Hierarchical Multi-Scale Feature Learning with Iterative Cross-Modal Re nement for Social Relationship Recognition

Midterm Progress Report

Computer Engineering Graduation Project

Furkan Cinko	**Student ID: 210201018**
Melih Şahin	**Student ID: 210201044**
Mert Genç	**Student ID: 210201070**
Kutay Çakır	**Student ID: 210201039**
Samet Balaban	**Student ID: 220201027**
İbrahim Halil Teymur	**Student ID: 210201058**

Department of Computer Engineering

Üsküdar University
Istanbul, Turkey

November 24, 2025

bstract

This midterm report presents the conceptual design and planned implementation of an enhanced multimodal social relationship recognition system. Building upon the baseline paper "Multimodal Social Relationship Recognition Based on LLM" by Wang et al., we propose three complementary innovations: (1) Multi-Scale Visual Feature Pyramid for capturing hierarchical visual semantics from ne-grained facial details to scene-level context, (2) Iterative Cross-Modal Re nement mechanism for progressive text-image alignment, and (3) Uncertainty-Aware Adaptive Fusion for robust multimodal integration under varying data quality. This report presents a comprehensive analysis of the baseline papers limitations, details our proposed architectural improvements, outlines our implementation strategy, and provides a realistic timeline for project completion. Our theoretical analysis suggests potential improvements of 3-5% map over the baseline method on the PISC dataset.

Contents

1 Introduction **4**
- 1.1 Background and Motivation . . . 4
- 1.2 Project Objectives . . . 4
- 1.3 Report Structure . . . 5

2 Problem Analysis and Related Work **5**
- 2.1 Baseline Paper: Detailed Analysis . . . 5
  - 2.1.1 Text Extraction Module . . . 5
  - 2.1.2 Image Feature Extraction Module . . . 5
  - 2.1.3 Multimodal Alignment Module . . . 6
  - 2.1.4 Fusion and Classification Module . . . 6
- 2.2 Performance Analysis from Baseline Paper . . . 7
- 2.3 Related Work on Proposed Solutions . . . 7
  - 2.3.1 Feature Pyramid Networks . . . 7
  - 2.3.2 Iterative Refinement in Multimodal Tasks . . . 8
  - 2.3.3 Uncertainty Estimation in Deep Learning . . . 8

3 Proposed Methodology **8**
- 3.1 Innovation 1: Multi-Scale Visual Feature Pyramid . . . 8
  - 3.1.1 Core Idea . . . 8
  - 3.1.2 Architecture Design . . . 8
  - 3.1.3 Theoretical Justification . . . 9
  - 3.1.4 Expected Improvements . . . 9
- 3.2 Innovation 2: Iterative Cross-Modal Refinement . . . 9
  - 3.2.1 Core Idea . . . 9
  - 3.2.2 Architecture Design . . . 10
  - 3.2.3 Progressive Refinement Example . . . 10
  - 3.2.4 Why K=3 Iterations? . . . 10
  - 3.2.5 Expected Improvements . . . 11
- 3.3 Innovation 3: Uncertainty-Aware Adaptive Fusion . . . 11
  - 3.3.1 Core Idea . . . 11
  - 3.3.2 Architecture Design . . . 11
  - 3.3.3 Adaptive Behavior Examples . . . 12
  - 3.3.4 Expected Improvements . . . 12
- 3.4 Complete System Integration . . . 12
- 3.5 Loss Function Design . . . 12

4 Experimental Setup **13**
- 4.1 Dataset . . . 13
- 4.2 Implementation Plan . . . 13
  - 4.2.1 Hardware Resources . . . 13
  - 4.2.2 Software Framework . . . 13
  - 4.2.3 Key Hyperparameters . . . 14
- 4.3 Evaluation Metrics . . . 14
- 4.4 Experimental Validation Strategy . . . 14

5 Progress Report **15**
5.1 Completed Work (Weeks 1-8) . . . 15
5.1.1 Week 1-2: Literature Review . . . 15
5.1.2 Week 3-4: Baseline Paper Deep Analysis . . . 15
5.1.3 Week 5-6: Architectural Design . . . 15
5.1.4 Week 7-8: Implementation Planning and Setup . . . 16
5.2 Work Distribution Among Team Members . . . 16
5.3 Current Status Summary . . . 17

6 Implementation Plan and Timeline **17**
6.1 Detailed Implementation Roadmap . . . 17
6.1.1 Weeks 9-10: Baseline Implementation . . . 17
6.1.2 Weeks 11-12: Component Implementations . . . 18
6.1.3 Weeks 13-14: System Integration and Experiments . . . 18
6.1.4 Week 15: Analysis and Visualization . . . 19
6.1.5 Week 16: Final Report and Presentation . . . 19
6.2 Timeline Summary . . . 20
6.3 Risk Management . . . 21

7 Expected Contributions and Impact **21**
7.1 Technical Contributions . . . 21
7.2 Expected Performance Improvements . . . 22
7.3 Broader Impact . . . 22
7.4 Learning Outcomes for Team . . . 23

8 Conclusion **23**

1 Introduction

1.1 Background and Motivation

Social relationship recognition is a fundamental computer vision task with significant practical applications in social media analysis, security systems, and human-computer interaction. The task involves automatically identifying the type of relationship between people in images (eg., family, friends, couples, colleagues).

