Core-Structures-Guided Multi-Modal Classification Neural Architecture Search

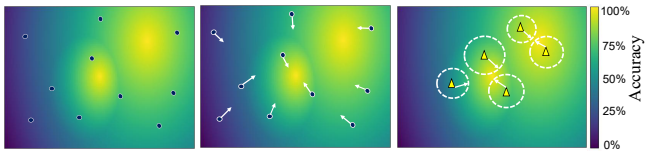
Pinhan Fu[1†], **Xinyan Liang**[1†], **Tingjin Luo**[2], **Qian Guo**[3], **Yayu Zhang**[1], **Yuhua Qian**[1*]
[1] Institute of Big Data Science and Industry, Shanxi University, Taiyuan 030006, China
[2] College of Science, National University of Defense Technology, Changsha, China 410073, China
[3] School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China
{fupinhan168, liangxinyan48, czguoqian}@163.com, {jinchengqyh, zhang_yayu93}@126.com , tingjinluo@hotmail.com

Abstract

The multi-modal classification methods based on neural architecture search (MMC-NAS) can automatically learn a satisfied classifier from a given multi-modal search space. However, as the number of multi-modal features and fusion operators increases, the complexity of search space has increased dramatically. Rapidly identifying the satisfied fusion model from this vast space is very challenging. In this paper, we propose an efficient MMC-NAS method based on an idea of shrink-and-expansion search space, called core-structures-guided neural architecture search (CSG-NAS). Specifically, an evolutionary algorithm is first used to find core structures from a shrunk space named as core structure search space determined by high-quality features and fusion operators. Then a local search algorithm is adopted to find the optimal MMC model from the expanded space determined by the discovered core structures and the rest features as well as fusion operators. Moreover, a knowledge inheritance strategy is introduced to further improve the overall performance and efficiency of the entire search process. Finally, extensive experimental results demonstrate the effectiveness of our CSG-NAS, attaining the superiority of classification performance, training efficiency and model complexity, compared to state-of-the-art competitors on several benchmark multi-modal tasks. The source code is available at https://github.com/fupinhan123/CSG-NAS.

1 Introduction

A satisfied multi-modal classification (MMC) model is located in the space determined by multiple features and fusion operators [Liang *et al.*, 2021; Liang *et al.*, 2024; Yin *et al.*, 2022]. According to the professional knowledge, experts can directly provide a solution such as [Xu *et al.*, 2024; Han *et al.*, 2023; Liang *et al.*, 2022a; Guo *et al.*, 2022]. Recently, some exciting results indicate that search-based techniques are able to obtain better solution than human designers. The random search strategy shown in Figure 1(a) may be feasible for querying all fusion architectures in a small search space. However, when the search space is larger, the strategy becomes computationally impractical.



(a) Random search (b) Randomly structures-guided search (c) Core-structures-guided search

Figure 1: Illustration of a 2D embedding of a high-dimensional search space, where black dots represent some *randomly* sampled multi-modal fusion structures from the entire fusion space, while yellow triangles depict ones with the high accuracy obtained from a pre-identified high-quality fusion space called core structures; white arrows indicate the search direction and the white dashed circles represent high-quality subspaces.

Later, researchers turned their attention to neural architecture search (NAS), proposing various MMC based on NAS methods (MMC-NAS) such as EDF [Liang *et al.*, 2021], BM-NAS [Yin *et al.*, 2022] and DC-NAS [Liang *et al.*, 2024]. As illustrated in Figure 1(b), these methods first randomly initialize some structures, and then use the directed search strategies such as evolutionary algorithm, reinforcement learning or gradient-based learning for the best multi-modal fusion architecture search. The existing MMC-NAS methods have achieved a significant success. However, the ever-expanding number of multi-modal features and fusion operations has resulted in an increasingly vast search space. Consequently, the training and evaluation processes become time-consuming, and the entire search process continues to face substantial challenges in terms of optimization and efficiency.

To improve search performance and mitigate challenges posed by large-scale search spaces, as illustrated in Figure 1(c), we propose a core-structures-guided multi-modal classification architecture search method (CSG-NAS). In comparison to existing methods, CSG-NAS rapidly narrows down the entire search space to a high-quality subspace and searches for the optimal fusion structure within this subspace. This avoids evaluating numerous underperforming structures, directly improving the performance of MMC-NAS. The motivation behind this method is that optimal multi-modal fusion

*Corresponding Author

architectures often consist of structures composed of superior features and fusion operators, which we refer to as core structures. By basing our search on these core structures, we can quickly locate regions with optimal multi-modal fusion architectures, thereby reducing the entire search space to a subspace with advanced performance. Subsequently, we conduct local search within these core structures to find the optimal multi-modal fusion architectures. To further improve search efficiency, we introduce a knowledge inheritance strategy on top of the evolutionary algorithm. In the crossbreeding and mutation of offspring, many methods typically only consider the encoding perspective, neglecting the weight parameters learned by the parents. We leverage these weight parameters to further improve the efficiency of the evolutionary process.

Our research has been validated on multiple multi-modal datasets, showcasing optimal performance in terms of accuracy and efficiency. Specifically, our contributions include:

- The optimal MMC architecture is definitely located in the subspace that is determined by high-quality features and fusion operators. Based on the finding, we give the definition of the core structures for the first time and propose a core-structures-guided multi-modal classification architecture search method, which can rapidly identify the optimal fusion architecture in a vast search space.
- The similar MMC architectures exhibit similar performance. With this idea, we design an innovative adaptive knowledge inheritance strategy for the evolutionary algorithm, which facilitates the sharing and reutilization of knowledge within the population, enhancing the learning capabilities.
- Extensive experimental comparisons across multiple multi-modal tasks demonstrate that CSG-NAS achieves competitive performance in terms of reduced search time and model parameter count compared to state-of-the-art multi-modal feature fusion methods.

