute value $|\tilde{w}_{c,h,w}|$ to get the saliency maps as the lower row in Figs. 5 and 6.

We utilize the visualization toolkit TF-KERAS-VIS 0.8.0 [119] to implement the saliency maps of our event-CNN tagger. In the following, we pick as examples a VBF event (Fig. 5) with a high CNN score (i.e., more VBF-like) and a GGF event (Fig. 6) with a low CNN score (i.e., more

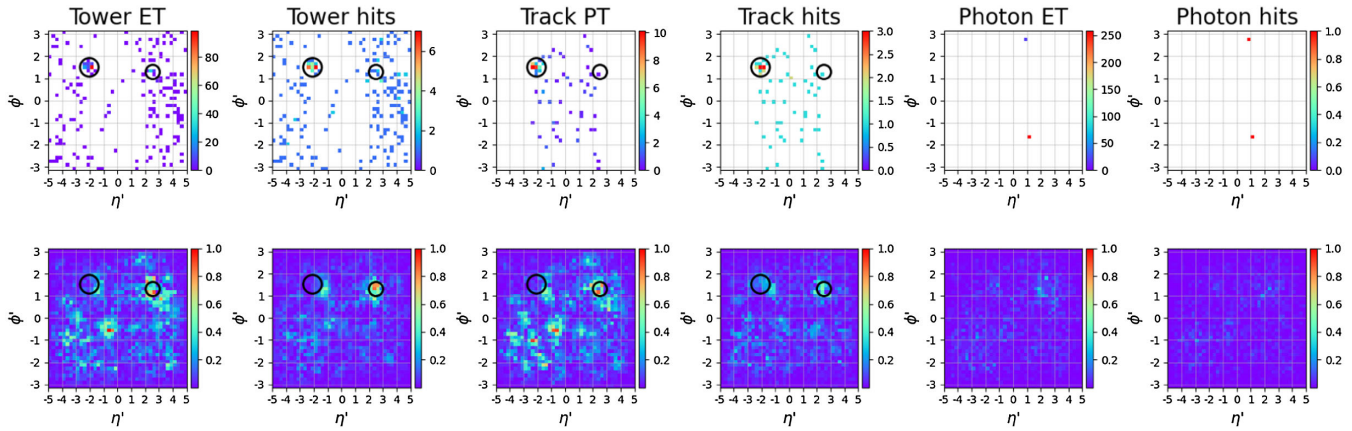

FIG. 5. A VBF event with a high event-CNN score. The upper six plots show the raw inputs of the model, while the lower counterparts are the saliency maps calculated by the corresponding normalized channels. The black circles show the locations of the clustered jets, with the circle size indicating the ordering in p_T. The color maps of the upper row indicate the actual input. The unit is GeV for Tower E_T, Track p_T, and isolated photon E_T, and counts for Tower hits, Track hits, and isolated photon hits. The color maps of the lower row indicate the relative saliency.

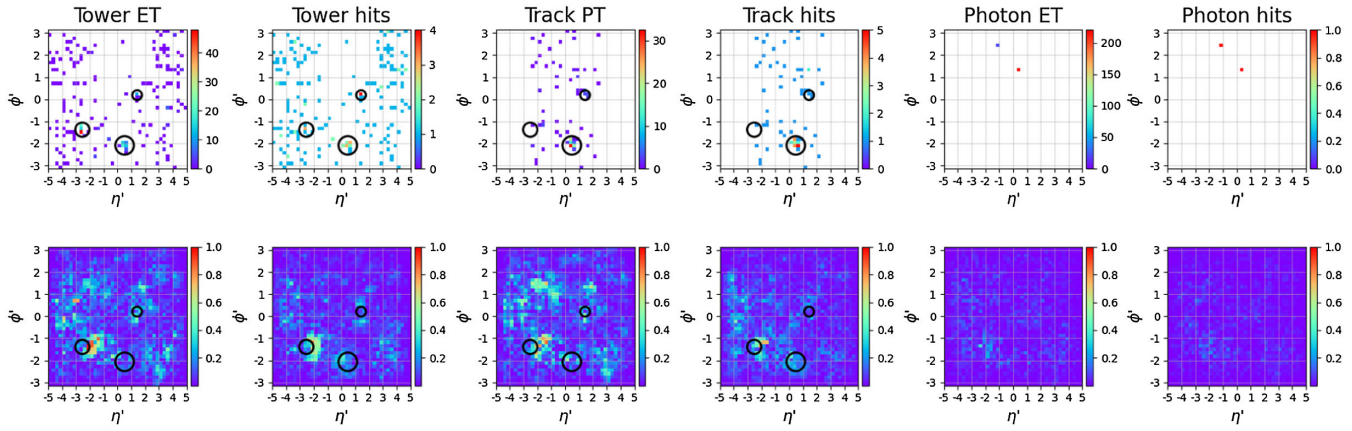

FIG. 6. Same as Fig. 5, but for a GGF event with a low event-CNN score.

GGF-like). In the plots, the clustered jets are marked by black circles, with their sizes indicating the jet's ordering in p_T. The color maps of the upper row indicate the actual value of the input, with the unit being GeV for Tower E_T, Track p_T, and isolated photon E_T and counts for Tower hits, Track hits, and isolated photon hits. In contrast, the color maps of the lower row indicate the relative saliency, i.e., the most salient pixel is scaled to one in plotting,

$$|\tilde{w}_{c,h,w}| \to \frac{|\tilde{w}_{c,h,w}|}{\max_{c,h,w}\{|\tilde{w}_{c,h,w}|\}}. \tag{4}$$

From the saliency maps, we observe that the CNN model generally focuses on the locations with more hadronic activities, as anticipated, because the jets contain crucial information for the classification of VBF and GGF events. In addition, the CNN is also seen to make use of lower p_T jets and hadronic activity that falls below the jet p_T threshold (set to 30 GeV in this work). This explains why the event-CNN performs better than the BDT. In our setup of the BDTs, we do not feed the information of the third jet into the model. Moreover, the input of the BDTs relies on our knowledge about what kind of high-level features is beneficial and hence cannot make use of unclustered energy in an event. Finally, we can observe that the event-CNN is much more focused on where jets are than the locations of photons, which indicates that the photon information is not crucial in the classification. This sheds light on the possibility of the Higgs-decay-mode-agnostic classifier which solely relies on the jet information. Details will be described in Sec. V.

C. Improvements of BDTs

In this subsection, we investigate more about how the BDTs, which rely on high-level kinematic variables as the features for training, can be further improved. Based on the study of the saliency maps in the previous subsection, we are motivated to consider information about additional hadronic activity in the event beyond the leading two jets. So we will study the benefits of including the 4-vector momentum of the third hardest jet, as well as inclusive kinematic variables that take all jets into account.

