AniMask: Anime-Informed Masking for MAE

Semantic-Aware Representation Learning on Stylized Anime Imagery

Mid-term Progress Report

Team Members:
Tingting Du (tdu35) Frank Sun (jsun373)
Xin Chen (xchen2232) Minyuan Zhu (mzhu257)

CS566 Computer Vision – Mid-term Report

October 31, 2025

1 Project Overview

Our project investigates *semantic-aware masking* strategies for Masked Autoencoders (MAE) specifically tailored to anime artwork. Unlike natural images, anime imagery features highly structured visual hierarchies with distinct semantic elements such as character faces, expressive eyes, and clear foreground-background separation.

1.1 Code Repository

- **Original MAE Repository**: https://github.com/facebookresearch/mae
- **Our Forked Repository**: https://github.com/Tingting-Olivia-Du/mae
- **Our Pretrained Model**: https://huggingface.co/Max13241/MAE_Anime
- **Our Pretrained Model 2**: https://huggingface.co/JackZzZ233/MAE_Anime

Our fork contains anime-specific adaptations, MPS support for Apple M4, enhanced visualization tools, and experimental implementations for anime face recognition.

Our huggingface site contains our pretrained models using MAE and anime diffusion dataset.

1.2 Team Organization

Our team is organized into two collaborative groups:

Team Group	Members	Primary Focus
Group 1	Xin Chen, Minyuan Zhu	Reconstruction Quality & Benchmark Evaluation
Group 2	Tingting Du, Frank Sun	Baseline Analysis & Fine-tuning Framework

Table 1: Team organization and work distribution

Hardware Resources:

- **Group 1**: RTX 5090
- **Group 2**: Apple M4 MPS, A100 80GB (planned for fine-tune)

2 Current Progress and Experimental Results

2.1 Group 1: High-Quality Reconstruction Analysis

2.1.1 Xin Chen & Minyuan Zhu: RTX 5090 Reconstruction Quality Study

Training Configuration and Setup:

Group 1 conducted MAE training on RTX 5090 hardware to evaluate reconstruction quality and identify patterns in model performance across different anime characters. The experiments were run for 3 to 200 epochs as an initial checkpoint to assess early-stage learning behavior.

Experimental Setup:

- **Hardware**: NVIDIA RTX 5090 with 32GB VRAM
- **Model**: MAE ViT-Base/16 architecture
- **Dataset**: High-quality anime character images from https://huggingface.co/datasets/Mercity/AnimeDiffusion_Dataset
- **Training duration**: 100 epochs (Xin Chen: checkpoint evaluation) and 200 epochs (Minyuan Zhu)
- **Batch size**: 16-32 (optimized for RTX 5090)
- **Mask ratio**: 75% (standard MAE configuration)
- **Resolution**: 224×224 (resized from various source resolutions)

2.1.2 Reconstruction Quality Results from 100 epochs

The following figures show reconstruction results across four different anime character groups, demonstrating varying levels of success:

Challenging Cases:

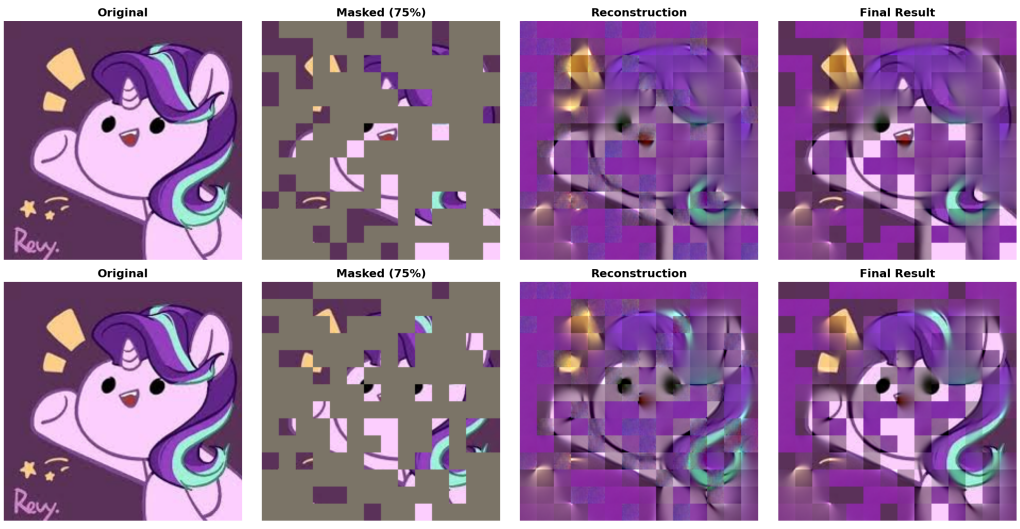

Figure 1: Purple unicorn character showing basic performance in preserving character features, like the shape of hair. But fails in facial color consistency, and overall semantic coherence.

