A Fast Neural Architecture Search Method for Multi-Modal Classification via Knowledge Sharing

Zhihua Cui[1] , **Shiwu Sun**[1] , **Qian Guo**[1*] , **Xinyan Liang**[2] , **Yuhua Qian**[2] , **Zhixia Zhang**[1]
[1]Shanxi Key Laboratory of Big Data Analysis and Parallel Computing, Taiyuan University of Science and Technology
[2]Institute of Big Data Science and Industry, Shanxi University
{cuizhihua, zhixiazhang}@tyust.edu.cn, {sswlat, czguoqian, liangxinyan48}@163.com, jinchengqyh@126.com

Abstract

Neural architecture search-based multi-modal classification (NAS-MMC) aims to automatically find optimal network structures for improving the multi-modal classification performance. However, most current NAS-MMC methods are quite time-consuming during the training process. In this paper, we propose a knowledge sharing-based neural architecture search (KS-NAS) method for multi-modal classification. The KS-NAS optimizes the search process by introducing a dynamically updated knowledge base to reduce the consumption of computational resource. Specifically, during the deep evolutionary search, individuals in the initial population acquire initial parameters from a knowledge base, and then undergo training and optimization until convergence is reached, avoiding the need for training from scratch. The knowledge base is dynamically updated by aggregating the parameters of high-quality individuals trained within the population, thus progressively improving the quality of the knowledge base. As the population evolves, the knowledge base continues to optimize, ensuring that subsequent individuals can obtain higher-quality initialization parameters, which significantly accelerates the training speed of the population. Experimental results show that the KS-NAS method achieves state-of-the-art results in terms of classification performance and training efficiency across multiple popular multi-modal tasks.

1 Introduction

With the advent of the big data era, the internet generates a vast amount of data in various modalities such as image, text, and sound. Compared to single-modal data, multi-modal data contain more diverse relationships and deeper semantic information. Multi-modal classification refers to the utilization of multi-modal data for understanding and categorizing targets, with its core lying in how to effectively fuse these multi-modal information [Liang *et al.*, 2022; Jiang *et al.*, 2023; Han *et al.*, 2022].

*Corresponding Author

In order to find the optimal network structure for multi-modal information fusion, researchers have focused on NAS. NAS-MMC can automatically search for more efficient and well-generalized multi-modal fusion strategies within a vast architectural space, such as DC-NAS [Liang *et al.*, 2024], CSG-NAS [Fu *et al.*, 2024], BM-NAS [Yin *et al.*, 2022], EDF [Liang *et al.*, 2021], and MFAS [Pérez-Rúa *et al.*, 2019]. Based on different search strategies, NAS is mainly divided into methods based on reinforcement learning (RL), evolutionary algorithms (EA), and gradient-based approaches. Each strategy possesses its unique strengths and limitations. Compared to RL, EA demonstrates certain advantages in training speed, however, it still entails considerable computational costs. As for gradient-based methods, although they excel in training efficiency, due to the adoption of batch training mode, their global search capability is relatively weak, making them prone to getting stuck in local optimal solutions. Therefore, it is highly necessary to seek a search strategy that is both fast and capable of producing excellent results.

In this paper, we innovatively propose a fast neural architecture search method for multi-modal classification via knowledge sharing, termed KS-NAS. Its core lies in leveraging a knowledge-sharing mechanism to optimize the search process and reduce computational resource consumption. Specifically, we first define a dynamically updated knowledge base, which exists in the form of a supernet and encompasses potential representations of all possible network architectures within the search space. Subsequently, the supernet is trained using a gradient-based optimization strategy to initialize the knowledge base. During the subsequent search process, individuals in each generation of the population acquire initialization parameters from the current knowledge base, thereby avoiding the time-consuming process of training from scratch. These individuals are then trained and optimized in the deep evolutionary search process until convergence. Once this generation of individuals converges, parameters of the outstandingly performing individuals are selected and fed back into the knowledge base for updating. In this manner, the quality of the knowledge base continues to improve as the population evolves, ensuring that subsequent individuals can obtain more excellent initialization parameters. By continuously iterating this process, the KS-NAS method achieves dual enhancements in population quality and training speed. The quality of the initial population in each gen-

eration becomes increasingly high, resulting in progressively faster training speeds in each subsequent generation.

