Enhancing Precision in Dermoscopic Imaging using TransUNet and CASCADE

Mahdi Niknejad
Department of Computer Engineering
Amirkabir University of Technology
Tehran, Iran
emahdi.niknejad@gmail.com

Mahdi Firouzbakht
Department of Computer Engineering
Amirkabir University of Technology
Tehran, Iran
mahdifiruzbakht23@gmail.com

Maryam Amirmazlaghani
Department of Computer Engineering
Amirkabir University of Technology
Tehran, Iran
mazlaghani@aut.ac.ir

***Abstract*—Medical imaging is advancing at a rapid pace, revolutionizing medicine. Automated skin lesion segmentation is vital for early skin cancer diagnosis, yet segmenting lesions in dermoscopic images presents considerable challenges. Despite recent good performance, CNN-based algorithms are unable to effectively learn explicit global and long-range semantic information because of the fundamental locality of convolution operations. The first medical image segmentation framework, TransUNet, was presented employing Vision Transformer as a strong encoder in a U-shaped architecture, in light of the growing interest in self-attention mechanisms in computer vision and their potential to address this issue. Using hierarchical vision transformers' multi-scale features, CASCADE is a novel attention-based decoder. The components of CASCADE are a convolutional attention module that improves the local and long-range context by suppressing background information and an attention gate that combines features. Using TransUNet's Encoder and CASCADE as the Decoder, the TransCASCADE model efficiently makes use of the global context stored by Transformers and detailed, high-resolution spatial information from CNN features. In this study, we utilized a novel optimizer called Lion to train our model, which improved memory efficiency, reduced complexity, and required fewer hyperparameters. Employing Lion resulted in superior performance compared to the base models. Experimental results demonstrate the high efficiency of the proposed method on PH2 dataset.**

***Keywords*—Deep Learning, Dermoscopic Imaging, Medical Image Segmentation, UNet, Vision Transformer**

I. Introduction

One of the most significant stages of pre-diagnostic, in-treatment, and post-treatment evaluations for a variety of diseases is the segmentation of medical images. It can be viewed as a prediction problem that generates segmentation maps of lesions. The development and growing application of medical imaging techniques (MRI, PET, CT scan, X-ray, endoscopy, and many more) have made it important to have tools for the automatic extraction of this data. Deep learning techniques have become more practical for these jobs nowadays thanks to advancements in hardware, and deep learning forms the basis of the most widely used approaches [1].

Many medical image segmentation research projects have extensively used convolutional neural networks (CNNs). Due to its ability to generate high-resolution segmentation maps, U-Net has shown extreme performance in the segmentation of medical images [2]. Various alternative architectures, including U-Net++, U-Net+, and 3D U-Net, have demonstrated outstanding performance in medical image segmentation, due to the efficient encoder-decoder architecture of U-Net [1].

Despite the powerful representation capabilities and reasonable performance of convolutional neural network-based methods, these architectures have limitations in learning long-range dependencies between image pixels. While convolutional neural networks have demonstrated excellent performance, learning meaningful context across large distances is hindered by the intrinsic locality of convolutional processes. This implies that these networks won't perform well if images contain structural information with notable variations in texture and shape. Some architectures use attention mechanisms in the architecture to improve feature maps for more accurate medical image segmentation to overcome this limitation. Extracting Long-range dependencies is still challenging for attention-based methods, despite their increased performance.

The latest advances in vision transformers have tackled limitations related to long-range dependencies, especially in medical image segmentation. Transformers rely on attention mechanisms, initially introduced for sequence-to-sequence prediction in natural language processing. Transformers can learn long-range dependencies by using self-attention to identify correlations between all of the input tokens. Vision transformers split an image into non-overlapping patches and feed them into the transformer unit together with positional embeddings, deriving inspiration from the success of transformers in natural language processing. The TransUNet architecture, which increases the extraction of global semantic information and spatial features, is an example of how vision transformers are applied. This architecture includes a cascaded decoder to capture local pixel-wise relationships and a transformer encoder to extract long-range dependencies [4]. In light of these challenges, we enhance the base model by utilizing a novel attention-based decoder known as "CASCADE" [1]. A hierarchical representation derived from vision transformers is used by CASCADE. Using attention gates and attention pooling modules, this decoder learns the semantic and spatial relationships between pixels to improve feature maps.

