Gumbel Rao Monte Carlo based Bi-Modal Neural Architecture Search for Audio-Visual Deepfake Detection

Aravinda Reddy PN[1]

Raghavendra Ramachandra[2]

Krothapalli Sreenivasa Rao[1]

Pabitra Mitra[1]

[1] Indian Institute of Technology Kharagapur, India

[2] Norwegian University of Science and Technology, Gjøvik Norway

Abstract

Deepfakes pose a critical threat to biometric authentication systems by generating highly realistic synthetic media. Existing multimodal deepfake detectors often struggle to adapt to diverse data and rely on simple fusion methods. To address these challenges, we propose Gumbel-Rao Monte Carlo Bi-modal Neural Architecture Search (GRMC-BMNAS), a novel architecture search framework that employs Gumbel-Rao Monte Carlo sampling to optimize multimodal fusion. It refines the Straight through Gumbel Softmax (STGS) method by reducing variance with Rao-Blackwellization, stabilizing network training. Using a two-level search approach, the framework optimizes the network architecture, parameters, and performance. Crucial features are efficiently identified from backbone networks, while within the cell structure, a weighted fusion operation integrates information from various sources. By varying parameters such as temperature and number of Monte carlo samples yields an architecture that maximizes classification performance and better generalisation capability. Experimental results on the FakeAVCeleb and SWAN-DF datasets demonstrate an impressive AUC percentage of 95.4%, achieved with minimal model parameters.

1 Introduction

The advancement of deep generative models [12] has brought about highly convincing synthetic audio and visuals, posing significant security risks. These technologies can potentially circumvent biometric systems that depend on unique individual characteristics. For instance, visual deepfakes employ techniques to change facial features, simulate wrongful acts, and modify appearances. Moreover, the latest developments in deepfake technology have made it possible to replicate human voices in real-time [4]. Techniques for cloning voices use neural networks to create speech that sounds strikingly similar to a specific person, which complicates the reliability of authentication systems and opens up possibilities for impersonating public figures and committing financial deception.

Neural Architecture Search (NAS) [14] identifies optimal neural network designs within a predefined architecture space. Recent multimodal NAS (MMNAS) [24] explores attention

mechanisms but relies on static network topologies. Recently [23] proposed a multimodal DNN architecture that combined feature-level fusion with individual feature selection using Softmax. he impact of varying sample numbers on medical image analysis has been explored using the Gumbel-Softmax distribution within the NAS framework [2]. Recently [19] used Straight through Gumbel-Softmax based estimator-based bimodal NAS for audio-visual fake detection with reduced model parameters. However, the STGS-BMNAS suffers from high variance introduced by the Gumbel noise often suffers from unstable training dynamics.

The aim of this work is to develop a highly stable automatic architecture for audio-visual deepfake detection. So we propose Gumbel-Rao Monte carlo based bi-modal neural architecture search (GRMC-BMNAS) which adaptively learns architectures from a pool of operations for audio-visual deepfake detection and trains faster and offers better results on the test set performance. GRMC-BMNAS adopts a two-level search similar to [19] where it learns unimodal features from the backbone network by sampling the search space by varying the temperature parameter and Monte Carlo samples. In the second-level search, we utilize the weighted fusion strategy by varying the temperature and Monte Carlo samples. Increasing Monte Carlo samples expands the search space of primitive operations, allowing for a more accurate selection based on softmax probabilities as shown in Figure 1. As illustrated in Figure 2, the average entropy of GRMC-BMNAS consistently outperforms both STGS-BMNAS and the standard Softmax baseline [23]. This indicates that GRMC-BMNAS achieves a lower entropy, suggesting a faster convergence during training. Our proposed framework matches the performance of STGS-BMNAS [19] while requiring significantly less training time and computational resources (GPU days). Moreover, our model demonstrates superior generalization on test data. The main contributions of this paper are as follows:

