Published for SISSA by Springer

Received: January 13, 2024
Revised: February 2, 2024
Accepted: March 6, 2024
Published: March 26, 2024

Multi-scale cross-attention transformer encoder for event classification

A. Hammad,[a] **S. Moretti**[b,c] **and M. Nojiri**[a,d,e]

[a]*Theory Center, IPNS, KEK,*
1-1 Oho, Tsukuba, Ibaraki 305-0801, Japan
[b]*School of Physics and Astronomy, University of Southampton,*
Highfield, Southampton, U.K.
[c]*Department of Physics & Astronomy, Uppsala University,*
Box 516, SE-751 20 Uppsala, Sweden
[d]*The Graduate University of Advanced Studies (Sokendai),*
1-1 Oho, Tsukuba, Ibaraki 305-0801, Japan
[e]*Kavli IPMU (WPI), University of Tokyo,*
5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8583, Japan

E-mail: hamed@post.kek.jp, S.Moretti@soton.ac.uk, nojiri@post.kek.jp

Abstract: We deploy an advanced Machine Learning (ML) environment, leveraging a multi-scale cross-attention encoder for event classification, towards the identification of the $gg \to H \to hh \to b\bar{b}b\bar{b}$ process at the High Luminosity Large Hadron Collider (HL-LHC), where h is the discovered Standard Model (SM)-like Higgs boson and H a heavier version of it (with $m_H > 2m_h$). In the ensuing boosted Higgs regime, the final state consists of two fat jets. Our multi-modal network can extract information from the jet substructure and the kinematics of the final state particles through self-attention transformer layers. The diverse learned information is subsequently integrated to improve classification performance using an additional transformer encoder with cross-attention heads. We showcase that our approach surpasses current alternative methods used to establish sensitivity to this process in performance, whether solely based on kinematic analysis or combining this with mainstream ML approaches. Then, we employ various interpretive methods to evaluate the network results, including attention map analysis and visual representation of Gradient-weighted Class Activation Mapping (Grad-CAM). Finally, we note that the proposed network is generic and can be applied to analyse any process carrying information at different scales. Our code is publicly available for generic use.[1]

Keywords: Higgs Production, Jets and Jet Substructure, Multi-Higgs Models

ArXiv ePrint: 2401.00452

[1]https://github.com/AHamamd150/Multi-Scale-Transformer-Encoder.

https://doi.org/10.1007/JHEP03(2024)144

Contents

1 Introduction **1**

2 Transformer encoder **2**
2.1 Attention mechanism 3

3 Physics example **5**
3.1 The 2HDM 6
3.2 Analysis strategy 8
3.3 Data pre-processing 10

4 Results **14**
4.1 The influence of cross-attention 15

5 Interpretation of the transformer encoder results **17**
5.1 Attention maps 18
5.2 Grad-CAM 20

6 Conclusion **22**

A Networks structure **23**

1 Introduction

Information about jet identification provides powerful insights into collision events and can help to separate different physics processes originating these. This information can be extracted from the elementary particles localized inside a jet. Recently, various methods have been used to exploit the substructure of a jet to probe new physics signatures using advanced Machine Learning (ML) techniques [1–5].

Conversely, using the reconstructed kinematics from the final state jets for event classification spans the full phase space and exhibits large classification performance [6–18]. Such high-level kinematics (i.e., encoding the global features of the final state particles), possibly together with the knowledge of the properties of (known or assumed) resonant intermediate particles, remains blind to the information encoded inside the final state jets.

A possible way to extract information from both jet substructure and global jet kinematics is to concatenate the information extracted from a multi-modal network [19–24]. If the kinematics and jet substructure have different performances in event discrimination, the network assigns higher weight values to the layers associated with kinematics as input, while allocating lower values to layers associated with local jet information as input. Such a simple concatenation leads to an imbalance of the extracted information, within which the kinematic information generally dominates [25].

In this paper, we present a novel method for incorporating different-scale information extracted from both global kinematics and substructure of jets via a transformer encoder

JHEP03(2024)144

with a cross-attention layer. The model initially extracts the most relevant information from each dataset individually using self-attention layers before incorporating these using a cross-attention layer. The method demonstrates a larger improvement in classification performance compared to the simple concatenation method.

To assess our results, we analyze the learned information by the transformer layers through the examination of the attention maps of the self- and cross-attention layers. Attention maps provide information about the (most) important embedded particles the model focuses on when classifying signal and background events. However, they cannot highlight the region in the feature (e.g., phase) space crucial for model classification. For this purpose, we utilize Gradient-weighted Class Activation Mapping (Grad-CAM) to highlight the geometric region in the $\eta - \phi$ (detector) plane where the model focuses on classifying events.

We test our approach for the dominant decay channel of Higgs boson pairs with Standard Model properties (hh) produced at the LHC, that is, into four b-(anti)quarks. This signal has historically proved to be extremely challenging to extract owing to a significant QCD background. Lately, there have been several attempts to tackle this signature using both standard [26–28] and ML [29] approaches. Furthermore, in the case that the hh intermediate state emerges from the (resonant) decay of a heavier Higgs state (H), each of the two would be (slim) b-jets produced by the two h decays actually merge into one (fat) jet, as the two h states can be very boosted. The final states in the detectors little resemble the primary parton kinematics of the underlying physics in such case.

The plan of this paper is as follows. In the next section, we describe the basics of a transformer encoder. Then, in section 3, we introduce the physics process that we use as an example. In section 4, we present our numerical results. In section 5, we interpret the classification results using various methods. The section 6 is for conclusions. The details of our network structure can be found in the appendix.

2 Transformer encoder

Transformers were originally proposed as sequence-to-sequence models for machine translation [30]. The main ingredient of the original transformer model is the encoder-decoder block. However, the models using encoder block only often appear for event classification analysis at the LHC [31–33].

Inherited by the word tokens in the original transformer model, transformer encoders are used to analyze events in terms of clouds of particles for High Energy Physics (HEP) analysis [34, 35]. Particle clouds represent the final state particles as a permutation invariant sequence of particles. Such a representation has the ability to share the advantages of particle based representations, especially the flexibility to include arbitrary features for each particle.

The motivation to apply transformer encoders to particle clouds stems from their inherent ability to model interactions between particles irrespective of their spatial proximity. By leveraging self-attention mechanisms, transformer encoders enable each particle to dynamically weigh the influence of other particles within the entire cloud, thus capturing both local and global dependencies. This can potentially revolutionize the analysis of HEP systems, particularly by offering a more holistic understanding of their behavior and interactions.

JHEP03(2024)144

Understanding the scientific operation of transformer encoders in the context of particle clouds requires diving into the core components of these models. At the heart of the transformer architecture is the attention mechanism, an algorithm that allows the model to focus on different parts of the input sequence when making predictions. An attention mechanism operates by assigning attention weights to different particles based on their relevance to the current particle being processed. This allows the model to consider global relationships and dependencies, enabling it to capture emergent behaviors, interactions, and patterns that may not be apparent in filter based methods, e.g., Convolutional Neural Networks (CNNs), which mainly extract the local information.

2.1 Attention mechanism

The attention mechanism is an essential component of transformer models, playing a crucial role in capturing information and dependencies amongst particles. In the transformer architecture, the attention mechanism enables the model to focus selectively on different parts of the input sequence, allowing for the modelling of complex relationships and dependencies. In general, the attention mechanism operates by assigning different weights to different elements in the input sequence, emphasizing the more relevant parts while downplaying the less relevant ones.[1] The attention mechanisms broadly span two types, as follows.

- **Self-attention** is a more advanced form of attention where the model attends to different positions in the input sequence to weight their importance concerning the current position. In the context of the transformer model, self-attention allows each element in the sequence to attend to all other elements, capturing both local and global dependencies. Attention scores are calculated and used to combine the values associated with different positions linearly. The self-attention mechanism enables the model to consider the entire context, making it particularly effective for tasks where long-range dependencies are crucial.

- **Cross-attention** extends the self-attention mechanism to handle input sequences from different sources. In the transformer architecture, it is often used when processing pairs of sequences of different structures. Cross-attention allows each element in the first sequence to attend to all other elements in the subsequent sequence. This facilitates modelling the relationships between different modalities or extracting the relevant information from sequences with different scales.

Consider the input data sets (x_i, x_j) that have first been passed by a linear fully connected NN layer to generate the weight matrices as follows:

$$Q_i = W^Q \cdot x_i \,, \quad K_j = W^K \cdot x_j \,, \quad V_j = W^V \cdot x_j \,, \tag{2.1}$$

where K, Q and V are called key, query, and value vectors, respectively, and used to compute the attention to the whole data set.

[1] Dropping the less informative instances from the data can rectify the sparsity problem when using CNNs to analyze jet images.

JHEP03(2024)144

Scaled dot-product attention can then be defined as

$$\alpha_{ij} = \text{softmax}\left(\frac{Q_i \cdot K_j^T}{\sqrt{d}}\right) = \frac{\exp(Q_i \cdot K_j^T/\sqrt{d})}{\sum_j \exp(Q_i \cdot K_j^T/\sqrt{d})}\,, \tag{2.2}$$

while the attention output is computed as a weighted sum of the attention scores as

$$\mathcal{Z}_i = \sum_j \alpha_{ij} V_j\,. \tag{2.3}$$

This is called self-attention if the attention is computed for the same data set, i.e., $x_i = x_j$. The weights matrices have the dimensions of $W_Q^{i\times i}, W_K^{i\times i}, W_V^{i\times i}$ which mixes the features of the input data and retain the dimension of the embedded input to the original one. In contrast, if the two input data sets differ, i.e., $x_i \neq x_j$, cross attention is needed. In this case, the weight matrices should have different dimensions with $W_Q^{i\times i}, W_K^{j\times i}, W_V^{j\times i}$ in order to calculate attention. Attention output is used to scale the input data set via a skip connection as

$$\widetilde{x_i} = x_i + \mathcal{Z}_i\,. \tag{2.4}$$

The newly transformed data set $\widetilde{x_i}$ indicates the attention importance of each element in the data set to the whole elements in the set. Although the attention output mixes the input and feature tokens, the skip connection keeps the reference to the order of the original input data set. In the subsequent sections of the paper, we use the modified particle tokens to provide a straightforward interpretation of the results.

At its basic level, each transformer layer includes a multi-head attention, which combines different attention heads, allowing for parallel multi-dimensional processing of the inputs. Multi-head attention is a key innovation in the transformer model architecture, enhancing expressive power and capturing complex patterns in data by allowing the model to attend to different aspects of the input sequence simultaneously. Therefore, this mechanism eases the understanding of varied and subtle connections within the data, offering a more thorough representation.

As explained, a single set of attention weights is computed for the entire input sequence. Multi-head attention extends this concept by employing multiple attention heads, each responsible for learning different aspects of the relationships within the data. Each attention head independently processes the input sequence, producing a set of output values. These outputs are then linearly combined to form the final output of the multi-head attention layer.

Mathematically, if h represents the number of attention heads, and head_i denotes the ith attention head, the output $\mathcal{O}$ is obtained by concatenating the outputs of each attention head and linearly transforming these:

$$\mathcal{O} = \text{CONCAT}\left(\text{head}_1, \text{head}_2, \cdots \text{ head}_h\right) W^O\,, \tag{2.5}$$

with W^O the learnable linear transformation matrix which has the dimensions of $W_O^{(h*i)\times i}$ to retain the same dimensions as the original input. This enables the model to capture different aspects of relationships and dependencies simultaneously.

The choice of the number of attention heads, denoted as h_i, is a crucial hyperparameter in designing a transformer model. Increasing the number of attention heads has several

JHEP03(2024)144

implications, such as enhancing the model's capacity to capture complex relationships. It is also important to mention that a higher number of attention heads also increases computational complexity. Training and inference times and memory requirements could increase. Therefore, the number of attention heads should be balanced based on the task requirements and available computational resources.

In this particular context, we present an innovative methodology aimed at integrating inputs characterized by distinct scales within a multi-modal transformer model featuring cross-attention layers. The schematic representation of the network architecture is shown in figure 1. Considering the specific case of the HEP process to be studied at the LHC, $gg \to H \to hh \to b\bar{b}b\bar{b}$, the model dynamically adjusts three streams through self-attention transformer layers, each devoted to analyzing the leading jet, second-leading jet and the reconstructed kinematics, respectively. At this juncture, the model independently extracts pivotal information from each data set, leveraging self-attention mechanisms before their collective processing through a cross-attention layer.