Skin cancer is one of the most general cancer types in over the world and Skin lesion segmentation has a critical

979-8-3503-7634-0/24/$31.00

role in the early diagnosis of skin cancer by computer-aided systems. However, automatic segmentation of skin lesions in dermoscopic images is a challenging task due to difficulties including artifacts (hairs, gel bubbles, ruler markers), indistinct boundaries, low contrast and varying sizes and shapes of the lesion images [5].

In this paper, we leverage the CASCADE decoder architecture to improve upon the TransUNet model, which is used as our base model. We introduce a new loss function as one of the modifications we make to the model. Better output was obtained by optimizing the model and adding the Cascade decoder.

This paper is organized as follows: Section II provides an overview of some related works on skin lesion segmentation and in general medical image segmentation based on CNNs and Vision Transformers. Section III explains the dataset used. The proposed method is described in Section IV. Section V explains the experimental results. Finally, Section VI concludes the paper.

II. Related Work

In this section, we intend to briefly discuss related works conducted in this field. We first explain traditional and deep learning-based methods of segmenting skin lesions in images. Next, we provide an overview of the most widely used convolutional neural network-based methods for segmenting medical images. In conclusion, we examine the latest use of vision transformers in the domain of image segmentation.

A. Skin Lesion Segmentation

Before the advent of deep learning, thresholding, and active contour models—two of the most well-established techniques—were frequently the foundation of skin lesion segmentation approaches. Even if the era of deep learning has changed, traditional feature extraction techniques have gradually given way to deep neural networks. This shift is characterized by a growing preference for end-to-end methods to effectively address the complexities associated with skin image segmentation [6].

B. Medical Image Segmentation based on CNNs

In medical image segmentation, convolutional neural networks—particularly the U-Net architecture with its encoder-decoder structure and several versions—have shown outstanding performance. Owing to the U-shaped structure's ease of use and efficient operation, new kinds of U-Net-like methods are always being developed in the area. Before concatenation, for instance, **U-Net++** [7] adds a set of densely interconnected skip connections to fill in semantic gaps between the encoder and decoder feature maps. **Attention U-Net** [8] offers attention gates, a revolutionary innovation that allows the model to focus on targets of different sizes and forms. Last but not least, **U2-net** [9] uses a two-level layered structure with the U-Net structure applied at each level, utilizing Residual U-blocks (RSU). By combining receptive fields of various sizes, this architecture can collect more meaningful information.

C. Vision Transformer

Inspired by the success of Transformer in various Natural Language Processing tasks, more and more Transformer-based methods appear in Computer Vision tasks. Among the recent vision transformers, **ViT** [10] is the first attempt that proves pure Transformer-based architecture can achieve SOTA performance on image recognition when pre-training on large datasets such as ImageNet-22K and JFT-300M. Transformer-specific teacher-student strategy is introduced by **DeiT** [11]. It includes the process of knowledge transfer and is based on a distillation identifier. Through the attention mechanism, the student learns from the teacher model. This study shows that the DeiT architecture requires significantly less data and computational resources to perform as well as the ViT, even when it comes to image classification. It has demonstrated strong performance after being successfully trained on a smaller dataset, such as ImageNet-1k.

ViT's static and non-multiscale feature maps are one of its challenges, as they result in a notable loss of spatial information. The **PVT** [12] architecture was designed to address this issue by including a multiscale mode into the transformer architecture for the first time. The main difference between this type of architecture and convolutional neural networks is the incorporation of global attention throughout the entire process. The addition of a new attention mechanism called spatial reduction attention, which lowers computational cost based on a specified reduction factor, is another important advantage that PVT has over ViT.