(i) 4-vector momentum of the third jet in p_T ordering, which is denoted as "j3vec;"
(ii) $HT = \sum_{j\in\text{jets}} p_T^j$, which characterizes the p_T distribution of the jets;
(iii) $\tilde{\eta} = \sum_{j\in\text{jets}} |\eta^j|$, which characterizes the positional distribution of the jets; and
(iv) number of jets.

We will call the set of features including HT, $\tilde{\eta}$, and the number of jets as a "jet-profile." The normalized distributions of HT, $\tilde{\eta}$, and the number of jets are already shown in the last row of Fig. 1.

We are interested in how the additional information improves the best BDT we have so far, so we will add these extra variables to "BDT: baseline + jet-CNN." The ROC curves of the BDTs trained with further inputs of these additional variables are plotted in Fig. 7. From the AUCs,

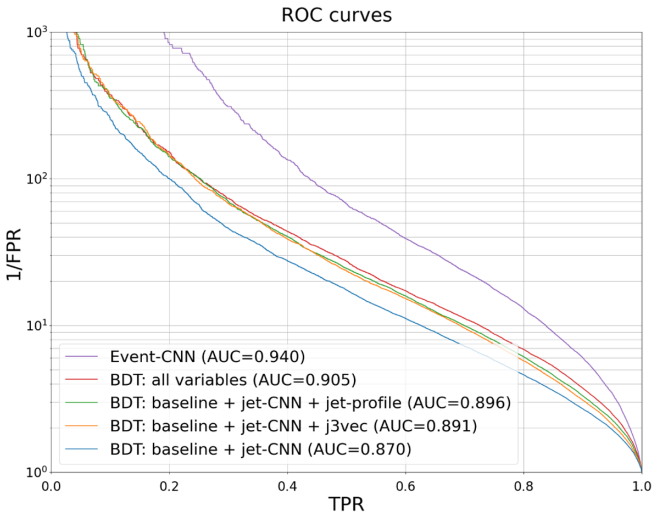

FIG. 7. ROC curves of BDT trained on additional high-level features.

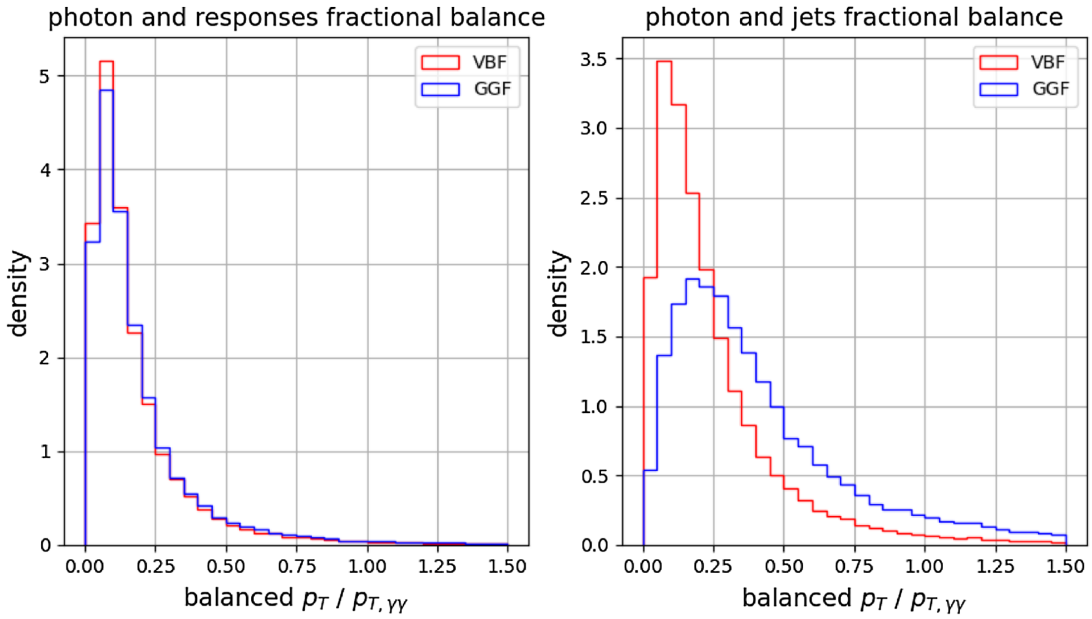

FIG. 8. The fractional p_T-balance of the leading diphoton and other objects, nonphoton responses on the left and up to leading three jets on the right. To calculate the balance, we first vector-sum the momenta of the diphoton and other objects, and then take its transverse momentum. Finally, the balanced p_T is divided by the p_T of the diphoton.

we can see both the additional 4-vector momentum of the third jet and the jet-profile can improve the performance of classification. Their improvements are comparable to each other, as seen from the ROC curves as well as the similar AUCs. We have also checked that adding the additional 4-vector momentum and the jet-profile together into the BDT does not further improve the AUC, which is a piece of direct evidence that these two sets of variables provide equivalent information to the BDTs. The reason is that the crucial information contained in both sets is the existence of the third jet. A characteristic of GGF events is that they tend to have more than two jets, which can be seen in the distribution of the number of jets in Fig. 1. By examining the actual trees in the BDTs trained by the 4-vector momentum of the third jet and the jet-profile, respectively, we indeed find that the existence of the third jet provides a clear separation between VBF and GGF events and therefore plays an important role in both cases.

Finally, in "BDT: all variables", we consider all the high-level features, including event-related characteristics (i.e., m_{jj}, $\Delta\eta_{jj}$, ϕ^*, $p_{Tt}^{\gamma\gamma}$, $\Delta R_{\gamma j}^{\min}$, η^*, HT, $\tilde{\eta}$, and the number of jets), and jet-related information of the three leading jets (i.e., 4-vector momenta, jet-CNN scores, and the girth, central/sided integrated jet shape of each jet without taking summation or average). This BDT achieves the best AUC, 0.905, among all the other BDTs and improve the baseline significantly. However, despite this sizable improvement from the baseline, the event-CNN still outperforms "BDT: all variables" with an even larger AUC.

V. A HIGGS-DECAY-AGNOSTIC VBF VS GGF CLASSIFIER

In the Introduction, we noted that event-level classifiers trained on low-level inputs could potentially be agnostic to the Higgs decay mode, due to the scalar nature of the Higgs and the p_T-balance of the whole event. Here we will explore this idea further, by seeing to what extent p_T balance allows the Higgs momentum to be predicted from the hadronic activity in the event, and then to what extent our classifiers suffer when the Higgs decay products are removed from the event.