The main role of the cross-attention layer is to extract the local jet substructure information effectively and incorporate it into the extracted kinematic information. Notably, the adaptability of the cross-attention layer in merging information from one data set into another affords flexibility in determining how to integrate the extracted information, providing the option to accentuate jet information for enhancing kinematics. Once the most relevant information from the data sets is extracted and combined via the cross-attention layer, we feed the output to fully connected NNs to analyze the captured information and compute the classification probability. The inclusion of self-attention layers in the model holds significance, as it allows for the independent extraction of the most relevant information from each data set before their amalgamation using the cross-attention mechanism. This characteristic makes the model proficient in analyzing multi-scale data characterized by intricate structures.

3 Physics example

We analyse SM-like di-Higgs boson (hh) production at the HL-LHC (with an integrated luminosity of $3000\,\text{fb}^{-1}$) within the framework of the 2HDM. In the boosted regime, where the di-Higgs boson is produced from an on-shell heavy Higgs, H, the final state features two fat jets, as illustrated in figure 2 by the two red cones therein. Currently, ATLAS analysis [36] shows the limit on the production cross section of heavy scalar decaying to two fat Higgs jets. The given limit on the production cross section is prominent with 95% C.L. to be $\sigma(pp \to H \to hh) < 50\,\text{fb}$ for $m_H = 1\,\text{TeV}$, and we expect the bound to further improve at the HL-LHC.

Therefore, to start with, in this section, we provide a brief review of the 2HDM with type-II Yukawa couplings, focusing on the aspects that are relevant to our analysis. We then describe the strategy behind our numerical analysis, together with its constituent elements, i.e., the event generation and detector simulation procedures, as well as the signal and background properties, in terms of the overall kinematics and internal dynamics of jets. We adopt different transformer encoder configurations to analyze the kinematics and jet substructure individually and efficiently combine the information from both of these.

JHEP03(2024)144

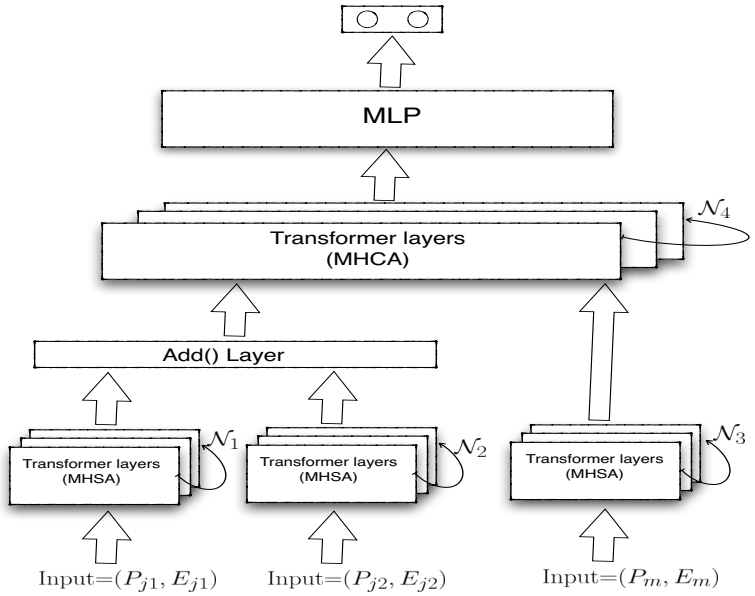

Figure 1. Structure of the transformer model used. Here, P_{j1}, P_{j2} are the number (dimnesion) of the leading and second leading jet constituents while the P_m's are the number of the reconstructed particles, namely, j_1, j_2, and H; E_{j1}, E_{j2}, and E_m are number the corresponding features of each dataset. Also, MHSA stands for multi-heads self-attention layers, and MHCA stands for multi-heads cross-attention layers. Finally, the $\mathcal{N}_i$'s are the number of the used transformer layers. The transformer layers are stacked and work sequentially, as pointed out by the black arrow.

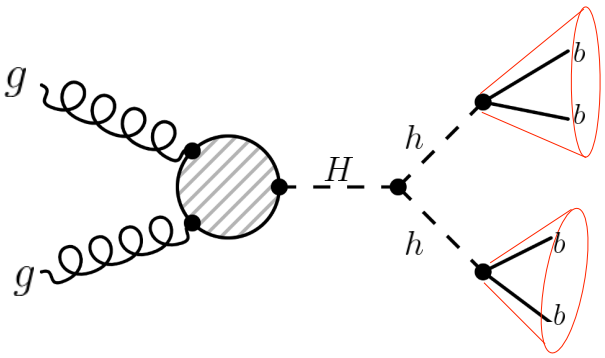

Figure 2. Feynman diagram for the signal process.

3.1 The 2HDM

The 2HDM is an extension of the SM through a second $\mathrm{SU}(2)_L$ Higgs doublet with the same quantum numbers under the SM symmetry gauge group [37, 38]. The most general 2HDM Higgs potential is given by

$$\begin{aligned} V_\phi = {} & m_{11}^2(\phi_1^\dagger\phi_1) + m_{22}^2(\phi_2^\dagger\phi_2) - \left[m_{12}^2(\phi_1^\dagger\phi_2) + \text{h.c.}\right] \\ & + \lambda_1(\phi_1^\dagger\phi_1)^2 + \lambda_2(\phi_2^\dagger\phi_2)^2 + \lambda_3(\phi_1^\dagger\phi_1)(\phi_2^\dagger\phi_2) + \lambda_4(\phi_1^\dagger\phi_2)(\phi_2^\dagger\phi_1) \\ & + \frac{1}{2}\left[\lambda_5(\phi_1^\dagger\phi_2)^2 + \left[\lambda_6(\phi_1^\dagger\phi_1) + \lambda_7(\phi_2^\dagger\phi_2)\right](\phi_1^\dagger\phi_2) + \text{H.c.}\right] . \end{aligned} \tag{3.1}$$

JHEP03(2024)144

The potential structure allows for Flavor Changing Neutral Currents (FCNCs) at the tree level, which is strongly constrained by experimental measurements. Applying a global Z_2 symmetry to the scalar potential, with $(\phi_1, \phi_2) \to (\phi_1, -\phi_2)$ transformations, prevents the existence of such FCNC sources [39]. However, the most general Yukawa interaction violates such a Z_2 symmetry, thus leading to potential FCNCs at the tree level, as pointed out in ref. [40]. Therefore, to tame the latter, only specific Yukawa structures, known as Types [37], are allowed. Yet, to enable Electro-Weak Symmetry Breaking (EWSB) consistent with the measured particle spectrum of the SM, a softly broken Z_2 symmetry should eventually be enabled by requiring a small but non-vanishing term $m_{12}^2(\phi_1^\dagger \phi_2)$ and setting $\lambda_6 = \lambda_7 = 0$. (Herein, softly means that the model still respects the Z_2 symmetry at small distances through all orders of perturbation theory.) The soft mass m_{12}^2 and λ_5 are in general complex, though [41, 42]. In the following, we will consider a real potential that thus preserves the CP symmetry, $\mathrm{IM}(m_{12}^2) = \mathrm{Im}(\lambda_5) = 0$. In such a configuration of the 2HDM, then 7 independent parameters remain, which are the λ_i's, with $i = 1, \ldots 5$, $\tan\beta = v_2/v_1$[2] and m_{12}^2, from which the physical parameters, i.e., Higgs boson masses and couplings, are obtained, with the constraint that one of the two CP-even Higgs fields should be the discovered one with mass of 125 GeV or so (which in our case is the h field). Finally, as mentioned already, we restrict our study to the Type-II among the possible Yukawa structures.

The tree level mass matrix squared for the Higgs fields can be obtained as

$$\left(\mathcal{M}^2\right)_{ij} = \left.\frac{\partial V_\phi}{\partial h_i \partial h_j}\right|_{h_{i,j}=0} , \tag{3.2}$$

where the h_i's ($i = 1, \ldots, 4$) are the four components of the complex doublet fields. Upon EWSB, three physical neutral scalars are obtained after diagonalizing the corresponding mass matrices, as intimated, two CP-even (scalar) ones (h, H) and a CP-odd (pseudoscalar) one (A), with masses given by

$$m_{h,H}^2 = \frac{1}{2}\left[\chi_{11}^2 + \chi_{22}^2 \mp \sqrt{(\chi_{11}^2 - \chi_{22}^2)^2 + 4(\chi_{12}^2)^2}\right] , \tag{3.3}$$

$$m_A^2 = \frac{2m_{12}^2}{\sin 2\beta} - \lambda_5 v^2 , \tag{3.4}$$

with

$$\chi_{11}^2 = m_{12}^2 \tan\beta + 2\lambda_1 v^2 \cos^2\beta , \tag{3.5}$$

$$\chi_{22}^2 = m_{12}^2 \cot\beta + 2\lambda_2 v^2 \sin^2\beta , \tag{3.6}$$

$$\chi_{12}^2 = -m_{12}^2 + \frac{1}{2}(\lambda_3 + \lambda_4 + \lambda_5) v^2 \sin 2\beta , \tag{3.7}$$

where the VEVs satisfy the relation $v = \sqrt{v_1 + v_2}$ (with v being the SM one).[3]

To stay with the neutral Higgs sector, the imposed CP conservation only allows for tree level couplings between two massive gauge bosons and the CP-even Higgs states, while the CP-odd Higgs state can only couple to a gauge boson and a CP-even Higgs one. Furthermore,

[2] With v_1 and v_2 being the Vacuum Expectation Values (VEVs) of the two Higgs doublets.

[3] The other two Higgs states emerging from the 2HDM after EWSB are charged and are denoted by $H^\pm$.

JHEP03(2024)144

m_H[GeV]	λ_1	λ_2	λ_3	λ_4	λ_5	$\tan\beta$	m_{12}^2[TeV^2]	$\cos(\beta-\alpha)$	σ_{prod}[fb]
600	1.80	0.23	1.75	−2.06	−1.09	5.00	−78.97	0.31	0.86
800	3.20	0.25	1.75	−2.06	−1.29	4.00	−128.91	0.33	0.375
1000	1.0	0.16	3.50	−2.06	−1.49	3.00	−302.92	0.37	0.11
2000	1.0	0.14	2.75	−1.06	−1.97	5.00	−889.05	0.32	0.024

Table 1. Input parameters for our four BPs. The last column shows the production cross section for the process $gg \to H \to hh$.

all neutral Higgs states can couple to fermions. The coupling strength of the neutral Higgs bosons to both matter and forces are parameterized in terms of $\tan\beta$ and another parameter, α, which is the mixing angle between the CP-even Higgs states [37]. Furthermore, the triple scalar coupling is independent of the Yukawa structure and is given by [43]

$$\lambda_{(H,h,h)} = -\frac{e\ c_{\beta-\alpha}}{2m_W s_W s_{2\beta}^2}\left[(2m_h^2 + m_H^2)s_{2\alpha}s_{2\beta} + (3s_{2\alpha} - s_{2\beta})m_{12}^2\right], \tag{3.8}$$

where e is the electric charge and s, c are the sin and cos of the given angle.

The 2HDM free parameters are constrained from various theoretical considerations and experimental observations, as described in [44]. Thus, profiting from the scan results performed therein, we adopt four Benchmark Points (BPs), with $m_H = 600, 800, 1000$, and 2000 GeV, that satisfy all the current bounds. In table 1, we show the parameters values of these points while the last column shows the production cross section σ_{prod} of our target process (prior to the two $h \to b\bar{b}$ decays) at $\sqrt{s} = 14$ TeV.

3.2 Analysis strategy

With the theoretical setup clarified, we now proceed to a phenomenological study of di-Higgs boson production, focusing on final states with two boosted fat jets. We align our analysis with the boosted analysis presented in the latest ATLAS paper [36]. The primary background contamination arises from QCD multijet processes, specifically $pp \to jjjj$, contributing an estimated 90% of the total background, while the di-top process $\bar{t}t$ contributes at the 10% level. Other background processes, including SM h, hh, and EW di-boson production, have been assessed to make negligible contributions to the selected event yields. Therefore, they are not included in our analysis. Given that BSM di-Hggs events suffer from huge background contamination and it is not trivial to extract the signal information, we employ various configurations of transformer encoders for this analysis.

Commencing with the analysis of the global information encoded in the high-level reconstructed kinematics of both the signal and relevant background events, we employ a transformer encoder with multi-head self-attention to optimize the separation power between signal and background events. However, the presence of similar (to the signal) kinematic structures in some background processes poses a challenge to the classification efficiency of this network. Additionally, the substantial cross section of the background events diminishes signal significance, even after optimizing the cut on the output score.