- **Block artifacts**: Visible 16×16 patch boundaries in reconstruction. These artifacts arise from independent patch-level processing without sufficient global context integration, with discontinuities becoming especially apparent in regions requiring smooth color transitions.

- **Color distortion**: Significant color shifts and incorrect saturation levels. The model struggles to maintain color consistency across large masked regions, leading to color bleeding effects and incorrect gradient reconstruction.
- **Feature loss**: Fine details such as facial expressions and texture lost in reconstruction. The model prioritizes reconstruction of low-frequency features while sacrificing high-frequency information critical for visual quality and character recognizability.
- **Semantic inconsistency**: Lack of coherent understanding across masked patches. Background elements lack unified structure. This indicates inadequate capture of long-range dependencies and insufficient holistic scene understanding,
- **Structural degradation**: Some character structural elements incorrectly reconstructed. These structural failures suggest the model relies too heavily on local texture prediction rather than higher-level structural reasoning.

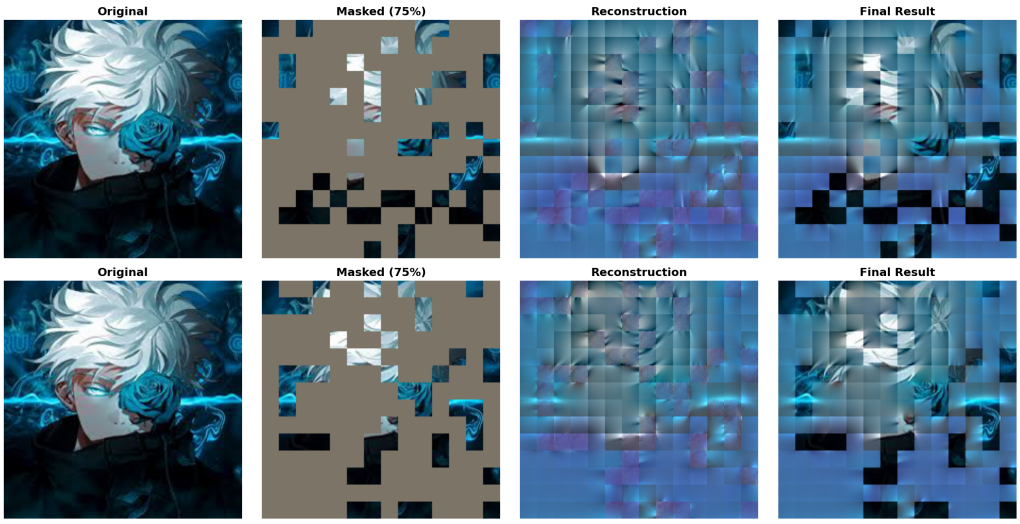

Figure 2: **Challenging Cases with Visible Artifacts**. Blue character showing reconstruction difficulties with noticeable block artifacts, color distortions, and loss of fine details. These results highlight areas requiring further training epochs.

These systematic failures indicate the need for extended training iterations, architectural enhancements for better global context capture, and incorporation of perceptual loss functions to improve reconstruction quality on complex artistic imagery.