2 Related Work

Multi-Modal Fusion: It fuses the relevant information from different modalities, achieving better performance [Jiang *et al.*, 2024; Liang *et al.*, 2022b].In context of deep neural networks, multi-modal fusion techniques are generally classified into three types: early fusion, late fusion, and intermediate layer fusion. Early fusion combines low-level features, late fusion combines high-level features such as predictions from the output layer of the network. Additionally, research on intermediate feature fusion indicates its benefits for learning gains and can enhance later fusion to improve performance. Therefore, some studies propose fusion at multiple intermediate layers. For example, CentralNet [Vielzeuf *et al.*, 2019] and MMTM [Vaezi Joze *et al.*, 2020] connect latent representations at each layer and pass them as auxiliary information into deeper layers. However, such approaches significantly increase the parameters of multi-modal fusion models.

Neural Architecture Search: NAS [Liu *et al.*, 2019] has been introduced to automate the design of neural models, aiming to discover efficient architectures with competitive performance. The initial approach [Zoph *et al.*, 2018] iteratively generated and trained candidate architectures with a reinforcement learning controller, but incurred high computational costs. Subsequent research adopted methods such as genetic algorithms [Real *et al.*, 2019], Bayesian optimization [Mendoza *et al.*, 2016], and predictors [Wei *et al.*, 2023].The introduction of supernetworks was a significant milestone; for example, SMASH [Brock *et al.*, 2018] utilized a one-shot network, training multiple candidate architectures simultaneously through shared weights, significantly reducing training time. To address limitations in exploratory performance, some methods like OFA [Cai *et al.*, 2020] employed novel path sampling and optimization techniques. Overall, one-shot networks have proven to be efficient, achieving state-of-the-art NAS performance.

NAS-based Multi-Modal Fusion: The integration of NAS into multi-modal learning has attracted significant interest. Evolutionary algorithms like EDF [Liang *et al.*, 2021] and DC-NAS [Liang *et al.*, 2024] maintained a set of architectures by generating new multi-modal fusion architectures through genetic operations such as mutation and crossover. Sequential model-based optimization algorithms are used in MFAS to search for given single-modal to multi-modal fusion architectures. Differentiable methods such as MMIF [Peng *et al.*, 2020], 3D-CDC-NAS2 [Yu *et al.*, 2021], and BM-NAS [Yin *et al.*, 2022] optimize shared weights and architectural parameters to significantly reduce computational resource requirements and improve its search efficiency. However, they still require a considerable amount of time when dealing with complex and extensive multi-modal fusion search spaces.

3 Methods

In this paper, we propose a core structures-guided neural architecture search (CSG-NAS) for finding the optimal multi-modal fusion architecture. CSG-NAS consists of two steps: (1) core structure search (CSS) and (2) core structures-guided optimal fusion architecture search. The main framework of CSG-NAS is shown in Figure 2.

To avoid confusion, we provide precise definitions for certain terms here. A population consists of individuals, where each individual corresponds to a MMC model encoded in the form of a tree. All representations extracted from different modalities are collectively referred to as features. The space composed of high-quality features and fusion operators is termed the core structure search space, while the remaining constitutes the non-core structure search space.

3.1 Retionality

Let S be the entire search space of multi-modal fusion, one partitions it into the architecture set with better performance S^1 and the architecture set with worse performance S^2 via one partition strategy. Our goal is to focus on searching the small portion of space (called the shrunk space in this paper) where S^1 resides and eliminate the space where S^2 is located. This can be achieved using a space shrunk method that can remove architectures with relatively poorer performance from the search space. There are two obvious advantages when one search algorithm works on the shrunk search space than S that 1) the shrunk search space with smaller cardinality leads to an improved training efficiency; 2) the higher