We conducted experimental validations on multiple multi-modal datasets, and the results indicate that the aforementioned virtuous cycle design strategy not only reduces the consumption of computational resources but also successfully alleviates the prevalent inefficiency issue faced by current EA-based NAS methods. This achievement fully demonstrates the exceptional performance of our method in terms of both efficiency and accuracy. Specifically, our contributions are as follows:

- Individuals can obtain initial parameters from a dynamically updated knowledge base, avoiding the need for training from scratch, which accelerates the Neural NAS process. In traditional NAS methods, each candidate architecture requires training from scratch, which is an extremely time-consuming and resource-intensive process. However, in KS-NAS, by introducing a dynamically updated knowledge base, individuals can directly obtain pre-trained, high-quality initialization parameters, significantly reducing the training time for each generation of individuals.
- The knowledge base is dynamically updated by aggregating the parameters of high-quality individuals in the population, gradually improving its quality. As a core component of the KS-NAS method, the quality of the knowledge base directly affects search efficiency and the accuracy of the final results. After each generation of search, outstanding individuals are selected from the population, and their parameters are integrated into the knowledge base to ensure that it continues to improve its quality as the population evolves. This dynamic update mechanism allows the knowledge base to continuously accumulate excellent network architecture features, providing more valuable initialization parameters for subsequent searches.
- As the population evolves, both the initialization parameters of the population and the quality of the knowledge base improve, forming a virtuous cycle that promotes each other. Compared with existing EA-based NAS methods, KS-NAS maintains high accuracy while achieving faster search speeds and reduced computational load, providing new ideas and methods for addressing the computational challenges in the NAS field.

2 Related Work

Multi-Modal Fusion: Multi-modal classification integrates information from different modalities or views to enable models to better understand data and make accurate classifications [Liang *et al.*, 2025; Guo *et al.*, 2024]. Among them, multi-modal fusion plays a crucial role. Within the framework of deep neural networks, multi-modal fusion techniques can primarily be categorized into three major types based on the fusion stage: early fusion, intermediate fusion, and late fusion. Early fusion focuses on integrating low-level features from different modalities at the early stage of data processing [Wang *et al.*, 2018; Yu *et al.*, 2018]. Intermediate fusion refers to the integration of features from different modalities at the intermediate levels of a model [Joze *et al.*, 2020; Vielzeuf *et al.*, 2019]. Late fusion emphasizes the fusion of information from different modalities at the decision level [Han *et al.*, 2020]. In practical applications, the choice of fusion strategy depends on the specific application scenario, data characteristics, and the design objectives of the model.

NAS: NAS aims to automatically discover an optimal network structure [Elsken *et al.*, 2019]. Based on different search strategies, NAS can be mainly divided into the following three categories: RL-based, EA-based, and gradient-based method. (1) The RL-based search strategy [Jaafra *et al.*, 2019; Balaprakash *et al.*, 2019] frames the neural architecture search problem within a Markov decision process framework, employing reinforcement learning algorithms to acquire optimal search policies. A notable disadvantage of these strategy is the high computational cost. (2) The EA-based search strategy [Fu *et al.*, 2024; Han *et al.*, 2024] models neural architecture search as an evolutionary optimization problem, where candidate network structures are regarded as individuals, their performance as fitness, and new individuals are generated through genetic operators, with the optimal individuals being preserved. Although EA are generally faster than RL methods, they still face considerable computational costs. (3) Gradient-based methods [Yu *et al.*, 2022; Ye *et al.*, 2022] require the prior establishment of a supernet, updating parameters by computing gradients of the loss function with respect to the network architecture parameters. They generally exhibit high computational efficiency. Compared to EA, gradient-based methods typically possess weaker global search capabilities and are prone to getting trapped in local optima.

Unlike the aforementioned methods, individuals in KS-NAS can acquire initial parameters from a dynamic knowledge base that is initially obtained via training a supernet, thereby avoiding training from scratch; The knowledge base is dynamically updated by aggregating high-quality individual parameters from the population, leading to gradual quality improvement. As the population evolves, the initial parameters of the new individuals become better and better, and so does the quality of the knowledge base; the two reinforce each other. Ultimately, this significantly reduces the computational load and addresses the inefficiency issue of existing population-based NAS methods.

3 Methods

In this paper, we propose the KS-NAS for multi-modal classification to find the optimal multi-modal fusion network structure. The main idea of this method is to construct a dynamic knowledge base for knowledge sharing. By retrieving initial parameters from the knowledge base, we avoid training from scratch, addressing the issue of high computational cost in EA-based NAS. Meanwhile, high-quality individual parameters within the population are aggregated through a fair selection process, dynamically updating the knowledge base and gradually improving its quality, which facilitates the search for the optimal architecture. The overall framework of the KS-NAS is shown in Figure 1.

