PHYSICS-INSPIRED FEATURE FUSION FOR PLATE PARAMETER ESTIMATION FROM ACOUSTIC IMPULSE RESPONSES

Zhenyu Guo[1], Yun Zhang[1], Liangming Chen[2], Wei Liu[1], and Gongping Huang[1]

[1]Electronic Information School, Wuhan University, Wuhan, Hubei 430072, China
[2]College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China
zhenyuguo@whu.edu.cn, yun_zhang@whu.edu.cn, clm267432@gmail.com,
wei_liu@whu.edu.cn, gongpinghuang@whu.edu.cn

ABSTRACT

Estimating physical plate parameters from impulse responses is a challenging inverse problem. Task A of the first Digital Audio Effects Parameter Estimation Challenge requires the recovery of six identifiable parameters from displacement impulse responses. In this work, we propose a physics-inspired feature fusion network (PIFFN) that combines a pretrained convolutional backbone with a 15-dimensional physics-inspired feature vector computed from the impulse response. These physics-inspired features describe amplitude scale, temporal decay, and spectral structure without relying on modal-distribution priors. The proposed model is evaluated on the official validation set, achieving an overall normalized mean squared error of 0.00362. Compared with the official particle swarm optimization baseline and backbone-only model, PIFFN shows a clear performance improvement, demonstrating its effectiveness for plate parameter estimation.

1. INTRODUCTION

Plate reverberation is a classic form of artificial reverberation generated through the vibration of a thin metal plate [1, 2, 3]. Its impulse response is governed by a dense set of resonant modes whose frequencies, decay characteristics, and modal gains depend on the plate material, geometry, tension, and observation position [4]. Although forward simulation of plate vibration is well established [5], recovering the underlying physical parameters from an observed impulse response remains a challenging inverse problem. Task A of the Digital Audio Effects Parameter Estimation Challenge addresses this problem [6]. Given a displacement impulse response, the objective is to estimate six identifiable parameters: the mass per unit area μ, normalized bending stiffness D/μ, normalized tension T_0/μ, plate length L_y, and output position (x_o, y_o).

Previous studies have investigated inverse parameter estimation for thin plates using boundary measurements. For linear, inhomogeneous, isotropic thin plates, Young's modulus and Poisson's ratio can be recovered from boundary information [7]. Global recovery results have also been established for bending stiffness and other elastic parameters of Kirchhoff-Love plates using boundary

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62471340 and in part by China Scholarship Council (CSC) under Grant 202506270089.

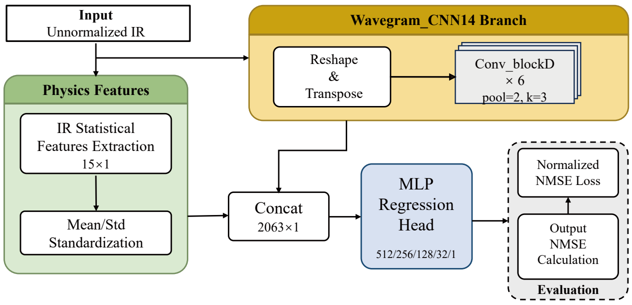

Figure 1: An overview of the proposed PIFFN architecture.

Cauchy data [8]. In addition, mechanical parameters of fiber metal laminates have been identified by matching modeled and measured frequency response functions [9]. Differentiable modal modeling has been extended to strings, membranes, and plates, enabling the estimation of tension, stiffness, and geometric parameters [10]. Related differentiable audio models further support physical parameter estimation, geometric reasoning, and impact position prediction from acoustic observations [11]. More recently, physics-informed neural networks have provided an alternative framework for related inverse problems [12]. Physics-informed constraints have been applied to structural stiffness identification [13], while the governing equation of Kirchhoff-Love plates has been embedded into such networks for plate damage identification from measured flexural guided wavefields [14].

To address this parameter estimation problem, we propose a Physics-Inspired Feature Fusion Network (PIFFN). The proposed method combines a pretrained audio backbone with physics inspired impulse response features that characterize amplitude scale, temporal decay, and spectral structure. Experimental results show that PIFFN achieves lower normalized mean squared errors than both the official optimization based baseline and the backbone-only baseline.