- To achieve faster, generalizable design of automatic architecture for bi-modal learning (extendable to multimodal learning), we propose an automatic approach named Gumbel-Rao Monte Carlo approximation based NAS for audio-visual deepfake detection which adopts two level schema.
- The GRMC-BMNAS is an end-to-end framework which is fully searchable using two level schema. The Gumbel-Softmax trick uses Gumbel noise to approximate categorical samples in a differentiable manner. In GRMC, multiple Gumbel noise samples are drawn to enhance this approximation. Monte Carlo estimation averages these samples to approximate expectations, while Rao-Blackwellization conditions on discrete outcomes to reduce variance, improving the estimator's efficiency.
- Our study assessed the GRMC-BMNAS model for audio-visual deepfake detection through extensive experiments. Empirical evidence indicates that our model trains faster and has fewer parameters compared to existing state-of-the-art models.

The rest of the paper is organized as follows. Section 2 discusses about the related work, Section 3 presents proposed work, Section 4 discusses about datasets, experimental protocol, architecture search and evaluation and results, and Section 6 concludes the paper.

2 Related work

In recent years, deep learning techniques have been extensively used to create convincing fake videos by altering both visual and auditory components. Notable research includes methods like the Siamese Network proposed in Emotions Don't Lie, [16] which compares affective cues from both modalities within a video. The Modality Dissonance Score (MDS) network, introduced in Not Made for Each Other [5], highlights differences in audio-visual

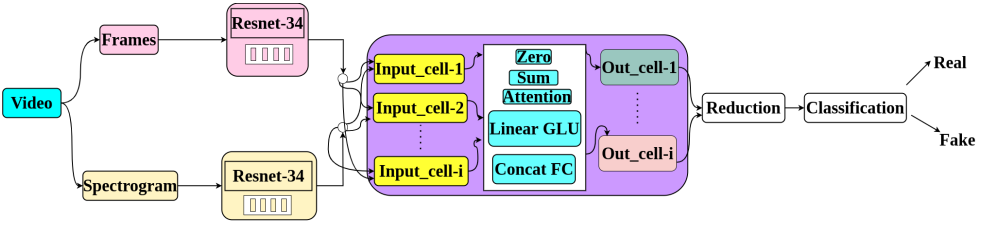

Figure 1: Block diagram showing the proposed GRMC-BMNAS employs a two-stage search to optimize bimodal fusion. The first stage identifies crucial features, while the second stage determines the optimal architecture using a pool of operations.

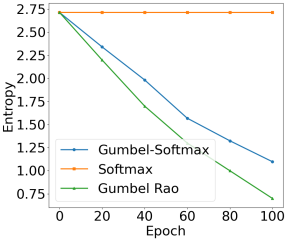

Figure 2: Average entropy plot for two learnable parameters for the proposed GRMC-BMNAS, and the existing STGS-BMNAS [19] and Softmax [23].

pairs using contrastive loss. Additionally, approaches focusing on phoneme-viseme [1] mismatch reveal how deepfake techniques struggle to accurately replicate mouth shape dynamics corresponding to specific sounds. Techniques like multimodal trace extracts [20] and cross-modal learning further enhance deepfake detection by analyzing audio-visual correspondences. Some methods even rely on self-supervised learning [8] and anomaly detection using unlabeled real data. Overall, these advancements contribute to better identifying inconsistencies in deepfake videos. Recently [17] employs a two-stage approach. First, self-supervised learning extracts features from real videos. Subsequently, these features are fine-tuned for deepfake classification using supervised learning.

3 Proposed method: GRMC-BMNAS

This work introduces a new framework called GRMC-BMNAS for deepfake detection, which optimizes network exploration by sampling from the Gumbel distribution and using the Monte Carlo approximation to average these samples. Rao-Blackwellized conditions on discrete outcomes reduce variance and improve estimator efficiency. At the first level, features from the backbones are sampled and cells are explored within a directed acyclic graph (DAG). A cell is a DAG consisting of ordered sequences on a node, where each node is a latent representation with directed edges linked to primitive operations that transform the node. The second level involves a DAG of nodes within a cell, each representing an operation chosen from a predefined pool.