Shown in the left plot of Fig. 8 are histograms of the p_T balance of the whole event, $|\sum_{i\in\text{reconstructed particles}} \vec{p}_{Ti}|$, normalized by the p_T of the Higgs. We see that the p_T is well-balanced among the low-level features, so the Higgs transverse momentum can be well-reconstructed from the nonphoton reconstructed particles. Meanwhile, the right plot of Fig. 8 depicts the p_T balance between the photons and the leading three jets, again normalized by the p_T of the Higgs. Here we see that while the leading three jets can capture the p_T information of the photons to some extent, it is not as informative as the responses and, therefore, the balance is not as complete.

Figure 9 shows the impact of removing the Higgs decay products (in our case, the two photons) from the event before training the VBF vs GGF classifiers. We see that removing photon information from the event-CNN hardly changes the AUC. On the other hand, removing photon information from the high-level features for the BDT

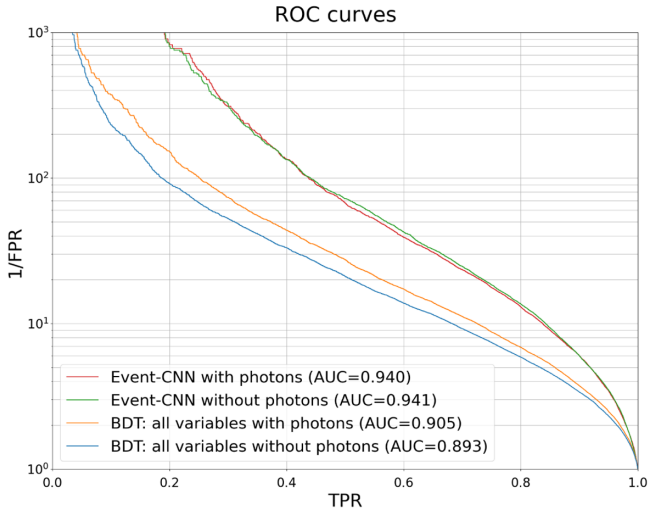

FIG. 9. ROC curves of event-level classifiers with and without the photon information.

reduces the AUC from 0.905 to 0.893.[3] The degradation in performance in the BDT is still not very large, but it is larger than that of the event-CNN. This is completely in line with the histograms shown in Fig. 8.

All in all, we confirm here that due to the p_T balance of the events, the performance of VBF vs GGF classification does not depend much on the Higgs decay products, especially for the whole-event CNN that is based on low-level inputs. This raises the intriguing possibility that one could train a single VBF vs GGF classifier that is agnostic to the Higgs decay mode, and could be applied equally optimally to a variety of Higgs decay channels in a uniform way. This could have benefits for data-driven calibration and reducing systematic uncertainties associated with VBF tagging.

VI. CONCLUSIONS

In this paper, we have studied machine learning approaches to event-level classification of Higgs production modes, focusing on the important problem of VBF vs GGF discrimination. Building on previous studies [12,13], we have shown that full-event deep learning classifiers that utilize low-level inputs (full-event images, sequences of particle 4-vectors) significantly outperform classifiers based on high-level physics features (kinematics, jet shapes). We have explored both CNNs trained on full-event images, and permutation-invariant self-attention networks trained on sequences of particle 4-vectors. Although the full-event CNN achieved the best performance in our studies—improving beyond the baseline shallow network by more than a factor of 4–15 in background rejection across a wide range of signal efficiencies—perhaps our work provides a useful starting point for further optimization of the attention-based approach.

We also studied why the event-level CNNs perform so much better than the shallow networks based on high-level features. Using saliency maps we saw how additional jets in the event beyond the first two contribute to the CNN classification, as well as unclustered hadronic activity (jets below the p_T threshold). By adding high-level features derived from these additional jets, we have confirmed that the performance of the shallow networks can indeed be improved and be brought somewhat closer to the event-level CNN.

Finally, in this work we have gone beyond previous approaches and explored the possibility of a VBF vs GGF classifier that is agnostic to the Higgs decay mode. A classifier trained on the low-level information of the full event should be able to reconstruct the Higgs transverse momentum from p_T balance and, since the Higgs is a scalar, its decay products (at least when it decays to electroweak states) should be well-factorized from the rest of the event. Therefore, a full-event, low-level classifier should be largely independent of the Higgs decay channel. We have taken the first steps toward verifying that in this work, by showing how the performance of the full-event CNN is virtually unchanged when trained on the events with and without the diphotons from the Higgs decay.

Some future directions of our work include: generalizing our work to more Higgs production modes (e.g., ZH, WH and ttH) using a multiclass classifier; further fleshing out our idea of a decay-agnostic classifier by studying other Higgs decay modes besides $H \to \gamma\gamma$; studying even more recent deep learning classifiers such as graph networks; adding particle interaction information into the event-level self-attention model (as was inspired by Ref. [86]); and incorporating symmetries such as Lorentz invariance into the architecture of the neural network to achieve even better performance (as was done recently for top-tagging in Ref. [84]).

ACKNOWLEDGMENTS

We are grateful to Tae Min Hong and Yi-Lun Chung for helpful discussions. We also thank Kai-Feng Chen and Iftah Galon for their participation in this project at the early stage. In addition, we thank Alexander Karlberg for pointing out the importance of the local dipole recoil scheme for VBF. C. C. and S. W. were supported in part by the Ministry of Science and Technology of Taiwan under Grants No. MOST-108-2112-M-002-005-MY3 and No. 111-2112-M-002-018-MY3. The work of D. S. was supported by DOE Grant No. DOE-SC0010008.

APPENDIX A: ARCHITECTURES OF VARIOUS NEURAL NETWORKS

In this appendix, we show explicitly the architectures of the neural networks constructed in this work. The shape

[3]In more detail, in Fig. 9, "all variables with photons" refers to the feature set used in "BDT: all variables" in Fig. 7, while "all variables without photons" refers to the same feature set but with the photon-related variables (i.e., ϕ^*, $p_{Tt}^{\gamma\gamma}$, $\Delta R_{\gamma j}^{\min}$, η^*) all excluded.

of the input and output tensors are shown as numbers inside the parentheses. The first argument "None" represents the batch size, which is a separate hyperparameter independent of the model architecture.