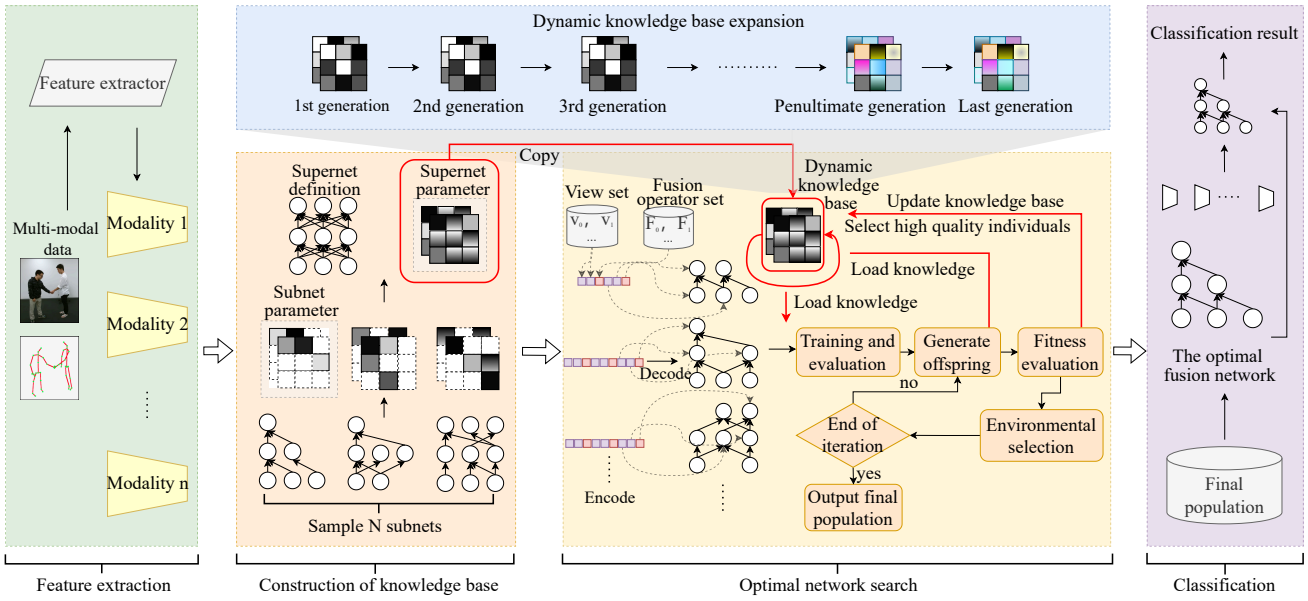

Figure 1: The framework of the proposed method.

3.1 Concept Description

In this section, we explain some of the terms in the paper to avoid conceptual confusion.

Population is a collection of several individuals.

Individual is an integer encoding of a multi-modal fusion network that includes multiple views and fusion operators.

Supernet is a network that contains potential representations of all possible network architectures in the search space and is a concrete representation of the knowledge base.

Subnet is sampled from the supernet and represents the decoded form of an individual.

3.2 Construction of Knowledge Base

The KS-NAS utilizes an evolutionary search approach to find the optimal fusion network. During the evolutionary process, we construct a dynamically updatable knowledge base with the purpose of enabling knowledge sharing among individuals in the population. In this paper, the knowledge base is implemented in the form of a supernet, which contains potential representations of all possible network architectures within the search space. Initially, the knowledge base contains no knowledge, and a gradient-based optimization algorithm is used to train the supernet to initialize the knowledge base.

As the evolutionary process continues, individuals in the population retrieve initial parameters from the knowledge base, thus avoiding the need to train from scratch. In each generation, all individuals begin with the same starting point. After training and evaluation, high-quality individuals are fairly selected by comparison with the worst fitness of the previous generation, the network parameters of high-quality individuals are integrated to update the knowledge base. The corresponding layer parameters from all high-quality individuals' networks are aggregated through averaging and subsequently integrated with the homologous layers of the supernetwork, thereby ensuring that the quality of knowledge stored in the library is continuously improved. This process facilitates the dynamic updating of the knowledge base with high-quality information. Furthermore, individuals within the same generation undergo fair competition, as they all share the same starting conditions so that good individuals will not be selected because of different starting points. The knowledge from previous generations serves as a foundation for the subsequent generations, helping to accelerate their training processes and enabling them to find better solutions. Their synergistic relationship contributes to reducing computational costs and enhancing classification performance.

3.3 Optimal Network Search

We accelerate the training of individuals in the population by sharing knowledge through the knowledge base, adopt the fair selection method to ensure the fairness of individual selection and improve the quality of knowledge in the knowledge base.

First, we define and initialize the supernet as the knowledge base. N subnets are sampled from the supernet, with parameter sharing among them. Features are extracted from multi-modal data using a feature extractor, and then subnets are trained using batch training to update the parameters of the supernet, which also serves as the knowledge base. In the subsequent deep evolutionary search process, subnets are regarded as individuals. Individuals of the same generation acquire knowledge from the same knowledge base for parameter initialization, followed by training. After training for a few generations, evaluation is performed, and the parameters of high-quality individuals are selected to update the knowledge base, facilitating the initialization of individuals in the next generation. The key steps of KS-NAS include population initialization, fitness evaluation, selection, offspring generation, and environmental selection.

Algorithm 1 Pseudo-code of KS-NAS.

Input: Training data D_{train}, validation data D_{valid}, feature extractor E.
Parameter: Population size N, maximum number of generations T.
Output: Last generation population P_T.