JHEP03(2024)144

We use MadGraph5 [45] to estimate multi-parton amplitudes and to generate signal and background events for subsequent processing. Background processes $pp \to bbbb$[4] and $pp \to t\bar{t}$ are computed at Leading Order (LO) while the Higgs production from gluon-gluon fusion is calculated at Next-to-LO (NLO) in QCD using an effective coupling calculated by SPheno [46, 47]. PYTHIA [48] is used for parton shower, hadronization, heavy flavor decays, and for adding the soft underlying event. The $\bar{t}t$ background is simulated at LO and up to two more jets with the matching scale 20 GeV via the MLM method [49, 50].

Following the event selection in ATLAS analysis [36], we use the DELPHES package [51] for detector simulation. DELPHES parameterizes the detector response, including tracks, calorimeter deposits, and high level objects such as isolated electrons, jets, taus, and Missing E_T (MET). We use the default ATLAS card for resolution, but the (fat)jets are reconstructed from the Eflow objects, combining tracks and calorimeter information. Fat jets are clustered using the anti-k_T algorithm [52, 53] with cone radius $R = 1$ and, to ensure further suppression of pile-up noise, jet trimming is performed [54]. To enhance the network performance, one may consider applying initial cuts on certain variables before inputting the distributions into the network, aiming to amplify the signal and suppress the backgrounds. We follow the pre-selection cuts outlined in [36], requiring two fat jets with a double b-tagging each. Moreover, each event must have at least two fat jets with radius $R = 1.0$ and $pT > 450$ GeV for the leading jet and $pT > 250$ GeV for the second leading jet. Each of the two fat jets is required to have a pseudorapidity $|\eta(J)| < 2.5$, a lower mass cut of $m(J) > 50$ GeV, and a mass window of 200 GeV is applied for the m_H reconstruction for $m_H \leq 1$ TeV and relaxed for higher masses to allow for more statistics. Unlike the ATLAS analysis, we do not consider pile-up effects in this analysis. Moreover, ATLAS analysis uses a new neural-network-based b-tagging algorithm DL1r [55], while we use the Delphes b-tagging with flate effeciency of 80%.

In addition to the global kinematical variables, we can utilize jet substructure to distinguish between signal and background events. This naturally arises from the fact that jets initiated by different particles exhibit distinct characteristics. While heavy boosted particles, such as $W^{\pm}$, Z and Higgs bosons, can result in jets with a distinctive multi-prong structure, quark and gluon jets are unlikely to have such structure. Furthermore, the boosted colour singlet particle is isolated in colour flow. Therefore, two b jets from Higgs decay are colour-connected only among themselves, unlike QCD jets.

Consequently, the features of the parent particles can be inferred from the structure of the jet constituents. This information enables the recovery of various local details about events from different processes, serving as a discriminating variable between signal and background events. The study of jet substructure to identify the parent particle initiating a jet, thereby distinguishing jets initiated from heavy boosted particles from QCD jets was introduced in [56–70] (and references herein). Recently, improvement on jet identification continued by using ML methods for jet image analysis [71–79], graph based analysis [80–82] or sequence based analysis [83–88]. In this paper, we especially employ a multi-modal transformer encoder with self-attention multi-heads to analyze the jet contents. The different modalities are designed to extract information from the leading and second-leading jet contents in parallel before a simple concatenation is performed for classification purposes. Without cross attention

[4]We simulate the multi-jets QCD processes in Madgraph as $pp \to b\bar{b}b\bar{b}$ to speed up the event generation.

JHEP03(2024)144

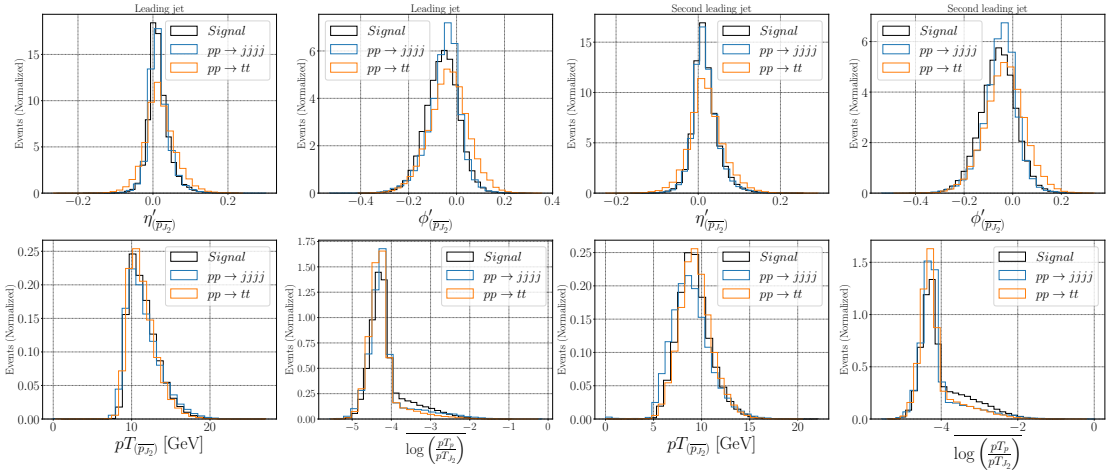

Figure 3. Left: distributions of 10000 leading jets features averaged over the jet constituents. Right: distributions of 10000 second leading jets features averaged over the jet constituents. Signal distributions are for the BP with $m_H = 1\,\text{TeV}$.

to the high-level kinematical variable discussed next, the classification performance is based solely on the information localized inside the fat jet. Integrating inputs of varying scales encompassing both kinematics and jet substructure information, we utilize a multi-modal transformer encoder equipped with three streams and cross-attention head. The first and second streams process information from the leading and second-leading jet contents. Each of them features a transformer encoder with self-attention heads. Once important features are extracted from the jets, they are aggregated in an addition layer. The third stream, dedicated to high-level kinematics, employs a transformer encoder with self-attention heads. The output from the addition layer and the final layer of the third transformer are fed into a cross-attention layer. This cross-attention layer is pivotal in connecting information extracted from the jet constituents to the corresponding kinematics, enhancing the overall classification performance. To elucidate the impact of the cross-attention layer, we introduce a fourth model wherein we substitute the cross-attention layer with a straightforward concatenation layer.

3.3 Data pre-processing

Particle clouds enable configuring diverse data into the network, emphasizing the permutation symmetry of inputs to yield a promising representation of jets. Initially, we pre-process the data sets for the leading and second-leading jet contents up to 50 constituents each. The particles are arranged in descending order based on their transverse momentum. For events with fewer constituents, the remaining positions are padded with zeros, ensuring conformity with the stipulated count.[5] For each instance of the jet contents we store 4 features: pT, η, ϕ and $\log \frac{pT}{pT_{(\text{jet})}}$ [35]. Figure 3 shows the four features averaged over the number of jet contents for 10000 events of the leading jet (left) and second leading jet (right).

[5]We stress here that we use an attention mask such that the network performance is not affected by the padded events [30].

JHEP03(2024)144

To optimize the network discriminative accuracy, it is imperative to pre-process the jet contents, ensuring the manifestation of a multi-prong structure specific to signal events. We use the preprocessing steps that were introduced for jet image analysis. The preprocessing allows learning from small input data and considerably speeds up the learning process.[6] For this purpose, the following transformations are applied before inputting the data into the network.[7]

- **Translation** Jet contents are shifted in the $\eta - \phi$ directions such that the jet axis is at the center of the $\eta - \phi$ plane.

- **Rotation** Rotation is executed to mitigate the stochastic nature inherent in the decay angle concerning the $(\eta - \phi)$ coordinate system. This alignment is achieved comprehensively by ascertaining the principal axis of the original data and subsequently rotation around the jet-energy centroid. This rotation ensures that the principal axis consistently aligns vertically. The rotation transformation is performed by first computing the leading eigenvector of the covariance matrix as the principle axis of the jet. A rotating angle, θ, is then defined as $\arctan2\left(\frac{x_1}{x_2}\right)$, with x_1, x_2 are the first and second components of the eigenvector respectively. Finally, the rotating angle is used to rotate the $(\eta - \phi)$ coordinates of the jet constituents to new non-physical coordinates, $(\eta' - \phi')$, in which the principle axis of the jet is always vertical.

- **Flipping** Jet constituents are reflected over the vertical axis such that the right side of η' always has the highest momentum. This ensures that the hardest radiation always appears in similar locations, which can be exploited to enhance the classification performance.

After pre-processing transformations, input data sets for the leading and second leading jets have the dimensions of $(n, 50, 4)$, where n is the number of events, 50 is the number of jet constituents, and 4 is the number of pre-processed features.

Figure 4 presents the cumulative average of 50000 p_T distributions for both the leading (upper row) and second-leading (lower row) jets. The impact of pre-processing transformations is evident in revealing the multi-prong structure characteristic of signal events, wherein the leading and second-leading subjets are localized in specific regions within the $(\eta' - \phi')$ plane. In contrast, subjets from QCD multi-jets exhibit a broad energy range, lacking a discernible prong structure. Conversely, $\bar{t}t$ events show a distinct three-prong structure attributable to the fully hadronic decays of the top quark. Notably, despite the multi-prong structure in $\bar{t}t$ background events, their contribution to the overall background is merely 10%. We will see later that background rejection efficiency is high, therefore $t\bar{t}$ background can be important to estimate the accessibility of the signal.

The kinematics data sets have dimension $(n, 3, 6)$ with n as the number of events, 3 as the number of reconstructed particles, leading, second leading jet, and heavy Higgs, and 6

[6] In principle, we can use the input data without the preprocessing, but this needs a large input data set and training for a long time [33].

[7] Other than the rotation and flipping as proposed below, it is also possible to recluster the fat jet into subjets and define the rotation and flipping based on the subjet locations.

JHEP03(2024)144

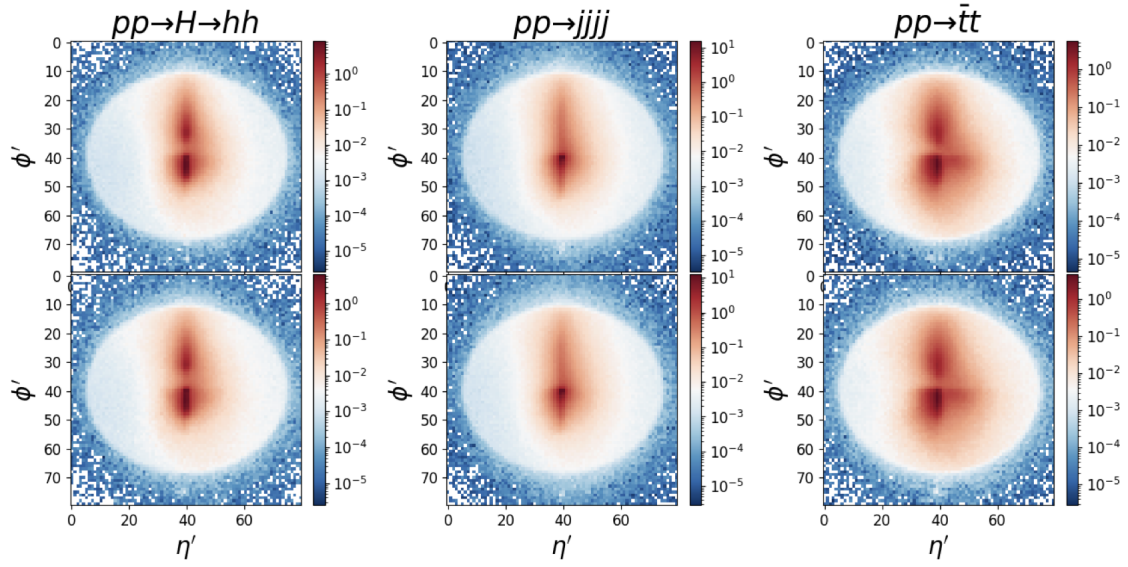

Figure 4. For illustration purposes, we show the averag p_T distribution of 50000 events of the leading (second leading) fat jet contents in the upper raw (lower raw) after pre-processing steps for both signal and backgrounds. The $\eta' - \phi'$ plane are divided into 80 bins in each direction between -1 to 1. The signal events (left) are simulated for the BP with $m_H = 1\,\text{TeV}$ and shown against the yield of the $bbbb$ (center) and $\bar{t}t$ (right) background events. Here, X and Y ticks indicate the bin in η and ϕ direction.

as the number of the kinematic variables for each reconstructed particle. The 6 kinematic variables are mass m, p_T, η, ϕ, energy of the jet (E) and the rotation angle of the jet (θ). Note that we assign 5 inputs corresponding to the 4 momenta of the jet. Because of the kinematical constraints $p^2 = m^2$ and $p_H = p_{J1} + p_{J2}$, there are only 8 physically independent observable among 15 kinematical inputs. These additional inputs help the network to figure out relevant features for the classification.