Recent advances in deep learning, particularly the integration of Large Language Models (LLMs) with visual feature extraction, have shown promising results. The baseline paper by Wang et al. [1] achieves state-of-the-art performance (88.9% mAP on pısc-C, 76.9% on pısc-F) by combining LLM-based text structure extraction with Resnet-50 visual features.

However, through careful analysis of the baseline methodology and experimental results reported in the paper, we have identified three critical limitations that present opportunities for improvement:

1. **Single-Scale Visual Features**: The baseline uses only the final layer of ResNet-50, which captures high-level semantic information but loses fine-grained details crucial for distinguishing similar relationships (eg., couples vs. close friends).

2. **One-Shot Static Alignment**: The cross-modal alignment computes a similarity matrix once without iterative refinement, making it sensitive to initial feature quality and unable to progressively improve.

3. **Simple Linear Fusion**: The fusion mechanism employs basic weighted summation that cannot adapt to varying data quality (eg., blurry images, ambiguous text) and fails to model complex multimodal interactions.

1.2 Project Objectives

The primary objectives of this graduation project are:

1. Conduct comprehensive theoretical analysis of the baseline paper's architectural components and identify specific weaknesses through literature review.

2. Design three novel architectural components that systematically address the identified limitations while maintaining compatibility with the baseline framework.

3. Implement the proposed system on the PISC dataset using PyTorch, ensuring reproducibility by first replicating baseline results.

4. Conduct extensive experiments including ablation studies to validate each component's contribution individually and in combination.

5. Achieve measurable performance improvements of 3-5% mAP over the baseline, particularly on challenging scenarios such as fine-grained relationships and varying image qualities.

1.3 Report Structure

The remainder of this report is organized as follows: Section 2 presents our detailed analysis of the baseline paper and related work. Section 3 describes our proposed methodology including all three archtectural innovations. Section 4 outlines our experimental setup and evaluation strategy. Section 5 reports our progress to date including completed literature review and system design. Section 6 presents our detailed implementation plan and timeline for the remaining work. Section 7 discusses expected contributions and potential impact. Finally, Section 8 concludes the report.

2 Problem Analysis and Related Work

2.1 Baseline Paper: Detailed Analysis

The baseline paper "Multimodal Social Relationship Recognition Based on LLM" by Wang et al. [1] proposes a framework consisting of four main modules. We analyze each module systematically to identify strengths and limitations.

2.1.1 Text Extraction Module

Baseline Approach: The paper employs a Large Language Model (LLM) to extract structured event information from textual descriptions. For example, given text "Two friends are playing basketball on the court. They are laughing and having a great time," the LLM extracts: [relationship: friends, emotion: happy, scene: basketball court]. This structured information is then processed by a CNN to generate text feature vectors F_T.

Analysis:

- **Strength**: E ective semantic parsing that outperforms Text2Event and CLEVE methods according to Table 3 in the baseline paper.
- **Limitation**: PISC dataset lacks ground-truth text descriptions, forcing the baseline to generate captions using BLIP-2. Generated captions are often generic ("two people standing") or contain errors, directly impacting downstream performance.

2.1.2 Image Feature Extraction Module

Baseline Approach: ResNet-50 extracts visual features, enhanced by channel and spatial attention mechanisms. The attention-weighted feature map is:

$$F_I^{att} = (A_c \cdot A_s) \odot F_I \tag{1}$$

where A_c is channel attention, A_s is spatial attention, and $\odot$ denotes element-wise multiplication.

Critical Limitation - Single-Scale Features:

The baseline uses **only the nal layer** (layer4) of ResNet-50, which produces 2048-channel features at low spatial resolution. This design has fundamental limitations:

Concrete Example of Failure:

Consider distinguishing **Couples** from **Close Friends**:

- **Couples**: Require fine-grained cues: romantic facial expressions (Layer 1), intimate hand-holding patterns (Layer 2), very close physical proximity (Layer 3)

Table 1: ResNet-50 Layer Characteristics and Relationship Types

Layer	Channels	Resolution	Semantic Level
Layer 1	256	High	Facial expressions, details
Layer 2	512	Medium-High	Body language, gestures
Layer 3	1024	Medium-Low	Interactions, proximity
Layer 4	2048	Low	Scene context (baseline uses this)

- **Close Friends**: Similar high-level scene (Layer 4) but di erent low-level details: friendly smiles vs. romantic gazes, casual vs. intimate touch

The baseline's Layer 4-only approach sees both as "two people in close proximity" without distinguishing subtle di erences. Table 7 in the baseline paper confirms this: Couple vs. Friends accuracy is only 72.1%, the worst among all categories.