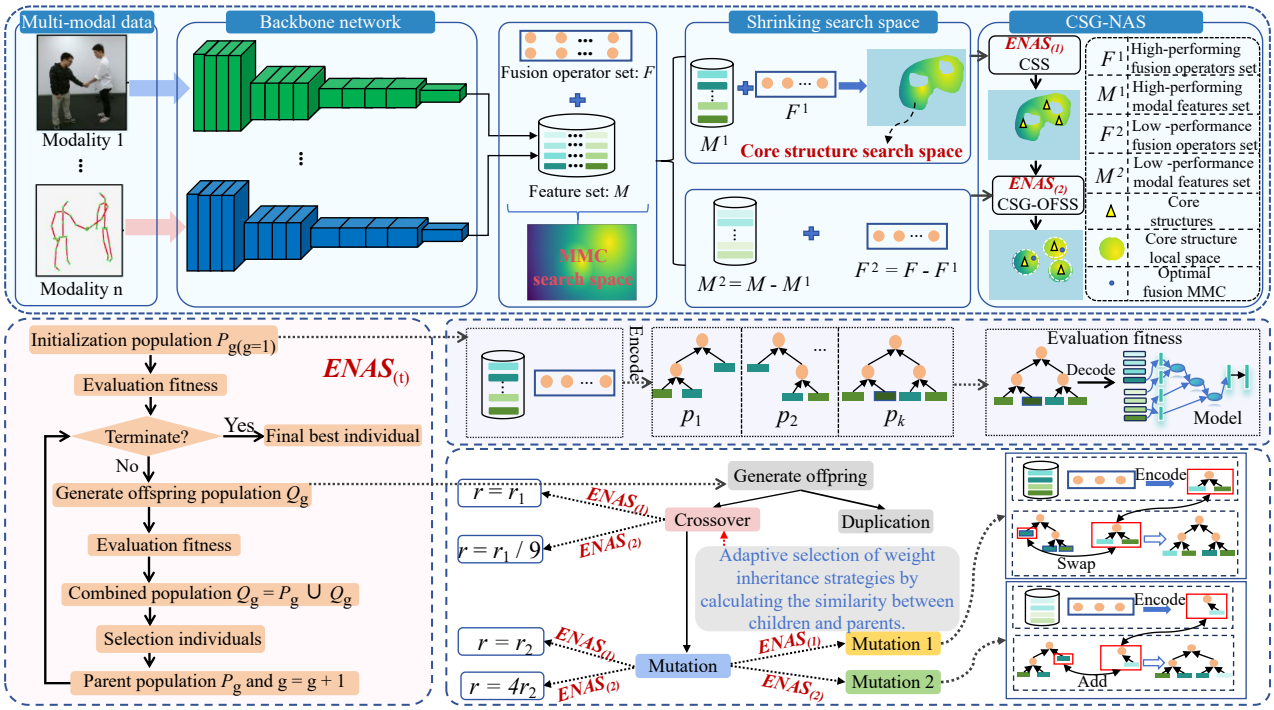


Figure 2: The framework of CSG-NAS, where M^1 and F^1 denote the high-quality feature set and fusion operator set, respectively.

proportion of optimal fusion architectures allows the search algorithm to concentrate training and exploration in more favorable regions of the search space. This leads to a better performance ranking. Therefore, our objective is to use the algorithm to locate the space where S^1 resides and allocate computational resources to enhance the efficiency of the algorithmic search, while disregarding the space where S^2 is situated. In this paper, we achieve this objective by proposing the Core Structures Search algorithm (CSS) detailed in Section 3.2.

3.2 CSS: Core Structure Search

By examining previously advanced multi-modal NAS methods such as MFAS [Perez Rua *et al.*, 2019], EDF [Liang *et al.*, 2021], BM-NAS [Yin *et al.*, 2022] and DC-NAS [Liang *et al.*, 2024], it is observed that the optimal multi-modal fusion architectures often contain certain structures that consist of the features and fusion operators with good performance, which we refer to as *core structures*. Hence, it can be assumed that the core structures are located in the subspace that is determined by the high-quality features and fusion operators. The subspace is referred to as *core structure search space*.

Based on the assumption, we can narrow down the entire search space to the core structure search space by evaluating the performance of each feature and fusion operator at a relatively low cost. Specifically, given n features denoted as M_1, M_2, ..., M_n and a single-modal classifier f. And then we obtain each feature M_i classification performance by passing it f, and select the first k_1 higher-performance features to the high-quality feature set M^1. Given m fusion operators denoted as F_1, F_2, ..., F_m and a multi-modal classifier h. And then, we obtain the classification performance of each fusion operator F_i by replacing the fusion way of h with F_i, and select the first k_2 higher-performance fusion operators to high-quality fusion operator set F^1.

The space composed of M^1 and F^1 is termed the core structure search space. Next, in the core structure search space we employ the proposed enhanced evolutionary algorithm ENAS_1 for the search. The final result after ENAS_1 iterations is referred to as the core structures. The detailed process of the evolutionary algorithm can be found in Section 3.4. The remaining features and fusion operators are put into two sets M^2 and F^2, i.e., $M^2 = M - M^1$ and $F^2 = F - F^1$. Both of them will be used in CSG-OFSS in Section 3.3.

3.3 CSG-OFSS: Core Structures-Guided Optimal Fusion Architectures Search

Due to the fact that optimal multi-modal fusion architectures often contain core structures, we can leverage the core structures obtained in the first-stage search to rapidly determine a high-quality search sub-space containing the optimal multi-modal fusion architecture. This subspace is comprised of the neighborhood surrounding the core structures.

The core structures identified through the ENAS_1 search become the focus in this stage, where we concentrate on exploring the neighborhoods of these core structures to find the optimal multi-modal fusion architecture. Initially, the neighborhood of a core structure refers to the multi-modal fusion architectures formed by continuously adding substructures composed of the remaining features M^2 and fusion operator set F^2 to the core structures. To obtain the optimal multi-modal fusion architecture, we employ ENAS_2 to search the

neighborhood of each core structure. Through the evolutionary algorithm, we can adaptively evaluate the fusion architectures around each core structure, achieving the goal of searching the neighborhood. The detailed process of the evolutionary algorithm can be found in Section 3.4.

3.4 Search Strategy

Two armed evolutionary algorithms ENAS_1 and ENAS_2 with our proposed adaptive knowledge inheritance (AKI) are used as search strategy to core structures search and the optimal multi-modal fusion architecture, respectively. AKI will be detailed in Section 3.6.

Initially, we initialize a population from the core structure search space and iteratively discover the core structures through the search process. Subsequently, evolutionary algorithms are utilized to search within the neighborhoods of these structures. Specifically, we treat the core structures as a new population and focus on expanding individuals in this population using mutation operations during the iterative search process. This expansion is aimed at exploring the surrounding neighborhoods. For instance, when an individual undergoes a mutation operation, we generate a substructure from the feature set M^2 and the fusion operator set F^2, and expand the individual through the addition operation of the mutation, as illustrated in the lower right corner of Figure 2. The key steps of CSG-NAS include population initialization, fitness evaluation, offspring generation, and selection.

Population Initialization: Generate a population P consisting of K individuals randomly from the core structure search space.

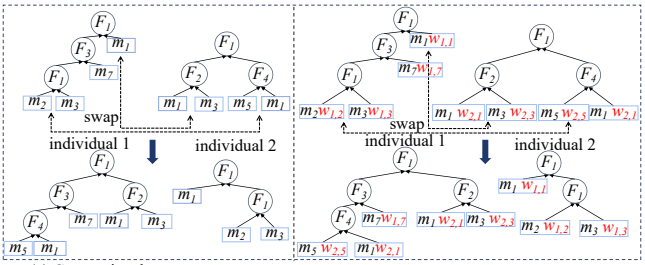
Fitness Evaluation: Each individual is decoded into a multi-modal classification model, with detailed decoding processes outlined in Section 3.5. Subsequently, the model is trained and evaluated using the corresponding dataset to obtain the fitness value for each individual.