2. PROPOSED METHOD

An overview of the proposed PIFFN architecture is shown in Fig. 1, followed by details on the preprocessing, network design and training strategy.

2.1. Preprocessing

For Task A, the proposed system estimates the six identifiable plate parameters from the official impulse response. The displacement

Table 1: Physics-inspired impulse-response features.

Group	Feature	Definition
Amplitude scale	Logarithmic peak amplitude	$\log(\max_n \lvert x[n]\rvert + \epsilon)$
	Logarithmic RMS amplitude	$\log(r+\epsilon)$, $r = \sqrt{N^{-1}\sum_n x^2[n] + \epsilon}$
	Logarithmic signal energy	$\log\left(\sum_{n=0}^{N-1} x^2[n] + \epsilon\right)$
Spectral structure	Spectral centroid	$\sum_k f_k M_k \,/\, \sum_k M_k$
	Spectral bandwidth	$\sqrt{\sum_k (f_k - f_c)^2 M_k \,/\, \sum_k M_k}$
	Spectral roll-off frequency	$f_r = \min\left\{ f_k \,\middle\vert\, \sum_{j:f_j \le f_k} P_j \ge 0.85 \sum_j P_j \right\}$
	Spectral flatness	$\mathrm{GM}(M_k+\epsilon) \,/\, (\mathrm{AM}(M_k)+\epsilon)$
	Dominant peak frequency	$f_{\text{peak}} = f_{k^*}$, where $k^* = \arg\max_{k>0} M_k$
	Low-band energy ratio	$E_{\mathcal{B}_l} \,/\, \sum_k P_k$
	Middle-band energy ratio	$E_{\mathcal{B}_m} \,/\, \sum_k P_k$
	High-band energy ratio	$E_{\mathcal{B}_h} \,/\, \sum_k P_k$
Temporal decay	Early-window energy ratio	$\sum_{n\in\mathcal{T}_e} x^2[n] \,/\, \sum_{n=0}^{N-1} x^2[n]$
	Middle-window energy ratio	$\sum_{n\in\mathcal{T}_m} x^2[n] \,/\, \sum_{n=0}^{N-1} x^2[n]$
	Late-window energy ratio	$\sum_{n\in\mathcal{T}_l} x^2[n] \,/\, \sum_{n=0}^{N-1} x^2[n]$
	Decay-slope coefficient	Slope a in $\log E(t) = at + b$, fitted to the short-time energy envelope

response is used on its original amplitude scale so that amplitude-related cues remain available to the feature branch and regression model. All waveforms are converted to 32 kHz before being passed to the model. The target vector is then defined as

$$\mathbf{y} = [y_1, \ldots, y_6]^\top = [\mu, D/\mu, T_0/\mu, L_y, x_o, y_o]^\top, \tag{1}$$

where y_i denotes the i-th component of $\mathbf{y}$. Since the first three target parameters span several orders of magnitude, we apply log-space min-max normalization:

$$y_i^{\text{norm}} = \frac{\log y_i - \min(\log y_i)}{\max(\log y_i) - \min(\log y_i)}, \quad i \in \{1,2,3\}, \tag{2}$$

where $\min(\cdot)$ and $\max(\cdot)$ denote fixed parameter-wise bounds derived from the challenge parameter ranges. For evaluation, the predictions of these three parameters are transformed back to their original physical scales before the NMSE is computed. The remaining geometry and output-position parameters use ordinary min-max normalization:

$$y_i^{\text{norm}} = \frac{y_i - \min(y_i)}{\max(y_i) - \min(y_i)}, \quad i \in \{4,5,6\}. \tag{3}$$

2.2. CNN14 Backbone

We employ CNN14, a 14-layer convolutional neural network (CNN) architecture introduced in the Pretrained Audio Neural Networks (PANNs) framework [15], as the acoustic feature extractor. The network takes the resampled displacement impulse response directly as waveform input and is initialized with weights pretrained on AudioSet [16]. The convolutional blocks extract hierarchical representations of the impulse response, and the resulting 2048-dimensional embedding $\mathbf{e}$ is passed to a regression head for physical parameter estimation.