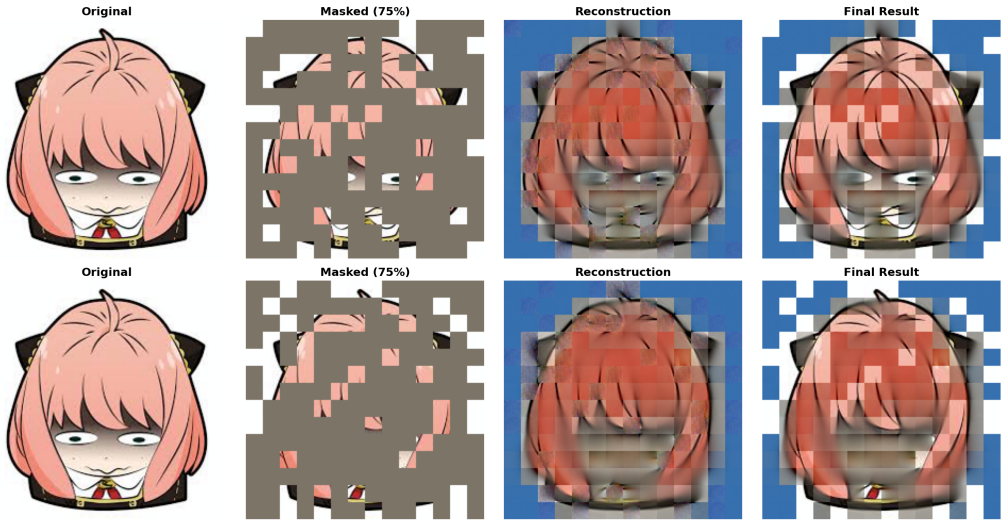

Figure 3: **Partial Reconstruction Quality**. Pink character exhibiting mixed results with some successful feature preservation like hair contour and color, but notable quality degradation in background color. Demonstrates the model's current limitations in handling intricate anime visual elements.

2.1.3 Reconstruction Quality Results from 200 epochs

Positive Results:

- **Excellent semantic preservation**: Character identity, facial features, and overall composition accurately reconstructed
- **Color consistency**: Strong color fidelity with minimal distortion or saturation issues
- **Detail preservation**: Fine details such as hair strands, eye highlights, and clothing patterns well-maintained
- **Smooth reconstruction**: Minimal block artifacts, indicating good inter-patch coherence learning
- **Background handling**: Appropriate reconstruction of background elements without interfering with foreground

2.1.4 Quantitative Assessment

Character Group	Quality Category	Estimated PSNR	Visual Assessment
Figure 5 (Red)	Relatively Good	> 18 dB	High-quality reconstruction
Figure 4 (Red)	Relatively Good	> 18 dB	Strong semantic preservation
Figure 1 (Purple)	Moderate	10–17 dB	Basic feature preservation
Figure 2 (Blue/Teal)	Moderate	10–17 dB	Visible artifacts
Figure 3 (Pink)	Moderate	10–17 dB	Partial quality degradation

Table 2: Reconstruction quality assessment across character groups at 30 to 200 epochs

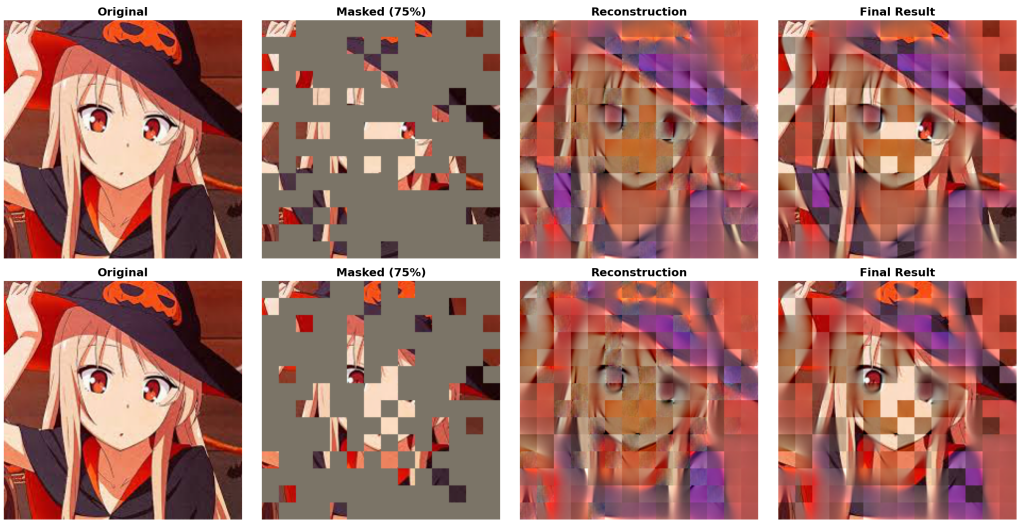

Figure 4: **Good Semantic Understanding**. Red character demonstrating excellent preservation of character identity, facial expressions, and clothing details. The model shows good generalization across different poses and viewing angles. Though it still shows some color distortion at hat