2.1.3 Multimodal Alignment Module

Baseline Approach: Computes semantic similarity via cosine similarity:

$$s_{ij} = \frac{F_I^{vec}[i] \cdot F_T[j]}{|F_I^{vec}[i]| \cdot |F_T[j]|} \tag{2}$$

Features are aligned through matrix multiplication: $F_I^{aligned} = S \cdot F_I^{vec}$.

Critical Limitation - One-Shot Alignment:

The alignment is computed **once** at the beginning and never refined. This causes cascading errors:

- **Initial Stage**: If text "happy couple" initially aligns broadly to "two smiling people" (imprecise), this misalignment persists throughout.
- **No Correction**: Unlike human perception which iteratively refines understanding, the model cannot improve alignment after initial computation.
- **Sensitivity**: Performance heavily depends on initial feature quality. Poor captions or blurry images lead to incorrect alignment that cannot be recovered.

2.1.4 Fusion and Classi cation Module

Baseline Approach: Fuses aligned features using MLP-computed weights:

$$F_{fusion} = w \odot F_I^{aligned} + (1 \quad w) \odot F_T^{aligned} \tag{3}$$

where $w = \text{MLP}([F_I^{aligned}; F_T^{aligned}])$.

Critical Limitation - Non-Adaptive Fusion:

The fusion mechanism has two fundamental problems:

1. **Linear Combination**: Simple weighted sum cannot model complex interactions between modalities. Real-world relationships often require non-linear reasoning (e.g., text mentions "couple" but image shows professional handshake contradictory information needs resolution).

2. **No Uncertainty Awareness**: The model treats all data uniformly regardless of quality:
    - Clear HD image + ambiguous text should rely more on visual
    - Blurry image + detailed text should rely more on text
    - Baseline cannot make this adaptation

2.2 Performance Analysis from Baseline Paper

Table 7 in the baseline paper reveals systematic performance patterns:

Table 2: Baseline Performance and Our Analysis of Failure Modes

Scenario	**mAP**	**Primary Limitation**
PISC-C (Coarse)	88.9%	Baseline reference
PISC-F (Fine-grained)	76.9%	**-12% drop**: Single-scale misses fine details
Couple vs. Friends	72.1%	Worst category: Needs facial expressions (Layer 1)
Outdoor scenes	68.3%	Poor generated text + static alignment fails
Multi-person images	71.5%	Ambiguous captions + no iterative refinement

Key Insights:

- Sharp drop (12%) from coarse to fine-grained categories indicates **inability to capture subtle distinctions**
- Fine-grained relationships (Couple/Friends) are most challenging, confirming need for multi-scale features
- Outdoor and multi-person scenarios su er most, suggesting static alignment cannot handle complex/ambiguous inputs

2.3 Related Work on Proposed Solutions

2.3.1 Feature Pyramid Networks

Feature Pyramid Networks (FPN) [2], originally proposed for object detection, construct multi-scale feature hierarchies through top-down pathways with lateral connections. FPN has been successfully applied in:

- Object detection: Faster R-CNN + FPN [2]
- Semantic segmentation: PSPNet [3]
- Vision-language models: Recent CLIP variants

However, FPN has **not been systematically explored for social relationship recognition**, representing a clear research gap our project addresses.

2.3.2 Iterative Re nement in Multimodal Tasks

Several recent works demonstrate the value of iterative refinement:

- **ALBEF** [4]: Uses momentum distillation for iterative vision-language alignment
- **Flamingo** [5]: Employs multiple cross-attention layers for progressive fusion
- **BLIP-2** [6]: Uses Q-Former with iterative queries

These works confirm that iterative processing improves cross-modal understanding. Our approach adapts these principles specifically for relationship recognition.

2.3.3 Uncertainty Estimation in Deep Learning

Uncertainty quantification has been explored in various domains:

- Bayesian Neural Networks for uncertainty estimation
- Ensemble methods for confidence calibration
- Multi-modal medical imaging: Uncertainty-weighted fusion

We pioneer the application of uncertainty-aware fusion specifically to social relationship recognition, where data quality varies significantly.

3 Proposed Methodology

Our proposed system extends the baseline framework with three synergistic innovations. Each component addresses one specific limitation while maintaining end-to-end trainability.

3.1 Innovation 1: Multi-Scale Visual Feature Pyramid

3.1.1 Core Idea

Instead of using only ResNet-50's final layer (layer4), we extract features from **all four layers** and combine them using Feature Pyramid Network architecture. This creates a hierarchical representation capturing both fine details and high-level semantics.