III. Dataset

In this section, our objective is to provide a brief overview of the PH2 dataset employed in the scope of skin lesion segmentation. The increasing incidence of melanoma has recently promoted the development of computer-aided diagnosis systems for the classification of dermoscopic images. The PH2 dataset was built up through a joint research collaboration between the Universidade do Porto, T´ecnico Lisboa, and the Dermatology Service of Hospital Pedro Hispano in Matosinhos, Portugal. This dataset contains a total of 200 dermoscopic images with a resolution of 768×560 pixels, containing 80 common nevi, 80 atypical nevi, and 40 melanomas [14]. In Figure 1, we observe some samples of the dataset.

The dataset is entirely labeled with masks and for training the models, we have divided all of them into an 80-20 ratio. The masks are binary images indicating which pixels in the original image belong to the skin lesion and which do not.

IV. Proposed Method

In this section, we will explain our proposed model. This model will be introduced as TransCascade-Li. As depicted in Figure 3, TransUNet is used as our model's encoder, and the CASCADE architecture employs upconv blocks for feature upsampling, attention gates for cascaded feature integration, and CAM blocks for enhancing feature maps. It initially uses attention gates to mix the upsampled features from the preceding decoder block with skip-connection features

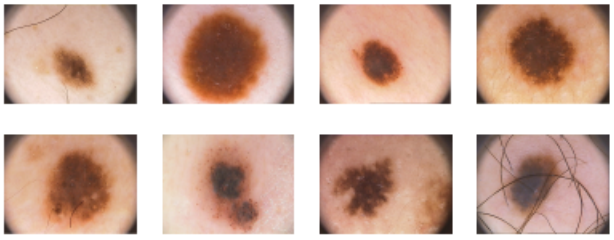

Fig. 1. An illustrative collection of images from PH2 dataset [14]

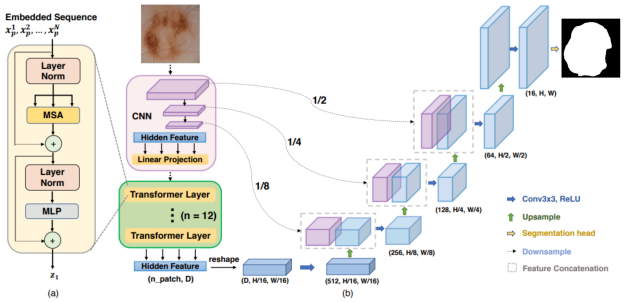

Fig. 2. The artichetecture of TransUNet [2]

for multi-scale feature aggregation. It then concatenates the upsampled features from the previous layer with the combined features. After that, it groups pixels with related characteristics in various image areas and reduces the influence of features that aren't significant by processing the concatenated features using CAM blocks. The final segmentation map is created by combining the four anticipated feature maps from the various scales that are obtained after feeding the output of each CAM block into a prediction head. In the following, We initially introduce the TransUNet and then, we describe the CASCADE.

A. TransUNet

The TransUNet model, a reliable substitute for medical image segmentation includes a transformer and a U-shaped convolutional network as part of its architecture. It can serve as an effective alternative for medical image segmentation. To extract global semantic information, the transformer first encodes the image patches that were taken from the feature map as a sequence input. However, to facilitate accurate localization, the decoder upsamples the encoded features.

Now that we are acquainted with the basic model in brief, let's see how it is structured: Assume that we have an image $x \in \mathbb{R}^{H\times W\times C}$ with dimensions $H \times W$ and C channels. Our objective is to forecast a mask image that has $H \times W$ dimensions. The most common approach involves directly training a convolutional neural network, like a U-Net, where the high-level features of the images are first fully represented by an encoder and then fully reconstructed using a decoder. But in contrast to current methods, we use the self-attention mechanism present in transformers in the encoder's design [2].