Extract V_{train}, V_{valid} from D_{train} and D_{valid} using the well-trained E;
$KS \leftarrow$ Build the knowledge base by training a supernet with gradient-based method;
Generate initial population P_0;
Initialize the individual network in P_0 with KS;
$F_0 \leftarrow$ Train and evaluate individual networks of P_0 using V_{train} and V_{valid};
$Pm \leftarrow$ Record individual networks parameters of P_0;
$t \leftarrow 0$
while $t < T$ **do**
  $Wst \leftarrow$ Find the worst fitness in F_t;
  $KS \leftarrow$ Calculate the mean of the Pm and update knowledge base;
  Select mating pool from P_t by the roulette wheel;
  $Q_t \leftarrow$ Generate offspring by crossing operator and mutation operator
  for each individual q in Q_t **do**
    $q \leftarrow$ Initialize by KS;
  end for
  Train and evaluate individual networks in Q_t using V_{train} and V_{valid};
  for each individual q in Q_t **do**
    if the fitness of $q > Wst$ **then**
      $Pm_q \leftarrow$ Save the parameters of q;
    end if
  end for
  $Pm \leftarrow Pm_q$;
  $P_{t+1} \leftarrow$ Select N individuals from P_t and Q_t by environment selection;
  $F_{t+1} \leftarrow$ Record the fitness of P_{t+1};
  $t \leftarrow t + 1$;
end while
return P_T.

Population initialization: Individuals are specifically implemented in the form of subnets, each of which is composed of input views and fusion operators arranged in a certain fusion order. There is a schematic diagram representing the subnet in Figure 2. N subnets are generated through random sampling to serve as individuals in the initial population. Each individual selects k non-repeating features from the available view features, thereby avoiding local optimal solutions caused by abnormal concentration of feature information in the initial population. The number of fusion operators used in each individual is $k - 1$, which corresponds to the number of fusion nodes for the selected feature.

Fitness evaluation: Each individual is decoded into a multi-modal fusion network. The detailed decoding process is shown in Subsection 3.4. Each network is initialized with the knowledge base and then trained with the dataset. The adaptive value is evaluated after the training.

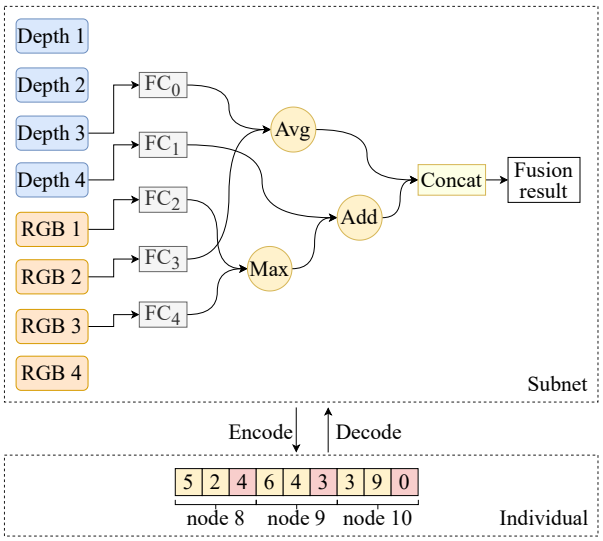

Figure 2: Encoding and decoding of individuals for EgoGesture dataset

Selection: We use roulette wheel selection to determine which individuals to mate with. The specific step is to generate a random number $r \in [0, 1]$, and then select the first individual whose cumulative probability exceeds the random number r. Through this approach, individuals with higher fitness have a greater chance of being selected, while also preserving the possibility of individuals with lower fitness being selected, helping to avoid falling into local optimal solutions.

Crossover and mutation: The CSG-NAS method inspires us [Fu *et al.*, 2024] to design an adaptive crossover rate and mutation rate, two parameters that are constantly adjusted over the course of evolution. The crossover rate gradually increases with evolutionary generations, while the mutation rate decreases accordingly. A higher initial mutation rate enhances population diversity and broadens the search space to prevent premature convergence. As evolution progresses, elevating the crossover rate facilitates effective gene recombination for fine-grained local optimization. This mechanism balances global exploration and local exploitation, ensuring both solution quality and convergence efficiency.

Environment selection: After combining parents and children, we adopted a meritocracy selection strategy to select the next generation of parent individuals to ensure that the best individuals reach the next generation and continue to pass on their good genes. The elite selection method effectively avoids the possible information loss in genetic manipulation by preferentially retaining the individuals with the highest fitness. At the same time, this strategy can accelerate the convergence process, reduce unnecessary fluctuations in the evolutionary process, and ensure the stable transmission of excellent genes.

The pseudo code of KS-NAS is shown in Algorithm 1.

3.4 Encoding and Decoding

In the population, each individual is encoded using integers and consists of multiple nodes. Each node is comprised of three digits, where the first two digits represent input features, and the third digit indicates the fusion operation performed on the two input features. The node itself represents the result of the fusion. The fusion operations include addition, multiplication, concatenation, maximize, and average. The relevant definitions can be found in work [Liang *et al.*, 2021]. Before decoding each individual into multi-modal fusion network, it is necessary to align the modal features in a fully connected way to facilitate fusion in the fusion network. Then, the corresponding modal features are fused according to the fusion network generated by decoding. After fusion, the final fusion results are transmitted to the FC and Softmax layers for the final output. The detailed explanation is shown in Figure 2.

