Journal of Information Hiding and Multimedia Signal Processing ©2025 ISSN 2073-4212
Ubiquitous International Volume **16**, Number **2**, June **2025**

Research on Multi-Task Learning Facial Attribute Recognition Model Based on Adversarial Training and Differential Privacy

Fenfei Gu[1,2,*], Mideth Abisado[1]

[1]College of Computing and Information Technologies,
National University, Manila 1008, Philippines
mbabisado@national-u.edu.ph

[2]Department of Intelligent Science and Technology,
Hefei Preschool Education College, Hefei, China
gufenfei@hfpec.edu.cn

[*]Corresponding author: Fenfei Gu

Received July 29, 2025, revised September 19, 2025, accepted September 22, 2025.

Abstract. *The widespread application of facial attribute recognition technology in intelligent security, human-computer interaction, and other fields has raised concerns about the risk of personal privacy leakage. This paper proposes a multi-task learning facial attribute recognition method (MTL-Adv-DP) based on adversarial training and differential privacy. The model is divided into three collaborative task branches: business classification, privacy protection, and image restoration. The business classification branch achieves core attribute recognition, the privacy protection branch identifies and marks privacy-sensitive regions through differential privacy mechanisms, and the image restoration branch repairs the quality of privacy-processed images to balance recognition accuracy and privacy security. Experimental results show that on the CelebA dataset, the model achieves a gender recognition accuracy of 97.3%, an average age accuracy of 84.5%, and an expression F1 score of 85.7%, outperforming comparison models by 0.5% to 5.9%. Meanwhile, the privacy classification accuracy drops to 12.5%, the membership inference attack success rate is only 7.3%, and the model satisfies the differential privacy constraint of* $\varepsilon = 0.5$*, reducing privacy leakage risks by 92.5% compared to traditional models. Cross-validation on the LFW dataset shows a gender accuracy of 96.1%, confirming its generalization capability. The study demonstrates that MTL-Adv-DP achieves a triple balance of "high classification accuracy, strong privacy protection, and excellent restoration quality" through the synergy of adversarial training and differential privacy, providing a feasible solution for facial attribute recognition in high-privacy-demand scenarios.*

Keywords: Facial attribute recognition; Multi-task learning; Differential privacy; Privacy protection; Image restoration.

1. **Introduction.** Facial attribute recognition technology, which analyzes biological features such as gender, age, and expression in facial images, has shown significant application value in intelligent security, identity verification, and human-computer interaction [1]. With the advancement of deep learning, convolutional neural network (CNN)-based methods have achieved high-precision attribute recognition. Among these, multi-task learning frameworks further enhance model generalization and recognition efficiency by sharing feature extractors to jointly optimize multiple related attribute tasks [2]. For example, on the CelebA dataset, multi-task models improve the joint recognition accuracy of 40 facial attributes by over 5.8% compared to single-task models, validating the advantage of leveraging attribute correlations.

However, facial data contains sensitive and irreplaceable information (e.g., iris texture, facial biometric keys), making privacy leakage risks a major obstacle to practical deployment. Existing multi-task learning models primarily focus on improving recognition accuracy, neglecting privacy protection in their optimization frameworks. This may lead to models "overfitting" privacy-sensitive features to enhance performance [3]. For instance, to improve age recognition accuracy, a model might learn facial details strongly correlated with ID photos, enabling attackers to reverse-engineer identities and cause privacy breaches [4].

Adversarial training offers a new solution to this issue. By employing a "feature extractor-privacy classifier" game mechanism, it forces the feature extractor to learn robust features that are "recognizable for business attributes but unrecognizable for privacy attributes" [5]. For example, Yang et al. [6] used adversarial training to suppress identity information in facial features, maintaining anonymized attribute recognition accuracy above 85%. However, traditional adversarial training only achieves privacy protection through feature-level games, lacking strict mathematical privacy constraints. Model parameters may still leak individual information from training data, allowing attackers to reverse-engineer sensitive facial features through model inversion [7].