Crossover and Mutation: The entire algorithm is divided into two stages: $ENAS_1$ for CSS and $ENAS_2$ for CSG-OFSS. Each stage has different crossover and mutation rates to meet its specific search objectives.

ENAS_1: Crossover rate is $r = r_1$, mutation rate is $r = r_2$ In this stage, following the traditional EA principles, the probability of crossover operation is significantly higher than the probability of mutation operation. The objective is to search for core structures, and the mutation operation generates a substructure from the core structure search space, replacing a specific substructure within the individual.

ENAS_2: Crossover rate is $r = r_1/9$, mutation rate is $r = 4r_2$. In contrast to the ENAS_1, ENAS_2 employs an extremely low crossover rate and a very high mutation rate. The goal is to explore the neighborhood around core structures; hence, the probability of mutation operation is high. The mutation operation generates a substructure from the non-core structure space and adds it to the individual, searching the neighborhood around the core structures. The low crossover rate aims to avoid disrupting the core structures themselves.

Selection: Binary tournament selection (BTS) [Liang *et al.*, 2021] is used.



(a) Conventional crossover operator (b)The novel crossover operator employs knowledge exchange for weight inheritance

Figure 3: The difference between original crossover and our one

3.5 Encoding and Decoding of Multi-Modal Classification Models

Each individual p within the population is encoded as a binary tree, where the leaf nodes denote features, and the branch nodes represent fusion operators. In this paper, fusion operators include concatenation, addition, multiplication, Maximize, and average. Detailed definitions can be found in work [Liang *et al.*, 2021]. For each individual, if the binary tree contains k features, then it must contain $k-1$ fusion operators. Each individual corresponds to a multi-modal classification model. The binary tree can be decoded into a multi-modal classification model through the following steps: 1) Channel the modality features, represented by the leaf nodes of the individual encoding tree, into fully connected layers (FC) for feature alignment to facilitate feature fusion; 2) Conduct feature fusion based on the fusion operators represented by the branch nodes; 3) Direct the fused features through FC and Softmax layers for the final prediction output.

3.6 AKI: Adaptive Knowledge Inheritance

In traditional evolutionary NAS, offspring are typically generated through crossover and mutation of the encoding structures, without considering the weights learned by the parent during training, as illustrated in Figure 3. However, in addition to the network structure, the initialization weights also play a crucial role in the performance of convolutional neural networks (CNNs). Well-known initialization methods, such as Xavier initializer and Kaiming initializer, are often used in neural architecture search methods. However, these initialization methods do not fully leverage the knowledge acquired from training CNNs.

Based on the analysis, we introduce a novel breeding strategy based on knowledge inheritance. Its core idea is that similar structures often demonstrate comparable performance. In other words, the more similarity exists between the structure of offspring and parent, the closer their performance tends to be. This perspective has been validated in numerous studies. Consequently, we determine whether to inherit knowledge from the parent by calculating the similarity between offspring and parent, specifically the weight knowledge acquired by the parent during training. When the similarity surpasses a threshold s, it needs to inherit the parent's knowledge, thereby facilitating the learning process of the offspring. This approach allows some offspring to avoid training from scratch and instead fine-tune on the weights learned by

the parent, leading to significant performance improvements and enhancing the efficiency of the entire search process.

To quantify the similarity between parent and offspring, the concept of tree edit distance [Gopal *et al.*, 2023] is introduced, representing the minimum number of changes of transforming one individual into another. The calculation is specifically performed using a dynamic programming algorithm. This strategy aims to improve learning performance and efficiency by sharing and reusing the weights learned by the parent. During the breeding process, in addition to manipulating the network structure, this strategy also transfers the weights learned by the parent to the offspring, which helps the offspring network to converge and learn faster while fully utilizing valuable information obtained in the parent network.

4 Experiments

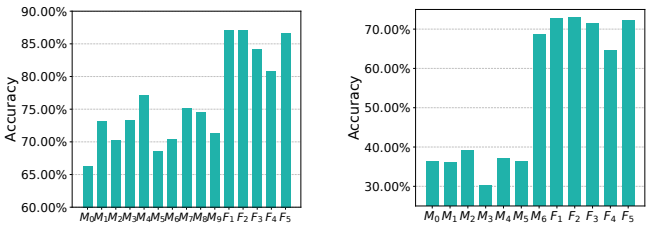
4.1 Multi-Modal Datasets

Our method is implemented using TensorFlow 2.0.3. The computational environment consists of Ubuntu 16.04.4, 512GB DDR4 RDIMM, 2X 40-Core Intel Xeon CPU E5-2698 v4 @ 2.20GHz, and NVIDIA Tesla P100. The used GPU configuration in this paper is the same as the MFAS.

We validated five popular multi-modal datasets: (1) ChemBook-10k (CB) [Liang *et al.*, 2021] dataset, designed for chemical structure image recognition in patent retrieval studies, which contains 100,000 chemical structure images distributed into 10,000 categories. (2) NUS-WIDE-128 (NUS) [Tang *et al.*, 2017] dataset, which contains 43,800 images divided into 128 categories. We chose a subset of 10 categories totalling 23,438 images from this dataset. (3) MM-IMDB [Arevalo *et al.*, 2017] dataset for the multi-label film genre classification task, which contains a total of 23 categories. The dataset is divided into a training set of 15,552 films, a validation set of 2,608 films, and a test set of 7,799 films. (4) NTU RGB-D [Shahroudy *et al.*, 2016] dataset for multimodal action recognition task containing 60 categories. The training, validation and test sets include 23,760, 2,519 and 16,558 samples, respectively. (5) EgoGesture [Zhang *et al.*, 2018] dataset for multimodal gesture recognition task containing 83 categories. The training set of this dataset includes 14,416 samples, the validation set includes 4,768 samples, and the test set includes 4,977 samples.