The architecture of the TransUNet can be seen in Figure 2. As evident, after obtaining the encodings from the transformer, $z_l \in \mathbb{R}^{\frac{HW}{P^2}\times D}$, we need to upsample them to obtain the mask. Here, $Y \in \mathbb{R}^{H\times W\times K}$ represents the number of classes. To restore the spatial order, the resized vector of the encoded feature, initially from $\mathbb{R}^{\frac{HW}{P^2}\times D}$, must be reshaped to $\mathbb{R}^{\frac{H}{P}\times\frac{H}{P}\times D}$.

B. CASCADE

The capacity of current transformer-based models to learn local information is restricted. Previous methods, which were unsuccessful in evaluations, tried to get around this restriction by embedding convolutional layers—which are effective at extracting spatial information—into the encoder or decoder units of the transformer. Convolution layers have recently been added to transformers to overcome this restriction in models like SegFormer, UFormer, and PVTv2. These structures still have difficulties even if they can learn some local (spatial) correlations between pixels. To address these issues, a novel decoder called Cascaded Attention Decoder (CASCADE) makes use of hierarchical representations of visual transformers. CASCADE integrates features with attention gates (AGs) and convolution attention modules (CAMs) to improve performance through skip connections. CASCADE captures both global and local (spatial) correlations between pixels because it uses attention-based convolution modules to aggregate multi-stage features and hierarchical transformers as the primary network.

Studies show that models incorporating CASCADE decoders significantly outperform models that are transformer-based, convolutional neural network-based, or a mix of these. When used with different hierarchical backbone networks, the suggested decoder is adaptable and easy to use. This architecture uses both a pyramid transformer and a hybrid CNN-transformer encoder (instead of only CNN) to provide adequate generalization and processing capacity for multi-scale feature analysis in medical image segmentation. PVTv2 continuously encodes spatial information using convolution operations as opposed to the traditional transformer patch embedding module. TransUNet concurrently captures the global and spatial relationships between features by layering a transformer on top of CNN. The second architecture, known as TransCASCADE, is used in this study.

The CAM block in CASCADE uses attention modules to enhance feature maps. Convolutional blocks, channel attention, and spatial attention are some of these modules. Channel attention selects what features to emphasize, while spatial attention informs where to focus within a feature map. Enhancing the features produced by channel and spatial attention is the final stage in the CAM process [1].

Section e of Figure 3 is an illustration of the channel attention. Both average pooling and max pooling procedures are used on the input feature map's spatial dimension to compute this attention. The output of each operation is then fed independently into a 1×1 convolutional layer that has

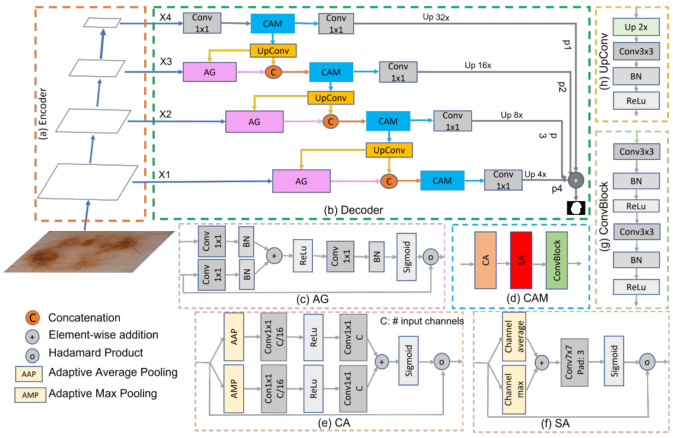

Fig. 3. The artichetecture of TransCASCADE [1]

the same number of channels. The outputs are fed into a 1×1 convolutional layer with $\frac{C}{16}$ channels after via a ReLU activation function. After element-wise summing, the outputs of the feature vectors are fused, and the resultant vector is then subjected to a sigmoid function.

Channel focus is followed by spatial attention, as seen in Figure 3, section f. Thus, the channel attention's output becomes the input for the spatial attention, which applies the max and average pooling operations on the input feature map's channel dimension. These two vectors are then concatenated, and the outcome is sent via an extra layer normalization and a 7×7 convolutional layer. The resultant vector is then subjected to a sigmoid function.