3.1.2 Architecture Design

Step 1: Multi-Layer Feature Extraction
Extract features from all ResNet-50 layers:

$$\begin{aligned} C_1 &= \text{ResNet-Layer1}(I) && \text{256 channels, high resolution} \\ C_2 &= \text{ResNet-Layer2}(C_1) && \text{512 channels} \\ C_3 &= \text{ResNet-Layer3}(C_2) && \text{1024 channels} \\ C_4 &= \text{ResNet-Layer4}(C_3) && \text{2048 channels, low resolution} \end{aligned} \tag{4}$$

Step 2: Feature Pyramid Network Construction
Build pyramid using top-down pathway:

$$
\begin{aligned}
P_4 &= \text{Conv}_{1\times1}(C_4) \\
P_3 &= \text{Upsample}(P_4) + \text{Conv}_{1\times1}(C_3) \\
P_2 &= \text{Upsample}(P_3) + \text{Conv}_{1\times1}(C_2) \\
P_1 &= \text{Upsample}(P_2) + \text{Conv}_{1\times1}(C_1)
\end{aligned} \tag{5}
$$

All pyramid levels P_i have uniform 256-channel dimensionality, enabling consistent processing.

Step 3: Attention Integration

Apply baseline's attention mechanisms to each pyramid level independently:

$$
P_i^{att} = (A_{c\,i} \cdot A_{s\,i}) \odot P_i, \quad i \in \{1, 2, 3, 4\} \tag{6}
$$

This allows the model to focus on di erent semantic levels adaptively.

3.1.3 Theoretical Justi cation

Di erent relationship types require di erent semantic levels:

Table 3: Relationship Types and Required Feature Scales

Relationship	**Key Visual Cues and Pyramid Levels**
Couples	Romantic facial expressions (P_1), intimate hand-holding (P_2), very close proximity (P_3)
Friends	Friendly smiles (P_1), casual gestures (P_2), comfortable distance (P_3)
Family	Varied ages visible in faces (P_1), protective postures (P_2), group cohesion (P_3)
Professional	Formal expressions (P_1), business attire (P_2), office environment (P_4)

By providing all scales, we enable the model to select appropriate semantic levels for each relationship type.

3.1.4 Expected Improvements

Based on theoretical analysis and related work:

- **Fine-grained relationships**: +2-3% (Layer 1-2 provide discriminative details)
- **Couple vs. Friends**: +4-5% (Currently 72.1%, major weakness)
- **Overall PISC-F**: +1.5-2% (Better fine-grained distinctions)

3.2 Innovation 2: Iterative Cross-Modal Re nement

3.2.1 Core Idea

Replace one-shot alignment with an iterative process that progressively refines text-image correspondence through multiple rounds of bidirectional attention.

3.2.2 Architecture Design

Algorithm: Iterative Re nement Process

Algorithm 1 Iterative Cross-Modal Refinement

Require: Visual features $V^{0)} = F_I^{multi}$, Text features $T^{0)} = F_T$
Require: Number of iterations $K = 3$
Ensure: Refined features $V^{K)}, T^{K)}$
for $k = 1$ to K **do**
  // Text attends to Visual
  $T_{att}^{k)} = \text{CrossAttention}(Q = T^{k\ 1)}, K = V^{k\ 1)}, V = V^{k\ 1)})$
  $T^{k)} = \text{LayerNorm}(T^{k\ 1)} + T_{att}^{k)})$

  // Visual attends to Text
  $V_{att}^{k)} = \text{CrossAttention}(Q = V^{k\ 1)}, K = T^{k)}, V = T^{k)})$
  $V^{k)} = \text{LayerNorm}(V^{k\ 1)} + V_{att}^{k)})$
end for
return $V^{K)}, T^{K)}$

Cross-Attention Mechanism:

$$\text{CrossAttention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

where $d_k = 256$ is the feature dimension.

3.2.3 Progressive Re nement Example

Consider text "happy couple" and an image showing two people:

- **Iteration 0**: Text broadly attends to both people (imprecise, 30% correct)
- **Iteration 1**: "happy" focuses on facial regions; "couple" begins distinguishing the pair (55% correct)
- **Iteration 2**: "happy couple" refines to hand-holding area + facial expressions (78% correct)
- **Iteration 3**: Final precise alignment to all relationship indicators (92% correct)

The residual connections ($T^{k\ 1)} + T_{att}^{k)}$) ensure that information from previous iterations is preserved while new refinements are added incrementally.

3.2.4 Why K=3 Iterations?

Based on literature review:

- ALBEF uses 2-3 iterations
- Flamingo uses 3-4 cross-attention layers
- Empirical studies show diminishing returns after 3 iterations

We plan to validate this with ablation studies comparing $K \in \{1, 2, 3, 4, 5\}$.

3.2.5 Expected Improvements

- **Ambiguous text scenarios**: +2-3% (Iterative refinement corrects initial misalignments)
- **Complex multi-person images**: +2% (Progressive focus on target individuals)
- **Overall mAP**: +1.5-2%

3.3 Innovation 3: Uncertainty-Aware Adaptive Fusion

3.3.1 Core Idea

Introduce uncertainty estimation for each modality, then dynamically weight modalities based on estimated confidence levels. High-quality modalities receive higher weights; low-quality modalities are down-weighted.

3.3.2 Architecture Design

Step 1: Uncertainty Estimation

Learn to estimate uncertainty for each modality:

$$\begin{aligned} \sigma_V &= \text{Sigmoid}(\text{MLP}_V(V^{K)})) \in [0,1] \\ \sigma_T &= \text{Sigmoid}(\text{MLP}_T(T^{K)})) \in [0,1] \end{aligned} \tag{8}$$

Higher values indicate higher uncertainty (lower confidence).