Figure 5 shows the normalized kinematic distributions for the signal point with $m_H = 1\,\text{TeV}$ and backgrounds. In addition to the reconstructed high-level kinematics, we incorporate the θ_i distributions for the leading and second-leading jets (but not the heavy Higgs), which are the rotating angles of the leading and second leading jet contents.

We incorporate the data sets as input to the networks as the inputs to the first and second transformer encoders have the dimensions of $(n, 50, 4)$. Input to the third transformer encoder has the dimension of $(n, 3, 6)$. Once the data sets are pre-processed, we stack signal and background events in each data set separately, attaching labels of $Y = 1$ for the signal events and $Y = 0$ for the background events. During the training of the network, the model tries to minimize a categorical cross entropy loss function by minimizing the difference between the model prediction and the assigned labels. In this analysis, we use equal size data sets for signal and background events for training with 1 million events[8] and 100000 event for test.

[8]A major problem in any attention based transformer model which exhibits larger classification performance with larger size training set.

JHEP03(2024)144

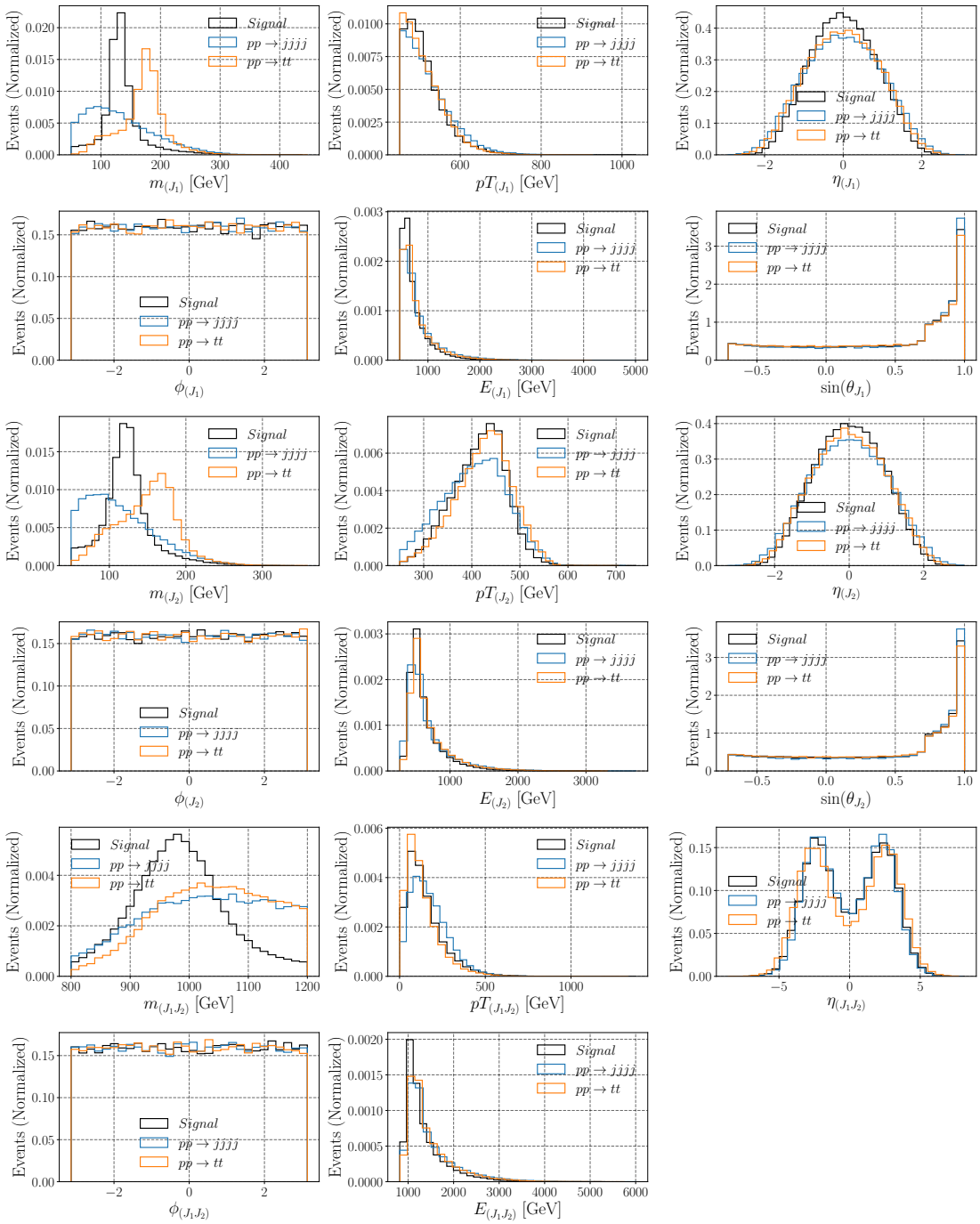

Figure 5. Kinematics distributions of 10000 events for the signal BP with m_H =1 TeV and the corresponding backgrounds after applying the pre-selection cuts.

JHEP03(2024)144

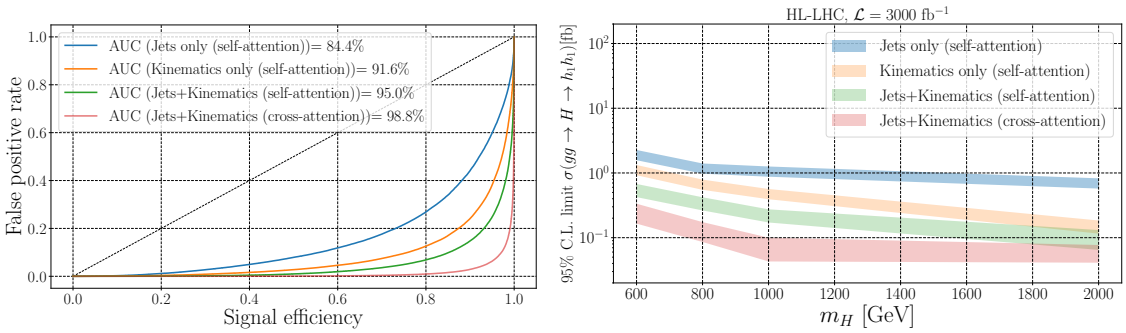

Figure 6. Left: the Receiver Operating Characteristic (ROC) curves for the four networks for the signal BP with $m_H = 1\,\text{TeV}$. Right: Estimated 95% upper limit, with $Z_A \geq 2$ from eq. (4.1), on the total cross section for the process $gg \to H \to hh$ (having factored out the SM-like $h \to b\bar{b}$ decays) at the HL-LHC (with an integrated luminosity of $3000\,\text{fb}^{-1}$) for different ML analyses. The band for each plot represents the upper and lower values for 5 independent training of different random number seeds.

4 Results

We now present the analysis results for probing the signature of the heavy scalar in the process of boosted di-Higgs boson production, $gg \to H \to hh \to b\bar{b}b\bar{b}$, at the HL-LHC (with an integrated luminosity of $3000\,\text{fb}^{-1}$). The discriminating power of each network measures how well the signal and background may be characterised through their different features, all entangled together into several kinematic distributions and jet substructure information. For this purpose, we utilize four different attention based transformer models which analyze the reconstructed high level kinematics or the jet substructures individually via transformer encoders with self-attention mechanisms. Alternatively, we adopt two multi-modals transformer encoders to analyze the combined information of kinematics and jets substructure. In the latter, we incorporate the different information using a simple concatenation layer or cross-attention layer. A full description of the used networks is in appendix A.

The classification performance of the utilized networks is presented in figure 6. In the left plot, we showcase the ROC for the employed networks for a signal with $m_H = 1\,\text{TeV}$. The multi-modal transformer encoder with cross-attention layers performs best, achieving an Area Under the Curve (AUC) of 98.8%. In contrast, the transformer encoder trained solely on the jet substructure information exhibits the lowest performance with an AUC of 84.4%. It is crucial to highlight the impact of the cross-attention layer, which enhances performance by 7% over the transformer model trained exclusively on kinematic information. Replacing the cross-attention layer with a simple concatenation layer results in a degradation of classification performance by approximately $\sim 4\%$, as depicted by the green line in the plot.

We now illustrate the impact of our classifier on $H \to hh$ exclusion and discovery. In order to optimize the signal-to-background yield, we enforce a cut on the network output score by keeping only 20 events of the background. With this choice, we alleviate the statistical errors that may occur for lower background [89]. The optimized signal and background events

JHEP03(2024)144

are used to derive the upper limit using the following formula [90]:

$$Z_A = \left[2\left((N_s+N_b)\ln\frac{(N_s+N_b)(N_b+\sigma_b^2)}{N_b^2+(N_s+N_b)\sigma_b^2} - \frac{N_b^2}{\sigma_b^2}\ln(1+\frac{\sigma_b^2 N_s}{N_b(N_b+\sigma_b^2)})\right)\right]^{1/2}, \tag{4.1}$$

with N_s and N_b being the number of signal and background events, respectively, and where σ_b characterizes the uncertainty in the background events chosen to be a conservative value of 10% [91]. In this approximation, one expects to exclude(discover) regions with a total significance of $Z_A > 2(5)$.

The network performance is subject to both training and statistical uncertainty from limited training and testing samples. For example, the network performance can be influenced by the random partitioning of the training and test data sets, and the network performance varies when repeating the training and test steps with new splits. We repeat the experiment for k times and report the results as bands between the highest and lowest values. In our results, we use $k = 5$, and the bands represent the values of the different represented experiments.

In figure 6 (right), we show the estimated 95% upper limit, namely $Z_A \geq 2$, on the production cross-section at the HL-LHC for heavy scalar mass ranges between $600 - 2000\,\text{GeV}$. For smaller mass, incorporating the kinematics and jet information via cross-attention layers yields the best performance among all other network configurations. For larger masses, the reconstructed kinematics of the signal show a clear difference from the background events. Therefore, the performance of the transformer models saturates. In fact, for the limit, e.g., $m_H = 2\,\text{TeV}$, the background events can be easily removed with a simple cut on the reconstructed distributions of the signal events, which exhibits a clear difference from the background distributions. The transformer network trained on the jet constituents only does not show a large impact with varying heavy scalar mass.

4.1 The influence of cross-attention

To evaluate the impact of the cross-attention layer on the classification performance, figure 7 presents the attention output, as defined in eq. (2.4), for both the multi-modal transformer with cross-attention trained on kinematics plus jet constituents and the transformer network trained on kinematics only. In both networks, the attention output has dimensions of $(3, 6)$, where 3 represents the reconstructed particles (leading, second-leading jet, and heavy Higgs), and the last dimension accounts for the utilized features. We stress that the input to the corss-attention layer in the network structure, shown in figure 2, is adjusted with the Query matrix encodes the jet constituents. In contrast, the Key and Value matrices encoding the kinematics. Accordingly, the output of the cross-attention layers has the same dimensions as the kinematics dataset. In principle, we have the freedom to choose whether to add the jet information to the kinematics by fixing the assigned Query, Key, and Value matrices, but we opted to incorporate the jet information into the high-level kinematics.