V. Experiments

In this section, we introduce our conducted experiments. First, we present the performance with the original hyperparameters. In addition, we investigated the performance using different loss functions, learning rates, and optimizers. We implemented the proposed model using PyTorch and all experiments were performed on Google Colab platform using a Tesla T4 GPU with 12 GB memory.

A. Parameters

We use a batch size of 16 and train each model maximum of 150 epochs. We use the input resolution and patch size P as 224×224 and 16, respectively. Additionally, we apply a learning rate and weight decay of 1e-4 to optimize the training process.

B. Evaluation Metrics

To compare the proposed new model with the baseline model, we utilized evaluation metrics, including the Dice and IoU indices, which are associated with four values: false negatives (FN), false positives (FP), true negatives (TN), and true positives (TP).

$$Dice = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{1}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{2}$$

C. Loss Function

In our proposed method loss function is defined as a weighted linear combination of multiple loss functions like Cross-Entropy, Dice, and IoU. The final performance heavily depends on selecting the correct (relative) weights for these loss functions.

Although the assignment of weights to each loss function depends on the specific application of the model, and using uniform weights is not always ideal, we experimented with both different and uniform coefficients for training our model. However, in both cases, we observed no improvement in performance. Therefore, we decided to use equal coefficients as the baseline.

$$loss = \alpha L_{dice} + \beta L_{ce} + \gamma L_{iou} \tag{3}$$

As we can see in Table I, the idea of adding the IoU loss function does not lead to improvement, and in this step, the base model without it performs better. Therefore, the IoU loss function will not be used further so we set the value of γ equal to zero.

TABLE I
Results of using different loss functions

		Base Model (TransUNet)		Proposed Model (TransCascade-Li)	
dataset	metric	w/o IoU	w/ IoU	w/o IoU	w/ IoU
PH2	DSC[a]	0.94	0.94	0.96	0.95
	IOU[b]	0.90	0.90	0.92	0.92

[a]Dice metric
[b]IoU metric

D. Lion Optimizer

Although academics still employ traditional optimizers like AdamW and SGD, some methods have been put out to automatically find more effective optimization techniques. An evolutionary machine learning method was used to find the "Lion" optimization algorithm, which is used to train neural networks, as reported by a research team from Google and the University of California in a recent study [16].

Using three to ten times less learning rate than Adam, the Lion optimizer is more memory-efficient. Its algorithm is also less complex and has fewer hyperparameters. Using Lion, researchers trained some models, including a vision transformer. We tried to integrate the recently released Lion Optimizer into our new model using these descriptions. The findings of the experiment are reported. Table II shows that using Lion in place of the optimizer improved both the original and the new models. Consequently, we will employ Lion as the main optimizer in the new model.

TABLE II
RESULTS OF USING DIFFERENT OPTIMIZERS

		Base Model (TransUNet)		Proposed Model (TransCascade-Li)	
dataset	metric	w/ SGD	w/ Lion	w/ SGD	w/ Lion
PH2	DSC	0.94	0.96	0.96	0.97
	IOU	0.90	0.92	0.92	0.94

It can be said that in the PH2 dataset, our model achieves an average of 0.97 in the Dice metric and 0.94 in the IoU metric, which are 3.2% and 4.4% higher than the base model, respectively. The qualitative result compared to the ground truth can be observed in Figure 4

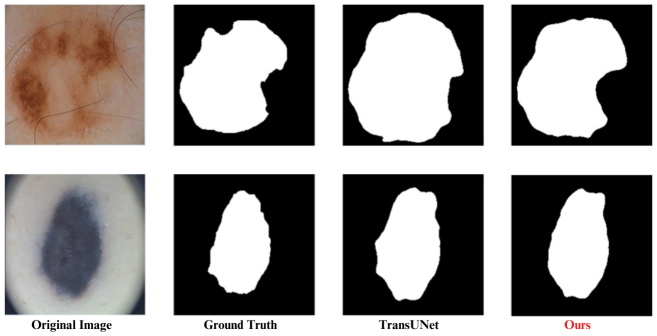

Fig. 4. The segmentation predictions of our method on PH2 dataset

E. Implement Final System

As a final system, we decided to create a web application using the Gradio library in Python. Gradio is an open-source Python library that facilitates the rapid development of user interfaces for machine learning models, providing a simple and attractive interface accessible through any browser. An advantage of Gradio is its ability to interact with models in web programs developed in Jupyter Notebooks or Colab. As seen in Figure 5, we observe the system output.