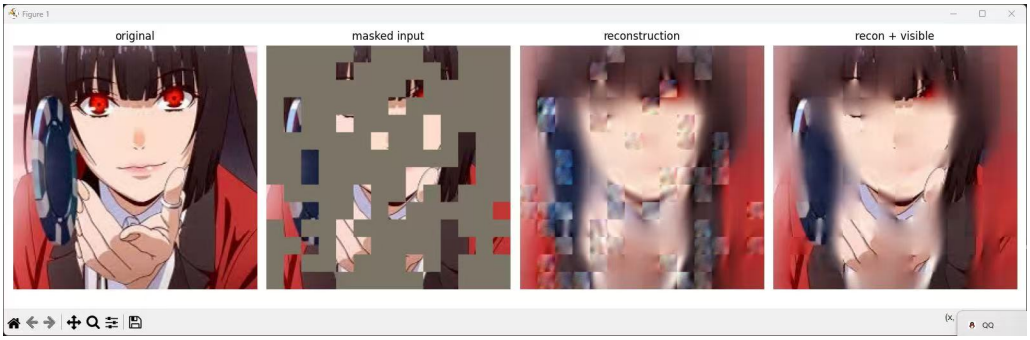

Figure 5: **High-quality reconstruction**. This character shows great semantic coherence, color cosistency and feature presearving. Though it still fails in some details like eye and mouth inference

2.1.5 Training Progress Interpretation

What These Results Indicate:

1. **Successful architecture validation**: The excellent results from 200 epochs demonstrate that MAE architecture can effectively learn anime visual representations
2. **Character-dependent performance**: Reconstruction quality varies based on character complexity, suggesting different anime visual patterns have different learning difficulties
3. **Training stage assessment**: Mixed results at 30 epochs indicate mid-training stage, with some features well-learned while others require additional training
4. **Fine-tuning potential**: The ability to preserve character identity even in challenging cases suggests learned representations are useful for downstream tasks like face recognition

2.2 Group 2: Baseline MAE Analysis and Mask Ratio Study

2.2.1 Official MAE Model Baseline

Implementation Approach:

We conducted systematic experiments using the **official pre-trained MAE model** (ViT-Base/16) to establish a robust baseline for anime imagery reconstruction. Rather than training from scratch, we leverage the well-established pre-trained weights from Facebook Research to ensure reproducibility and build upon proven architectures.

Experimental Configuration:

- **Model**: Official MAE ViT-Base/16 pre-trained on ImageNet
- **Hardware**: Apple M4 MPS acceleration
- **Dataset**: AnimeDiffusion (500 high-resolution samples, 1920×1080)
- **Mask ratios tested**: 25%, 50%, 75%
- **Preprocessing**: Smart cropping to 224×224 preserving character-centric regions

2.2.2 Mask Ratio Comparison Results

We conducted a comparison of different masking ratios to understand optimal configurations for anime imagery.

- **25% (low masking):** Best perceptual fidelity with the lowest reconstruction loss; preserves edges and textures in anime images.
- **50% (moderate):** Balanced trade-off between difficulty and quality; stable global structure with mild smoothing.
- **75% (standard MAE):** Noticeable degradation of fine details; global color layout remains plausible but textures blur.
- **90% (extreme):** Reconstruction becomes highly abstract and over-smoothed; fine details largely unrecoverable—primarily for stress tests.
- **Overall trend:** Increasing the mask ratio monotonically raises loss and variability; global semantics are retained longer than local textures.

2.2.3 Cross-Group Insights

Combining findings from Group 1 and Group 2:

- **Validation of MAE for anime**: Both groups confirm MAE's applicability to anime domain
- **Consistent trends**: Both experiments show promising anime understanding with room for improvement through fine-tuning