2.3. Physics-Inspired Features

A 15-dimensional feature branch is introduced as another feature extractor, which is motivated by hybrid audio representations that combine handcrafted and learned features [17]. These features encode complementary physical information: amplitude-related descriptors facilitate the estimation of mass and output position, temporal-decay descriptors characterize the response envelope, and spectral descriptors capture modal patterns governed by stiffness, tension, and plate geometry.

Let $x[n]$ be an impulse response of length N, and let $X[k]$ denote its discrete Fourier transform at frequency f_k. We write $M_k = \lvert X[k]\rvert$ for the magnitude spectrum and $P_k = M_k^2$ for the power spectrum. For a frequency band $\mathcal{B}$, define its spectral energy as $E_{\mathcal{B}} = \sum_{k\in\mathcal{B}} P_k$. The spectral descriptors follow standard audio-feature definitions [18]. For temporal energy descriptors, the early, middle, and late regions are defined as 0–0.05 s, 0.05–0.20 s, and after 0.20 s, respectively. For band energy descriptors, the low, middle, and high frequency bands are defined as 0–2000 Hz, 2000–8000 Hz, and above 8000 Hz, respectively. The features in Table 1 are extracted directly from the provided impulse response. The feature vector is standardized using the training-set mean and standard deviation before being concatenated with the PANNs embedding.

2.4. Feature Fusion and Regression

The fused representation is

$$\mathbf{z} = [\mathbf{e}; \mathbf{p}] \in \mathbb{R}^{2063}, \tag{4}$$

where $\mathbf{z}$ is the fused representation and $\mathbf{p} \in \mathbb{R}^{15}$ is the standardized handcrafted feature vector. The regression head is a multi-layer perceptron (MLP) with layer sizes $2063 \to 512 \to 256 \to 128 \to 6$, ReLU activations, and dropout. The output layer predicts six normalized regression values. No sigmoid activation is applied at the final layer; during evaluation, predictions are clipped to $[0,1]$ before inverse normalization.

Training is performed in the normalized target space using mean squared error (MSE):

$$\mathcal{L}_{\text{MSE}} = \frac{1}{6}\sum_{i=1}^{6} \left(\hat{y}_i^{\text{norm}} - y_i^{\text{norm}}\right)^2. \tag{5}$$

3. EXPERIMENT

3.1. Experimental Setup

The proposed PIFFN is trained and validated on synthetic plate-reverberation impulse responses generated by the official synthetic data-generation procedure. All waveforms are converted to 32 kHz before being passed to the CNN14 backbone, and the same impulse response is used to compute the 15 physics-inspired features in Table 1. Validation files are generated independently and excluded from model training and feature standardization. Statistics computed from the training set are applied unchanged to the validation set. Each response is treated as a single-channel displacement signal shared by the neural backbone and feature branch. The six target parameters are normalized by (2) and (3). Training is performed in the normalized target space using the MSE loss in (5). Training uses a batch size of 16, and the random seed is fixed to 50. The model is optimized using Adam [19] with an initial learning rate of 10^{-4}. A validation-loss-based learning-rate scheduler is configured with a patience of 100, a decay factor of 0.5, and a minimum learning rate of 10^{-6}. Early stopping is configured with a patience of 100 and a minimum improvement threshold of zero. The checkpoint with the lowest validation loss is selected for final evaluation. All experiments run on one NVIDIA GeForce RTX 3090 GPU with 24 GB memory. The server is equipped with Intel Xeon Platinum 8180 CPUs and 1.0 TiB RAM.

3.2. Estimation Performance

The Task A model is evaluated on 1,000 generated one-second impulse-response files that are separate from the 10,000 generated files used for training. All data are generated using the official synthetic data-generation procedure. We report normalized mean squared error (NMSE) in the normalized target space, matching the training objective in (5). The overall validation NMSE is calculated to be approximately 0.00362, indicating PIFFN's promising parameter estimation performance.