Figure 7 displays the distributions of the attention output for each transformed particle individually and averaged over the used features. The top row shows the attention output for signal and background events using a transformer encoder trained on kinematics only. Conversely, when the information of the jet constituents is included using the cross-attention layer, the attention output distributions for background events are broader, and the signal

JHEP03(2024)144

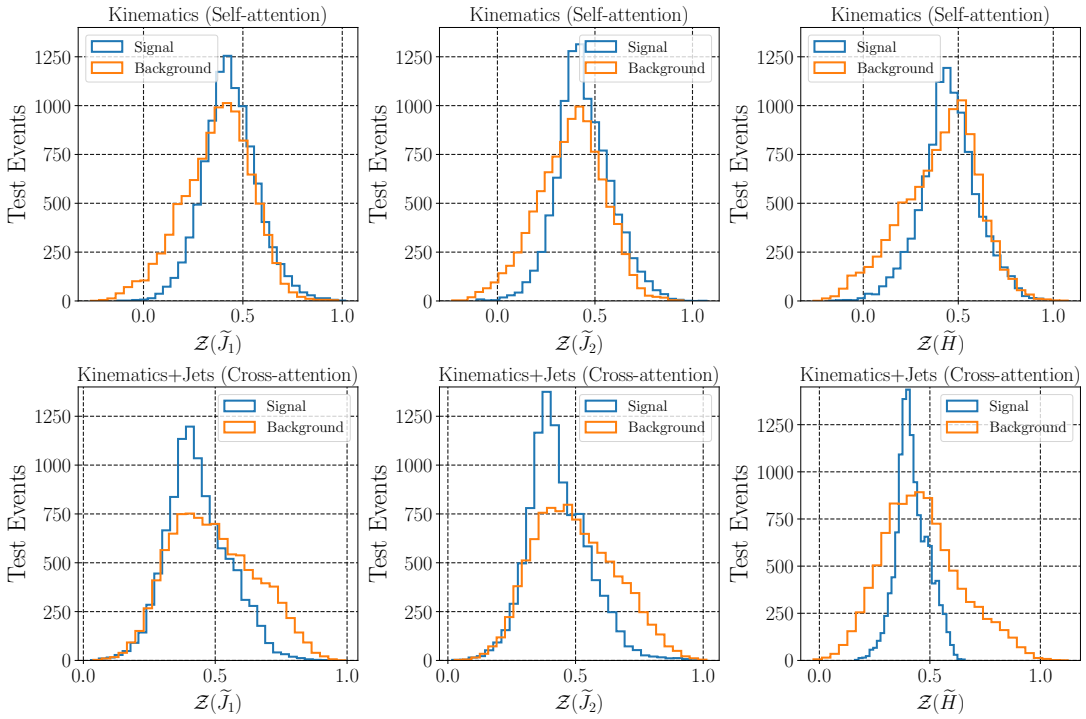

Figure 7. Top: output of the self-attention layer when trained on kinematics only. Bottom: output of the cross-attention layer when trained on kinematics and jet information. In both cases, attention output has the dimensions of (reconstructed particles × features), and for both plots we use 10000 test events and average over the features for the background and the signal point with $m_H = 1\,\text{TeV}$. $\widetilde{J_1}, \widetilde{J_2}$ and $\widetilde{H}$ represent the transformed particles as in eq. (2.4).

Kinematics	Kinematics + θ	Jet str. + Kinematics	Jet str. + Kinematics + θ
91.01%	91.60%	97.23%	98.68%

Table 2. Area Under the ROC (AUC) for the networks using Kinematics or Kinematics + Jet structure information with/without θ.

distributions are narrower. The fact that background jets lack a multi-prong structure with broader soft radiations influences the attention output for background events, increasing the output variations in the feature space.

Finally, we include, alongside the described kinematical information, also the rotation angle θ aligning the fat jet axis to the ϕ direction after shifting the jet η and ϕ to the centre of the $\eta - \phi$ plane. This information allows the network to reconstruct the full events and access the correlation of the jet shape to the other fat jet and the beam axis.

In table 2, we compare the AUC value of the network using Kinematical inputs (Kins), Kins + θ, Kins + jet substructure inputs (Jet str.), Kins + Jet str. + θ. Adding θ to Kins improves AUC by 0.59 while adding θ to Kins + Jet str. improves AUC by 1.45. This indicates the correlation between all inputs (Jet str., θ, and Kins) is contributing to the signal and background classification.

JHEP03(2024)144

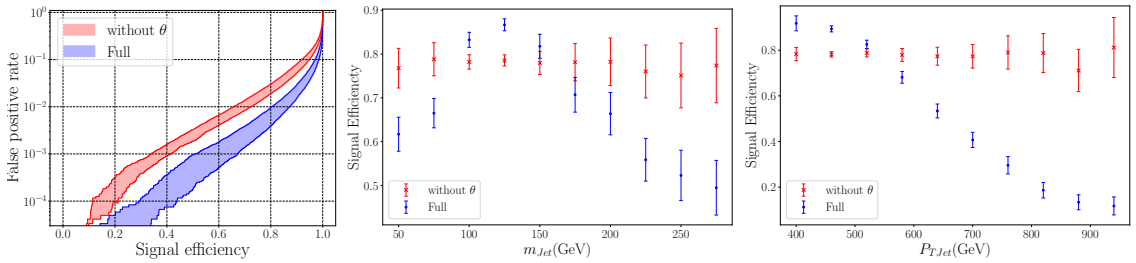

Figure 8. Left: the ROC curve and error band of the full model using θ input (blue) and the model without θ input (red). The ROC is obtained by using 20,000 signal and background testing events. The error is estimated as in figure 6. Middle(Right): the signal efficiency as varying m_{J_1} (p_{TJ_1}) for the best training results. The ratio is calculated with a score cut of 80% of the signal efficiency for 20,000 signal samples. The efficiency (without) using θ is shown by blue(red) bars indicating statistical errors without taking into training errors. The acceptance of the full model is higher than without θ input at $m_{J_1} \sim m_h(\ p_{J_1} \sim m_H/2)$.

In figure 8 Left, we show the ROC curve of the network trained without the θ inputs (red) compared to the ROC curve of our coss-attention model (blue). The improvement in the background rejection is a factor of four for a signal efficiency of 80%. Therefore, including θ results in a drastically increased performance. In figure 8 Middle and Right, we show the efficiency in rejecting background for the model with/without θ inputs. The model with θ has higher efficiency at $m_{J_1} \sim m_h$ and $p_T \sim \frac{m_H}{2}$. In short, the model can focus more on the $H \to hh$ kinematics with θ inputs. We also looked for simple correlations among θ and the other kinematical variables, such as η_J ϕ_J, but did not find any apparent ones contributing to the selection improvement, consistent with insignificant improvement by adding θ to the model using Kins. This indicates that the network is not merely utilizing the transformer output of the jet substructure individually. Instead, it tunes the jet substructure analysis under the condition of global kinematical information. The correlations within the internal structures of the jet will be investigated in future publications.

5 Interpretation of the transformer encoder results

In the following section, we discuss additional methods to interpret and analyze the results of the transformer encoder with cross-attention, which has the best classification performance in figure 6. The interpretation methods are generic and can be further applied to other networks to interpret their results. As attention-based transformer models excel in capturing intricate spatial relationships and global context within data, their interpretability becomes paramount. Interpretation methods for attention-based transformers aim to elucidate the visual cues, features, and regions that contribute significantly to the model's predictions. Common interpretation methods are

- **Attention Maps:** attention maps visualize the focus of the model by highlighting the particles in the cloud that receive higher attention. These maps provide a direct view into which particles are considered most relevant for prediction, facilitating an intuitive understanding of the model's decision-making process [92].

JHEP03(2024)144

- **Grad-CAM** It generates class-specific activation maps by weighting the gradients of the predicted class score with respect to the final transformer layer [93]. This technique highlights the regions in the feature space that are crucial for the model's classification decision and thus can provide a geometrical interpretation, $\eta - \phi$ plane, of the learned information by the network.

- **Saliency Maps** Saliency maps for transformer models are a form of interpretability technique used to understand and visualize the importance of different parts of the input sequence concerning the model's predictions. Saliency maps highlight the regions of the input that most significantly influence the model's output, providing insights into the model's decision-making process [94–96]. By examining the saliency map, users can gain insights into which parts of the input sequence are crucial for the model's predictions.

- **Layer-wise Relevance Propagation (LRP)** The primary goal of LRP is to assign relevance scores to input features, indicating their contribution to the model's output [97]. However, it's worth noting that LRP has limitations, and its effectiveness can vary depending on the specific neural network architecture and the nature of the task. Different variants of LRP have been proposed to address specific challenges and improve its applicability to various models.

The interpretation of attention-based transformer models is pivotal for unlocking their full potential and ensuring their responsible deployment in real-world applications. Among all the mentioned methods, we adopt the attention maps and Grad-CAM to interpret the learned information using the transformer model.

5.1 Attention maps

Attention maps serve as a bridge between the abstract nature of neural network computations and the desired interpretability. These maps visualize the attention scores assigned to each particle token in the input sequence, providing a representation of where the model focuses its attention during processing.

The analysis of the attention maps highlights the particle tokens that receive higher attention scores, indicating their significance in the model's decision. Also, it reveals how particle tokens relate to each other. For instance, it highlights the information extracted from the jet constituents relevant to the reconstructed objects. Importantly, examining attention maps can pinpoint areas where the model might struggle or make mistakes.

In this context, we utilize attention maps to analyze the acquired information from the last transformer layer of the transformed jet constituents. Our focus centres on the output of the network shown in figure 1. We begin by examining the attention maps of the Add() layer, which contains information about the jet substructure. In this case, the attention maps denoted as α_{ij} in eq. (2.2), have dimensions of $(n_{\text{heads}}, 50, 50)$, where 50 represents the number of constituents in the jet, and n_{heads} denotes the number of self-attention heads. We take $n_{\text{heads}} = 5$, see appendix A for detail.

Figure 9 displays the values of the attention maps for each attention head individually, with signal events in the top row and background events in the bottom row. Given that jet constituents are originally ordered by their momentum, the X and Y axes ticks represent the

JHEP03(2024)144

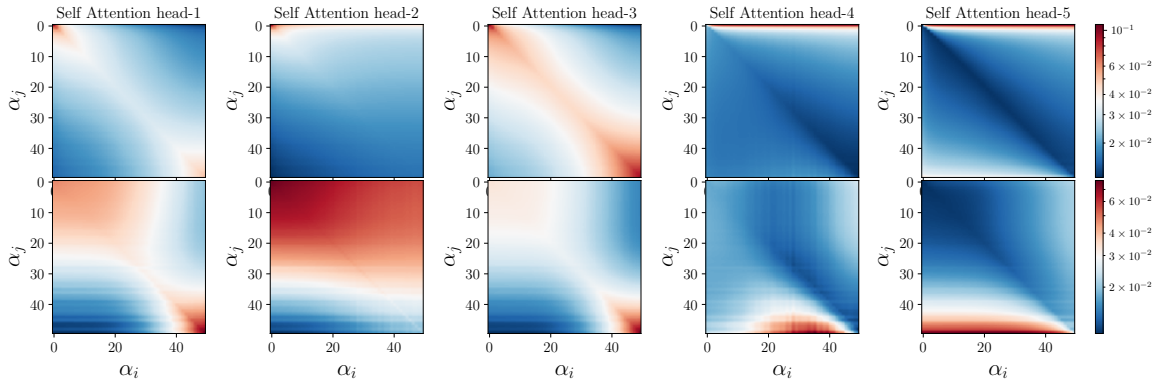

Figure 9. Attention maps of the last self-attention transformer layer, which processes the jet substructure for the signal (top) and backgrounds (bottom) for a 120K test event.

