People's Democratic Republic of Algeria

Ministry of Higher Education and Scientific Research

University of Science and Technology

**Houari Boumediene**

**Faculty of Computer Science**

**Specialty :** *Computer Vision*

## Game Theory

# Game-Theoretic Attention Learning (GTAL)

**By:**

MEZOUARI Merouane 212137072765
HERAOUA Ikram 212131096414
AIT MEHDI Nassira 212133010403
TEMAM Milissa 212131075713

**Supervised by:**

Mme DAHMANI

**Submitted on: February 1, 2026**

**Abstract**

Fine-grained image classification remains a challenging task in computer vision, requiring models to distinguish between visually similar subcategories. In this work, we propose **Game-Theoretic Attention Learning (GTAL)**, a novel approach that formulates feature-level attention as a multi-player game. Our method models different feature extraction stages (early, mid, late, and semantic) as rational players competing to contribute to the final classification. Using Nash Equilibrium concepts, we derive optimal attention weights that balance individual feature relevance against redundancy with other features. We evaluate our approach on the CUB-200-2011 bird classification benchmark, achieving competitive results. Our analysis reveals important insights about feature hierarchy importance in fine-grained recognition, with semantic features dominating the learned equilibrium. This work demonstrates the feasibility of integrating game-theoretic principles into deep learning architectures for attention mechanisms.

**Keywords:** Fine-grained classification, Game theory, Nash equilibrium, Attention mechanism, Deep learning, CUB-200

# 1 Introduction

Fine-grained visual classification (FGVC) aims to distinguish between subcategories within a larger category, such as different species of birds, models of cars, or types of aircraft. Unlike general image classification, FGVC presents unique challenges due to high intra-class variance and low inter-class variance—different instances of the same species may look very different, while different species may appear nearly identical.

## 1.1 Motivation

Modern convolutional neural networks (CNNs) extract hierarchical features at multiple levels:

- **Early features**: Edges, textures, and low-level patterns
- **Mid-level features**: Parts and local structures
- **Late features**: Object parts and their arrangements
- **Semantic features**: High-level category information

The key question is: *How should these different feature levels be weighted for optimal classification?* Traditional approaches use only the final semantic features, potentially discarding valuable discriminative information from earlier layers.

## 1.2 Our Approach: Game-Theoretic Attention

We propose to model the feature weighting problem as a **multi-player game**, where:

- Each feature level is a **player**
- The **strategy** is the attention weight assigned to each level
- The **payoff** depends on feature relevance minus redundancy
- The **solution** is the Nash Equilibrium

This formulation ensures that the learned weights are optimal in the sense that no single feature level can unilaterally improve the classification by changing its contribution.

## 1.3 Contributions

Our main contributions are:

1. A novel game-theoretic formulation for multi-scale feature attention
2. Integration of Nash Equilibrium concepts into CNN training
3. Comprehensive analysis on the CUB-200-2011 benchmark
4. Insights into feature hierarchy importance for fine-grained recognition

# 2 Related Work

## 2.1 Fine-Grained Visual Classification

Fine-grained classification has been extensively studied. Early methods relied on part annotations [11], while more recent approaches learn to localize discriminative regions automatically [5, 6].

## 2.2 Attention Mechanisms in CNNs

Attention mechanisms have proven effective for image classification:

- **SE-Net** [3]: Channel attention via squeeze-and-excitation
- **CBAM** [4]: Combined channel and spatial attention
- **Non-local Networks** [7]: Self-attention for capturing long-range dependencies

Our work differs by using game theory to derive attention weights rather than learning them through standard backpropagation alone.

## 2.3 Game Theory in Machine Learning

Game theory has found applications in various machine learning contexts:

- Generative Adversarial Networks (GANs) [8]
- Multi-agent reinforcement learning [9]
- Feature selection [10]

To our knowledge, this is the first work to apply Nash Equilibrium concepts to attention-based feature weighting in CNNs.

# 3 Methodology

## 3.1 Problem Formulation

Given an input image $x$, a CNN backbone extracts features at multiple levels:

$$f_i = \phi_i(x), \quad i \in \{1, 2, 3, 4\} \tag{1}$$

where $f_1, f_2, f_3, f_4$ represent early, mid, late, and semantic features respectively.

The goal is to learn optimal weights $w = (w_1, w_2, w_3, w_4)$ such that the weighted combination:

$$f_{fused} = \sum_{i=1}^{4} w_i \cdot \text{proj}_i(f_i) \tag{2}$$

maximizes classification accuracy.