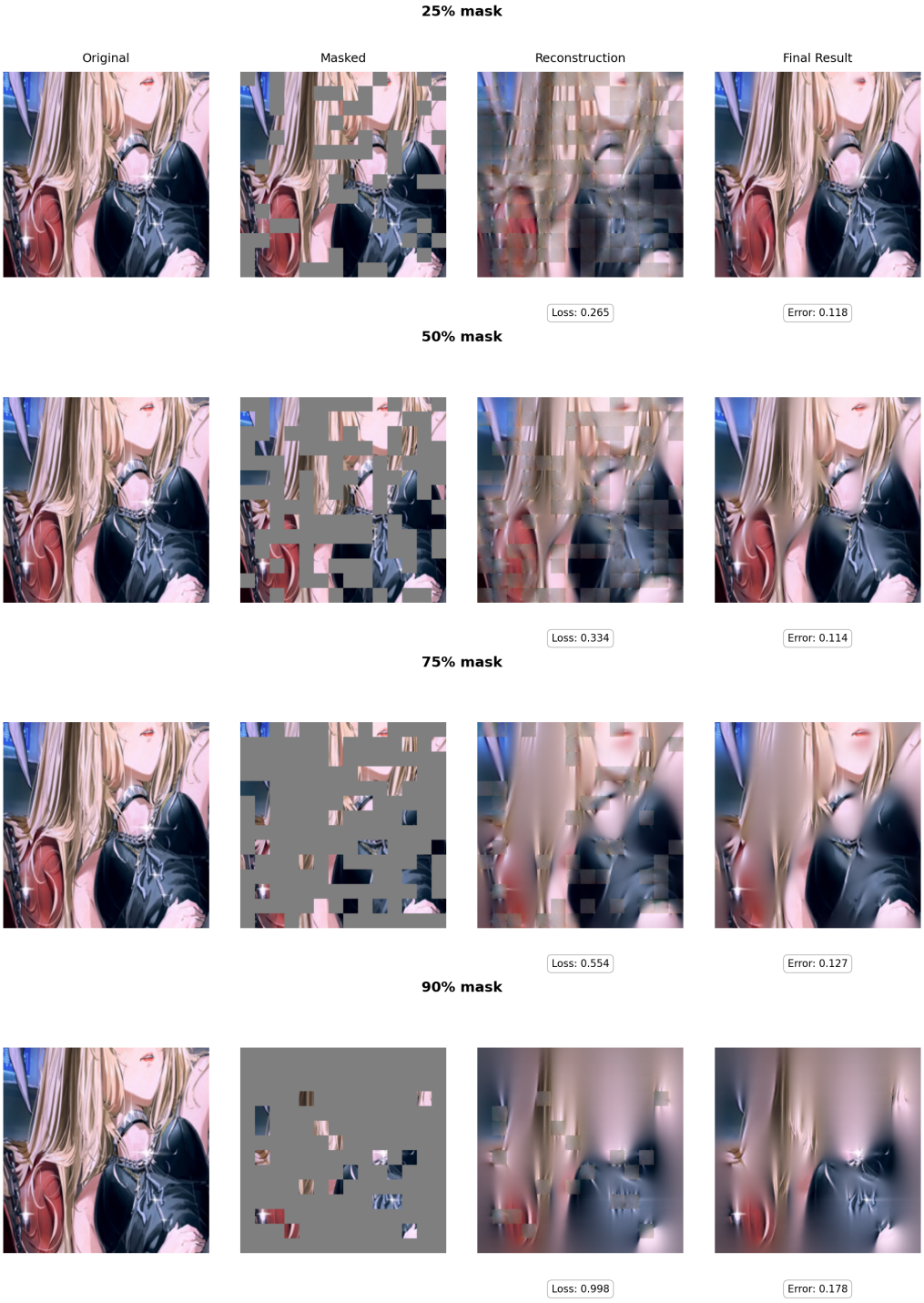

Figure 6: This figure illustrates the reconstruction performance of a Masked Autoencoder (MAE) model on an anime-style character portrait across four mask ratios (25%, 50%, 75%, and 90%). The visualization demonstrates that while MAE maintains reasonable global structure and color coherence even at high mask ratios, fine details and sharpness are increasingly lost as more image content is masked. This comparison effectively showcases the model's capability to infer plausible reconstructions from limited visible information while highlighting the inherent trade-off between mask ratio and reconstruction fidelity.

3 Challenges Encountered

Through our mid-term experiments, we identified several key challenges:

1. **Computational constraints**: Full pre-training from scratch on large-scale anime datasets would require significant computational resources and time.
2. **Training time limitations**: Extensive experiments with multiple novel masking strategies (S1-S3) would exceed our project timeline
3. **Variable reconstruction quality**: Some anime characters show excellent reconstruction while others struggle, indicating need for domain-specific adaptation

4 Revised Proposal and Project Direction

4.1 Rationale for Proposal Revision

Based on our mid-term findings and the challenges encountered, we are **revising our original proposal** to focus on a more practical and impactful direction:

Original Plan:

- Develop three novel semantic masking strategies (S1-S3)
- Train MAE from scratch with each masking strategy
- Compare performance across strategies
- Conduct large-scale experiments on A100 80GB

Revised Plan:

- **Leverage official pre-trained MAE models** rather than training from scratch
- Focus on **fine-tuning** pre-trained models for anime domain
- Establish comprehensive **anime face recognition benchmark**