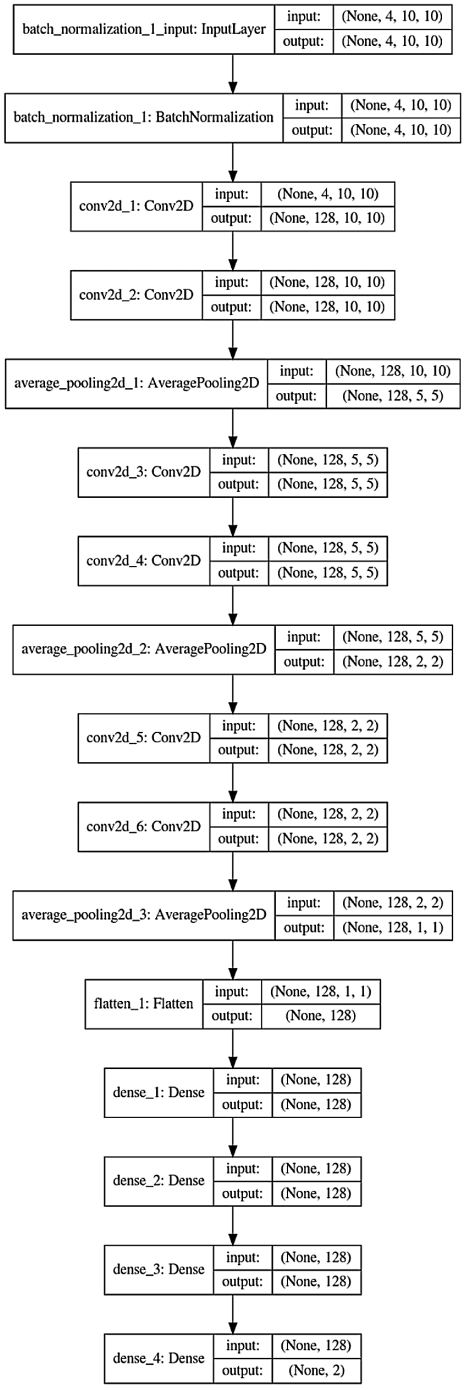

FIG. 10. The model structure of the jet-CNN.

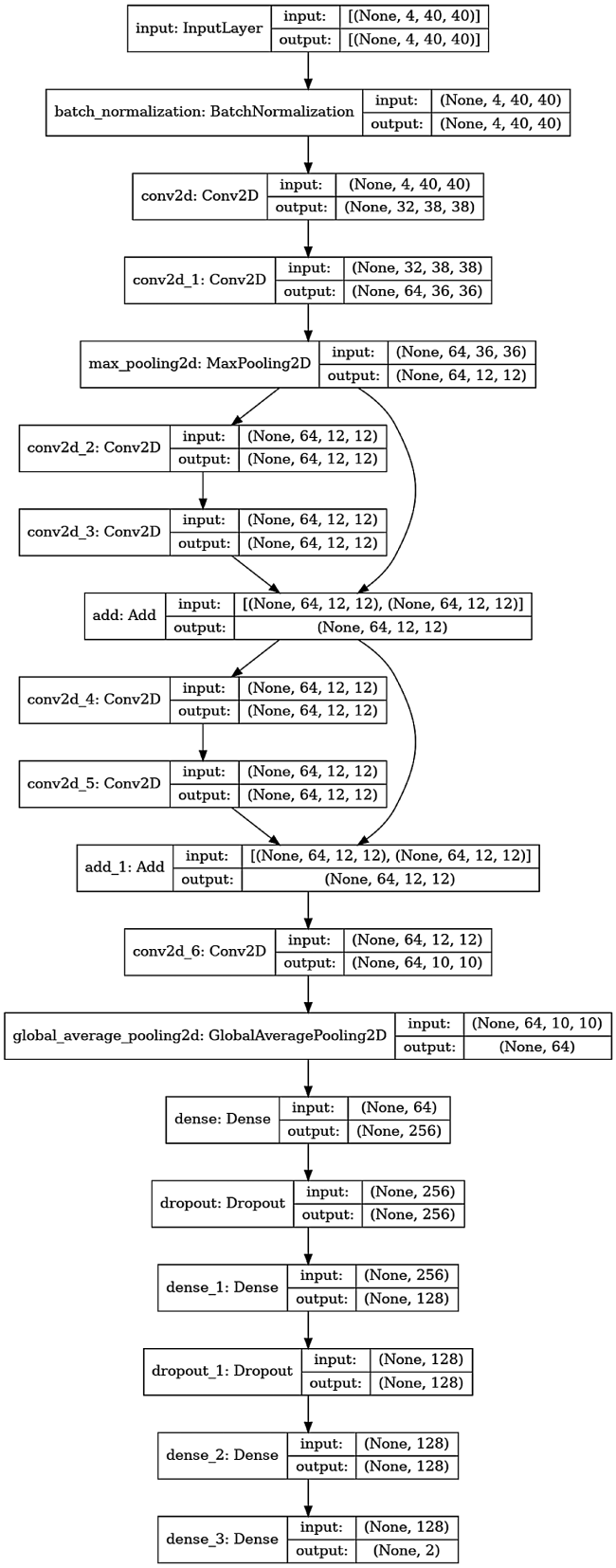

FIG. 11. The model structure of the event-CNN.

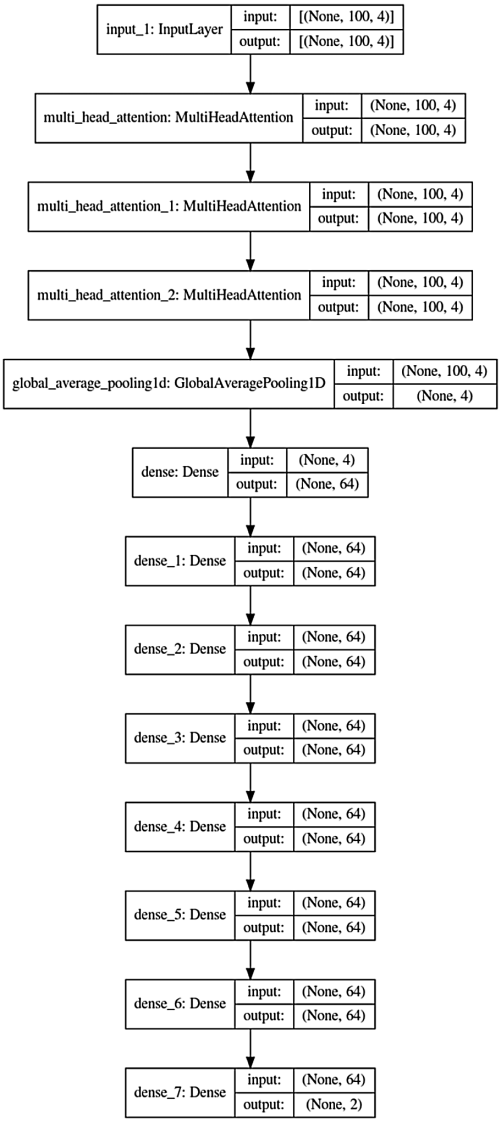

FIG. 12. The model structure of the self-attention model.