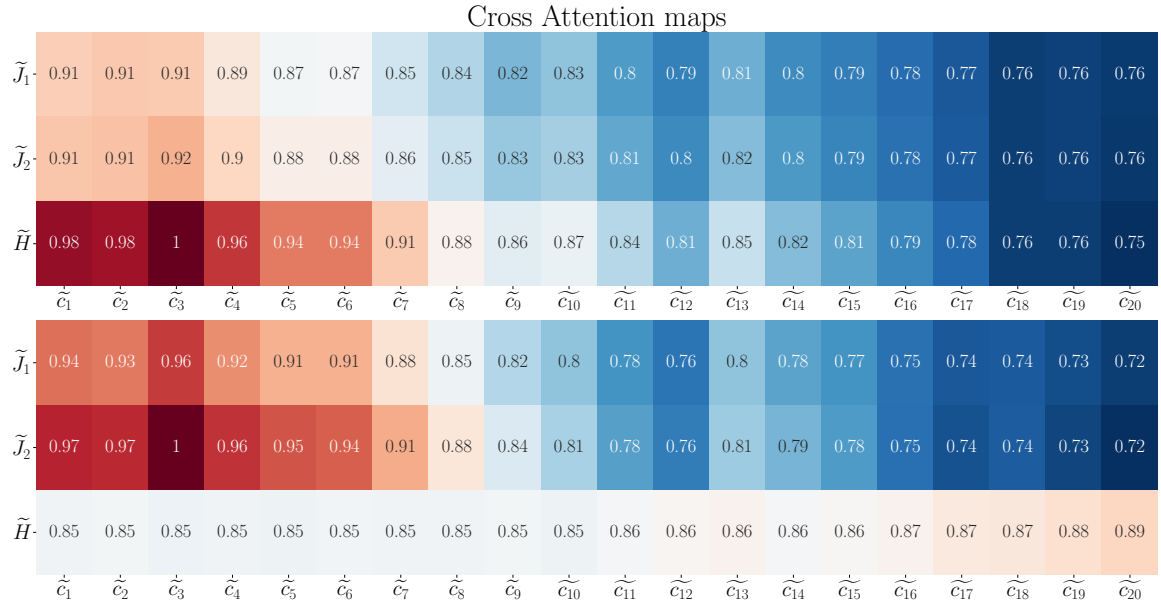

Figure 10. Cross-attention maps of the cross-attention transformer layer averaged over the 8 cross-attention heads, which processes the jet substructure and the event kinematics for the signal (top) and backgrounds (bottom) for a 120K test event. The X-axis shows the attention score for the first transformed 20$^{\text{th}}$ jet contents while the Y-axis shows the attention score for the transformed reconstructed final state particles.

attention values of the transformed jet constituents in descending order (where the zero tick represents the leading transformed jet constituent particle). The attention map values reveal that the model concentrates on the leading and second-leading jet constituents to identify events as signal-like, particularly evident in attention heads $1, 2, 4$, and 5. In fact, this reflects the efficiency of the network to capture the two-prong structure of the signal events. On the other hand, the network assigns high attention to the wide momentum orders of the jet constituents when the network identifies the input as a background event. We stress here

JHEP03(2024)144

that, while the transformer layers intermingle particle and feature tokens, the skip connection still preserves the order of the attention output in relation to the original input data.

The attention maps for background events exhibit significant agreement with the jet substructure of the background events presented in figure 4. To this end, through an analysis of the attention scores from the last transformer layer of the jet constituents, we confirm that the transformer model adeptly extracts the correct multi-prong structure of signal events. Meanwhile, for background events dominated by QCD processes, the model exhibits high attention across a wide momentum range of jet constituents.

The attention maps of the cross-attention layer illustrate the attention scores between the jet constituents and the reconstructed particles, including the leading and second-leading jets and the heavy Higgs. The dimension of the attention score in the cross-attention layer is $(n_{\text{heads}}, 3, 50)$, where 3 represents the number of reconstructed particles, 50 is the number of jet constituents, and n_{heads} is the number of cross-attention heads, set at 8. Figure 10 displays the cross-attention maps for signal events (top) and background events (bottom), averaged over the used cross-attention heads.

The cross-attention maps for signal events exhibit a stronger correlation between the highest momentum transformed jet constituents and the heavy Higgs. In contrast, the Heavy Higgs displays a flat attention pattern with jet constituents of different momenta for the background events. Indeed, the results from the cross-attention maps, along with the cross-attention output shown in figure 7, provide a comprehensive overview of the impact of the cross-attention layer. This layer effectively assigns information from the jet constituents to the kinematics of the reconstructed particles to enhance the classification performance.

5.2 Grad-CAM

Grad-CAM is a technique designed to visualize and interpret the decisions made by DNN models. It builds upon the idea of class activation maps (CAMs) [98, 99] but extends it to models with arbitrary architectures. The primary objective of Grad-CAM is to highlight the important regions in a transformed input features space, $\widetilde{\eta} - \widetilde{\phi}$ plane, that contribute to the prediction of a specific class [92].

Let $F_k(\widetilde{\eta}, \widetilde{\phi})$ represent the activation of the final transformer layer for the k^{th} event. The gradient of the predicted class score (Y_c) with respect to the activation output is computed as:

$$\frac{\partial Y_c}{\partial F_k} \tag{5.1}$$

This gradient is then globally averaged to obtain the importance weights (α) as

$$\alpha_k(\widetilde{\eta}, \widetilde{\phi}) = \frac{1}{Z} \sum \frac{\partial Y_c}{\partial F_k(\widetilde{\eta}, \widetilde{\phi}, \widetilde{p_T})} \tag{5.2}$$

where Z is the size of the feature activations, and the sum runs over the jet constituents. $\widetilde{\eta}$, $\widetilde{\phi}$ and $\widetilde{p_T}$ are the transformed features. The final Grad-CAM heatmap is a weighted sum of gradients as

$$\text{Grad-CAM}(\widetilde{\eta}, \widetilde{\phi}) = \frac{1}{k} \sum_k \alpha_k(\widetilde{\eta}, \widetilde{\phi}) F_k(\widetilde{\eta}, \widetilde{\phi}, \widetilde{p_T}) \tag{5.3}$$

JHEP03(2024)144

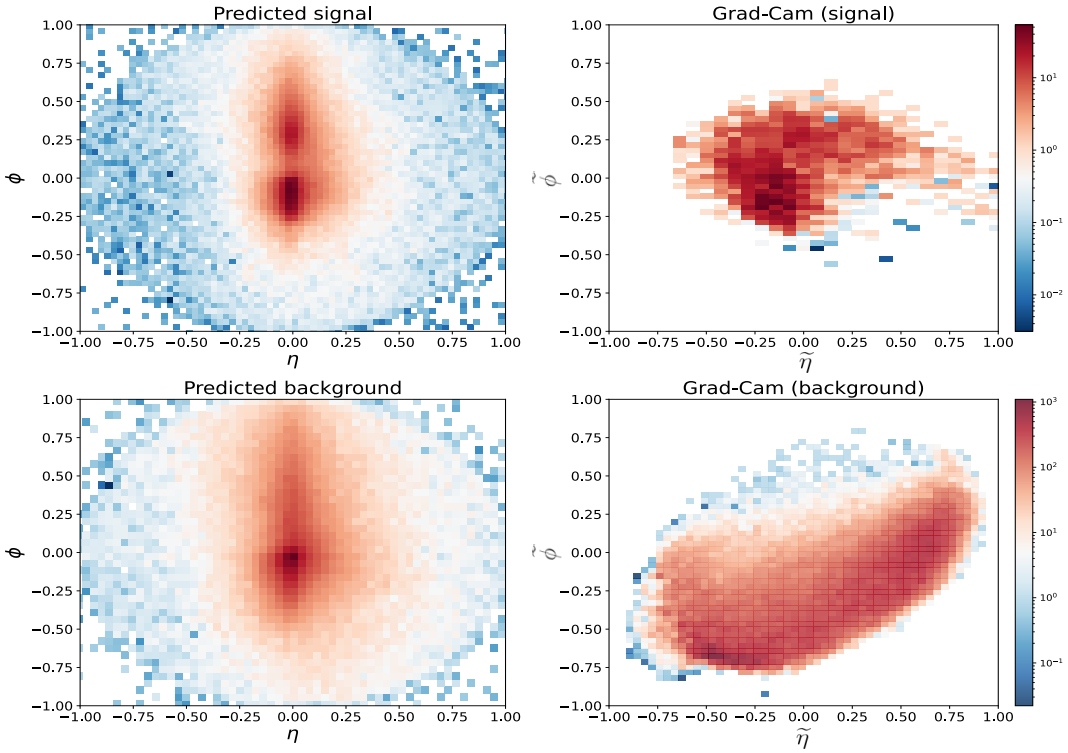

Figure 11. Grad-CAM results for 5000 test events of the transformer model with cross-attention. Left: p_T distribution of the jet constituents when events are predicted as signal events (top) and background events (bottom) in the η-ϕ plane. Note that the asymmetric pattern is due to the flipping transformation in the pre-processing steps in which all constituents with larger momentum are reflected in the positive η direction. Right: heat-map of the Grad-CAM results in transformed $\widetilde{\eta}$-$\widetilde{\phi}$ plane, as in eq. (2.4).

This heatmap highlights the regions of the input image that contribute the most to the prediction of the target class.

In general, it operates by utilizing the gradient information flowing into the final transformer layer in the following way: during the forward pass, the neural network processes the input particle cloud, and the activations of the final transformer layer are obtained. The gradients of the predicted class score with respect to the final transformer layer activation are computed during the backward pass. The gradients are then used to calculate the importance of the activation map. These importance scores are essentially the weights assigned to each spatial location of the final transformer layer. The weighted sum of the particle tokens is computed, creating the Grad-CAM. This map highlights the regions that contributed the most to the final prediction. Additionally, upsampling is often employed to match the Grad-CAM dimensions with the original input features.

To visualize the geometrical interpretation of the learned information from the jet constituents, we utilize Grad-CAM on the final self-attention layer of the jets, specifically, the Add() Layer depicted in figure 1. The results are shown in figure 11 for 5000 test images. The left panel illustrates the p_T distribution of the predicted events as signal (top) or background (bottom). Signal events are considered for the benchmark point with $m_H = 1\,\text{TeV}$. The right panel displays the heat map of the Grad-CAM output for the predicted signal (top) and the predicted background (bottom).

JHEP03(2024)144

The visualization of the heat map clarifies that the transformer model focuses on the two-prong structure to classify the input event as a signal. On the other hand, it relies on the soft-radiation pattern to classify the input event as a background event. Interestingly, we found that the result highlights that the model focuses on the positive η direction to make predictions, which is due to the flipping transformation done in the pre-processing step.

While Grad-CAM has the power to explain the considered regions in the feature space for the network predictions, one of its drawbacks is that it relies on gradient information from the final transformer layer. In cases where global context is crucial for decision-making, Grad-CAM may not capture long-range dependencies effectively. Moreover, Grad-CAM might be sensitive to small changes in the input, potentially making it less robust in the presence of adversarial examples.

6 Conclusion

In conclusion, this paper introduces an innovative method for enhancing event classification by effectively incorporating information from both global kinematics and substructure of jets in an event. Conventional approaches, using simple concatenation to combine the event information, have limitations, especially for scenarios where kinematical structures dominate. Specifically, the proposed method utilizes a transformer encoder with cross-attention layers, enabling the extraction of different scale information from both global kinematics and jet substructure. The results demonstrate a substantial improvement in classification performance compared to traditional concatenation methods. Indeed, the analysis of the learned information, conducted through attention maps and a Grad-CAM algorithm for visual representation, provides valuable insights into the model focus on important particles and geometric regions in the transformed $\widetilde{\eta} - \widetilde{\phi}$ plane that is crucial for event classification.

We have validated this approach by focusing on the dominant decay channel, i.e., into four b-jets, of SM-like Higgs boson pairs produced in the resonant decay of a heavier CP-even Higgs state, at the HL-LHC. This challenging scenario involves merging slim b-jets into a fatjet, due to the boosted nature of the lighter Higgs states so that the possibility of accessing partonic dynamics is apparently lost at the detector level. Furthermore, this occurs in an environment rich in tracks and calorimetric information not directly pertaining to the hard scattering sought, as typical of this CERN machine upgrade. Therefore, all these aspects add complexity to the classification task. Despite these challenges, the proposed method effectively addresses the intricacies of the final state in the detectors, ultimately outperforming mainstream signal selection procedures, whether based solely on kinematical analysis or less advanced ML tools. In the broader context, this research contributes to utilizing advanced jet identification techniques for global event reconstruction towards the understanding of collision events consisting of dynamics acting at various physics scales. Thus, the proposed method offers a promising avenue for improving the accuracy and efficiency of event classification in potentially many more complex scenarios encountered in high energy physics experiments.

JHEP03(2024)144

Acknowledgments

The work of SM is supported in part through the NExT Institute, the Knut and Alice Wallenberg Foundation under the Grant No. KAW 2017.0100 (SHIFT) and the STFC Consolidated Grant No. ST/L000296/1. AH and MN are funded by grant number 22H05113, "Foundation of Machine Learning Physics", Grant in Aid for Transformative Research Areas and 22K03626, Grant-in-Aid for Scientific Research (C).

A Networks structure

In this study, we employed four transformer encoders with distinct configurations to analyze various datasets. For all networks, we configured an output layer with two neurons and applied softmax activation. Additionally, we utilized the Adam optimizer [100] to minimize the sparse categorical cross-entropy loss function [101], setting the learning rate at 0.005. Our training dataset comprised one million samples, with 20,000 allocated for validation and 100,000 for testing. The training batches were adjusted to a size of 500.