Figure 2 summarizes the parameter-wise behavior of PIFFN. Both normalized material parameters, D/μ and T_0/μ, achieve the two lowest NMSE values, indicating that PIFFN effectively captures stiffness- and tension-dependent modal characteristics. In contrast, x_o has the largest NMSE, suggesting that output-position estimation remains comparatively more challenging.

To assess PIFFN's robustness across different individual validation samples, we also investigate the distribution of per-file NMSE values for PIFFN. As shown in Fig. 3, more than 900 validation samples are concentrated in the low-NMSE region (NMSE < 0.01), whereas fewer than 20 samples have NMSE values above 0.02. This distribution indicates that the reported overall NMSE is representative of the validation set.

3.3. Comparison with Baseline Methods

To assess PIFFN relative to alternative approaches, we compare it with the official particle swarm optimization (PSO) baseline [20] and a CNN14 model without the physics-inspired feature branch. As shown in Table 2, the PSO baseline and CNN14 model yield overall NMSEs of 0.06498 and 0.00786, respectively, whereas PIFFN attains an overall NMSE of 0.00362. Both CNN14 and PIFFN substantially reduce the overall NMSE relative to the PSO baseline, with PIFFN further improving on CNN14. At the

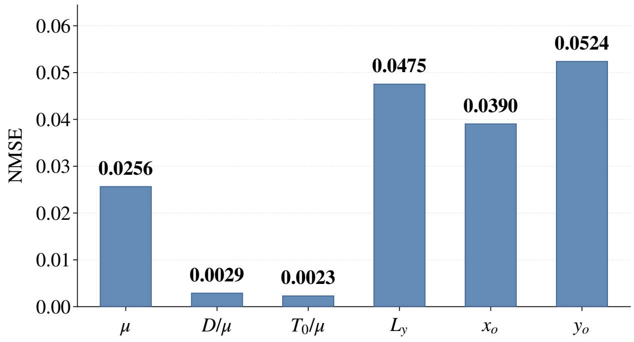

Figure 2: Per-parameter validation NMSE of PIFFN.

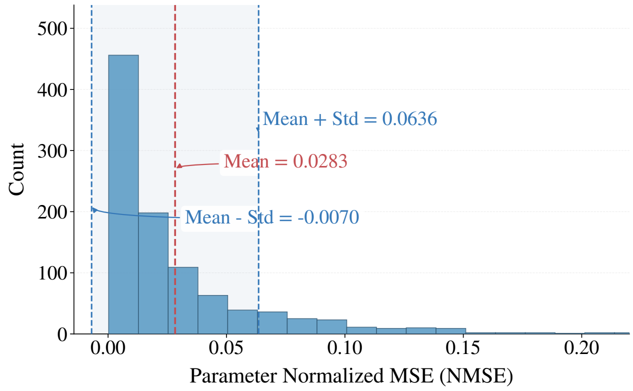

Figure 3: Distribution of per-file NMSE over the 1,000 validation impulse responses.

Table 2: NMSE comparison with the official PSO baseline and the backbone-only baseline.

Parameter	PSO	CNN14	PIFFN
Overall	0.06498	0.00786	**0.00362**
μ	0.03807	0.01536	**0.00477**
D/μ	0.01069	0.00316	**0.00101**
T_0/μ	0.00487	0.00244	**0.00181**
L_y	0.09770	0.00883	**0.00472**
x_o	0.12071	0.01180	**0.00618**
y_o	0.11782	0.00554	**0.00325**

parameter level, PIFFN achieves the lowest NMSE for all six targets, outperforming both the PSO baseline and the CNN14 model without physics-inspired feature fusion.