APPENDIX B: MULTISTREAM CNN

In this section, we examine an extension of CNN. In Ref. [13], an architecture called 2CNN extracts event-level and jet-level features simultaneously with two streams. One stream applies filters on event images, and the other deals with the leading non-Higgs jet images. Then the two streams are connected together and combined to form a single model. Inspired by this study, we want to investigate the possible improvement using this multi-stream architecture.

We adopt a three-stream CNN where one stream applies filters on the event images, and the other two process the images of the two leading jets respectively. Each stream is a toy ResNet model used in Sec. III C. We will call this architecture "event + 2jet-CNN." The model structure is shown in Fig. 13. All of the event images and jet images are pixelated into 40×40 pixels. The images are preprocessed as in Sec. III B and III C. The hyperparameters are the same as those in Table IV.

The ROC curves of our original event-CNN and this event + 2jet-CNN are plotted in Fig. 14. The curves almost overlap with each other and the AUCs are very similar, indicating that the event-CNN has already captured useful information for classification. The additional high-resolution jet images do not provide significant extra help to the performance.

APPENDIX C: DIPOLE SHOWER IN PYTHIA

In this section, we present the impact of different PYTHIA shower schemes applying to the VBF process. In contrast to the local dipole recoil scheme [103], which is the shower scheme we adopt in our study, we also show the kinematic distributions and training results based on samples showered by the default shower scheme in PYTHIA. As investigated in previous works [104–106], the default PYTHIA shower depicts the emission of additional jets in VBF poorly in the central region. This can be seen in Fig. 15, which shows an apparent discrepancy on the pseudorapidity distribution of the third jet.

Moreover, the classification performance is severely affected by the choice of the shower scheme. In Fig. 16, the performance of each classifier is improved when changing from the default shower to the local dipole shower. The reason of the lower AUC of the default shower is that it produces an incorrect pattern of the emission of the additional jets, which leads to the similarity between VBF and GGF. However, as Fig. 15 shows, the emission issue is solved by the local dipole recoil shower, and hence the performance becomes improved in the local dipole recoil scheme.

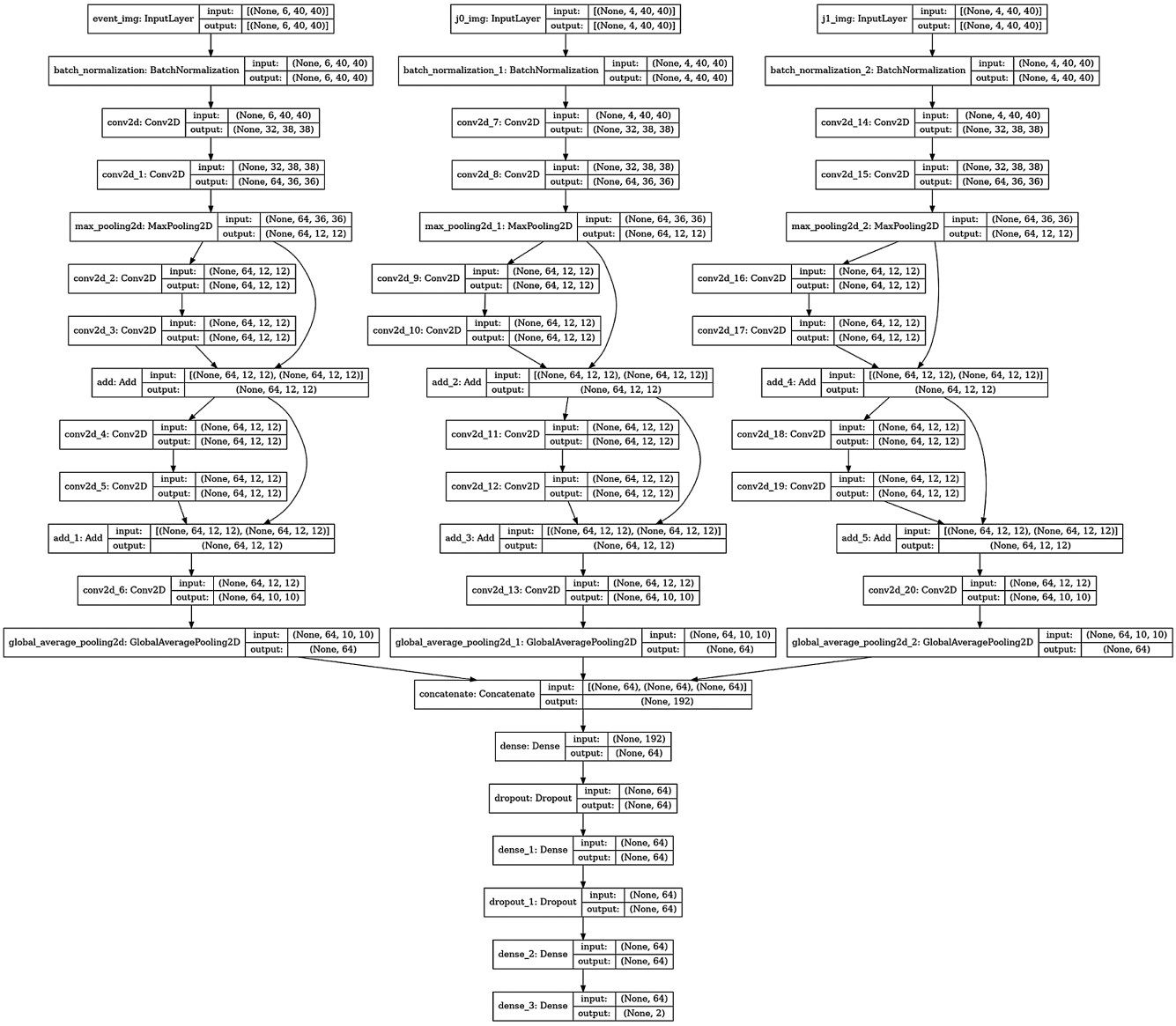

FIG. 13. The model structure of the multistream CNN model.