3.1 Gumbel distribution

The Gumbel distribution, also known as Type I within the generalized extreme value distributions, is tailored for modeling extreme events and anomalies. A ‘Gumbel’ random variable, which adheres to this distribution, is characterized by a duo of parameters: location parame-

ter $\mu \in \mathbb{R}$ and non-negative scale parameter $\beta \in \mathbb{R}_{\geq 0}$. The corresponding probability density and cumulative density functions are given by:

$$f(x) = \frac{1}{\beta} e^{-\frac{x-\mu}{\beta}} e^{-e^{-\frac{x-\mu}{\beta}}} \quad (1) \qquad F(x) = e^{-e^{-\frac{x-\mu}{\beta}}} \quad (2)$$

$$F^{-1}(u) = -\beta log(-log(u)) + \mu \quad (3)$$

The inverse cumulative density function (ICDF) is also called quantile function given by equation 3 and equation 3 is used in inverse transform sampling to transform sample from Uniform distribution $U(0,1)$ into a Gumbel via a double logarithmic relation.

3.2 Gumbel-max trick

The Gumbel-max technique is a strategy for drawing samples from a categorical random variable denoted by $I \sim Cat(\pi)$. It involves the addition of Gumbel-distributed noise, which is independent and identically distributed, to the log probabilities before normalization. More specifically $I = argmax_{i \in D} \{G^{(i)} + log\, \theta_i\} \sim Cat(\pi)$, where $G^{(1)}, G^{(2)}, G^{(3)} ... G^{(D)}$ are the i.i.d samples drawn from Gumbel distribution 3.

3.3 Gumbel-Softmax distribution

Instead of producing discrete or 'hard' samples from a categorical distribution that lacks structure, one can create 'soft' samples, which are especially beneficial for estimating gradients. To grasp the relationship between these hard and soft samples, it's essential to analyze the hard samples when they are expressed in their one-hot encoded form, that is, $⊮_\omega \in \{0,1\}^N$. Then $z = onehot(argmax_{i \in D} \{G^{(i)} + log\, \theta_i\})$

From [10, 15] we derive PDF of this distribution and denoted by $GS(\pi, \lambda)$. More specifically, the i^{th} index of soft sample $S_\lambda \in \{\mathbb{R}^N_{\geq 0} : |S_\lambda| = 1\}$ is defined in 4:

The temperature parameter λ in the Gumbel-Softmax distribution modulates its entropy and that of its samples. It serves as a measure of how much the soft sample S_λ deviates from a sample from $Cat(\pi)$. As λ trends towards zero, the distribution samples shift towards one-hot representations, aligning the Gumbel-Softmax distribution closely with the categorical distribution.

Continuous one-hot vector relaxation excels in learning representations and sequences. However, for tasks requiring discrete values, such as reinforcement learning actions, compressed data, or architecture search, we discretize the continuous output using argmax. In this family, the forward computation of f is unchanged, but backpropagation is computed "through" a surrogate. This surrogate is known as straight through gumbel softmax (STGS) defined in equation 5.

$$S_{i;\lambda} = \frac{exp((log\, \theta_i + G^{(i)})/\lambda)}{\sum_{j \in D} exp((log\, \theta_i) + G^{(j)})/\lambda)} \quad (4) \qquad \nabla_{STGS} := \frac{\partial f(S_\lambda)}{\partial S_\lambda} \frac{\partial S_\lambda}{\partial \phi} \quad (5)$$

where $S_\lambda = softmax(\phi + G)$ with $\phi = log\theta$ and G is the i.i.d Gumbel variable.