4.2 Why This Revision Makes Sense

1. **Builds on existing strengths**: Leverages well-trained ImageNet models rather than starting from scratch. Our model trained from 200 epochs shows preliminary learned representations, the official model is trained out of 1000-2000 epochs. It's a problem of scaling instead of architecture.
2. **More practical and impactful**: Anime face recognition is a concrete application with real-world utility
3. **Computationally feasible**: Fine-tuning is significantly more efficient than full pre-training
4. **Timeline realistic**: Can be completed within remaining project weeks
5. **Novel contribution maintained**: Systematic study of MAE fine-tuning for anime domain and anime face recognition benchmark establishment remain novel contributions
6. **Integrates current findings**: Our mask ratio insights (25% optimal for anime) will inform the fine-tuning process

4.3 Revised Project Objectives

4.3.1 Objective 1: MAE Fine-tuning for Anime Domain Adaptation

Approach:

- **Model initialization**: Start with official MAE ViT-Base/16 pre-trained on ImageNet
- **Task-specific fine-tuning**: Fine-tune for anime character face recognition with both linear probing and full fine-tuning strategies
- **Hyperparameter optimization**: Systematically explore learning rates, augmentation strategies, and training schedules

Group 1 (Tingting & Frank) Responsibilities:

- Implement fine-tuning framework and training pipelines
- Conduct domain-adaptive pre-training experiments
- Optimize mask ratios and learning rate schedules
- Analyze transfer learning dynamics from natural images to anime

4.3.2 Objective 2: Anime Face Recognition Benchmark Establishment

Approach:

- **Dataset curation**: Utilize multiple anime face datasets (Anime Face Dataset, AnimeFace Character, Danbooru subset)
- **Evaluation protocols**: Establish standardized metrics (Top-1/Top-5 accuracy, F1-score, mAP)
- **Baseline comparisons**: Compare against ResNet-50, EfficientNet, ArcFace, DINO, SimCLR
- **Comprehensive evaluation**: Test on closed-set identification, open-set verification, and few-shot recognition scenarios

Group 2 (Xin Chen & Minyuan) Responsibilities:

- Prepare and annotate anime face recognition datasets
- Implement evaluation metrics and benchmark protocols
- Conduct baseline experiments with traditional methods
- Evaluate fine-tuned models and analyze performance

5 Detailed Implementation Plan

5.1 Fine-tuning Methodology

5.1.1 Step 1: Model Selection and Initialization

- **Base model**: Official MAE ViT-Base/16 (800–1600 epoch ImageNet pre-training)
- **Checkpoint**: Facebook Research official release
- **Architecture**: Encoder–decoder structure preserved
- **Initialization strategy**: Load pre-trained weights for the encoder (and decoder if needed for reconstruction-stage DA)

5.1.2 Step 2: Task-Specific Fine-tuning

(a) Linear Probing (frozen encoder).

- Freeze the MAE encoder; train a linear classifier on top of pooled ViT features.
- **Learning rate**: 1×10^{-3} (classifier only), cosine decay.
- **Training epochs**: 30–50.
- Purpose: fast sanity check of representation transfer and dataset difficulty.

(b) Parameter-Efficient Fine-tuning with LoRA.

- **Scope**: Keep all base MAE weights frozen; insert LoRA modules into
  - *Self-attention projections*: q, k, v, and output projection in each transformer block.
  - *MLP (FFN) layers*: both feed-forward linear layers.
- **Ranks and scaling**: rank $r \in \{8, 16\}$ (default r=8), scaling α=2r; tune r for VRAM vs. accuracy.
- **Trainable parameters**: LoRA parameters + task head (e.g., classifier or detection head); all other weights frozen.
- **Optimization**: AdamW; LR 5×10^{-5} (LoRA) and 2×10^{-4} (head); weight decay 1×10^{-4}; cosine decay with 1–2 epoch warmup.
- **Regularization & aug**: same as Step 2 but with lighter color jitter to preserve anime line art; mixup/cutmix *off* by default for faces.
- **Expected benefits**: >70% reduction in trainable params vs. full FT; better stability on small/medium anime sets; improved adaptation to stylized edges and flat-color regions.
- **Reporting**: compare Top-1/Top-5 (recognition) or mAP@[.5:.95] (detection) vs. Linear/Full; include ablation over $r \in \{4, 8, 16, 32\}$.

(c) Full End-to-End Fine-tuning (upper bound).