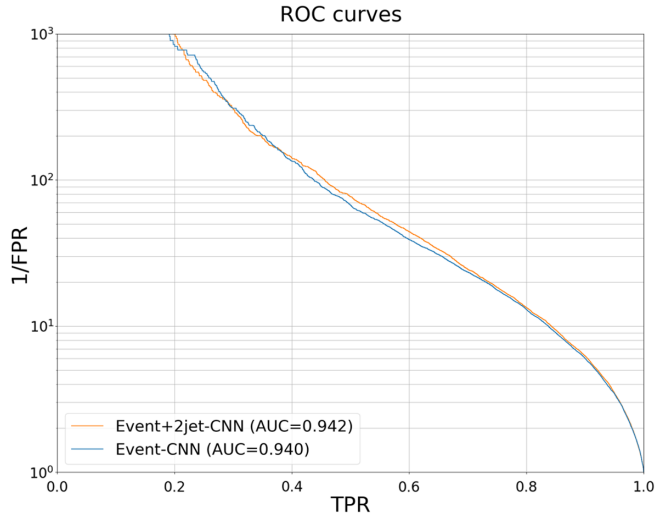

FIG. 14. ROC curves of the event-CNN and event + 2jet-CNN.

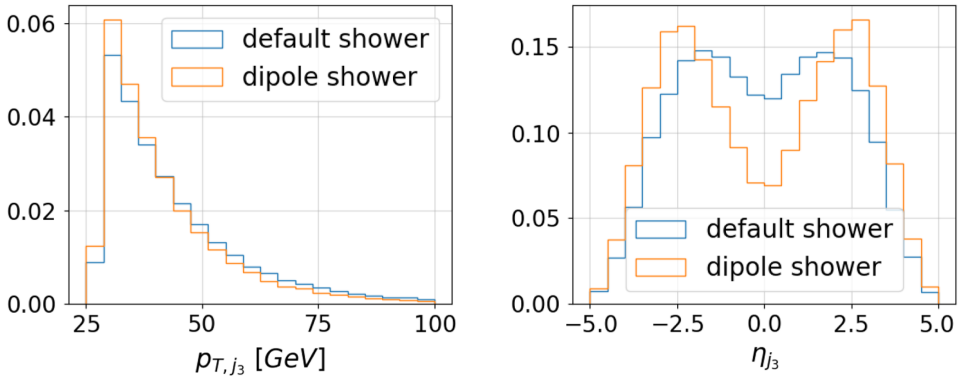

FIG. 15. Distributions of the transverse momentum and pseudorapidity of the third jet coming from the VBF events passing the VBF preselection. All histograms are normalized so that each area under the curve is one.

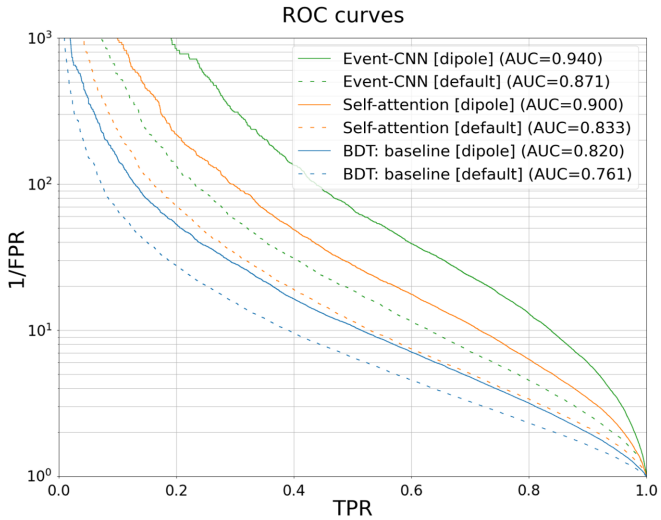

FIG. 16. ROC curves of several event-level classifiers in different shower schemes for VBF.