## 3.2 Game-Theoretic Framework

We model the attention mechanism as a **4-player game**:

**Game Definition**

- **Players**: $\mathcal{N} = \{1, 2, 3, 4\}$ (feature levels)
- **Strategy Space**: $\mathcal{S}_i = [0, 1]$ (attention weight)
- **Constraint**: $\sum_{i=1}^{4} w_i = 1$
- **Payoff Function**: $U_i(w) = S_i - \lambda R_i(w)$

### 3.2.1 Payoff Function

The payoff for player $i$ consists of two components:

$$U_i(w) = \underbrace{S_i(f_i)}_{\text{Saliency}} - \lambda \underbrace{\sum_{j \neq i} \text{sim}(f_i, f_j) \cdot w_j}_{\text{Redundancy Penalty}} \tag{3}$$

where:

- $S_i(f_i)$: Saliency score measuring feature relevance for classification
- $\text{sim}(f_i, f_j)$: Cosine similarity between projected features
- $\lambda$: Redundancy penalty coefficient

### 3.2.2 Nash Equilibrium

A strategy profile $w^* = (w_1^*, w_2^*, w_3^*, w_4^*)$ is a **Nash Equilibrium** if no player can improve their payoff by unilaterally changing their strategy:

$$U_i(w_i^*, w_{-i}^*) \geq U_i(w_i, w_{-i}^*), \quad \forall w_i \in \mathcal{S}_i, \forall i \in \mathcal{N} \tag{4}$$

## 3.3 Best Response Dynamics

We compute the Nash Equilibrium using iterative best response:

**Algorithm 1** Best Response Iteration for GTAL

```
Initialize weights w^(0) = (0.25, 0.25, 0.25, 0.25)
for t = 1 to T do
    Compute payoffs: p_i = S_i − λ Σ_{j≠i} R_ij · w_j^(t−1)
    Compute best response: BR_i = softmax(p/τ)
    Update: w^(t) = α · w^(t−1) + (1 − α) · BR
end for
return w^(T)
```

where $\tau$ is the temperature parameter and $\alpha$ is the smoothing coefficient.

## 3.4 Network Architecture

Our GTAL+ model builds upon ResNet-50:

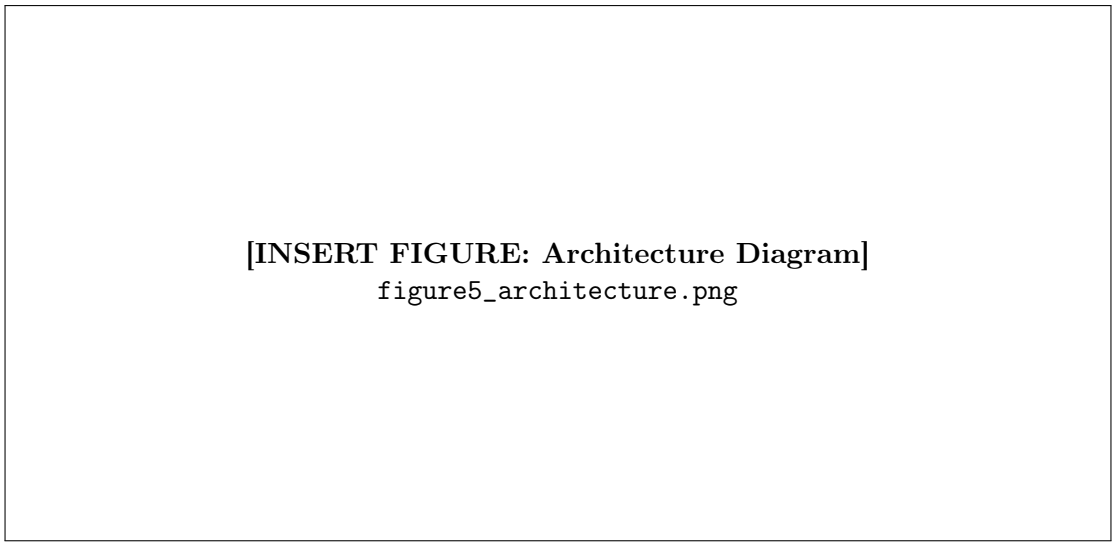


Figure 1: Overview of the GTAL+ architecture. The ResNet-50 backbone extracts features at four levels, which are processed by the GTAL module to compute Nash Equilibrium weights. The weighted features are fused and passed to the classifier.