Step 2: Precision-Weighted Fusion

Compute precision weights (inverse of uncertainty):

$$w_V = \frac{1}{\sigma_V + }, \quad w_T = \frac{1}{\sigma_T + } \tag{9}$$

Normalize:

$$\tilde{w}_V = \frac{w_V}{w_V + w_T}, \quad \tilde{w}_T = \frac{w_T}{w_V + w_T} \tag{10}$$

Basic weighted fusion:

$$F_{basic} = \tilde{w}_V \cdot V^{K)} + \tilde{w}_T \cdot T^{K)} \tag{11}$$

Step 3: Gated Non-Linear Fusion

Model complex interactions via gating:

$$\begin{aligned} \text{Gate} &= \text{Sigmoid}(\text{MLP}_{gate}([V^{K)}; T^{K)}])) \\ F_{fusion} &= \text{Gate} \odot F_{basic} + (1 \quad \text{Gate}) \odot V^{K)} \end{aligned} \tag{12}$$

The gate learns when to trust fused features vs. falling back to visual features.

3.3.3 Adaptive Behavior Examples

Scenario 1: Clear Image + Generic Text

- Visual: HD image, clear faces $\sigma_V = 0.15$ (high confidence)
- Text: "two people" $\sigma_T = 0.70$ (low confidence)
- Weights: $\tilde{w}_V = 0.82$, $\tilde{w}_T = 0.18$
- Decision: Rely primarily on visual (82%)

Scenario 2: Blurry Image + Detailed Text

- Visual: Blurry, poor lighting $\sigma_V = 0.68$ (low confidence)
- Text: "young couple holding hands" $\sigma_T = 0.22$ (high confidence)
- Weights: $\tilde{w}_V = 0.24$, $\tilde{w}_T = 0.76$
- Decision: Rely primarily on text (76%)

3.3.4 Expected Improvements

- **Blurry/low-quality images**: +3-4% (Down-weight unreliable visual, rely on text)
- **Ambiguous/generic text**: +2% (Down-weight unreliable text, rely on visual)
- **Overall robustness**: +1% across all scenarios

3.4 Complete System Integration

The three innovations integrate seamlessly into a unified framework:

1. **Text Processing**: LLM extraction + CNN encoding (from baseline) F_T
2. **Multi-Scale Visual**: FPN extracts P_1, P_2, P_3, P_4 F_I^{multi}
3. **Iterative Re nement**: $K = 3$ iterations refined $V^{K)}, T^{K)}$
4. **Uncertainty Fusion**: Estimate σ_V, σ_T adaptive fusion F_{fusion}
5. **Classi cation**: Fully connected layer + Softmax relationship prediction

3.5 Loss Function Design

Our training objective combines three components:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{refinement} + \lambda_2 \mathcal{L}_{uncertainty} \tag{13}$$

Cross-Entropy Loss (standard classification):

$$\mathcal{L}_{CE} = \sum_{c=1}^{C} y_c \log(p_c) \tag{14}$$

Re nement Consistency Loss (encourages consistent predictions across iterations):

$$\mathcal{L}_{refinement} = \sum_{k=1}^{K} \text{KL}(P^{k)} || P^{K)}) \tag{15}$$

Uncertainty Regularization Loss (prevents always-high uncertainty):

$$\mathcal{L}_{uncertainty} = |\sigma_V \quad 0.5| + |\sigma_T \quad 0.5| \tag{16}$$

We plan to set $\lambda_1 = 0.1$ and $\lambda_2 = 0.05$ based on literature, with ablation studies to validate.

4 Experimental Setup

4.1 Dataset

We will use the PISC (People in Social Context) dataset [7], identical to the baseline:

- **Total Images**: 22,670
- **Relationship Pairs**: 96,568
- **PISC-C** (Coarse-grained): 3 categories (Intimate, Non-intimate, No relation)
- **PISC-F** (Fine-grained): 6 categories (Friends, Family, Couple, Professional, Commercial, No relation)

Data Split (following baseline):

- Training: 80% (18,136 images)
- Validation: 10% (2,267 images)
- Testing: 10% (2,267 images)

4.2 Implementation Plan

4.2.1 Hardware Resources

We have access to:

- GPU: NVIDIA RTX 3080/3090 (via university lab or cloud services)
- Alternative: Google Colab Pro with A100 GPU

4.2.2 Software Framework

- PyTorch 2.0+ (deep learning framework)
- Transformers library (for LLM components)
- Timm library (for pre-trained ResNet-50)
- Weights & Biases (experiment tracking)

Table 4: Planned Hyperparameters

Parameter	**Value**
Batch Size	32
Learning Rate	1e-4
Optimizer	AdamW
Epochs	50
FPN Channels	256
Refinement Iterations (K)	3
λ_1 (Refinement Loss)	0.1
λ_2 (Uncertainty Loss)	0.05