Following data preprocessing, we obtained three datasets: one for the leading jet contents with dimensions of $(50, 4)$, where 50 represents the number of jet constituents and 4 denotes the considered features (p_T, η, ϕ, $\log(p_T)$). We also use the same information for the second leading jet contents. The last dataset for event kinematics with dimensions of $(3, 6)$, where 3 corresponds to the leading jet, the second leading jet and the (reconstructed) heavy Higgs boson while the 6 represents the used kinematics as in figure 5. The structure of the different networks is the following:

- A two-stream self-attention transformer encoder is employed for jet substructure. The network takes two separate data sets for the leading and second-leading jet constituents as input, processed through two distinct transformer layers. Each transformer layer is repeated three times. These transformer layers consist of five self-attention heads operating in parallel. The output from the attention heads is then integrated with the original input data via a skip connection layer [102]. The resulting output from the skip connection is flattened and forwarded to two fully connected layers with 128 and 4 neurons, respectively, using the GELU activation function [103]. The output from the final fully connected layer is subsequently combined with the self-attention output through a second skip connection layer.[9] The final output of the transformer layer undergoes a normalization layer and has the same dimension as the input dataset. The normalized output from each transformer layer is combined through an addition layer. This output then passes through a Multi-Layer Perceptron (MLP) comprising two fully connected layers with dimensions 128 and 64, employing the GELU activation function. Following each fully connected layer, a dropout layer with a dropout rate of 20% is applied. The output is then passed to the output layer for classification. The model is trained for 30 epochs with a batch of size 500 in 1421 seconds.

- A single-stream self-attention transformer encoder is employed for kinematics analysis. The network exclusively utilizes the kinematics dataset as input. To achieve this, we

[9] Skip connection is of the utmost importance to stabilize the gradient flow of the model.

JHEP03(2024)144

adopt the identical structure of the self-attention transformer encoder designed for jet substructure but with a singular stream. The model is trained for 30 epochs with a batch of size 500 in 1390 seconds.

- A three-stream transformer encoder is employed to analyze the leading and subleading jet constituents and the reconstructed kinematics. In this approach, we adjust the transformer layers for the leading and subleading jets from the first network, while the transformer layers for the kinematics are adapted from the latter network. The output of the self-attention transformer encoder layers for jet constituents is added via an addition layer. The resulting output from the addition layer, along with the output from the self-attention transformer layers of the kinematics, is then fed to a cross-attention transformer layer. This cross-attention transformer layer is repeated twice, and the output has the same dimensions as the input kinematics dataset, i.e., $(3, 6)$. Subsequently, this output passes through a MLP consisting of two fully connected layers with dimensions 128 and 64, utilizing the GELU activation function. After each fully connected layer, a dropout layer with a dropout rate of 20% is applied. The resulting output is then forwarded to the output layer for classification. The model is trained for 30 epochs with a batch of size 500 in 1576 seconds.

- The final network is configured to mirror the three-stream transformer encoder, with the only modification being the substitution of the cross-attention transformer layers with a single concatenation layer. The model is trained for 30 epochs with a batch of size 500 in 1282 seconds.

For training of all models, we use two NVIDIA RTX A6000 GPU cards using the Tensorflow mirror strategy with the utilization of 80% and 30% for the first and second cards, respectively, and memory consumption of 96% (48 GB) of both cards.

References

[1] A. Chakraborty, S.H. Lim and M.M. Nojiri, *Interpretable deep learning for two-prong jet classification with jet spectra*, *JHEP* **07** (2019) 135 [arXiv:1904.02092] [INSPIRE].

[2] Y.-L. Chung, S.-C. Hsu and B. Nachman, *Disentangling boosted Higgs boson production modes with machine learning*, 2021 *JINST* **16** P07002 [arXiv:2009.05930] [INSPIRE].

[3] J. Guo, J. Li, T. Li and R. Zhang, *Boosted Higgs boson jet reconstruction via a graph neural network*, *Phys. Rev. D* **103** (2021) 116025 [arXiv:2010.05464] [INSPIRE].

[4] C.K. Khosa and S. Marzani, *Higgs boson tagging with the Lund jet plane*, *Phys. Rev. D* **104** (2021) 055043 [arXiv:2105.03989] [INSPIRE].

[5] K. Datta, A. Larkoski and B. Nachman, *Automating the construction of jet observables with machine learning*, *Phys. Rev. D* **100** (2019) 095016 [arXiv:1902.07180] [INSPIRE].

JHEP03(2024)144

[6] D. Cogollo et al., *Deep learning analysis of the inverse seesaw in a 3-3-1 model at the LHC*, *Phys. Lett. B* **811** (2020) 135931 [arXiv:2008.03409] [INSPIRE].

[7] M. Grossi, J. Novak, B. Kersevan and D. Rebuzzi, *Comparing traditional and deep-learning techniques of kinematic reconstruction for polarization discrimination in vector boson scattering*, *Eur. Phys. J. C* **80** (2020) 1144 [arXiv:2008.05316] [INSPIRE].

[8] V.S. Ngairangbam, A. Bhardwaj, P. Konar and A.K. Nayak, *Invisible Higgs search through vector boson fusion: a deep learning approach*, *Eur. Phys. J. C* **80** (2020) 1055 [arXiv:2008.05434] [INSPIRE].

[9] C. Englert et al., *Sensing Higgs boson cascade decays through memory*, *Phys. Rev. D* **102** (2020) 095027 [arXiv:2008.08611] [INSPIRE].

[10] F.F. Freitas, J. Gonçalves, A.P. Morais and R. Pasechnik, *Phenomenology of vector-like leptons with deep learning at the Large Hadron Collider*, *JHEP* **01** (2021) 076 [arXiv:2010.01307] [INSPIRE].

[11] A. Stakia et al., *Advances in multi-variate analysis methods for new physics searches at the Large Hadron Collider*, *Rev. Phys.* **7** (2021) 100063 [arXiv:2105.07530] [INSPIRE].

[12] F. Jorge et al., *Top squark signal significance enhancement by different machine learning algorithms*, *Int. J. Mod. Phys. A* **37** (2022) 2250197 [arXiv:2106.06813] [INSPIRE].

[13] J. Ren et al., *Detecting an axion-like particle with machine learning at the LHC*, *JHEP* **11** (2021) 138 [arXiv:2106.07018] [INSPIRE].

[14] D. Alvestad et al., *Beyond cuts in small signal scenarios: enhanced sneutrino detectability using machine learning*, *Eur. Phys. J. C* **83** (2023) 379 [arXiv:2108.03125] [INSPIRE].

[15] S. Jung, Z. Liu, L.-T. Wang and K.-P. Xie, *Probing Higgs boson exotic decays at the LHC with machine learning*, *Phys. Rev. D* **105** (2022) 035008 [arXiv:2109.03294] [INSPIRE].

[16] M. Drees, M. Shi and Z. Zhang, *Machine learning optimized search for the Z' from $\mathrm{U}(1)_{L_\mu - L_\tau}$ at the LHC*, arXiv:2109.07674 [INSPIRE].

[17] A.S. Cornell et al., *Boosted decision trees in the era of new physics: a smuon analysis case study*, *JHEP* **04** (2022) 015 [arXiv:2109.11815] [INSPIRE].

[18] X.C. Vidal, L.D. Maroñas and Á.D. Suárez, *How to use machine learning to improve the discrimination between signal and background at particle colliders*, *Appl. Sciences* **11** (2021) 11076 [arXiv:2110.15099] [INSPIRE].

[19] J. Lin, M. Freytsis, I. Moult and B. Nachman, *Boosting $H \to b\bar{b}$ with machine learning*, *JHEP* **10** (2018) 101 [arXiv:1807.10768] [INSPIRE].

[20] E.A. Moreno et al., *Interaction networks for the identification of boosted $H \to b\bar{b}$ decays*, *Phys. Rev. D* **102** (2020) 012010 [arXiv:1909.12285] [INSPIRE].

[21] Y.-L. Chung, K. Cheung and S.-C. Hsu, *Sensitivity of two-Higgs-doublet models on Higgs-pair production via $b\bar{b}b\bar{b}$ final state*, *Phys. Rev. D* **106** (2022) 095015 [arXiv:2207.09602] [INSPIRE].

[22] J.H. Kim et al., *Portraying double Higgs at the Large Hadron Collider*, *JHEP* **09** (2019) 047 [arXiv:1904.08549] [INSPIRE].

[23] L. Huang et al., *Portraying double Higgs at the Large Hadron Collider II*, *JHEP* **08** (2022) 114 [arXiv:2203.11951] [INSPIRE].

[24] W. Esmail, A. Hammad and S. Moretti, *Sharpening the $A \to Z^{(*)}h$ signature of the type-II 2HDM at the LHC through advanced machine learning*, *JHEP* **11** (2023) 020 [arXiv:2305.13781] [INSPIRE].

JHEP03(2024)144

[25] K. Ban, K. Kong, M. Park and S.C. Park, *Exploring the synergy of kinematics and dynamics for collider physics*, arXiv:2311.16674 [INSPIRE].

[26] A. Chakraborty et al., *Revisiting jet clustering algorithms for new Higgs boson searches in hadronic final states*, *Eur. Phys. J. C* **82** (2022) 346 [arXiv:2008.02499] [INSPIRE].

[27] A. Chakraborty et al., *Re-evaluating jet reconstruction techniques for new Higgs boson searches*, *PoS* **ICHEP2022** (2022) 503 [arXiv:2212.02246] [INSPIRE].

[28] A. Chakraborty et al., *Fat b-jet analyses using old and new clustering algorithms in new Higgs boson searches at the LHC*, *Eur. Phys. J. C* **83** (2023) 347 [arXiv:2303.05189] [INSPIRE].

[29] G. Cerro et al., *Spectral clustering for jet reconstruction*, *PoS* **ICHEP2022** (2022) 771 [arXiv:2211.10164] [INSPIRE].

[30] A. Vaswani et al., *Attention is all you need*, in the proceedings of the 31st *international conference on neural information processing systems*, (2017) [arXiv:1706.03762] [INSPIRE].

[31] B. Käch, D. Krücker and I. Melzer-Pellmann, *Point cloud generation using transformer encoders and normalising flows*, arXiv:2211.13623 [INSPIRE].

[32] T. Finke, M. Krämer, A. Mück and J. Tönshoff, *Learning the language of QCD jets with transformers*, *JHEP* **06** (2023) 184 [arXiv:2303.07364] [INSPIRE].

[33] H. Qu, C. Li and S. Qian, *Particle transformer for jet tagging*, arXiv:2202.03772 [INSPIRE].

[34] P.T. Komiske, E.M. Metodiev and J. Thaler, *Energy flow networks: deep sets for particle jets*, *JHEP* **01** (2019) 121 [arXiv:1810.05165] [INSPIRE].

[35] H. Qu and L. Gouskos, *ParticleNet: jet tagging via particle clouds*, *Phys. Rev. D* **101** (2020) 056019 [arXiv:1902.08570] [INSPIRE].

[36] ATLAS collaboration, *Search for resonant pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state using pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Phys. Rev. D* **105** (2022) 092002 [arXiv:2202.07288] [INSPIRE].

[37] G.C. Branco et al., *Theory and phenomenology of two-Higgs-doublet models*, *Phys. Rept.* **516** (2012) 1 [arXiv:1106.0034] [INSPIRE].

[38] T.D. Lee, *A theory of spontaneous T violation*, *Phys. Rev. D* **8** (1973) 1226 [INSPIRE].

[39] S.L. Glashow and S. Weinberg, *Natural conservation laws for neutral currents*, *Phys. Rev. D* **15** (1977) 1958 [INSPIRE].

[40] I.F. Ginzburg and M. Krawczyk, *Symmetries of two Higgs doublet model and CP violation*, *Phys. Rev. D* **72** (2005) 115013 [hep-ph/0408011] [INSPIRE].

[41] S. Antusch, O. Fischer, A. Hammad and C. Scherb, *Testing CP properties of extra Higgs states at the HL-LHC*, *JHEP* **03** (2021) 200 [arXiv:2011.10388] [INSPIRE].

[42] S. Antusch, O. Fischer, A. Hammad and C. Scherb, *Explaining excesses in four-leptons at the LHC with a double peak from a CP violating two Higgs doublet model*, *JHEP* **08** (2022) 224 [arXiv:2112.00921] [INSPIRE].

[43] A. Arhrib et al., *Double neutral Higgs production in the two-Higgs doublet model at the LHC*, *JHEP* **08** (2009) 035 [arXiv:0906.0387] [INSPIRE].

[44] A. Hammad, M. Park, R. Ramos and P. Saha, *Exploration of parameter spaces assisted by machine learning*, *Comput. Phys. Commun.* **293** (2023) 108902 [arXiv:2207.09959] [INSPIRE].