4 Experiments

4.1 Datasets

All of our experiments are implemented using Torch 2.4.1 on Ubuntu 16.04.4. It is equipped with 512GB DDR4 RDIMM memory, two 40-core Intel Xeon CPU E5-2698v4 @ 2.20GHz processors, and uses an NVIDIA Tesla P100 GPU. The effectiveness of our method is verified on five popular multi-modal datasets. The following is the introduction of the datasets: (1) ChemBook-10k (CB) [Liang *et al.*, 2021] dataset, which is a chemical structure image recognition dataset for patent search research. The dataset consists of 1 million images of chemical structures belonging to 10,000 categories. (2) The NUS-Wide-128 (NUS) [Tang *et al.*, 2016] dataset contains 43,800 single-label images from 128 categories. In our experiment, we used a subset of this dataset containing 23,438 images from 10 categories, each associated with a label, with at least 1,500 images per category. (3) The MM-IMDB dataset [Arevalo *et al.*, 2017] is a multi-modal dataset collected from the Internet Movie Database, including 25,959 movies and their associated posters, plots, genres, and other metadata, comprising a total of 27 non-mutually exclusive genres. Due to a severe class imbalance, we chose to use only 23 types for the classification task, and the dataset was divided into 15,552 movies for training, 2,608 movies for validation, and 7,799 movies for testing purposes. (4) NTU RGB-D [Shahroudy *et al.*, 2016] is a large-scale multi-modal action recognition dataset with 56,880 samples, divided into 60 categories in total. The training, validation, and test sets included 23,760, 2,519, and 16,558 samples, respectively. (5) EgoGesture [Zhang *et al.*, 2018] is a multi-modal gesture recognition task dataset containing 24,161 gesture samples, with a total of 83 categories. There were 14,416 samples in the dataset for training, 4,768 samples for validation, and 4,977 samples for testing.

4.2 Comparison Methods

To better validate the effectiveness of KS-NAS, we compared it to a variety of methods including several SOTA algorithms.

(1) Single-mode method: Inflated ResNet-50 [Baradel *et al.*, 2018], Co-occurrence [Li *et al.*, 2018], 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], Maxout MLP [Goodfellow *et al.*, 2013], VGG Transfer [Simonyan, 2014].

(2) Traditional multi-modal methods and advanced fusion operators: TMC [Han *et al.*, 2022], TMOA [Liu *et al.*, 2022], AWDR [Yang *et al.*, 2019], RAMC [Jiang *et al.*, 2022], Two-stream [Simonyan and Zisserman, 2014], GMU [Arevalo *et al.*, 2017], CentralNet [Vielzeuf *et al.*, 2019], MMTM [Joze *et al.*, 2020], MTUT [Gupta *et al.*, 2019], MBL [Kim *et al.*, 2016], MFB [Yu *et al.*, 2018], TFN [Zadeh *et al.*, 2017], LMF [Liu *et al.*, 2018], PTP [Hou *et al.*, 2019].

(3) Multi-modal fusion method based on NAS: EDF [Liang *et al.*, 2021], MFAS [Pérez-Rúa *et al.*, 2019], BM-NAS [Yin *et al.*, 2022], 3D-CDC-NAS2 [Yu *et al.*, 2021], DC-NAS [Liang *et al.*, 2024] and CSG-NAS [Fu *et al.*, 2024].

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	89.20±0.06	74.52±0.40
KS-NAS(ours)	**93.21±0.15**	**75.64±0.48**

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

4.3 Experiments Results and Analysis

For the CB and NUS datasets, we employ a five-fold cross-validation approach to partition the data into training and testing sets, allowing a more accurate estimation of the model's performance and generalization ability. To ensure a fair comparison with other multi-modal fusion methods, we follow the setup of EDF [Liang *et al.*, 2021], using the same feature extractors for data processing and the same fusion operator for constructing the search space. We compare KS-NAS with several advanced multi-modal fusion operators and methods, among which only EDF, DC-NAS and CSG-NAS are NAS-based. The results are shown in Table 1, indicating that KS-NAS achieves state-of-the-art classification performance. On the CB dataset, our method outperforms the advanced NAS-based multi-modal methods EDF, DC-NAS, and CSG-NAS by 4.75%, 4.69%, and 4.01%, respectively. On the NUS dataset, it surpasses them by 1.97%, 1.44%, and 1.12%, respectively. This suggests that knowledge sharing can indeed improve the performance of multi-modal classification tasks.