4.2.3 Key Hyperparameters

4.3 Evaluation Metrics

Following the baseline paper, we will use:

1. **Mean Average Precision (mAP)**: Primary metric for overall performance
2. **Per-Class Accuracy**: To analyze performance on each relationship type
3. **F1-Score**: Macro-averaged across all classes
4. **Confusion Matrix**: To identify specific error patterns

4.4 Experimental Validation Strategy

Our experiments will proceed in three phases:

Phase 1: Baseline Reproduction

- Implement baseline model exactly as described in the paper
- Reproduce reported results (88.9% PISC-C, 76.9% PISC-F)
- Validate our implementation and understanding

Phase 2: Component-Wise Evaluation (Ablation Studies)

- Test each innovation individually:
  - Baseline + FPN only
  - Baseline + Iterative Refinement only
  - Baseline + Uncertainty Fusion only
- Quantify each component's individual contribution

Phase 3: Complete System Evaluation

- Test full model with all three innovations
- Compare against baseline and other methods from the paper (GR2N, MT-SRR, GA-GCN)
- Analyze performance on specific challenging scenarios

5 Progress Report

5.1 Completed Work (Weeks 1-8)

5.1.1 Week 1-2: Literature Review

We conducted comprehensive literature review focusing on:

- **Social Relationship Recognition**: Studied 15+ papers including GR2N, GA-GCN, MT-SRR, and recent multimodal approaches
- **Feature Pyramid Networks**: Reviewed original FPN paper and applications in object detection, segmentation, and vision-language tasks
- **Iterative Re nement**: Analyzed ALBEF, Flamingo, BLIP-2, and other iterative cross-modal methods
- **Uncertainty Estimation**: Studied Bayesian approaches, ensemble methods, and uncertainty-aware fusion in medical imaging

Key Findings Documented:

- FPN has not been applied to social relationship recognition (research gap)
- Iterative refinement consistently improves performance in vision-language tasks
- Uncertainty-aware fusion is underexplored in this domain

5.1.2 Week 3-4: Baseline Paper Deep Analysis

We performed detailed analysis of the baseline paper:

- Studied all four modules (text extraction, image feature extraction, alignment, fusion)
- Analyzed mathematical formulations (Equations 1-7 in baseline)
- Examined experimental results (Tables 1-7, Figures 1-5 in baseline)
- Identified three critical limitations with supporting evidence from the paper

Analysis Documented in Section 2 of this report.

5.1.3 Week 5-6: Architectural Design

We designed our three innovations in detail:

Innovation 1 - Multi-Scale FPN:

- Determined FPN architecture: 4-level pyramid with 256 channels per level
- Designed integration with attention mechanisms
- Planned aggregation strategy for multi-scale features

Innovation 2 - Iterative Re nement:

- Designed bidirectional cross-attention mechanism
- Determined K=3 iterations based on literature
- Planned residual connections and layer normalization

Innovation 3 - Uncertainty Fusion:

- Designed uncertainty estimator architecture (MLP-based)
- Formulated precision-weighted fusion mechanism
- Designed gated fusion for non-linear interactions

Documentation: Created detailed architectural diagrams and mathematical formulations (Sections 3.1-3.3).

5.1.4 Week 7-8: Implementation Planning and Setup

Development Environment Setup:

- Installed PyTorch 2.0.1, CUDA 11.8, Python 3.9
- Set up virtual environment with all required libraries
- Configured GPU access (university lab / cloud resources)
- Created GitHub repository for code version control

Dataset Acquisition:

- Downloaded PISC dataset (22,670 images)
- Studied dataset structure and annotation format
- Prepared data loading pipeline design

Project Organization:

- Divided team into three pairs (2 members per innovation)
- Established communication channels (WhatsApp group, weekly meetings)
- Created shared documentation (Google Docs, Overleaf for report)
- Set up experiment tracking plan (Weights & Biases account)

5.2 Work Distribution Among Team Members

Our 6-member team is organized as follows:

Collaboration Strategy:

- Weekly team meetings (Saturdays, 2 hours)
- Daily progress updates via WhatsApp
- Code reviews before merging (minimum 2 approvals)
- Shared documentation updated continuously

Table 5: Team Work Distribution

Team Members	Responsibility	Timeline
Members 1 & 2	Innovation 1: Multi-Scale FPN implementation, visual feature extraction	Weeks 9-12
Members 3 & 4	Innovation 2: Iterative Refinement implementation, cross-attention mechanism	Weeks 9-12
Members 5 & 6	Innovation 3: Uncertainty Fusion implementation, system integration, experiments	Weeks 9-14

Table 6: Project Completion Status

Phase	Completion
Literature Review	100%
Baseline Analysis	100%
Architectural Design	100%
Environment Setup	100%
Dataset Preparation	100%
Implementation	0% (starts Week 9)
Experiments	0% (starts Week 13)
Final Report	30% (this midterm report)
Overall Project	**45%**

5.3 Current Status Summary

6 Implementation Plan and Timeline

6.1 Detailed Implementation Roadmap

6.1.1 Weeks 9-10: Baseline Implementation

Objective: Reproduce baseline results to validate our understanding.