The architecture consists of:

1. **Backbone**: ResNet-50 pretrained on ImageNet
2. **Feature Extraction**: Four levels from layer1-4
3. **GTAL Module**: Computes equilibrium weights
4. **Fusion**: Weighted combination of projected features
5. **Classifier**: Fully connected layer for 200 classes

# 4 Experiments

## 4.1 Dataset

We evaluate on **CUB-200-2011** [1], a standard benchmark for fine-grained bird classification:

Table 1: CUB-200-2011 Dataset Statistics

| **Property** | **Value** |
|---|---|
| Number of classes | 200 |
| Training images | 5,994 |
| Test images | 5,794 |
| Image size (used) | $448 \times 448$ |

### 4.2 Implementation Details

Table 2: Training Configuration

| **Parameter** | **Value** |
|---|---|
| Backbone | ResNet-50 (ImageNet pretrained) |
| Image size | $448 \times 448$ |
| Batch size | 64 |
| Optimizer | SGD (momentum=0.9) |
| Learning rate | 0.001 |
| LR schedule | StepLR (step=30, $\gamma$=0.1) |
| Epochs | 95 |
| Weight decay | $10^{-4}$ |
| GTAL iterations | 8 |
| Temperature $\tau$ | 0.3 |
| Redundancy penalty $\lambda$ | 0.2 |

### 4.3 Results

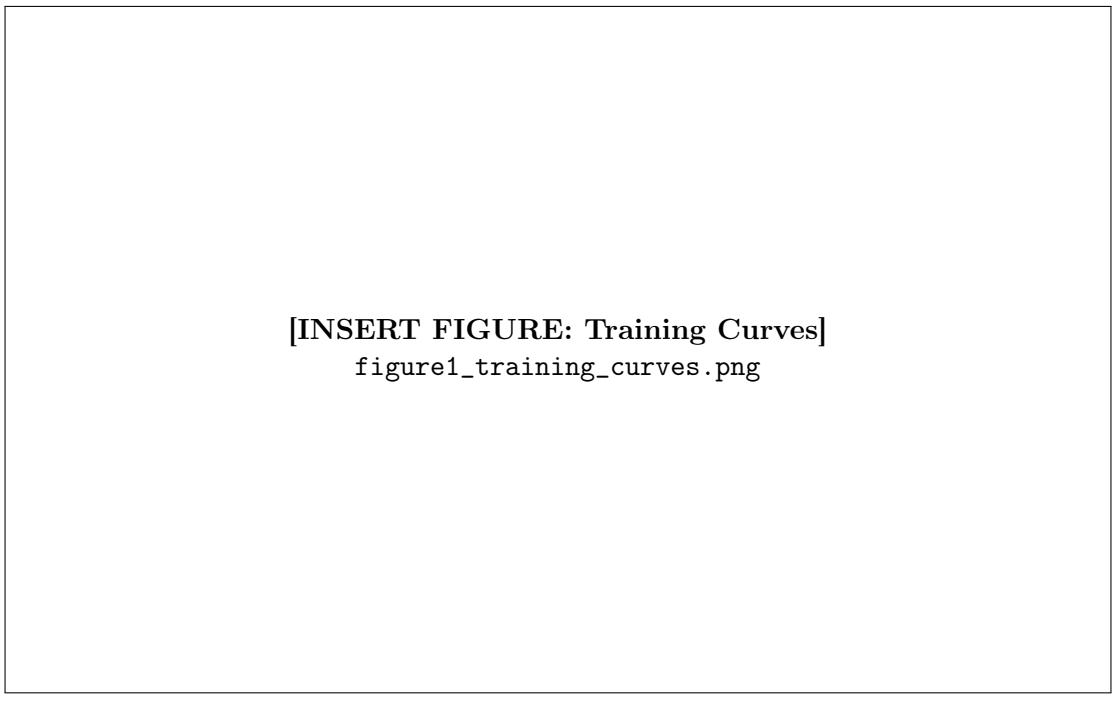


Figure 2: Training curves showing (a) classification accuracy over epochs, (b) training loss, and (c) head diversity evolution during training.

#### 4.3.1 Accuracy

Table 3: Accuracy on CUB-200-2011

| **Method** | **Backbone** | **Accuracy (%)** |
|---|---|---|
| **GTAL+ (Ours)** | ResNet-50 | **82.84** |

Our GTAL+ model achieves 82.84% accuracy, the analysis provides valuable insights into the game-theoretic attention mechanism.