4.2 Comparison Methods

To validate the effectiveness and efficiency of the CSG-NAS, we compared it with several state-of-the-art algorithms. These peer competitors can be broadly categorized based on whether the architecture is manually designed. The first category is MMC whose fusion architectures are designed by human experts, including MBL [Kim *et al.*, 2017], MFB [Yu *et al.*, 2018], TFN [Zadeh *et al.*, 2017], LMF [Liu *et al.*, 2018], PTP [Hou *et al.*, 2019], TMC [Han *et al.*, 2023], TMOA [Liu *et al.*, 2022], AWDR [Yang *et al.*, 2019], RAMC [Jiang *et al.*, 2022], Maxout MLP [Goodfellow *et al.*, 2013] , VGG Transfer [Simonyan and Zisserman, 2015], GMU [Arevalo *et al.*, 2017], CentralNet [Vielzeuf *et al.*, 2019], Inflated ResNet-50 [Baradel *et al.*, 2018], Co-occurrence [Li *et al.*, 2018], MMTM [Vaezi Joze *et al.*, 2020], VGG-16 + LSTM [Yang and Tian, 2014], C3D + LSTM + RSTTM [Molchanov *et al.*, 2016], I3D [Carreira and Zisserman, 2017], ResNext-101 [Köpüklü *et al.*, 2019], and MTUT [Gupta *et al.*, 2019]. The second category is NAS-based MMC methods including EDF [Liang *et al.*, 2021], MFAS [Perez Rua *et al.*, 2019], BM-NAS [Yin *et al.*, 2022], 3D-CDC-NAS2 [Yu *et al.*, 2021] and DC-NAS [Liang *et al.*, 2024].



(a) CB (b) NUS

Figure 4: The accuracies of single feature and fusion operation

4.3 Performance Comparison

Results on CB and NUS. To mitigate the randomness introduced by data splitting and network initialization, each dataset is divided into training and testing sets via 5-fold cross-validation.

From Figure 4, the performance of each feature and fusion operation can be observed. Following the aforementioned algorithm, we first select high-quality features and fusion operations to form the core structure search space. For example, for the CB dataset, we chose features M_1, M_3, M_4, M_7, M_8, and fusion operations F_1, F_2, F_5. For the NUS dataset, we selected features M_2, M_4, M_6, and fusion operations F_1, F_2, F_5. We then search for the core structures and, based on the core structures, use a local algorithm to search for the optimal multi-modal fusion architecture.

To thoroughly demonstrate the advancement of MMC-NAS, our experimental settings follow the EDF [Liang *et al.*, 2021]. We compared MMC-NAS with some advanced multi-modal fusion operators and existing advanced multi-modal fusion methods. From the results in Table 1, we can conclude that, compared to advanced fusion operators, our use of basic fusion operators with our search strategy achieves a significant lead. Among multi-modal methods, except for EDF and DC-NAS, all others are non-NAS methods. It is evident that the performance of MMC-NAS methods is superior to manual selection. Due to our ability to rapidly locate core structures, locally search for optimal fusion architectures, and utilize a knowledge inheritance mechanism, our performance surpasses EDF and DC-NAS.

Results on MM-IMDB. To ensure a fair comparison with other explicit multi-modal fusion methods, we adopted the same neural network backbone models as BM-NAS and DC-NAS to extract various modality features, with the weighted F1 score as the evaluation metric. The parameters are set as follows: the population size N is 20, the number of population iterations T is 8, the dimension of the fusion vector FD is 128, and modality features are repeatable. According to Table 2, the weighted F1 score of CSG-NAS is better than the

Method	CB	NUS
Advanced fusion operators		
MBL	82.38±0.32	70.60±0.29
MFB	87.94±0.32	71.34±0.40
TFN	73.45±0.30	63.66±1.22
LMF	82.81±0.18	71.74±0.70
PTP	85.08±0.11	71.83±0.50
Multi-modal methods		
TMC	77.88±0.20	72.73±0.30
TMOA	86.81±0.09	72.60±0.48
EmbraceNet	85.85±0.09	72.43±0.38
AWDR	86.66±0.16	72.44±0.66
RAMC	85.36±0.46	72.51±0.67
EDF	88.46±0.27	73.67±0.64
DC-NAS	88.52±0.13	74.20±0.32
CSG-NAS(ours)	**89.20±0.06**	**74.52±0.40**

Table 1: The accuracy on the CB and NUS dataset are reported

Method	Modality	F1-W(%)
Unimodal methods		
Maxout MLP (ICML13)	Text	57.54
VGG Transfer (ICLR15)	Image	49.21
Multi-modal methods		
Two-stream (NIPS14)	Image + Text	60.81
GMU (ICLR17)	Image + Text	61.70
CentralNet (ECCV18)	Image + Text	62.23
MFAS (CVPR19)	Image + Text	62.50
BM-NAS (AAAI22)	Image + Text	62.92±0.03
DC-NAS (AAAI24)	Image + Text	63.70±0.11
CSG-NAS (ours)	Image + Text	**64.12±0.12**

Table 2: Multi-label genre classification results on MM-IMDB dataset. Weighted F1 (F1-W) is reported.

existing multi-modal classification methods, surpassing the state-of-the-art MFAS, BM-NAS, and DC-NAS by 1.96%, 1.22%, and 0.42%, respectively.

Results on NTU RGB-D. Followed the data preprocessing pipelines of BM-NAS and DC-NAS to ensure the fairness of the experimental results. Specifically, Inflated ResNet-50 [Baradel *et al.*, 2018] and Co-occurrence [Li *et al.*, 2018] are adopted as the feature extractor for the skeleton and video modality, respectively, extracting eight features denoted as skeleton_i and video_i, where $i = 1, 2, 3, 4$. Moreover, the population size is 28, iteration times is 8, fusion modality dimension is 64, and modalities are not reused. In Table 3, our method achieved a cross-subject accuracy of 91.12%, outperforming recent methods using video and skeleton modalities. By examining the optimal architectures obtained by MFAS, BM-NAS and DC-NAS, it can be seen that they both contain high-quality features, such as skeleton_4 and video_4, validating the necessity of searching from the core structure space.