3.4 Gumbel-Rao Monte Carlo Estimator (GRMC)

The GR estimator aims to reduce the variance associated with GS-ST by introducing a local expectation.

$$\nabla_{GR} = \frac{\partial f(D)}{\partial D} E\left[\frac{dsoftmax_\tau(\Theta+G)}{d\Theta}|D\right] \quad (6) \qquad \theta_j + G_j|D = \begin{cases} -log(E_j) + logZ(\theta) & \text{if j=i} \\ -log(\frac{E_j}{exp(\theta_j)} + \frac{E_i}{Z(\theta)}) & \text{o.w.} \end{cases} \quad (7)$$

However, this expectation is in multiple variables and is not analytically tractable. Hence [18] used Monte Carlo integration with K samples from $G|D$.

$$\nabla_{GRMCK} = \frac{\partial f(D)}{\partial D}\left[\frac{1}{K}\sum_{k=1}^{K}\frac{dsoftmax_\tau(\theta+G^k)}{d\theta}\right] \quad (8)$$

where $G^k \sim \theta + G|D$ i.i.d. using the reparameterization and K is the number of similar distributions or sampling size. Note that the total derivative $dsoftmax_\tau(\theta+G^k)/d\theta$ is taken through both θ and G^k.

3.5 Analysing our proposed GRMC-BMNAS

To better understand our proposed method, we establish three propositions:

1. **Proposition 1:** For the two learnable parameters α (first level search) and γ (second level search) the Gumbel-Rao Monte Carlo (GRMCK) estimator yields lower variance compared to the Straight-Through Gumbel-Softmax (STGS) estimator then:

   Let the estimators denoted by (6) and (7) be ∇_{STGS} and ∇_{GRMCK}. Let $\nabla_\alpha = dE[f(D)]/d\alpha$, $\nabla_\gamma = dE[f(D)]/d\gamma$ represent the actual gradient we are attempting to estimate. Then for all values $K \geq 1$, we have

   $$E\left[\|\nabla_{GRMCK} - \nabla_\alpha\|^2\right] \leq E\left[\|\nabla_{STGS} - \nabla_\alpha\|^2\right] \quad (9) \qquad E\left[\|\nabla_{GRMCK} - \nabla_\gamma\|^2\right] \leq E\left[\|\nabla_{STGS} - \nabla_\gamma\|^2\right] \quad (10)$$

2. **Proposition 2:** Increasing the number of Monte Carlo samples K in the GRMCK estimator further reduces mean squared error compared to STGS. Let θ^K_{GRMCK} be the GRMCK estimator with K samples, θ_{STGS} be the STGS estimator, θ be the true gradient, $Var(\theta^K_{GRMCK})$ be the variance of GRMCK with K samples, $Var(\theta_{STGS})$ be the variance of STGS estimator. Then, the mean squared error (MSE) for both estimators can be expressed as follows: $MSE(\theta^K_{GRMCK}) = \mathbb{E}(\theta^K_{GRMCK} - \theta^2)$, $MSE(\theta_{STGS}) = \mathbb{E}[\theta_{STGS} - \theta^2]$, then by Jensen inequality we have

   $$\mathbb{E}[(\theta^K_{GRMCK} - \theta^2)] \leq Var(\theta^K_{GRMCK}) \quad (11)$$

3. **Proposition 3:** Let θ_{STGS} be the gradient estimator using the STGS estimator and θ^K_{GRMCK} be the gradient obtained using GRMC estimator with K samples. Let θ denote the true gradient. The asymptotic biases of two estimators are given by

   (a) **Asymptotic bias of STGS estimator:**
   The STGS estimator generally retains a non-zero bias as the number of samples from gumbel distribution increases

   (b) **Asymptotic bias of GRMC estimator:**
   The GRMC estimator designed to reduce variance and improve the gradient estimation, has an aymptotic bias that approches zero as the number of Monte Carlo sample K increases. Formally,

   $$\lim_{K\to\infty} \mathbb{E}[\theta_{STGS}] - \theta \neq 0 \quad (12) \qquad \lim_{K\to\infty} \mathbb{E}[\theta_{STGS}] - \theta = 0 \quad (13)$$

   (c) **Comparative statement:** $|\lim_{K\to\infty} \mathbb{E}[\theta_{GRMC}] - \theta| < |\lim_{K\to\infty} \mathbb{E}[\theta_{STGS}] - \theta|$

3.6 Modality feature extraction

Similar to [19] we employ pre-trained ResNet-34 models for both image (facial) and speech feature extraction. Instead of using the final output, we extract features from intermediate layers of the neural network to capture richer and more abstract representations of the input data. Instead of using the final output, we extract features from intermediate layers of the neural network to capture richer and more abstract representations of the input data.