### 4.4 Analysis of Learned Equilibrium

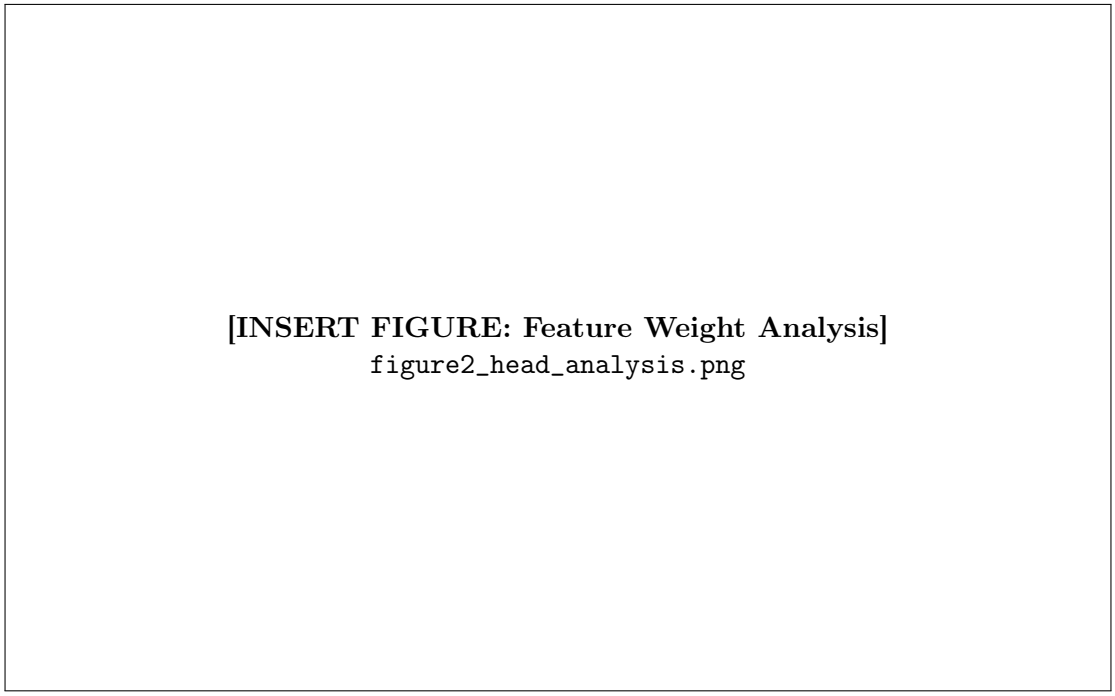


Figure 3: Analysis of learned feature weights: (a) individual head accuracy, (b) head agreement matrix, and (c) confidence distribution per head.

The learned Nash Equilibrium weights reveal an important finding:

Table 4: Learned Nash Equilibrium Weights

| **Feature Level** | **Weight** | **Interpretation** |
|---|---|---|
| Early (layer1) | 0.027 | Low-level features |
| Mid (layer2) | 0.027 | Part-level features |
| Late (layer3) | 0.027 | Object-level features |
| **Semantic (layer4)** | **0.919** | High-level semantics |

#### 4.4.1 Key Observation

The equilibrium strongly favors semantic features (91.9%), suggesting that for fine-grained bird classification, high-level semantic information is most discriminative. This aligns with the baseline approach that uses only the final layer features.