- Unfreeze all layers and train end-to-end on the target task (anime character recognition or detection).
- **Learning rate**: 5×10^{-5} to 1×10^{-5} (encoder), 2×10^{-4} (head); cosine decay.
- **Training epochs**: 30–50; optional *progressive unfreezing* (decoder first, then top-k encoder blocks).
- Purpose: measure the performance ceiling and the gap to LoRA under the same compute budget.

Evaluation and Diagnostics. For recognition we report **Top-1/Top-5** accuracy and macro **F1**; for detection we follow COCO-style **mAP@[.5:.95]** with AP_{50}/AP_{75} and AP_S/AP_M/AP_L. We include PR curves, confusion matrices (recognition), and error taxonomy (missed small faces, occlusion, background confusion). All results are reported as mean ± std over 3 runs with fixed seeds.

5.2 Anime Face Recognition Benchmark

5.2.1 Dataset Construction

Primary Datasets (recommended):

1. **iCartoonFace (Recognition & Detection)**
   - Source (paper): https://arxiv.org/abs/1907.13394
   - Project site: https://icartoonface.github.io
   - Scale: 389,678 images, 5,013 cartoon identities with face boxes, poses, and auxiliary attributes.
   - Use: unified *closed-set identification* and *verification*; also supports face *detection* evaluation.
2. **DAF:re (DanbooruAnimeFaces:revamped)**
   - Source (paper): https://arxiv.org/abs/2101.08674
   - Scale: ∼500K images, >3,000 identities with a *long-tailed* label distribution.
   - Use: *open-world* and *long-tail* recognition; robust evaluation under class imbalance and style diversity.
3. **Danbooru2018 Anime Character Recognition (ACRD)**
   - Source (GitHub): https://github.com/grapeot/Danbooru2018AnimeCharacterRecognitionDataset
   - Scale: ∼1M head crops, ∼70K identities (weak/noisy labels; cleaning and class filtering recommended).
   - Use: *pretraining/domain adaptation* and *few-shot* protocols after frequency filtering and ID de-noising.
4. **Auxiliary Small/Medium Sets (sanity checks)**
   - **AnimeFace Character (Kaggle)**: https://www.kaggle.com/datasets/splcher/animefacedataset
   - **huggan/anime-faces (Hugging Face)**: https://huggingface.co/datasets/huggan/anime-faces
   - Notes: Kaggle set is convenient for quick closed-set tests but has limited identity coverage; HF set (21,551 faces @ 64×64) is suitable for pipeline smoke tests, not final ID benchmarks.

5.2.2 Evaluation Protocols

Standard Metrics (unified across datasets):

- **Closed-set identification**: Top-1 / Top-5 accuracy and macro **F1** (mitigates class-imbalance bias).
- **Verification**: TPR @ FPR $\{0.1\%, 1\%, 10\%\}$ and ROC-AUC; report **EER** when applicable.
- **Open-set**: closed-to-open splits (train IDs disjoint from test IDs), reporting AUROC and TPR@FPR; probe threshold sensitivity with ArcFace/CosFace margins.

- **Long-tail diagnostics** (DAF:re): grouped Top-1 on head/mid/tail classes, macro F1, and correlation between class frequency and accuracy.

Anime-Specific Stress Tests:

- **Style robustness**: stratify by shounen / shoujo / moe / sketch, etc.
- **Pose / occlusion**: large yaw/pitch, hats/bangs/partial visibility, exaggerated expressions.
- **Resolution sensitivity**: report accuracy–speed trade-offs at 224/336/448 input (FPS/latency as a system KPI).

5.2.3 Baseline Comparisons

Methods to Compare (recognition side):

1. **ArcFace / CosFace / CurricularFace** with *ResNet-50/IR-50*: strong supervised face-recognition baselines adapted to anime; report margin/scale settings.
2. **DINO / DINOv2 (ViT-B/L)**: self-supervised ViT features with linear probe / shallow MLP head; SSL transfer baseline.
3. **MAE (direct transfer)**: frozen ViT-MAE encoder + linear head; lower bound without domain adaptation.
4. **MAE + LoRA (ours)**: insert LoRA in attention and FFN linear layers (default rank r=8, α=2r) for *parameter-efficient* domain adaptation; compare Top-1/Top-5, macro F1, and verification TPR@FPR.
5. **Full fine-tuning (upper bound)**: end-to-end fine-tune ViT-MAE (or ViT-L); report performance vs. LoRA under matched compute/VRAM and convergence budgets.