For the MM-IMDB dataset, to ensure a fair comparison with other multi-modal fusion methods, we follow the setup of BM-NAS [Yin *et al.*, 2022], using Maxout MLP as the

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	61.7
CentralNet (ECCV18)	Image + Text	62.23
MFAS (CVPR19)	Image + Text	62.5
BM-NAS (AAAI22)	Image + Text	62.92±0.03
DC-NAS (AAAI24)	Image + Text	63.70±0.11
CSG-NAS (IJCAI24)	Image + Text	64.12±0.12
KS-NAS(ours)	Image + Text	**66.57±0.08**

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

Method	Modality	Acc (%)
Unimodal methods		
Inflated ResNet-50 (CVPR18)	Video	83.91
Co-occurrence (IJCAI18)	Pose	85.24
Multi-modal methods		
Two-stream (NIPS14)	Video + Pose	88.6
GMU (ICLR17)	Video + Pose	85.8
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 (IJCAI24)	Video + Pose	**91.12±0.03**
KS-NAS(ours)	Video + Pose	91.11±0.04

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

backbone model for the text modality, and VGG Transfer as the backbone model for the RGB image modality. The evaluation metric used is the weighted F1 score, which is a reliable indicator for measuring multi-label classification performance due to the highly imbalanced nature of the dataset, rather than other types of F1 scores. The use of the weighted F1 score is also consistent with previous methods for easy comparison. For the parameters of our architecture, we set the population size to N = 20, the number of iterations to T = 20, the dimension of the fusion vector FD = 256, and the modality features are all reproducible. As shown in Table 2, KS-NAS outperforms existing multi-modal classification methods in terms of the weighted F1 score, surpassing MFAS, BM-NAS, DC-NAS, and CSG-NAS by 4.07%, 3.65%, 2.87%, and 2.45%, respectively.

For the NTU RGB-D dataset, to ensure a fair comparison of methods, we follow the setup of BM-NAS, using the dilated ResNet-50 as the backbone model for the video modality and Co-occurrence as the backbone model for the skeleton modality. This design ensures that all methods in the experiment share the same backbone network. For KS-NAS, the experimental settings are a population size of 20, 20 iterations, a fusion dimension of 256, and reusable modality features. As shown in Table 3, our method achieves a cross-subject accuracy of 91.11%. Compared to other methods, KS-NAS performs slightly worse than CSG-NAS but still outperforms most advanced multi-modal methods, surpassing MFAS, BM-NAS, and DC-NAS by 1.61%, 0.63%, and 0.26%, respectively.

Method	Modality	Acc (%)
Unimodal methods		
ResNext-101 (FG19)	RGB	93.75
VGG-16+LSTM (CVPR14)	Depth	77.7
C3D+LSTM+RSTTM	Depth	90.6
I3D (CVPR17)	Depth	89.47
ResNeXt-101 (FG19)	Depth	94.03
Multi-modal methods		
VGG-16+LSTM (CVPR17)	RGB + Depth	81.4
C3D+LSTM+RSTTM	RGB + Depth	92.2
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 (IJCAI24)	RGB + Depth	95.25±0.04
KS-NAS(ours)	RGB + Depth	**95.27±0.06**

Table 4: Gesture recognition results on EgoGesture dataset

For the EgoGesture dataset, to ensure a fair comparison of methods, we follow the setup of BM-NAS, using ResNeXt-101 as the backbone model for both the RGB and depth video modalities. This design ensures that all methods in the experiment share the same backbone network. We compare KS-NAS with various unimodal and multi-modal methods. The experimental settings for KS-NAS include a population size of 20, 20 iterations, a fusion dimension of 256, and reusable modality features. As shown in Table 4, our method achieves a cross-subject accuracy of 95.27%. Compared to other methods, KS-NAS achieves state-of-the-art classification performance.

In summary, KS-NAS, by sharing knowledge through a knowledge base, has improved model performance to a certain extent and provided new perspectives and implementation strategies for multi-modal classification tasks.

4.4 Search Efficiency Comparison

We analyze KS-NAS from three aspects: model parameters, search efficiency, and classification performance. To better validate the capability of KS-NAS, we compare it with various powerful multi-modal fusion methods, with experimental results shown in Table 5. From the table, it can be observed that although KS-NAS has a larger model size, it achieves superior classification performance while requiring the least search time.

On the CB and NUS datasets, the search efficiency of KS-NAS is nearly three and four times faster than the state-of-the-art method CSG-NAS, with better performance. On the MM-IMDB dataset, KS-NAS outperforms state-of-the-art methods in terms of search efficiency and classification performance, and has a small number of parameters. Although

Method	Dataset	Parameters	Time	CP (%)
EDF	NUS	0.31M	11.43	73.67
DC-NAS	NUS	0.53M	4.61	74.2
CSG-NAS	NUS	0.37M	2.71	74.52
KS-NAS(ours)	NUS	0.78M	**0.88**	**75.64**
EDF	CB	2.28M	78.01	88.48
DC-NAS	CB	2.41M	61.88	88.45
CSG-NAS	CB	2.47M	24.68	89.2
KS-NAS(ours)	CB	6.32M	**6.08**	**93.21**
BM-NAS	MM-IMDB	0.65M	1.24	62.94
DC-NAS	MM-IMDB	0.42M	1.19	63.7
CSG-NAS	MM-IMDB	0.56M	0.98	64.12
KS-NAS(ours)	MM-IMDB	0.50M	**0.81**	**66.57**
MMTM	NTU	8.61M	-	88.92
MFAS	NTU	2.16M	603.64	89.5
BM-NAS	NTU	0.98M	53.68	90.48
DC-NAS	NTU	0.26M	13.63	90.85
CSG-NAS	NTU	0.19M	5.19	**91.12**
KS-NAS(ours)	NTU	0.94M	**2.65**	91.11
BM-NAS	Ego	0.61M	20.67	94.96
DC-NAS	Ego	0.19M	4.57	95.22
CSG-NAS	Ego	0.20M	3.27	95.25
KS-NAS(ours)	Ego	2.71M	**1.30**	**95.27**