Results on EgoGesture. The settings of the BM-NAS method are followed where ResNeXt-101 [Köpüklü *et al.*, 2019] as the backbone network for RGB and depth video modalities. CSG-NAS is compared with various single-modal and multi-modal methods. The experimental settings

Method	Modality	Acc (%)
Unimodal methods		
Inflated ResNet-50 (CVPR18)	Video	83.91
Co-occurence (IJCAI18)	Pose	85.24
Multi-modal methods		
Two-stream (NIPS14)	Video + Pose	88.60
GMU (ICLR17)	Video + Pose	85.80
MMTM (CVPR20)	Video + Pose	88.92
CentralNet (ECCV18)	Video + Pose	89.36
MFAS (CVPR19)	Video + Pose	89.50±0.60
BM-NAS (AAAI22)	Video + Pose	90.48±0.24
DC-NAS (AAAI24)	Video + Pose	90.85±0.05
CSG-NAS (ours)	Video + Pose	**91.12±0.03**

Table 3: Action recognition results on NTU RGB-D dataset

Method	Modality	Acc (%)
Unimodal methods		
ResNext-101 (FG19)	RGB	93.75
VGG-16 + LSTM (CVPR14)	Depth	77.70
C3D + LSTM + RSTTM	Depth	90.60
I3D (CVPR17)	Depth	89.47
ResNeXt-101 (FG19)	Depth	94.03
Multi-modal methods		
VGG-16 + LSTM (CVPR17)	RGB + Depth	81.40
C3D + LSTM + RSTTM	RGB + Depth	92.20
I3D (CVPR17)	RGB + Depth	92.78
MMTM (CVPR20)	RGB + Depth	93.51
MTUT (3DV19)	RGB + Depth	93.87
3D-CDC-NAS2 (TIP21)	RGB + Depth	94.38
BM-NAS (AAAI22)	RGB + Depth	94.96±0.07
DC-NAS (AAAI24)	RGB + Depth	95.22±0.05
CSG-NAS (ours)	RGB + Depth	**95.25±0.04**

Table 4: Gesture recognition results on EgoGesture dataset

for CSG-NAS included a population size of 28, 10 iterations, no modal reuse, and a fusion dimension of 32. Table 4 presents the experimental results on the EgoGesture dataset. Compared to other unimodal/multimodal methods, CSG-NAS achieves state-of-the-art fusion performance, indicating that CSG-NAS can effectively enhance gesture recognition performance on the EgoGesture dataset.

4.4 Search Efficiency Comparison

The goal of this section is to compare CSG-NAS with several powerful MMC baseline methods, including MFAS, EDF, BM-NAS, DC-NAS, and MMTM, focusing primarily on search efficiency and performance to demonstrate its advanced capabilities. The research results are comprehensively summarized in Table 5. From the table, it can be observed that on five complex datasets, CSG-NAS finds the optimal fusion architecture in the shortest time. For example, on NUS and CB, we outperform EDF and DC-NAS in terms of performance and achieve nearly four times the efficiency of the EDF method and twice that of the state-of-the-art DC-NAS method. On NTU RGB-D and EgoGesture, we achieve significant advantages in both performance and efficiency, with

Method	Dataset	Parameters	Time	CP (%)
EDF	NUS	0.31M	11.43	73.67
DC-NAS	NUS	0.53M	4.61	74.20
CSG-NAS(ours)	NUS	0.37M	**2.71**	**74.52**
EDF	CB	2.28M	78.01	88.48
DC-NAS	CB	2.41M	61.88	88.45
CSG-NAS(ours)	CB	2.47M	**24.68**	**89.20**
BM-NAS	MM-IMDB	0.65M	1.24	62.94
DC-NAS	MM-IMDB	0.42M	1.19	63.70
CSG-NAS(ours)	MM-IMDB	0.56M	**0.98**	**64.12**
MMTM	NTU	8.61M	-	88.92
MFAS	NTU	2.16M	603.64	89.50
BM-NAS	NTU	0.98M	53.68	90.48
DC-NAS	NTU	0.26M	13.63	90.85
CSG-NAS(ours)	NTU	0.19M	**5.19**	**91.12**
BM-NAS	Ego	0.61M	20.67	94.96
DC-NAS	Ego	0.19M	4.57	95.22
CSG-NAS(ours)	Ego	0.20M	**3.27**	**95.25**

Table 5: Comparison of model size, time (GPU hours) and classification performance (CP) of generalized multi-modal NAS methods.

Version	CS	KI	Time	Acc (%)
CSG-NAS$_1$	False	False	67.96±5.65	88.45±0.22
CSG-NAS$_2$	True	False	36.75±1.09	88.57±0.03
CSG-NAS	True	True	**24.68±3.41**	**89.20±0.06**

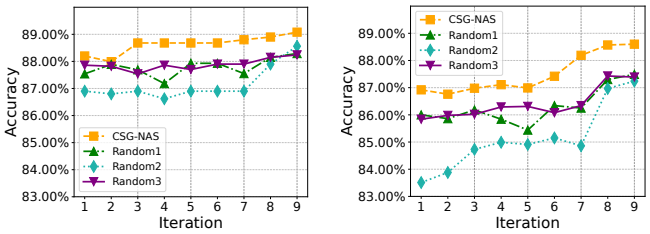
Table 6: Ablation study of CSG-NAS

the time consumption for searching the optimal fusion model reduced by almost ten times compared to the state-of-the-art BM-NAS and twice that of DC-NAS. This is attributed to our core structures-guided neural architecture search method, which significantly narrows down the search space, effectively avoiding the evaluation of a large number of poorly performing models, and focusing on assessing models with superior performance.