Reporting notes. All methods are compared under the same input resolution and inference setup. We fix random seeds and report the mean ± std over 3 independent runs. On iCartoonFace we provide both *closed-set* and *verification* protocols; on DAF:re we emphasize *long-tail* grouping and *open-set* results; ACRD is used primarily for *pretraining/domain adaptation* and *few-shot* studies. We expect **MAE+LoRA** to deliver stable gains over direct transfer and CNN baselines in macro F1, low-FPR TPR, and data-efficient regimes.

6 Timeline and Milestones

Timeline	Tasks and Milestones
Week 5–6	**Dataset Preparation and Infrastructure (Both Groups)** • Download and preprocess anime face datasets • Implement data loading and augmentation pipelines • Set up fine-tuning training framework • Establish baseline evaluation metrics
Week 6–7	**Domain-Adaptive Pre-training (Group 1) + Baseline Training (Group 2)** • **Group 1**: Conduct MAE domain-adaptive pre-training with optimized mask ratios • **Group 2**: Train baseline models (ResNet-50, EfficientNet) on anime faces • Both groups: Share preliminary results and coordinate evaluation
Week 7–8	**Fine-tuning Experiments (Group 1) + SOTA Comparison (Group 2)** • **Group 1**: Implement linear probing and full fine-tuning on AnimeFace Character dataset • **Group 2**: Implement and evaluate SOTA methods (ArcFace, DINO, SimCLR) • Both groups: Conduct ablation studies and optimize hyperparameters
Week 8–9	**Comprehensive Evaluation and Analysis** • Evaluate all methods on benchmark protocols (closed-set, open-set, few-shot) • Compare our MAE fine-tuning against all baselines • Conduct detailed error analysis and failure case investigation • Generate visualizations and interpretation of learned representations
Week 9–10	**Final Report and Deliverables** • Write comprehensive final report with all results • Create demo application for anime character recognition • Prepare presentation materials • Release code, models, and benchmark protocols

Table 3: Revised project timeline with clear group responsibilities

7 Expected Contributions

7.1 Academic Contributions

1. **First systematic MAE fine-tuning study for anime**: Comprehensive investigation of how to adapt masked autoencoder models to stylized anime domain
2. **Anime face recognition benchmark**: Standardized datasets, protocols, and baselines for future research
3. **Transfer learning analysis**: Understanding of natural image to anime domain adaptation characteristics
4. **Performance improvements**: Demonstrate superior anime character recognition compared to existing methods

7.2 Practical Impact

1. **Improved anime applications**: Better performance for real-world anime character identification systems
2. **Accessible implementation**: Consumer hardware experiments (M4, RTX 5090) enable wider adoption
3. **Open-source release**: Code, pre-trained models, and benchmarks publicly available
4. **Practitioner guidelines**: Actionable recommendations for applying MAE to anime visual tasks
5. **Benchmark platform**: Standardized evaluation enabling future research comparisons

8 Conclusion

Our mid-term progress has successfully established a strong foundation for anime-domain self-supervised learning research. Through systematic experiments, we have:

1. **Validated MAE's applicability**: Demonstrated that masked autoencoders can effectively learn anime visual representations when using official pre-trained models
2. **Discovered mask ratios differences**: Found that 25% masking significantly outperforms standard 75% for anime imagery, but 75% has better generalizable learning representations.
3. **Assessed reconstruction quality**: Identified both successes (200 epochs) and challenges (30 epochs) in anime reconstruction
4. **Established collaborative workflow**: Successfully coordinated two-group team structure with complementary focuses

Revised Direction: Based on our findings and practical considerations, we have revised our proposal to focus on:

- **Fine-tuning official pre-trained MAE models** rather than large-scale training from scratch
- **Anime face recognition benchmark** establishment with comprehensive SOTA comparisons

- **Practical anime character identification** demonstrating real-world applicability

This revised direction maintains strong scientific contributions while ensuring feasibility within our timeline and computational resources. By leveraging pre-trained models and focusing on practical applications, we deliver both academic insights and real-world impact.

Next Steps: We will proceed with dataset preparation, implement fine-tuning frameworks, establish benchmark protocols, and conduct comprehensive evaluations comparing our approach against existing methods.

Code Repository: All implementations, experiments, and findings will be made available at: https://github.com/Tingting-Olivia-Du/mae

References

[1] He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked Autoencoders Are Scalable Vision Learners. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009.

[2] Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4690–4699.

[3] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9650–9660.

[4] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 1597–1607.

[5] Jin, Y., & Takiguchi, T. (2018). Anime Character Identification Using Multi-Task Learning. *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2268–2272.