JHEP03(2024)144

[45] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, *JHEP* **07** (2014) 079 [arXiv:1405.0301] [INSPIRE].

[46] W. Porod, *SPheno, a program for calculating supersymmetric spectra, SUSY particle decays and SUSY particle production at e^+e^- colliders*, *Comput. Phys. Commun.* **153** (2003) 275 [hep-ph/0301101] [INSPIRE].

[47] W. Porod and F. Staub, *SPheno* 3.1*: extensions including flavour, CP-phases and models beyond the MSSM*, *Comput. Phys. Commun.* **183** (2012) 2458 [arXiv:1104.1573] [INSPIRE].

[48] T. Sjostrand, S. Mrenna and P.Z. Skands, *PYTHIA* 6.4 *physics and manual*, *JHEP* **05** (2006) 026 [hep-ph/0603175] [INSPIRE].

[49] J. Alwall et al., *Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions*, *Eur. Phys. J. C* **53** (2008) 473 [arXiv:0706.2569] [INSPIRE].

[50] M.L. Mangano, M. Moretti, F. Piccinini and M. Treccani, *Matching matrix elements and shower evolution for top-quark production in hadronic collisions*, *JHEP* **01** (2007) 013 [hep-ph/0611129] [INSPIRE].

[51] DELPHES 3 collaboration, *DELPHES* 3*, a modular framework for fast simulation of a generic collider experiment*, *JHEP* **02** (2014) 057 [arXiv:1307.6346] [INSPIRE].

[52] M. Cacciari, G.P. Salam and G. Soyez, *The anti-k_t jet clustering algorithm*, *JHEP* **04** (2008) 063 [arXiv:0802.1189] [INSPIRE].

[53] S. Catani, Y.L. Dokshitzer, M.H. Seymour and B.R. Webber, *Longitudinally invariant K_t clustering algorithms for hadron hadron collisions*, *Nucl. Phys. B* **406** (1993) 187 [INSPIRE].

[54] D. Krohn, J. Thaler and L.-T. Wang, *Jet trimming*, *JHEP* **02** (2010) 084 [arXiv:0912.1342] [INSPIRE].

[55] ATLAS collaboration, *ATLAS flavour-tagging algorithms for the LHC run* 2 *pp collision dataset*, *Eur. Phys. J. C* **83** (2023) 681 [arXiv:2211.16345] [INSPIRE].

[56] J.M. Butterworth, A.R. Davison, M. Rubin and G.P. Salam, *Jet substructure as a new Higgs search channel at the LHC*, *Phys. Rev. Lett.* **100** (2008) 242001 [arXiv:0802.2470] [INSPIRE].

[57] D.E. Kaplan, K. Rehermann, M.D. Schwartz and B. Tweedie, *Top tagging: a method for identifying boosted hadronically decaying top quarks*, *Phys. Rev. Lett.* **101** (2008) 142001 [arXiv:0806.0848] [INSPIRE].

[58] Y. Cui, Z. Han and M.D. Schwartz, *W-jet tagging: optimizing the identification of boosted hadronically-decaying W bosons*, *Phys. Rev. D* **83** (2011) 074023 [arXiv:1012.2077] [INSPIRE].

[59] T. Plehn, M. Spannowsky and M. Takeuchi, *How to improve top tagging*, *Phys. Rev. D* **85** (2012) 034029 [arXiv:1111.5034] [INSPIRE].

[60] D.E. Soper and M. Spannowsky, *Finding top quarks with shower deconstruction*, *Phys. Rev. D* **87** (2013) 054012 [arXiv:1211.3140] [INSPIRE].

[61] C. Anders et al., *Benchmarking an even better top tagger algorithm*, *Phys. Rev. D* **89** (2014) 074047 [arXiv:1312.1504] [INSPIRE].

[62] G. Kasieczka et al., *Resonance searches with an updated top tagger*, *JHEP* **06** (2015) 203 [arXiv:1503.05921] [INSPIRE].

[63] J. Thaler and K. Van Tilburg, *Identifying boosted objects with N-subjettiness*, *JHEP* **03** (2011) 015 [arXiv:1011.2268] [INSPIRE].

JHEP03(2024)144

[64] J. Thaler and K. Van Tilburg, *Maximizing boosted top identification by minimizing N-subjettiness*, *JHEP* **02** (2012) 093 [arXiv:1108.2701] [INSPIRE].

[65] A.J. Larkoski, G.P. Salam and J. Thaler, *Energy correlation functions for jet substructure*, *JHEP* **06** (2013) 108 [arXiv:1305.0007] [INSPIRE].

[66] I. Moult, L. Necib and J. Thaler, *New angles on energy correlation functions*, *JHEP* **12** (2016) 153 [arXiv:1609.07483] [INSPIRE].

[67] A.J. Larkoski, S. Marzani, G. Soyez and J. Thaler, *Soft drop*, *JHEP* **05** (2014) 146 [arXiv:1402.2657] [INSPIRE].

[68] A. Abdesselam et al., *Boosted objects: a probe of beyond the Standard Model physics*, *Eur. Phys. J. C* **71** (2011) 1661 [arXiv:1012.5412] [INSPIRE].

[69] A. Altheimer et al., *Jet substructure at the Tevatron and LHC: new results, new tools, new benchmarks*, *J. Phys. G* **39** (2012) 063001 [arXiv:1201.0008] [INSPIRE].

[70] A. Altheimer et al., *Boosted objects and jet substructure at the LHC. Report of BOOST*2012, *held at IFIC Valencia,* 23^{rd}–27^{th} *of July* 2012, *Eur. Phys. J. C* **74** (2014) 2792 [arXiv:1311.2708] [INSPIRE].

[71] J. Cogan, M. Kagan, E. Strauss and A. Schwartzman, *Jet-images: computer vision inspired techniques for jet tagging*, *JHEP* **02** (2015) 118 [arXiv:1407.5675] [INSPIRE].

[72] L.G. Almeida et al., *Playing tag with ANN: boosted top identification with pattern recognition*, *JHEP* **07** (2015) 086 [arXiv:1501.05968] [INSPIRE].

[73] L. de Oliveira et al., *Jet-images — deep learning edition*, *JHEP* **07** (2016) 069 [arXiv:1511.05190] [INSPIRE].

[74] P. Baldi et al., *Jet substructure classification in high-energy physics with deep neural networks*, *Phys. Rev. D* **93** (2016) 094034 [arXiv:1603.09349] [INSPIRE].

[75] J. Barnard, E.N. Dawe, M.J. Dolan and N. Rajcic, *Parton shower uncertainties in jet substructure analyses with deep neural networks*, *Phys. Rev. D* **95** (2017) 014018 [arXiv:1609.00607] [INSPIRE].

[76] P.T. Komiske, E.M. Metodiev and M.D. Schwartz, *Deep learning in color: towards automated quark/gluon jet discrimination*, *JHEP* **01** (2017) 110 [arXiv:1612.01551] [INSPIRE].

[77] G. Kasieczka, T. Plehn, M. Russell and T. Schell, *Deep-learning top taggers or the end of QCD?*, *JHEP* **05** (2017) 006 [arXiv:1701.08784] [INSPIRE].

[78] S. Macaluso and D. Shih, *Pulling out all the tops with computer vision and deep learning*, *JHEP* **10** (2018) 121 [arXiv:1803.00107] [INSPIRE].

[79] S. Choi, S.J. Lee and M. Perelstein, *Infrared safety of a neural-net top tagging algorithm*, *JHEP* **02** (2019) 132 [arXiv:1806.01263] [INSPIRE].

[80] F. Mokhtar, R. Kansal and J. Duarte, *Do graph neural networks learn traditional jet substructure?*, in the proceedings of the 36^{th} *conference on neural information processing systems: workshop on machine learning and the physical sciences*, (2022) [arXiv:2211.09912] [INSPIRE].

[81] F. Ma, F. Liu and W. Li, *Jet tagging algorithm of graph network with Haar pooling message passing*, *Phys. Rev. D* **108** (2023) 072007 [arXiv:2210.13869] [INSPIRE].

[82] S. Gong et al., *An efficient Lorentz equivariant graph neural network for jet tagging*, *JHEP* **07** (2022) 030 [arXiv:2201.08187] [INSPIRE].

JHEP03(2024)144

[83] D. Guest et al., *Jet flavor classification in high-energy physics with deep neural networks*, *Phys. Rev. D* **94** (2016) 112002 [arXiv:1607.08633] [INSPIRE].

[84] J. Pearkes, W. Fedorko, A. Lister and C. Gay, *Jet constituents for deep neural network based top quark tagging*, arXiv:1704.02124 [INSPIRE].

[85] S. Egan et al., *Long Short-Term Memory (LSTM) networks with jet constituents for boosted top tagging at the LHC*, arXiv:1711.09059 [INSPIRE].

[86] K. Fraser and M.D. Schwartz, *Jet charge and machine learning*, *JHEP* **10** (2018) 093 [arXiv:1803.08066] [INSPIRE].

[87] A. Butter, G. Kasieczka, T. Plehn and M. Russell, *Deep-learned top tagging with a Lorentz layer*, *SciPost Phys.* **5** (2018) 028 [arXiv:1707.08966] [INSPIRE].

[88] G. Kasieczka, N. Kiefer, T. Plehn and J.M. Thompson, *Quark-gluon tagging: machine learning vs detector*, *SciPost Phys.* **6** (2019) 069 [arXiv:1812.09223] [INSPIRE].

[89] G. Cowan, K. Cranmer, E. Gross and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, *Eur. Phys. J. C* **71** (2011) 1554 [*Erratum ibid.* **73** (2013) 2501] [arXiv:1007.1727] [INSPIRE].

[90] LHC Dark Matter Working Group collaboration, *LHC dark matter working group: next-generation spin-0 dark matter models*, *Phys. Dark Univ.* **27** (2020) 100351 [arXiv:1810.09420] [INSPIRE].

[91] E. Arganda, A. Delgado, R.A. Morales and M. Quirós, *LHC search strategy for squarks in Higgsino-LSP scenarios with leptons and b-jets in the final state*, *Particles* **5** (2022) 265 [arXiv:2206.05977] [INSPIRE].

[92] H. Chefer, S. Gur and L. Wolf, *Transformer interpretability beyond attention visualization*, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (2021), p. 782 [arXiv:2012.09838].

[93] R.R. Selvaraju et al., *Grad-CAM: visual explanations from deep networks via gradient-based localization*, in *Proceedings of the IEEE international conference on computer vision*, (2017), p. 618 [arXiv:1610.02391].

[94] Y. Huang et al., *SSiT: saliency-guided self-supervised image transformer for diabetic retinopathy grading*, arXiv:2210.10969.

[95] N. Duong-Trung, D.-M. Nguyen and D. Le-Phuoc, *Temporal saliency detection towards explainable transformer-based timeseries forecasting*, arXiv:2212.07771.

[96] C. Lu, H. Zhu and P. Koniusz, *From saliency to DINO: saliency-guided vision transformer for few-shot keypoint detection*, arXiv:2304.03140.

[97] A. Binder et al., *Layer-wise relevance propagation for neural networks with local renormalization layers*, in *Artificial neural networks and machine learning — ICANN* 2016*:* 25th *international conference on artificial neural networks, Barcelona, Spain,* 6–9 *September* 2016*, Proceedings, part II* 25, Springer, (2016), p. 63 [arXiv:1604.00825].

[98] I. Cherepanov, A. Ulmer, J.G. Joewono and J. Kohlhammer, *Visualization of class activation maps to explain AI classification of network packet captures*, in 2022 *IEEE symposium on visualization for cyber security (VizSec)*, IEEE (2022), p. 1 [DOI:10.1109/VizSec56996.2022.9941392] [arXiv:2209.02045]

[99] B. Zhou et al., *Learning deep features for discriminative localization*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), p. 2921 [arXiv:1512.04150].

JHEP03(2024)144

[100] D.P. Kingma and J. Ba, *Adam: a method for stochastic optimization*, arXiv:1412.6980 [INSPIRE].

[101] J. Terven, D.M. Cordova-Esparza, A. Ramirez-Pedraza and E.A. Chavez-Urbiola, *Loss functions and metrics in deep learning*, arXiv:2307.02694.

[102] Z. Lai et al., *Rethinking skip connections in encoder-decoder networks for monocular depth estimation*, arXiv:2208.13441.

[103] D. Hendrycks and K. Gimpel, *Gaussian Error Linear Units (GELUs)*, arXiv:1606.08415 [INSPIRE].

JHEP03(2024)144