4.5 Ablation Study

To provide a more in-depth analysis for the proposed CSG-NAS, we conducted ablation experiments on the CB dataset.

Effectiveness of Core Structures (CS) and Knowledge Inheritance (KI): To delve into the impact of CS and KI on CSG-NAS, we conducted a performance analysis in three scenarios of CSG-NAS. From Table 6, we can draw the following conclusions: Compared to searching the entire space, using core structures to guide neural architecture search allows achieving the same performance level in a shorter time. Looking at the standard deviations of CSG-NAS$_1$ and CSG-NAS$_2$, CSG-NAS$_2$ exhibits higher stability, indicating consistent discovery of the optimal fusion architecture. On the other hand, CSG-NAS$_1$ shows larger fluctuations and may occasionally converge to suboptimal solutions. The results clearly demonstrate that the use of knowledge inheritance architectures can effectively enhance search efficiency and improve performance. For instance, the time was reduced from 36 hours to 25 hours and the performance increased from 88.57% to 89.20%.



(a) Top1 accuracy (b) Top28 accuracy

Figure 5: Comparison between core structures selection strategies

Cases	Inheritance conditions	Acc (%)
1	Not inheriting	88.56
2	All inheriting	89.07
3	Random inheriting	88.78
4	Inheritance with $s < 0.5$	88.64
5	Inheritance with $s > 0.5$	**89.33**

Table 7: Effect of Inheritance Architectures on CSG-NAS

Analysis of Core Structures Selection Strategies: To investigate the impact of core structures selection on subsequent search, four experiments were conducted. The first one experiments involved searching for core structures using high-quality features and fusion operators, while the next three experiments randomly selected features and fusion operators for core structures search. The experimental results in Figure 5 clearly indicate that searching for core structures using high-quality features and fusion operations leads to significantly superior outcomes compared to the case of randomly selecting features and fusion operations.

Knowledge Inheritance: To investigate the impact of knowledge inheritance on CSG-NAS, we conducted five experiments including no inheritance, complete inheritance, random inheritance, and inheritance occurs based on empirical observation when similarity is greater than 0.5 and less than 0.5. The results in Table 7 show that inheriting from offspring with a similarity greater than 0.5 leads to optimal performance. This indicates that determining knowledge inheritance based on the similarity between parent and offspring has a significant positive impact on CSG-NAS.

5 Conclusion

This paper has investigated a rapid and adaptive method for searching multi-modal fusion architectures, utilizing a core structure to swiftly narrow down the entire search space to a compact subset with state-of-the-art performance. The core structures have been employed for local search in its vicinity to identify the optimal multi-modal fusion architecture. Additionally, a novel knowledge inheritance strategy has been proposed to further enhance performance. Extensive experiments have validated CSG-NAS's advantages. The use of CSG-NAS holds the promise of significantly alleviating the challenges posed by vast multi-modal fusion search spaces and remarkably improved search efficiency.

Contribution Statement

Pinhan Fu and Xinyan Liang contributed equally to this work.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Nos. 62136005, 62306171, 62376281, 62106132, 62306170, 62376146), the National Key Research and Development Program of China (No. 2021ZD0112400), the Science and Technology Major Project of Shanxi (No. 202201020101006), and Young Scientists Fund of the Natural Science Foundation of Shanxi (No. 202203021222183).

References

[Arevalo *et al.*, 2017] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and FabioA. González. Gated multimodal units for information fusion. *Cornell University - arXiv*, 2017.

[Baradel *et al.*, 2018] Fabien Baradel, Christian Wolf, Julien Mille, and Graham W. Taylor. Glimpse clouds: Human activity recognition from unstructured feature points. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 469–478, 2018.

[Brock *et al.*, 2018] Andrew Brock, Theodore Lim, James Millar Ritchie, and Nicholas J. Weston. SmaSH: One-shot model architecture search through hypernetworks. In *International Conference on Learning Representations*, pages 1–11, 2018.

[Cai *et al.*, 2020] Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, and Song Han. Once for all: Train one network and specialize it for efficient deployment. In *International Conference on Learning Representations*, pages 1–15, 2020.

[Carreira and Zisserman, 2017] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4733, 2017.

[Goodfellow *et al.*, 2013] Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. In *Proceedings of the 30th International Conference on International Conference on Machine Learning*, volume 28, page 1319–1327, 2013.

[Gopal *et al.*, 2023] Bhavna Gopal, Arjun Sridhar, Tunhou Zhang, and Yiran Chen. Lissnas: Locality-based iterative search space shrinkage for neural architecture search. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023.

[Guo *et al.*, 2022] Qian Guo, Yuhua Qian, and Xinyan Liang. GLRM: Logical pattern mining in the case of inconsistent data distribution based on multigranulation strategy. *International Journal of Approximate Reasoning*, 143:78–101, 2022.

[Gupta *et al.*, 2019] Vikram Gupta, Sai Kumar Dwivedi, Rishabh Dabral, and Arjun Jain. Progression modelling for online and early gesture detection. In *2019 International Conference on 3D Vision*, pages 289–297, 2019.

[Han *et al.*, 2023] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2551–2566, 2023.

[Hou *et al.*, 2019] Ming Hou, Jiajia Tang, Jianhai Zhang, Wanzeng Kong, and Qibin Zhao. *Deep Multimodal Multilinear Fusion with High-Order Polynomial Pooling*. 2019.

[Jiang *et al.*, 2022] Bingbing Jiang, Junhao Xiang, Xingyu Wu, Yadi Wang, Huanhuan Chen, Weiwei Cao, and Weiguo Sheng. Robust multi-view learning via adaptive regression. *Information Sciences*, 610:916–937, 2022.