3.6.1 First level search: GRMC relaxation over the cells

We extract single-modal features from pre-trained backbone networks for both image (I) and speech (S) cues. These extracted features are denoted as I^i and S^i respectively. Then we formulate the first level nodes in a sequence. Then $\mathbb{F} = \{I^{(1)}, I^{(2)}..., I^{(N_A)}, S^{(1)}, S^{(2)}..., S^{(N_B)}... Cell^{(1)}..., Cell^{(N)}\}$ Let $\mathbb{F}^a$, $\mathbb{F}^b$ be any two nodes from $\mathbb{F}$. Let α be the weight parameter connecting between $\mathbb{F}^{(a)}$, $\mathbb{F}^{(b)}$ then each edge is selected based on the unary operation. Let $\mathbb{O}^F$ be the set of candidate operations

$$\mathbb{O}^F = \begin{cases} Identity(x) = x & \text{selecting an edge} \\ Zero(x) = 0 & \text{discarding an edge} \end{cases}$$

where each operation refers to a function o. to be applied on the $cell^{(a)}$ then by applying the gumbel rao.

$$\bar{o}(a,b)_\lambda = \sum_{o \in O} \left[\frac{1}{K} \sum_{k=1}^{K} \frac{exp(\alpha_o^{(a,b)} + G_{(a)}^k | D)}{\sum_{o' \in O} exp(\alpha_o + G_{(b)}^k | D)} \right] o(x) \tag{14}$$

$$\nabla_{GRMCK} = \frac{\partial f(\bar{o}(a,b))}{\partial \bar{o}(a,b))} \frac{\partial \bar{o}(a,b)}{\partial \phi} \tag{15}$$

where K = sampling size which influences the entropy of the Gumbel Rao distribution and $\phi = log(\alpha)$.

A cell is densely connected and receives input from all its predecessors $o^v = \sum_{u<v} \bar{o}^{(u,v)}(o^{(i)})$. In the evaluation stage, since we want deterministic predictions, the probabilities obtained from the Gumbel Rao distribution can be directly used without the need for sampling or argmax operation as, $(a,b) = \alpha_o(a,b)$. Using softmax probabilities during evaluation provides deterministic predictions without relying on sampling. Softmax offers probabilistic interpretations and is more robust to noise compared to the deterministic argmax approach.

3.6.2 Second level: Weighted fusion

Following the approach in [19], we employ the same predefined candidate operations. Each operation takes two input tensors x, y, and produces an output tensor z, all of which have dimensions $\mathbb{R}^{N \times C \times L}$. These operations are detailed in Table 1.

The second level of GR-BMNAS optimizes a weighted fusion strategy within a cell structure. A cell is a directed acyclic graph composed of nodes representing latent representations and edges representing operations. The cell's architecture is defined by edge and operation configurations, while weight parameters are learned during optimization.

Weighted fusion strategy: In this stage, the inner structure of $Cells^{(n)}$ is an ordered sequence of $\mathbb{C}_n$ then $\mathbb{C}_n = I, S, N^{(1)}, ...N^{(M)}$

A cell comprises three nodes: an input node $in_c^{(i)}$ and two intermediate nodes $c^{(j)}$, $c^{(l)}$. The input node processes the backbone network's output and generates two intermediate representations using a weighted fusion of candidate operations determined by the Gumbel-

Operation	Function
Zero(x,y)	The Zero operation, eliminates an entire node, effectively discarding its contribution .
Sum(x, y):	The DARTS framework [14], introduced , employs a method to combine two features using summation. $Sum(x,y) = X + Y$
Attention(x,y)	The Attention operation, as described in [21], employs scaled dot-product attention, where a query x and key-value pairs y are used. $Attention(x,y) = Softmax(xy^T/(\sqrt{C} \times y))$
LinearGLU(x,y)	The LinearGLU operation combines two inputs x,y, using a linear layer followed by the gated linear unit (GLU) activation [6]. $LinearGLU(x,y) = xW_1 \bigodot Sigmoid(yW_2)$
ConcatFC(x,y)	The ConcatFC operation involves concatenating two inputs, (x,y) and and passing the concatenated vector through a fully connected (FC) layer with ReLU activation. $ConcatFC(x,y) = ReLU((x,y).W + b)$