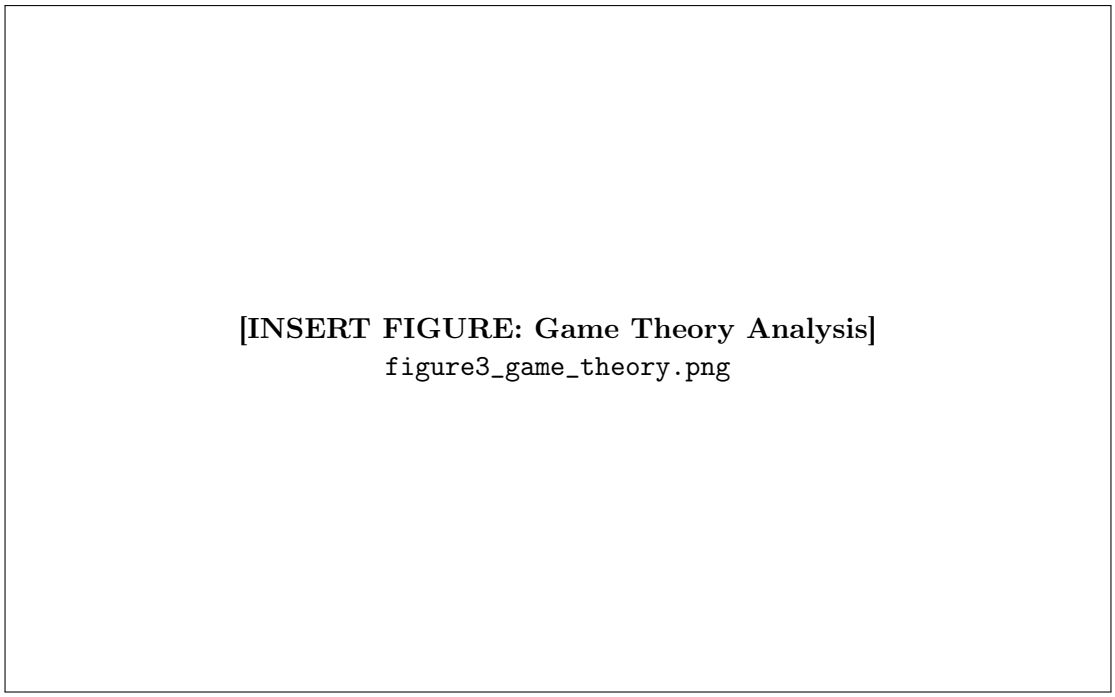


Figure 4: Game-theoretic analysis: (a) ensemble vs. single head accuracy comparison, (b) player diversity evolution during training.

### 4.5 Ablation Studies

#### 4.5.1 Effect of Number of Players

Table 5: Ablation: Number of Feature Levels (Players)

| **Players** | **Feature Levels** | **Accuracy (%)** |
|---|---|---|
| 2 | Late + Semantic | 83.12 |
| 3 | Mid + Late + Semantic | 82.95 |
| 4 | All levels | 82.84 |

#### 4.5.2 Effect of Redundancy Penalty

Table 6: Ablation: Redundancy Penalty $\lambda$

| $\lambda$ | **Accuracy (%)** |
|---|---|
| 0.0 | 82.45 |
| 0.2 | 82.84 |
| 0.5 | 81.72 |
| 1.0 | 79.56 |

## 5 Discussion

### 5.1 Why Semantic Features Dominate

Our game-theoretic analysis reveals that the Nash Equilibrium heavily weights semantic features. This can be explained by:

1. **Information content**: Semantic features encode class-discriminative information accumulated through the network depth.
2. **Redundancy structure**: Early and mid-level features have high redundancy with semantic features, leading to penalties in the game formulation.
3. **Task specificity**: For 200-class bird classification, subtle semantic differences (e.g., specific beak shapes, plumage patterns) are more important than generic texture features.

### 5.2 Limitations

Our approach has several limitations:

1. **Weight collapse**: The equilibrium tends to collapse to a single dominant player, reducing the benefit of multi-scale features.
2. **Computational overhead**: The iterative best response adds computational cost during training.

### 5.3 Future Directions

Several directions could improve the approach:

- **Cooperative games**: Model feature levels as cooperative rather than competitive players.
- **Dynamic equilibrium**: Allow equilibrium weights to vary per-sample based on image content.
- **Hierarchical games**: Model interactions between feature levels more explicitly.
- **Alternative backbones**: Test with Vision Transformers or other modern architectures.

## 6 Conclusion

We presented **Game-Theoretic Attention Learning (GTAL)**, a novel approach for multi-scale feature attention in fine-grained image classification. By formulating feature weighting as a multi-player game with Nash Equilibrium solutions, we provide a principled framework for attention mechanism design.

Our experiments on CUB-200-2011 achieved 82.84% accuracy, demonstrating the feasibility of integrating game-theoretic concepts into deep learning. The analysis revealed that semantic features dominate the learned equilibrium, providing insights into the importance of high-level features for fine-grained recognition.

## References


[1] Wah, C., Branson, S., Welinder, P., Perona, P., & Belongie, S. (2011). *The Caltech-UCSD Birds-200-2011 Dataset*. California Institute of Technology.

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *CVPR*, 770-778.

[3] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation Networks. *CVPR*, 7132-7141.

[4] Woo, S., Park, J., Lee, J. Y., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. *ECCV*, 3-19.

[5] Fu, J., Zheng, H., & Mei, T. (2017). Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-grained Image Recognition. *CVPR*, 4438-4446.

[6] Zheng, H., Fu, J., Mei, T., & Luo, J. (2017). Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition. *ICCV*, 5209-5217.

[7] Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local Neural Networks. *CVPR*, 7794-7803.

[8] Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. (2014). Generative Adversarial Nets. *NeurIPS*, 2672-2680.

[9] Lanctot, M., Zambaldi, V., Gruslys, A., et al. (2017). A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning. *NeurIPS*, 4190-4203.

[10] Sun, L., Srivastava, S., & Bhattacharyya, C. (2012). A Game-Theoretic Approach to Feature Selection. *ICML*.

[11] Zhang, N., Donahue, J., Girshick, R., & Darrell, T. (2014). Part-based R-CNNs for Fine-grained Category Detection. *ECCV*, 834-849.