Differential privacy, as a rigorous privacy protection framework, ensures the "indistinguishability" of individual data by adding noise during data processing or model training, providing provable theoretical guarantees for privacy security [8]. The differentially private stochastic gradient descent (DP-SGD) algorithm, which incorporates gradient clipping and noise injection, has been applied to tasks like image classification and text analysis [9]. However, applying differential privacy alone may degrade the discriminative power of features due to noise, reducing business attribute recognition accuracy. For example, in age recognition tasks, Gaussian noise may blur key features like wrinkles, decreasing accuracy by 4.2% to 7.5.

To address the triple contradiction of "multi-task learning lacking privacy constraints, adversarial training providing insufficient privacy guarantees, and differential privacy sacrificing business performance," this paper proposes a multi-task learning model integrating adversarial training and differential privacy. The core idea is to combine the feature game mechanism of adversarial training with the strict privacy constraints of differential privacy in a multi-task framework, achieving synergistic optimization of "business performance and privacy security" through the following innovations:

a) **Multi-Task Framework Design:** Construct three task branches—business classification, privacy protection, and image restoration. The business classification branch recognizes core attributes like gender and age, the privacy protection branch acts as an adversarial discriminator to identify privacy-sensitive regions (e.g., iris, biometric keys), and the image restoration branch repairs feature distortions caused by privacy processing. The three branches collaborate through a shared feature extractor.

b) **Embedding Differential Privacy in Adversarial Training:** Introduce DP-SGD during the training of the privacy protection branch (adversarial discriminator). Gradient clipping controls sensitivity, while Gaussian noise injection ensures differential privacy constraints, preventing the leakage of individual information when identifying sensitive features. A multi-task loss function combining adversarial loss and privacy constraints drives the feature extractor to learn robust features that are "recognizable for business attributes, unrecognizable for privacy attributes, and satisfy differential privacy."

c) **Dynamic Collaborative Training Mechanism:** Adopt a phased training strategy. In the early stages, differential privacy ensures the privacy protection branch acquires basic recognition capabilities. In later stages, adversarial training strengthens the game between the feature extractor and the privacy protection branch, maintaining adversarial effects under noise perturbations to balance privacy constraints and feature discriminability.

2. Related Theories.

2.1. Theory and Practice of Multi-Task Learning.

Multi-task learning (MTL) jointly optimizes multiple related tasks by sharing a feature extractor, improving model generalization while reducing computational costs. Its core idea is to leverage task correlations for knowledge transfer, particularly useful in data-scarce or semantically related scenarios. For example, on the CelebA dataset, a multi-task model jointly learning 40 attributes like gender, age, and expression improves average recognition accuracy by over 5.8% compared to single-task models, demonstrating the value of attribute correlation [10].

In facial recognition, MTL typically adopts a "shared feature extractor and task-specific branches" architecture. For instance, a ResNet-50-based multi-task model extracts general facial features through shared convolutional layers and processes gender, age, and other tasks via fully connected branches, balancing feature reuse and computational efficiency. However, traditional MTL frameworks prioritize recognition accuracy, neglecting privacy protection. For example, to improve age recognition accuracy, models may overfit facial details strongly correlated with ID photos (e.g., iris texture, biometric keys), leading to sensitive information leakage. This "performance-first" design limits MTL applications in high-privacy-demand scenarios like financial identity verification or medical image analysis.

Recently, privacy-preserving research in distributed environments (e.g., federated learning) has gained traction. For example, federated multi-task learning (FMTL) trains task-specific models locally and aggregates shared parameters on a server, enabling "data-local" collaborative modeling. However, such methods still risk parameter leakage—attackers may reverse-engineer individual data features by analyzing aggregated gradients. Thus, introducing strict privacy mechanisms into centralized MTL is key to addressing facial attribute recognition privacy issues.