Table 1: Candidate operations used in the second level search

Rao Monte Carlo method.

$$= \sum_{o^s \in \mathbb{O}^s} \left[\frac{1}{K} \sum_{k=1}^{K} \frac{exp((\gamma^{(i)} + G^k_{(i)})/D)}{\sum_{o' \in \mathbb{O}^s} exp((\gamma^{(i)}_{\bar{o}^s} + G^k_{(i)})/D)} \right] \times w_i(f(c^{(j)}, c^l)) \tag{16}$$

The weights, γ determine the contribution of candidate operations. During evaluation, we directly use the probabilities from the Gumbel-Rao distribution for decision-making, eliminating the need for sampling or selecting the maximum value $o^{(i)} = \gamma^{(i)}_c$. The edge weights (β) are also relaxed using straight-through gumbel rao similar to the first level. The output node combines the results from all transformation nodes.

3.7 Optimizing Neural Architectures through Parameter learning

We employ our proposed GRMC method to jointly optimize both weight parameters and architecture in an end-to-end training process. The objective of architecture search is to minimize the loss function, $\mathbb{L}_\omega$ while simultaneously reducing the number of model parameters i.e., $\min_{\omega,\alpha,\gamma} \mathbb{E}_{\mathbb{A} \sim p_{(\alpha,\gamma)_{\lambda,K}}} |\mathbb{L}_\omega(\mathbb{A}))|$. The primary objective is to minimize the expected performance of architectures sampled from the search space i,e., $p_{(\alpha,\beta,\gamma)_{\lambda,K}}(\mathbb{A})$. Our method involves sampling network architectures from a distribution parameterized by α, β, and γ, controlled by a temperature parameter λ and Monte carlo samples K. The loss is computed for the sampled architecture, and gradients are calculated with respect to both architecture parameters and network weights using a straight-through estimator. By optimizing these parameters, we aim to find an optimal architecture with minimal parameters.

4 Experiments and Results

4.1 Datasets

FakeAVCeleb: We evaluate our method on the FakeAVCeleb dataset [11], containing 19,500 fake and 500 real videos of 500 celebrities.
SWAN-DF dataset [13] is the first publicly available collection of high-quality audio-visual deepfakes, built upon the SWAN database of real-world videos. It contains 24,000 fake and 2,800 real video samples. More details regarding the dataset split is given in the supplementary material.

4.2 Evaluation methodology

We employ a two-pronged evaluation strategy. First, we assess model performance on a combined dataset of FakeAVCeleb and SWAN-DF. Second, we evaluate generalization by training on one dataset and testing on the other. Since both databases are biased towards fake videos than the real videos.To mitigate dataset bias, we apply 36 different data augmentation techniques, resulting in 50,742 training, 10,718 validation, and 8,963 test samples similar to [19]. Detailed information regarding dataset partitioning and augmentation techniques can be found in the supplementary material.

4.3 Architecture search and evaluation

Our experiments involve a two-stage operation selection process. Initially, edges are selected or discarded from a pool $\mathbb{O}^F$. Subsequently, operations from a different pool $\mathbb{O}^S$ are considered. The number of monte carlo samples (K) is varied across different temperature parameter λ settings for larger and optimal search space for lower training loss (see training and validation loss graphs for different monte carlo samples and different temperature values in supplementary material). The algorithm stops when the selection of operations within the neural cell stabilizes for both learnable parameters, α and γ which is measured as $E(\alpha) = -\sum_{a,b}\sum_{o\in\mathbb{O}^F} \alpha^o_{(a,b)}\, log(\alpha^o_{ab})$ and $E(\gamma) = -\sum_{a,b}\sum_{o\in\mathbb{O}^S} \alpha^o_{(a,b)}\, log(\gamma^o_{ab})$. Experiments were conducted on V100 Tesla GPUs with 16GB memory. The model was trained using PyTorch with Adam optimizer, a batch size of 8 for 100 epochs, and specific hyperparameters for learning rates, weight decay, and momentum (more details in the supplementary material). The optimal architecture determined through the search process, is trained for 100 epochs with a batch size of 64. Its performance is then evaluated on a held-out test set.