Tasks:

- Implement LLM text extraction module
- Implement ResNet-50 + attention for image features
- Implement cosine similarity alignment
- Implement MLP-based fusion and classification
- Train on PISC dataset
- Validate results match paper (88.9% PISC-C, 76.9% PISC-F)

Deliverable: Working baseline model with reproduced results.
Responsibility: All team members collaborate on understanding baseline.

6.1.2 Weeks 11-12: Component Implementations

Pair 1 (Members 1 & 2): Multi-Scale FPN

- Modify ResNet-50 to output features from layers 1-4
- Implement FPN architecture with top-down pathway
- Integrate attention mechanisms at each pyramid level
- Test with baseline (FPN replaces single-scale features)
- Target: +1-2% mAP over baseline

Pair 2 (Members 3 & 4): Iterative Re nement

- Implement cross-attention mechanism
- Build iterative refinement loop with K=3 iterations
- Add residual connections and layer normalization
- Test with baseline (replaces one-shot alignment)
- Target: +1.5-2% mAP over baseline

Pair 3 (Members 5 & 6): Uncertainty Fusion

- Implement uncertainty estimator MLPs
- Implement precision-weighted fusion
- Implement gated non-linear fusion
- Test with baseline (replaces simple fusion)
- Target: +1% mAP over baseline

Deliverables: Three independent component implementations, each validated individually.

6.1.3 Weeks 13-14: System Integration and Experiments

Week 13: Integration

- Integrate all three components into unified system
- Debug integration issues
- Optimize memory usage and training speed
- Initial training runs on full system

Week 14: Ablation Studies

- Experiment 1: Baseline only
- Experiment 2: Baseline + FPN
- Experiment 3: Baseline + Refinement
- Experiment 4: Baseline + Uncertainty
- Experiment 5: Baseline + FPN + Refinement
- Experiment 6: Baseline + FPN + Uncertainty
- Experiment 7: Baseline + Refinement + Uncertainty
- Experiment 8: Full model (all three components)

Deliverable: Complete experimental results with ablation analysis.

6.1.4 Week 15: Analysis and Visualization

- Generate attention visualizations for FPN levels
- Visualize iterative refinement progression
- Analyze uncertainty estimates vs. actual errors
- Create confusion matrices
- Identify failure cases and error patterns
- Compare with baseline and other methods (GR2N, MT-SRR, GA-GCN)

Deliverable: Comprehensive analysis with visualizations and insights.

6.1.5 Week 16: Final Report and Presentation

- Write final report following Turkish Journal of EE&CS template
- Include all sections: Introduction, Related Work, Methodology, Experiments, Results, Conclusion
- Create presentation slides (20-25 slides)
- Prepare demo (if time permits)
- Finalize code repository with documentation
- Prepare supplementary materials

Deliverable: Complete final submission package.

6.2 Timeline Summary

Table 7: Detailed Project Timeline

Week	Milestone
	Completed Weeks 1-8)
1-2	Literature review completed
3-4	Baseline analysis completed
5-6	Architectural design completed
7-8	Environment setup and dataset preparation completed
	Planned Weeks 9-16)
9-10	Baseline implementation and reproduction
11-12	Three component implementations
13	System integration
14	Ablation studies and experiments
15	Analysis, visualization, comparison
16	Final report and presentation
Jan 25, 2026	**Final Submission Deadline**

6.3 Risk Management

Table 8: Identified Risks and Mitigation Strategies

Risk	Probability	Mitigation Strategy
Baseline reproduction fails	Medium	Allocate 2 full weeks (9-10), contact paper authors if needed, use open-source implementations as reference
Component integration issues	Medium	Modular design with clear interfaces, extensive testing, 1 week bu er (Week 13)
Insufficient performance gains	Low	Conservative targets (+3% mAP), each component independently validated, theoretical analysis supports improvements
GPU resource constraints	Medium	Use Google Colab Pro as backup, optimize batch size, implement mixed-precision training
Team member availability	Low	Pair programming ensures knowledge sharing, clear documentation, regular meetings
Dataset issues	Very Low	PISC is well-established and publicly available, already downloaded and verified

Contingency Plans:

- If baseline reproduction takes longer: Reduce ablation experiments, focus on full model
- If GPU resources insufficient: Use smaller batch sizes, reduce number of training epochs
- If performance gains insufficient: Emphasize qualitative analysis and architectural contributions
- If integration fails: Submit individual components with theoretical integration plan

7 Expected Contributions and Impact

7.1 Technical Contributions

Our project will contribute:

1. **First application of Feature Pyramid Networks to social relationship recognition**

- Demonstrates importance of multi-scale visual features
- Shows di erent relationships require di erent semantic levels
- Provides architectural blueprint for future work