[1] G. Aad *et al.* (ATLAS Collaboration), Phys. Lett. B **716**, 1 (2012).
[2] S. Chatrchyan *et al.* (CMS Collaboration), Phys. Lett. B **716**, 30 (2012).
[3] M. Cepeda *et al.*, CERN Yellow Rep. **7**, 221 (2019).
[4] M. Grazzini, Sm Higgs production cross sections at $\sqrt{S} = 13$, 14 and 27 Tev (update in cern hl-lhc yr 2019) (2019), https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CERNHLHE2019.
[5] A combination of measurements of Higgs boson production and decay using up to 139 fb^{-1} of proton–proton collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS experiment, Tech. Rep. (CERN, Geneva, 2020) all figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2020-027.
[6] G. Aad *et al.* (ATLAS Collaboration), Eur. Phys. J. C **80**, 957 (2020); **81**, 29(E) (2021); **81**, 398(E) (2021).
[7] Measurement of the properties of Higgs boson production at $\sqrt{s} = 13$ TeV in the $H \to \gamma\gamma$ channel using 139 fb^{-1} of pp collision data with the ATLAS experiment, Technical Report (CERN, Geneva, 2020) all figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2020-026.
[8] Measurements of gluon fusion and vector-boson-fusion production of the Higgs boson in $H \to WW^* \to e\nu\mu\nu$ decays using pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, Technical Report (CERN, Geneva, 2021) all figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2021-014.
[9] A. M. Sirunyan *et al.* (CMS Collaboration), J. High Energy Phys. 07 (2021) 027.
[10] Measurements of properties of the Higgs boson in the W boson pair decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV, Technical Report (CERN, Geneva, 2022).
[11] A. M. Sirunyan *et al.* (CMS Collaboration), Eur. Phys. J. C **81**, 488 (2021).
[12] C.-H. Chan, K. Cheung, Y.-L. Chung, and P.-H. Hsu, Phys. Rev. D **96**, 096009 (2017).
[13] Y.-L. Chung, S.-C. Hsu, and B. Nachman, J. Instrum. **16**, P07002 (2021).
[14] V. Rentala, N. Vignaroli, H.-n. Li, Z. Li, and C. P. Yuan, Phys. Rev. D **88**, 073007 (2013).
[15] J. Lin, M. Freytsis, I. Moult, and B. Nachman, J. High Energy Phys. 10 (2018) 101.
[16] G. Aad *et al.* (ATLAS Collaboration), Eur. Phys. J. C **82**, 622 (2022).
[17] G. Aad *et al.* (ATLAS Collaboration), Phys. Lett. B **812**, 135980 (2021).
[18] CMS Collaboration, Search for Higgs boson decays to a z boson and a photon in proton-proton collisions at $\sqrt{s} = 13$ tev, Report No. CMS-HIG-19-014, CERN-EP-2022-019, CMS-HIG-19-014-003 (2022), http://cds.cern.ch/record/2807790.
[19] A. M. Sirunyan *et al.* (CMS Collaboration), Phys. Lett. B **805**, 135425 (2020).
[20] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, arXiv:1703.03130.
[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Adv. Neural Inf. Process. Syst. **30** (2017).
[22] J. Pumplin, Phys. Rev. D **44**, 2025 (1991).
[23] J. Cogan, M. Kagan, E. Strauss, and A. Schwartzman, J. High Energy Phys. 02 (**2015**) 118.
[24] L. G. Almeida, M. Backović, M. Cliche, S. J. Lee, and M. Perelstein, J. High Energy Phys. 07 (2015) 086.
[25] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman, and A. Schwartzman, J. High Energy Phys. 07 (2016) 069.
[26] P. Baldi, K. Bauer, C. Eng, P. Sadowski, and D. Whiteson, Phys. Rev. D **93**, 094034 (2016).
[27] P. T. Komiske, E. M. Metodiev, and M. D. Schwartz, J. High Energy Phys. 01 (**2017**) 110.
[28] M. Kagan, L. d. Oliveira, L. Mackey, B. Nachman, and A. Schwartzman, EPJ Web Conf. **127**, 00009 (2016).
[29] D. Guest, J. Collado, P. Baldi, S.-C. Hsu, G. Urban, and D. Whiteson, Phys. Rev. D **94**, 112002 (2016).
[30] J. Barnard, E. N. Dawe, M. J. Dolan, and N. Rajcic, Phys. Rev. D **95**, 014018 (2017).
[31] P. T. Komiske, E. M. Metodiev, and J. Thaler, J. High Energy Phys. 04 (2018) 013.
[32] Quark versus gluon jet tagging using jet images with the ATLAS detector, Technical Report (CERN, Geneva, 2017) all figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2017-017.
[33] J. Pearkes, W. Fedorko, A. Lister, and C. Gay, arXiv:1704.02124.
[34] G. Kasieczka, T. Plehn, M. Russell, and T. Schell, J. High Energy Phys. 05 (2017) 006.
[35] K. Datta and A. Larkoski, J. High Energy Phys. 06 (2017) 073.
[36] A. Butter, G. Kasieczka, T. Plehn, and M. Russell, SciPost Phys. **5**, 028 (2018).
[37] K. Datta and A. J. Larkoski, J. High Energy Phys. 03 (2018) 086.
[38] S. Egan, W. Fedorko, A. Lister, J. Pearkes, and C. Gay, arXiv:1711.09059.
[39] S. Schramm (ATLAS Collaboration), EPJ Web Conf. **182**, 02113 (2018).
[40] G. Louppe, K. Cho, C. Becot, and K. Cranmer, J. High Energy Phys. 01 (2019) 057.
[41] T. Cheng, Comput. Softw. Big Sci. **2**, 3 (2018).
[42] A. M. Sirunyan *et al.* (CMS Collaboration), J. Instrum. **13**, P05011 (2018).
[43] P. T. Komiske, E. M. Metodiev, B. Nachman, and M. D. Schwartz, Phys. Rev. D **98**, 011502 (2018).
[44] S. Choi, S. J. Lee, and M. Perelstein, J. High Energy Phys. 02 (2019) 132.
[45] S. Macaluso and D. Shih, J. High Energy Phys. 10 (2018) 121.
[46] P. T. Komiske, E. M. Metodiev, and J. Thaler, J. High Energy Phys. 01 (2019) 121.
[47] G. Kasieczka, N. Kiefer, T. Plehn, and J. M. Thompson, SciPost Phys. **6**, 069 (2019).
[48] F. A. Dreyer, G. P. Salam, and G. Soyez, J. High Energy Phys. 12 (2018) 064.
[49] K. Fraser and M. D. Schwartz, J. High Energy Phys. 10 (2018) 093.