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

Version	KS	FS	Time	Acc(%)
KS-NAS$_1$	False	False	228.59	93.15±0.18
KS-NAS$_2$	True	False	7.26	93.12±0.32
KS-NAS	True	True	**6.08**	**93.21±0.15**

Table 6: Ablation study of KS-NAS (CB).

Version	KS	FS	Time	Acc(%)
KS-NAS$_1$	False	False	11.61	75.46±0.62
KS-NAS$_2$	True	False	1.04	74.61±0.65
KS-NAS	True	True	**0.88**	**75.64±0.48**

Table 7: Ablation study of KS-NAS (NUS).

on the NTU RGB-D dataset, our method's performance is slightly lower than CSG-NAS, its search efficiency is twice as fast, and its performance is still better than other advanced multi-modal fusion methods. On the EgoGesture dataset, KS-NAS also achieves the highest search efficiency and excellent classification performance. These results demonstrate the superiority of KS-NAS in terms of efficiency.

4.5 Ablation Study

To better analyze KS-NAS, we conducted ablation experiments on the CB and NUS datasets. The results are shown in Tables 6 and 7. They reveal that, when comparing the classification performance of KS-NAS$_1$ and KS-NAS$_2$, the knowledge-sharing (KS) is effective in improving search efficiency and reducing search time. As shown in Figure 3, under the same conditions, with the KS from the knowledge base, the number of training iterations required by individuals gradually decreases as the population evolves, leading to an improvement in search efficiency. However, due to the variability in the quality of stored knowledge, this approach can lead to local optima, resulting in suboptimal classification performance. The standard deviations indicate that KS-NAS$_2$ has lower stability, as the uneven quality of knowledge causes unfair training evaluations for different structures.

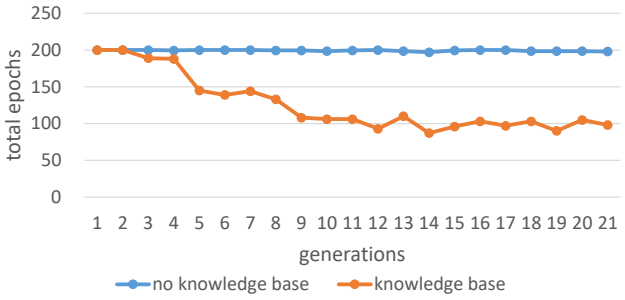

Figure 3: Comparison of epochs

Comparing the data between KS-NAS$_2$ and KS-NAS shows that the inclusion of a fair selection (FS) method improves search efficiency, classification results, and stability. This improvement is due to fair selection ensuring the quality of the knowledge in the knowledge base, allowing different structures to be trained to their specified limits in a short time with similar starting points, increasing the chances of finding better structures and enhancing classification performance. Finally, it is found that under the joint action of knowledge sharing and fair selection, the search efficiency is improved by ten times or even greater, and the classification performance is also improved to some extent.

5 Conclusion

In this paper, we propose a knowledge sharing-based neural architecture search (KS-NAS) method to reduce the time-consuming nature of current NAS-MMC methods during the training process. The KS-NAS method constructs a dynamically updatable knowledge base for knowledge sharing, enabling individuals to obtain initial parameters from the knowledge base and avoiding the need for training from scratch. High-quality individuals, in turn, contribute to enhancing the quality of the knowledge base, creating a mutually beneficial relationship between the two. This method significantly reduces computational overhead and improves the efficiency of existing population-based NAS-MMC methods. Finally, the experimental results on multiple popular multimodal tasks demonstrate that our method achieves state-of-the-art results in terms of classification performance and training efficiency. In the future, we will conduct in-depth research on improving the quality of the knowledge base, such as narrowing the search space based on core structures to control the direction of knowledge base updates and designing metrics to quantify the quality of the knowledge base numerically.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62406218, 62306171), Fundamental Research Program of Shanxi Province (No. 202203021222183), Science and Technology Major Project of Shanxi (No. 202201020101006), Open Project Foundation of Intelligent Information Processing Key Laboratory of Shanxi Province (No. CICIP2023005), Taiyuan University of Science and Technology Scientific Research Initial Funding (No. 20222106), Reward funds for outstanding doctor of work in coming to Jin (No. 20232029).