2. **Novel iterative cross-modal re nement mechanism**
   - First iterative approach specifically for relationship recognition
   - Demonstrates progressive alignment improves over one-shot methods
   - Provides insights on optimal number of iterations (K=3)
3. **Uncertainty-aware adaptive fusion for relationship recognition**
   - Pioneers application of uncertainty estimation in this domain
   - Demonstrates adaptive behavior improves robustness
   - Provides interpretable confidence scores
4. **Comprehensive experimental validation**
   - Extensive ablation studies quantifying each component
   - Demonstrates synergistic e ects of combined innovations
   - Provides insights for future research directions

7.2 Expected Performance Improvements

Based on theoretical analysis and related work, we expect:

Table 9: Expected Performance Improvements

Metric/Scenario	Baseline	Our Method (Expected)
PISC-C (Coarse-grained)	88.9%	92.0-92.5% (+3.1-3.6%)
PISC-F (Fine-grained)	76.9%	80.5-81.0% (+3.6-4.1%)
Couple vs. Friends	72.1%	78.0-79.0% (+5.9-6.9%)
Outdoor scenes	68.3%	73.0-74.0% (+4.7-5.7%)
Multi-person images	71.5%	76.0-77.0% (+4.5-5.5%)

Conservative Target: +3% mAP overall (achievable with existing designs)
Optimistic Target: +5% mAP overall (if all components synergize well)

7.3 Broader Impact

Our work has potential applications in:

- **Social Media Platforms**: Automatic relationship tagging in photos, privacy setting recommendations
- **Security and Surveillance**: Identifying suspicious relationship patterns, crowd behavior analysis

- **Assistive Technologies**: Describing social contexts for visually impaired users
- **Digital Photo Organization**: Smart album creation based on relationships
- **Human-Robot Interaction**: Robots understanding social dynamics to interact appropriately

7.4 Learning Outcomes for Team

This project provides valuable learning experiences:

- Deep understanding of multimodal learning and cross-modal fusion
- Hands-on experience with PyTorch and modern deep learning practices
- Research methodology: literature review, problem identification, solution design
- Collaboration and project management in a team setting
- Scientific writing and presentation skills

8 Conclusion

This midterm report presents our progress on developing an enhanced multimodal social relationship recognition system. Through comprehensive analysis of the baseline paper by Wang et al., we identified three critical limitations: single-scale visual features, one-shot static alignment, and simple linear fusion. We proposed three synergistic innovations to address these limitations:

1. **Multi-Scale Visual Feature Pyramid** using FPN to capture hierarchical semantic information from fine-grained details to scene-level context
2. **Iterative Cross-Modal Re nement** mechanism for progressive text-image alignment through bidirectional attention
3. **Uncertainty-Aware Adaptive Fusion** for robust multimodal integration that adapts to varying data quality

Our progress to date includes:

- **Completed**: Comprehensive literature review (Weeks 1-2)
- **Completed**: Detailed baseline analysis identifying specific weaknesses (Weeks 3-4)
- **Completed**: Complete architectural design with mathematical formulations (Weeks 5-6)
- **Completed**: Development environment setup and dataset preparation (Weeks 7-8)

We have a clear and realistic implementation plan for the remaining 8 weeks (Weeks 9-16), with well-defined milestones, risk mitigation strategies, and deliverables. Our team of 6 members is organized into three pairs, each responsible for one innovation, ensuring efficient parallel development.

Based on theoretical analysis and related work, we expect our complete system to achieve 3-5% mAP improvement over the baseline, with particularly strong gains on challenging scenarios such as fine-grained relationship distinctions (Couple vs. Friends) and varying image qualities. Beyond performance improvements, our work will contribute novel architectural insights and demonstrate the value of multi-scale features, iterative refinement, and uncertainty-aware fusion for social relationship recognition.

We are confident that with our detailed plan, strong theoretical foundation, and committed team, we will successfully complete this project and make meaningful contributions to the field of social relationship recognition.

Acknowledgments

We thank our project advisor Dr.Gamze USLU for guidance and valuable feedback throughout this project. We also acknowledge the computational resources that will be provided by Üsküdar University Computer Engineering Department and cloud computing services for model training.

References

[1] H. Wang, Z. Zhang, M. Xia, D. Huang, R. Chang, and S. Guo, "Multimodal Social Relationship Recognition Based on LLM," *IEEE Access*, vol. 13, pp. 149247-149259, 2025.

[2] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR)*, 2017, pp. 2117-2125.

[3] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid Scene Parsing Network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. CVPR)*, 2017, pp. 2881-2890.

[4] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before Fuse: Vision and Language Representation Learning with Momentum Distillation," in *Proc. Adv. Neural Inf. Process. Syst. NeurIPS)*, 2021, pp. 9694-9705.

[5] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al., "Flamingo: A Visual Language Model for Few-Shot Learning," in *Proc. Adv. Neural Inf. Process. Syst. NeurIPS)*, 2022, pp. 23716-23736.

[6] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," *arXiv preprint arXiv:2301.12597*, 2023.

[7] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "People in Social Context (PISC) Dataset," Technical Report, 2017.