4.4 Performance comparision

We assess model performance using standard metrics: Area Under the Curve (AUC) and classification accuracy (ACC) similar to [19].

SOTA models: We compare our method to state-of-the-art audio-visual deepfake detection techniques, including Not made for each other [5], Voice-Face [3], Audio-Visual Anomaly detection [8], ID-Reveal [6], Multimodal-trace [20], Ensemble learning [9], and POI-AV [21] and STGS-BMNAS [19].

Training: To ensure fair comparison, all models were trained on our dataset using identical pre-processing steps and adhering to strict data partitioning to prevent overlap between training, validation, and testing sets.

Table 2 presents a comparison of our proposed GRMC-BMNAS model with SOTA methods using a combined dataset. Our model surpases the recent SOTA models POI-AV [21], Multimodaltrace [20], and ID-Reveal [6], STGS-BMNAS [19] on combined datasets in terms of accuracy (ACC) and area under the curve (AUC) metrics, while using fewer model parameters. Our model exhibits lower variance compared to STGS-BMNAS, supporting Proposition 1 (Figure given in the supplementary material). Additionally, our model achieves lower mean squared error (MSE) for learnable parameters α and γ, corroborating Proposition 2 (For figure see supplementary material). Optimal architecture with minimal GPU days is obtained with K=100 and $\lambda = 0.1$ is shown in Figure 5 . Receiver operating characteristic (ROC) curves for $\lambda = 0.1$ and K=100, distinguishing between real and fake data, are provided in the supplementary material.

Model performance on unseen data: Table 3 presents the performance of our model on unseen data. Our model significantly outperforms STGS-BMNAS demonstrating its superior

Method	AUC(%)	ACC(%)	Params(M)	GPU days	Search
Voice-face [8]	82	86	174	-	Gradient
Audio-visual anamoly detection [8]	93	-	41	-	Gradient
Not made for each other [8]	81	84.56	122	-	Gradient
ID-Reveal [8]	78	80.1	7.3	-	Gradient
MultimodalTrace [21]	84	91.26	15	-	Gradient
Ensemble-learning [8]	84	86	12	-	Gradient
POI-AV [21]	93.9	90.9	-	-	-
BM-NAS [23]	92.26	91.4	0.62	4	Gradient
STGS-BMNAS[23]	94.4	95.5	0.26	2	Straight through estimator
GRMC-BMNAS (Ours)	**95.5**	**96.5**	**0.20**	1.5	**Straight through estimator**

Table 2: Comparison of our proposed GRMC-BMNAS with SOTA approaches tested on our test data

Trained on ↓	Tested on ↓							
	GRMC-BMNAS				STGS-BMNAS			
	FakeAVCeleb		SWAN-DF		FakeAVCeleb		SWAN-DF	
	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC
FakeAVCeleb	94.7	93.5	91.6	91.2	92.7	91.8	85.6	84.7
SWAN-DF	90.8	91.2	95.1	94.8	84.8	83.2	93.1	92.8

Table 3: Generalisation of our proposed model to the seen and unseen data

Temperature (λ)		No of Samples (K)		
		10	100	1000
$\lambda = 0.1$	AUC	92.16	95.5	96.96
	Model parameters	341760	205565	189574
$\lambda = 0.5$	AUC	91.75	94.04	90.45
	Model parameters	322456	192452	175642
$\lambda = 1.0$	AUC	91.16	93.95	90.2
	Model parameters	299490	187852	167845

Table 4: Evaluation of searched architecture with different temperature values and with varying Monte carlo samples

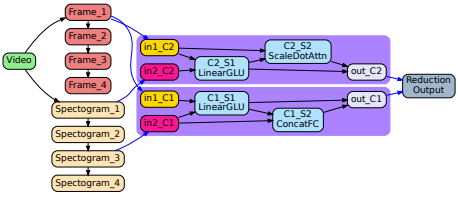

Table 5: Best architecture obtained wit K=100 and $\lambda = 0.1$

generalization capabilities. This improved performance can be attributed to the model's reduced variance and lower mean squared error as established in previous sections.