To further examine parameter specific estimation performance, Fig. 4 compares the normalized predictions of CNN14 and PIFFN. Although the T_0/μ samples appear dispersed, their NMSE remains low because the actual parameter range is narrower than that derived from the independent challenge bounds of T_0 and μ. Compared with CNN14, PIFFN produces predictions that are more concentrated around the identity line, particularly for μ, D/μ, and x_o, consistent with the NMSE reductions reported in Table 2. PIFFN also reduces the larger deviations observed for CNN14 near the boundaries of these parameter ranges. These

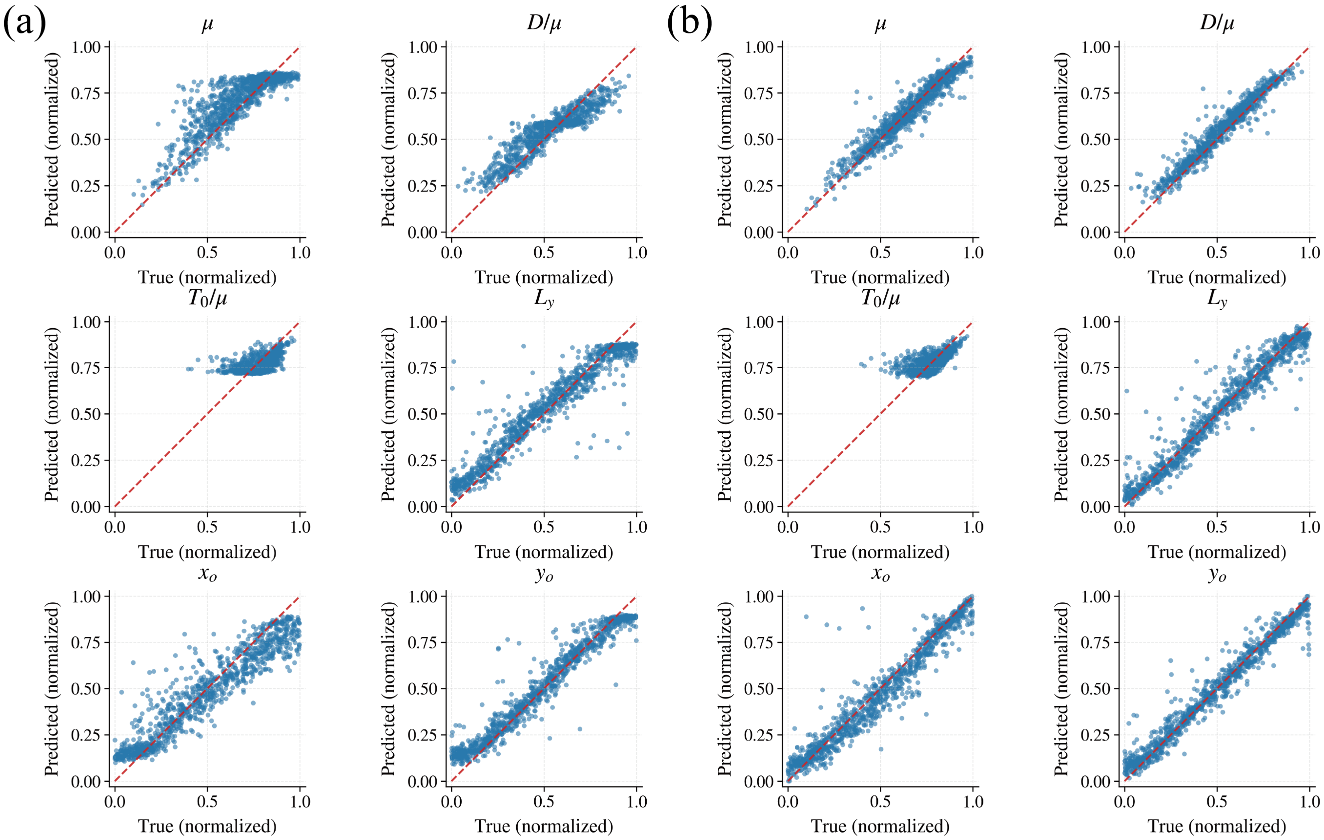

Figure 4: Normalized prediction-versus-target scatter plots for the six parameters on the validation set. The six left-hand panels form (a) and show CNN14 without physics-inspired feature fusion, whereas the six right-hand panels form (b) and show PIFFN. Each point represents one of the 1,000 validation impulse responses. In each panel, the red dashed line indicates ideal predictions.

results demonstrate that physics inspired feature fusion improves both overall estimation accuracy and robustness across different parameter ranges.