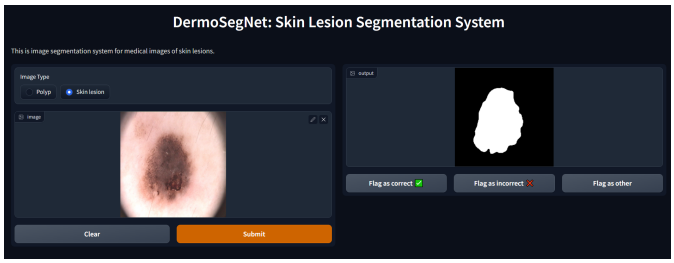

Fig. 5. The figure of final system

VI. CONCLUSION

In this paper, we introduced a novel structure for skin lesions and we demonstrated that using the proposed structure has a great effect on improving deep learning performance in skin lesion segmentation tasks. Additionally, experiments demonstrate that CASCADE effectively enhances transformer features.

REFERENCES

[1] Rahman, M.M. and Marculescu, R., "Medical image segmentation via cascaded attention decoding," In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 6222-6231).

[2] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L. and Zhou, Y., "Transunet: Transformers make strong encoders for medical image segmentation," arXiv preprint arXiv:2102.04306.

[3] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. and Wang, M., "Swin-unet: Unet-like pure transformer for medical image segmentation," In European conference on computer vision (pp. 205-218). Cham: Springer Nature Switzerland.

[4] Pan, S., Liu, X., Xie, N. and Chong, Y., "EG-TransUNet: a transformer-based U-Net with enhanced and guided models for biomedical image segmentation," BMC bioinformatics, 24(1), p.85.

[5] Ünver, H.M. and Ayan, E., "Skin lesion segmentation in dermoscopic images with combination of YOLO and grabcut algorithm," Diagnostics, 9(3), p.72.

[6] Chen, J., Chen, J., Zhou, Z., Li, B., Yuille, A. and Lu, Y., "MT-TransUNet: Mediating multi-task tokens in transformers for skin lesion segmentation and classification," arXiv preprint arXiv:2112.01767.

[7] Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N. and Liang, J., "Unet++: A nested u-net architecture for medical image segmentation," In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4 (pp. 3-11). Springer International Publishing.

[8] Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B. and Glocker, B., "Attention u-net: Learning where to look for the pancreas," arXiv preprint arXiv:1804.03999.

[9] Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R. and Jagersand, M., "U2-Net: Going deeper with nested U-structure for salient object detection," Pattern recognition, 106, p.107404.

[10] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. and Uszkoreit, J., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929.

[11] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A. and Jégou, H., "Training data-efficient image transformers & distillation through attention," In International conference on machine learning (pp. 10347-10357). PMLR.

[12] Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P. and Shao, L., "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," In Proceedings of the IEEE/CVF international conference on computer vision (pp. 568-578).

[13] Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P. and Shao, L., "Pvt v2: Improved baselines with pyramid vision transformer," Computational Visual Media, 8(3), pp.415-424.

[14] Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R. and Rozeira, J., "PH 2-A dermoscopic image database for research and benchmarking," In 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC) (pp. 5437-5440). IEEE.

[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., "Attention is all you need," Advances in neural information processing systems, 30.

[16] Chen, X., Liang, C., Huang, D., Real, E., Wang, K., Liu, Y., Pham, H., Dong, X., Luong, T., Hsieh, C.J. and Lu, Y., "Symbolic discovery of optimization algorithms," arXiv preprint arXiv:2302.06675.