4.5 Ablation study

Table 4 presents the outcomes of an ablation study examining the influence of temperature and Monte Carlo samples on model performance. Consistent with Proposition 2, increasing the number of Monte Carlo samples generally leads to smaller model sizes and higher AUC values, albeit at the expense of increased computational cost. Based on these findings, an optimal architecture was determined with $\lambda = 0.1$ and K=100. Respective architectures produced using different parameter settings can be found in the supplementary material.

5 Conclusion

This paper introduces GRMC-BMNAS, a novel architecture search method for audio-visual deepfake detection. Our approach leverages a two-stage Gumbel-Rao Monte Carlo sampling process to efficiently discover optimal architectures. By reducing variance and mean squared error, GRMC-BMNAS surpasses existing methods like STGS-BMNAS in both training efficiency and generalization performance.

References

[1] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 660–661, 2020.

[2] Jianlong Chang, Xinbang Zhang, Yiwen Guo, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Differentiable architecture search with ensemble gumbel-softmax. *arXiv preprint arXiv:1905.01786*, 2019.

[3] Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Xiaojun Chang, and Liqiang Nie. Voice-face homogeneity tells deepfake. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–22, 2023.

[4] Akash Chintha, Bao Thai, Saniat Javid Sohrawardi, Kartavya Bhatt, Andrea Hickerson, Matthew Wright, and Raymond Ptucha. Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1024–1037, 2020.

[5] Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other-audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM international conference on multimedia*, pages 439–447, 2020.

[6] Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15108–15117, 2021.

[7] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pages 933–941. PMLR, 2017.

[8] Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10491–10503, 2023.

[9] Ammarah Hashmi, Sahibzada Adil Shahzad, Wasim Ahmad, Chia Wen Lin, Yu Tsao, and Hsin-Min Wang. Multimodal forgery detection using ensemble learning. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1524–1532. IEEE, 2022.

[10] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[11] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S Woo. Fakeavceleb: A novel audio-video multimodal deepfake dataset. *arXiv preprint arXiv:2108.05080*, 2021.

[12] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018.

[13] Pavel Korshunov, Haolin Chen, Philip N Garner, and Sébastien Marcel. Vulnerability of automatic identity recognition to audio-visual deepfakes. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2023.

[14] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

[15] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

[16] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2823–2832, 2020.

[17] Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. Avff: Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27102–27112, 2024.

[18] Max B Paulus, Chris J Maddison, and Andreas Krause. Rao-blackwellizing the straight-through gumbel-softmax gradient estimator. *arXiv preprint arXiv:2010.04838*, 2020.

[19] Aravinda Reddy PN, Raghavendra Ramachandra, Krothapalli Sreenivasa Rao, Pabitra Mitra, and Vinod Rathod. Straight through gumbel softmax estimator based bimodal neural architecture search for audio-visual deepfake detection. *arXiv preprint arXiv:2406.13384*, 2024.

[20] Muhammad Anas Raza and Khalid Mahmood Malik. Multimodaltrace: Deepfake detection using audiovisual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 993–1000, 2023.

[21] Chang-Sung Sung, Jun-Cheng Chen, and Chu-Song Chen. Hearing and seeing abnormality: Self-supervised audio-visual mutual learning for deepfake detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[23] 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.

[24] Zhou Yu, Yuhao Cui, Jun Yu, Meng Wang, Dacheng Tao, and Qi Tian. Deep multimodal neural architecture search. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3743–3752, 2020.