[50] Y.-C. J. Chen, C.-W. Chiang, G. Cottin, and D. Shih, Phys. Rev. D **101**, 053001 (2020).
[51] K. Datta, A. Larkoski, and B. Nachman, Phys. Rev. D **100**, 095016 (2019).
[52] H. Qu and L. Gouskos, Phys. Rev. D **101**, 056019 (2020).
[53] A. Chakraborty, S. H. Lim, and M. M. Nojiri, J. High Energy Phys. 07 (2019) 135.
[54] J. S. H. Lee, I. Park, I. J. Watson, and S. Yang, J. Korean Phys. Soc. **74**, 219 (2019).
[55] J. S. H. Lee, S. M. Lee, Y. Lee, I. Park, I. J. Watson, and S. Yang, J. Korean Phys. Soc. **75**, 652 (2019).
[56] S. Diefenbacher, H. Frost, G. Kasieczka, T. Plehn, and J. M. Thompson, SciPost Phys. **8**, 023 (2020).
[57] E. A. Moreno, T. Q. Nguyen, J.-R. Vlimant, O. Cerri, H. B. Newman, A. Periwal, M. Spiropulu, J. M. Duarte, and M. Pierini, Phys. Rev. D **102**, 012010 (2020).
[58] A. Andreassen, I. Feige, C. Frye, and M. D. Schwartz, Phys. Rev. Lett. **123**, 182001 (2019).
[59] E. A. Moreno, O. Cerri, J. M. Duarte, H. B. Newman, T. Q. Nguyen, A. Periwal, M. Pierini, A. Serikova, M. Spiropulu, and J.-R. Vlimant, Eur. Phys. J. C **80**, 58 (2020).
[60] J. Erdmann, J. Instrum. **15**, P01021 (2020).
[61] J. Li, T. Li, and F.-Z. Xu, J. High Energy Phys. 04 (2021) 156.
[62] E. Bols, J. Kieseler, M. Verzetti, M. Stoye, and A. Stakia, J. Instrum. **15**, P12012 (2020).
[63] A. Chakraborty, S. H. Lim, M. M. Nojiri, and M. Takeuchi, J. High Energy Phys. 07 (2020) 111.
[64] E. Bernreuther, T. Finke, F. Kahlhoefer, M. Krämer, and A. Mück, SciPost Phys. **10**, 046 (2021).
[65] S. H. Lim and M. M. Nojiri, Phys. Rev. D **105**, 014004 (2022).
[66] J. Guo, J. Li, T. Li, and R. Zhang, Phys. Rev. D **103**, 116025 (2021).
[67] M. J. Dolan and A. Ore, Phys. Rev. D **103**, 074022 (2021).
[68] V. Mikuni and F. Canelli, Eur. Phys. J. Plus **135**, 463 (2020).
[69] J. Li and H. Sun, arXiv:2009.00170.
[70] M. Kagan, arXiv:2012.09719.
[71] J. Erdmann, O. Nackenhorst, and S. V. Zeißner, J. Instrum. **16**, P08039 (2021).
[72] F. A. Dreyer and H. Qu, J. High Energy Phys. 03 (2021) 052.
[73] Y. Nakai, D. Shih, and S. Thomas, arXiv:2003.09517.
[74] S. Bhattacharya, M. Guchait, and A. H. Vijay, Phys. Rev. D **105**, 042005 (2022).
[75] A. M. Sirunyan *et al.* (CMS Collaboration), J. Instrum. **15**, P06005 (2020).
[76] M. Andrews *et al.*, EPJ Web Conf. **251**, 04030 (2021).
[77] J. Filipek, S.-C. Hsu, J. Kruper, K. Mohan, and B. Nachman, arXiv:2105.04582.
[78] V. Mikuni and F. Canelli, Mach. Learn. **2**, 035027 (2021).
[79] P. Konar, V. S. Ngairangbam, and M. Spannowsky, J. High Energy Phys. 02 (2022) 060.
[80] C. Shimmin, arXiv:2107.02908.
[81] F. Dreyer, G. Soyez, and A. Takacs, J. High Energy Phys. 08 (2022) 177.
[82] J. A. Aguilar-Saavedra, Eur. Phys. J. C **81**, 734 (2021).
[83] C. K. Khosa and S. Marzani, Phys. Rev. D **104**, 055043 (2021).
[84] S. Gong, Q. Meng, J. Zhang, H. Qu, C. Li, S. Qian, W. Du, Z.-M. Ma, and T.-Y. Liu, J. High Energy Phys. 07 (2022) 030.
[85] T. Kim and A. Martin, arXiv:2102.05124.
[86] H. Qu, C. Li, and S. Qian, in *Proceedings of the 39th International Conference on Machine Learning, PMLR* (2022), arXiv:2202.03772.
[87] Identification of jets containing *b*-Hadrons with recurrent neural networks at the ATLAS experiment, Technical Report (CERN, Geneva, 2017) all figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2017-003.
[88] Deep sets based neural networks for impact parameter flavour tagging in ATLAS, Technical Report (CERN, Geneva, 2020) all figures including auxiliary figures are available at https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PUBNOTES/ATL-PHYS-PUB-2020-014.
[89] T. Q. Nguyen, D. Weitekamp, D. Anderson, R. Castello, O. Cerri, M. Pierini, M. Spiropulu, and J.-R. Vlimant, Comput. Softw. Big Sci. **3**, 12 (2019).
[90] M. Andrews, M. Paulini, S. Gleyzer, and B. Poczos, Comput. Softw. Big Sci. **4**, 6 (2020).
[91] Y.-L. Du, K. Zhou, J. Steinheimer, L.-G. Pang, A. Motornenko, H.-S. Zong, X.-N. Wang, and H. Stöcker, Eur. Phys. J. C **80**, 516 (2020).
[92] B. Tannenwald, C. Neu, A. Li, G. Buehlmann, A. Cuddeback, L. Hatfield, R. Parvatam, and C. Thompson, arXiv:2009.06754.
[93] V. S. Ngairangbam, A. Bhardwaj, P. Konar, and A. K. Nayak, Eur. Phys. J. C **80**, 1055 (2020).
[94] G. Aad *et al.* (ATLAS Collaboration), Eur. Phys. J. C **80**, 942 (2020).
[95] A. M. Sirunyan *et al.* (CMS Collaboration), Phys. Rev. D **104**, 052004 (2021).
[96] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, J. High Energy Phys. 07 (2014) 079.
[97] H.-L. Lai, M. Guzzi, J. Huston, Z. Li, P. M. Nadolsky, J. Pumplin, and C. P. Yuan, Phys. Rev. D **82**, 074024 (2010).
[98] M. L. Mangano, M. Moretti, F. Piccinini, R. Pittau, and A. D. Polosa, J. High Energy Phys. 07 (2003) 001.
[99] A. Alloul, N. D. Christensen, C. Degrande, C. Duhr, and B. Fuks, Comput. Phys. Commun. **185**, 2250 (2014).
[100] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, Comput. Phys. Commun. **191**, 159 (2015).
[101] T. Sjöstrand, S. Mrenna, and P. Skands, J. High Energy Phys. 05 (2006) 026.
[102] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), J. High Energy Phys. 02 (2014) 057.
[103] B. Cabouat and T. Sjöstrand, Eur. Phys. J. C **78**, 226 (2018).
[104] S. Höche, S. Mrenna, S. Payne, C. T. Preuss, and P. Skands, SciPost Phys. **12**, 010 (2022).
[105] B. Jäger, A. Karlberg, S. Plätzer, J. Scheller, and M. Zaro, Eur. Phys. J. C **80**, 756 (2020).
[106] P. Konar and V. S. Ngairangbam, Phys. Rev. D **105**, 113003 (2022).
[107] M. Cacciari, G. P. Salam, and G. Soyez, Eur. Phys. J. C **72**, 1896 (2012).

[108] M. Cacciari, G. P. Salam, and G. Soyez, J. High Energy Phys. 04 (2008) 063.
[109] T. Chen and C. Guestrin, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, NY, 2016), pp. 785–794.
[110] M. Aaboud *et al.* (ATLAS Collaboration), Phys. Rev. D **98**, 052005 (2018).
[111] J. Shelton, *Theoretical Advanced Study Institute in Elementary Particle Physics: Searching for New Physics at Small and Large Scales* (World Scientific Publishing, 2013), pp. 303–340.
[112] S. Ioffe and C. Szegedy, in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37* (JMLR.org, Lille, France, 2015), pp. 448–456.
[113] V. Nair and G. Hinton, in *Proceedings of the 27th International Conference on International Conference on Machine Learning* (Omnipress, Madison, WI, 2010), pp. 807–814.
[114] M. Abadi *et al.*, TensorFlow: Large-scale machine learning on heterogeneous systems (2015), software available from tensorflow.org.
[115] F. Chollet *et al.*, KERAS (2015), https://github.com/keras-team/keras.
[116] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization (2014).
[117] K. He, X. Zhang, S. Ren, and J. Sun, in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2015), arXiv:1512.03385.
[118] K. Simonyan, A. Vedaldi, and A. Zisserman, in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings* (2014), arXiv:1312.6034.
[119] Y. Kubota, tf-keras-vis (2021), https://keisen.github.io/tf-keras-vis-docs/.