4. CONCLUSIONS

In this paper, we presented PIFFN, a physics-inspired feature fusion framework for estimating identifiable physical parameters of a plate from impulse responses. PIFFN integrates a pretrained CNN14 backbone with a compact set of handcrafted impulse-response descriptors, enabling the model to exploit both learned acoustic representations and explicit cues related to amplitude, temporal decay, and spectral characteristics. The proposed method achieves an overall NMSE of 0.00362 and consistently outperforms both the official PSO baseline and the backbone-only CNN14 model in overall and parameter-wise NMSE. These results demonstrate the effectiveness of incorporating physics-inspired features into the neural backbone. Moreover, the per-file NMSE distribution shows that most validation samples exhibit low estimation errors, indicating stable and consistent estimation performance.

5. REFERENCES

[1] U. Zölzer, Ed., *DAFx: Digital Audio Effects*, 2nd ed. Chichester, UK: John Wiley & Sons, 2011.

[2] K. Arcas and A. Chaigne, "On the quality of plate reverberation," *Appl. Acoust.*, vol. 71, no. 2, pp. 147–156, 2010.

[3] S. Bilbao, "A digital plate reverberation algorithm," *J. Audio Eng. Soc.*, vol. 55, no. 3, pp. 135–144, 2007.

[4] K. van den Doel and D. K. Pai, "The sounds of physical shapes," *Presence: Teleoperators Virtual Environ.*, vol. 7, no. 4, pp. 382–395, 1998.

[5] S. Bilbao, *Numerical Sound Synthesis: Finite Difference Schemes and Simulation in Musical Acoustics*. Chichester, UK: John Wiley & Sons, 2009.

[6] M. Ducceschi and L. Gabrielli, "The 1st DAFx parameter estimation challenge: A benchmark for plate reverb system identification," 2026.

[7] M. Ikehata, "An inverse problem for the plate in the love–kirchhoff theory," *SIAM J. Appl. Math.*, vol. 53, no. 4, pp. 942–970, 1993.

[8] S. Bhattacharyya and T. Ghosh, "Inverse problem for kirchhoff–love plate equation," *arXiv preprint arXiv:2102.00747*, 2021.

[9] Z. Xu, H. Li, W.-y. Wang, Y.-n. Liu, and B.-c. Wen, "Inverse identification of mechanical parameters of fiber metal laminates," *Proc. Inst. Mech. Eng., Part C: J. Mech. Eng. Sci.*, vol. 234, no. 8, pp. 1516–1527, 2020.

[10] R. Diaz and M. Sandler, "Fast differentiable modal simulation of non-linear strings, membranes, and plates," in *Proc. Int. Conf. Digital Audio Effects (DAFx25)*, 2025.

[11] X. Jin, C. Xu, R. Gao, J. Wu, G. Wang, and S. Li, "DiffSound: Differentiable modal sound rendering and inverse rendering for diverse inference tasks," in *Proc. ACM SIGGRAPH Conf. Papers*, 2024.

[12] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *J. Comput. Phys.*, vol. 378, pp. 686–707, 2019.

[13] X.-Y. Guo and S.-E. Fang, "Structural parameter identification using physics-informed neural networks," *Measurement*, vol. 220, p. 113334, 2023.

[14] W. Zhou and Y. F. Xu, "Damage identification for plate structures using physics-informed neural networks," *Mech. Syst. Signal Process.*, vol. 209, p. 111111, 2024.

[15] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 28, pp. 2880–2894, 2020.

[16] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "AudioSet: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2017, pp. 776–780.

[17] G. Elbanna, A. Biryukov, N. Scheidwasser-Clow, L. Orlandic, P. Mainar, M. Kegler, P. Beckmann, and M. Cernak, "Hybrid handcrafted and learnable audio representation for analysis of speech under cognitive and physical load," in *Proc. Interspeech*, 2022.

[18] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proc. Python Sci. Conf.*, 2015, pp. 18–25.

[19] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[20] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proc. IEEE Int. Conf. Neural Networks*, vol. 4, 1995, pp. 1942–1948.