References

[Arevalo *et al.*, 2017] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[Balaprakash *et al.*, 2019] Prasanna Balaprakash, Romain Egele, Misha Salim, Stefan Wild, Venkatram Vishwanath, Fangfang Xia, Tom Brettin, and Rick Stevens. Scalable reinforcement-learning-based neural architecture search for cancer deep learning research. In *Proceedings of the international conference for high performance computing, networking, storage and analysis*, pages 1–33, 2019.

[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 IEEE conference on computer vision and pattern recognition*, pages 469–478, 2018.

[Carreira and Zisserman, 2017] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[Elsken *et al.*, 2019] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. Neural architecture search: A survey. *Journal of Machine Learning Research*, 20(55):1–21, 2019.

[Fu *et al.*, 2024] Pinhan Fu, Xinyan Liang, Tingjin Luo, Qian Guo, Yayu Zhang, and Yuhua Qian. Core-structures-guided multi-modal classification neural architecture search. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, Kate Larson (Ed.). International Joint Conferences on Artificial Intelligence Organization*, pages 3980–3988, 2024.

[Goodfellow *et al.*, 2013] Ian Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. In *International conference on machine learning*, pages 1319–1327. PMLR, 2013.

[Guo *et al.*, 2024] Qian Guo, Xinyan Liang, Yuhua Qian, Zhihua Cui, and Jie Wen. A progressive skip reasoning fusion method for multi-modal classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 429–437, 2024.

[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 (3DV)*, pages 289–297. IEEE, 2019.

[Han *et al.*, 2020] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2020.

[Han *et al.*, 2022] 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, 2022.

[Han *et al.*, 2024] Xiaolong Han, Yu Xue, Zehong Wang, Yong Zhang, Anton Muravev, and Moncef Gabbouj. Sadenas: A self-adaptive differential evolution algorithm for neural architecture search. *Swarm and Evolutionary Computation*, 91:101736, 2024.

[Hou *et al.*, 2019] Ming Hou, Jiajia Tang, Jianhai Zhang, Wanzeng Kong, and Qibin Zhao. Deep multimodal multilinear fusion with high-order polynomial pooling. *Advances in Neural Information Processing Systems*, 32, 2019.

[Jaafra *et al.*, 2019] Yesmina Jaafra, Jean Luc Laurent, Aline Deruyver, and Mohamed Saber Naceur. Reinforcement learning for neural architecture search: A review. *Image and Vision Computing*, 89:57–66, 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.*, 2023] Bingbing Jiang, Chenglong Zhang, Yan Zhong, Yi Liu, Yingwei Zhang, Xingyu Wu, and Weiguo Sheng. Adaptive collaborative fusion for multi-view semi-supervised classification. *Information Fusion*, 96:37–50, 2023.

[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 IEEE/CVF conference on computer vision and pattern recognition*, pages 13289–13299, 2020.

[Kim *et al.*, 2016] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. *arXiv preprint arXiv:1610.04325*, 2016.

[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 *2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019)*, pages 1–8. IEEE, 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. *arXiv preprint arXiv:1804.06055*, 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.*, 2022] 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.*, 2024] Xinyan Liang, Pinhan Fu, Qian Guo, Keyin Zheng, and Yuhua Qian. DC-NAS: Divide-and-conquer neural architecture search for multi-modal classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13754–13762, 2024.

[Liang *et al.*, 2025] Xinyan Liang, Pinhan Fu, Yuhua Qian, Qian Guo, and Guoqing Liu. Trusted multi-view classification via evolutionary multi-view fusion. In *Proceedings of the 13th International Conference on Learning Representations*, pages 1–14, 2025.

[Liu *et al.*, 2018] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*, 2018.

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

[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 network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4207–4215, 2016.

[Pérez-Rúa *et al.*, 2019] Juan-Manuel Pérez-Rúa, Valentin Vielzeuf, Stéphane Pateux, Moez Baccouche, and Frédéric Jurie. Mfas: Multimodal fusion architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6966–6975, 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 IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[Simonyan, 2014] Karen Simonyan. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Tang *et al.*, 2016] Jinhui Tang, Xiangbo Shu, Guo-Jun 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, 2016.

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

[Wang *et al.*, 2018] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1430–1439, 2018.

[Yang and Tian, 2014] Xiaodong Yang and YingLi Tian. Super normal vector for activity recognition using depth sequences. In *Proceedings of the 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.

[Ye *et al.*, 2022] Peng Ye, Baopu Li, Yikang Li, Tao Chen, Jiayuan Fan, and Wanli Ouyang. b-darts: Beta-decay regularization for differentiable architecture search. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10874–10883, 2022.

[Yin *et al.*, 2022] Yihang Yin, Siyu Huang, and Xiang Zhang. Bm-nas: Bilevel multimodal neural architecture search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 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.

[Yu *et al.*, 2022] Hongyuan Yu, Houwen Peng, Yan Huang, Jianlong Fu, Hao Du, Liang Wang, and Haibin Ling. Cyclic differentiable architecture search. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):211–228, 2022.

[Zadeh *et al.*, 2017] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 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*, 20(5):1038–1050, 2018.