[Jiang *et al.*, 2024] Zhangqi Jiang, Tingjin Luo, and Xinyan Liang. Deep incomplete multi-view learning network with insufficient label information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(11):12919–12927, 2024.

[Kim *et al.*, 2017] Jin-Hwa Kim, KyoungWoon On, Woosang Lim, Jeong-Hee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *International Conference on Learning Representations*, pages 1–10, 2017.

[Köpüklü *et al.*, 2019] Okan Köpüklü, Ahmet Gunduz, Neslihan Kose, and Gerhard Rigoll. Real-time hand gesture detection and classification using convolutional neural networks. In *Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 1–8, 2019.

[Li *et al.*, 2018] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu. Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, page 786–792, 2018.

[Liang *et al.*, 2021] Xinyan Liang, Qian Guo, Yuhua Qian, Weiping Ding, and Qingfu Zhang. Evolutionary deep fusion method and its application in chemical structure recognition. *IEEE Transactions on Evolutionary Computation*, 25(5):883–893, 2021.

[Liang *et al.*, 2022a] Xinyan Liang, Yuhua Qian, Qian Guo, Honghong Cheng, and Jiye Liang. AF: An association-based fusion method for multi-modal classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9236–9254, 2022.

[Liang *et al.*, 2022b] Xinyan Liang, Yuhua Qian, Qian Guo, and Qin Huang. Multi-granulation fusion-driven method for many-view classification. *Journal of Computer Research and Development*, 59(8):1653–1667, 2022.

[Liang *et al.*, 2024] Xinyan Liang, Pinhan Fu, Qian Guo, Keyin Zheng, and Yuhua Qian. DC-NAS: Divide-and-conquer neural architecture search for multi-modal classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(12):13754–13762, 2024.

[Liu *et al.*, 2018] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2247–2256, 2018.

[Liu *et al.*, 2019] Hanxiao Liu, Karen Simonyan, and Yiming Yang. DARTS: Differentiable architecture search. In *Proceedings of the International Conference on Learning Representations*, pages 1–11, 2019.

[Liu *et al.*, 2022] Wei Liu, Xiaodong Yue, Yufei Chen, and Thierry Denoeux. Trusted multi-view deep learning with opinion aggregation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7585–7593, Jun. 2022.

[Mendoza *et al.*, 2016] Hector Mendoza, Aaron Klein, Matthias Feurer, Jost Tobias Springenberg, and Frank Hutter. Towards automatically-tuned neural networks. In *Proceedings of the Workshop on Automatic Machine Learning*, volume 64, pages 58–65, 2016.

[Molchanov *et al.*, 2016] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4207–4215, 2016.

[Peng *et al.*, 2020] Yige Peng, Lei Bi, Michael Fulham, Dagan Feng, and Jinman Kim. Multi-modality information fusion for radiomics-based neural architecture search. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention*, pages 763–771, 2020.

[Perez Rua *et al.*, 2019] JuanManuel Perez Rua, Valentin Vielzeuf, Stephane Pateux, Moez Baccouche, and Frederic Jurie. MFAS: Multimodal fusion architecture search. In *Proceedings of the 2019 Conference on Computer Vision and Pattern Recognition*, pages 6959–6968, 2019.

[Real *et al.*, 2019] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V. Le. Regularized evolution for image classifier architecture search. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[Shahroudy *et al.*, 2016] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1010–1019, 2016.

[Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the Third International Conference on Learning Representations*, page 1–14, 2015.

[Tang *et al.*, 2017] Jinhui Tang, Xiangbo Shu, Guojun Qi, Zechao Li, Meng Wang, Shuicheng Yan, and Ramesh Jain. Tri-clustered tensor completion for social-aware image tag refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1662–1674, 2017.

[Vaezi Joze *et al.*, 2020] Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzzolino, and Kazuhito Koishida. Mmtm: Multimodal transfer module for cnn fusion. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13286–13296, 2020.

[Vielzeuf *et al.*, 2019] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: A multilayer approach for multimodal fusion. In *European Conference on Computer Vision Workshops*, pages 575–589, 2019.

[Wei *et al.*, 2023] Chen Wei, Chuang Niu, Yiping Tang, Yue Wang, Haihong Hu, and Jimin Liang. Npenas: Neural predictor guided evolution for neural architecture search. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8441–8455, 2023.

[Xu *et al.*, 2024] Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. Reliable conflictive multi-view learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14):16129–16137, 2024.

[Yang and Tian, 2014] Xiaodong Yang and YingLi Tian. Super normal vector for activity recognition using depth sequences. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 804–811, 2014.

[Yang *et al.*, 2019] Muli Yang, Cheng Deng, and Feiping Nie. Adaptive-weighting discriminative regression for multi-view classification. *Pattern Recognition*, 88:236–245, 2019.

[Yin *et al.*, 2022] Yihang Yin, Siyu Huang, Xiang Zhang, and Dejing Dou. BM-NAS: Bilevel multimodal neural architecture search. In *Association for the Advancement of Artificial Intelligence*, pages 8901–8909, 2022.

[Yu *et al.*, 2018] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 29(12):5947–5959, 2018.

[Yu *et al.*, 2021] Zitong Yu, Benjia Zhou, Jun Wan, Pichao Wang, Haoyu Chen, Xin Liu, Stan Z. Li, and Guoying Zhao. Searching multi-rate and multi-modal temporal enhanced networks for gesture recognition. *IEEE Transactions on Image Processing*, 30:5626–5640, 2021.

[Zadeh *et al.*, 2017] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, 2017.

[Zhang *et al.*, 2018] Yifan Zhang, Congqi Cao, Jian Cheng, and Hanqing Lu. Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. *IEEE Transactions on Multimedia*, page 1038–1050, 2018.

[Zoph *et al.*, 2018] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning transferable architectures for scalable image recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8697–8710, 2018.