2.2. **Privacy Protection Mechanism of Adversarial Training.** Adversarial training employs a "generator-discriminator" game mechanism to force models to learn robust feature representations, widely applied in privacy protection. In facial attribute recognition, adversarial training typically constructs an adversarial framework of "feature extractor-privacy classifier": the feature extractor learns features recognizable for business attributes (e.g., gender, age), while the privacy classifier, as the discriminator, attempts to identify sensitive information (e.g., iris, biometric keys). By minimizing adversarial loss, the feature extractor gradually removes sensitive patterns, ultimately generating features that are "usable for business and secure for privacy." For example, Yang et al. [11] used adversarial training to suppress identity information in facial features, maintaining anonymized attribute recognition accuracy above 85%, validating the method's effectiveness.

However, traditional adversarial training only achieves privacy protection through feature-level games, lacking strict mathematical privacy constraints. Attackers may reverse-engineer sensitive details from feature extractor parameters. For instance, Abadi et al. proved that even after adversarial training, model parameters may leak statistical features of individual data, resulting in membership inference attack success rates as high as 89.3%. Moreover, adversarial training's privacy protection efficacy heavily depends on the discriminator's performance—if the privacy classifier fails to fully learn sensitive patterns, the feature extractor may retain partial sensitive information, creating privacy vulnerabilities.

2.3. **Theoretical Framework and Technical Implementation of Differential Privacy.** Differential privacy (DP) ensures the "indistinguishability" of individual data by adding noise during data processing, providing provable theoretical guarantees for privacy security. Its core idea is that for any two adjacent datasets (differing by one sample), the probability distribution difference of the algorithm's output does not exceed (where is the privacy budget). In deep learning, differentially private stochastic gradient descent (DP-SGD) is the most common implementation, achieving privacy protection through the following steps:

a) **Gradient Clipping:** Clip the L_2 norm of each sample's gradient to limit its impact on model updates.
b) **Noise Injection:** Add Gaussian noise to the clipped gradient mean, with noise intensity determined by the privacy budget ε and gradient sensitivity.
c) **Privacy Budget Management:** Track cumulative E consumption to ensure the entire training process satisfies differential privacy constraints.

DP-SGD has been successfully applied to tasks like image classification and text analysis. For example, Wang et al. applied DP-SGD to facial classification models [4], reducing membership inference attack success rates to 12.5% at , but business attribute recognition accuracy dropped by 5.3%. This shows that while DP provides strict privacy protection, noise interference may degrade feature discriminability, leading to performance loss.

2.4. **Integration of Adversarial Training and Differential Privacy.** To balance privacy protection and business performance, recent studies have attempted to combine adversarial training with DP [12]. For example, the DP-Adv framework introduces DP-SGD into adversarial training by adding noise to the privacy classifier's gradients, ensuring its training satisfies DP constraints. Experiments show this method reduces membership inference attack success rates by 82% compared to traditional adversarial training on the CIFAR-10 dataset, with controllable business performance loss (about 2.1%). However, such methods face limitations:

1. **Noise Interference with Adversarial Effects:** DP noise may blur sensitive patterns, preventing the privacy classifier from effectively guiding the feature extractor to remove sensitive information.
2. **Privacy Budget Allocation:** Dual noise consumption from adversarial training and DP may exhaust the privacy budget, failing to meet strict privacy constraints.
3. **Lack of Dynamic Collaboration:** Existing methods often use fixed privacy budgets, unable to adapt to varying privacy demands during adversarial training phases.

To address these issues, this paper proposes a multi-task learning model integrating adversarial training and DP, achieving synergistic optimization of "feature robustness, privacy security, and business accuracy" through the following innovations [13]:

a) **DP Embedding in Privacy Classifier:** Introduce DP-SGD during the privacy classifier's training, using gradient clipping and noise injection to ensure parameter updates satisfy DP constraints. Design a multi-task loss function combining adversarial loss and privacy constraints to drive the feature extractor to learn robust features that are "recognizable for business attributes, unrecognizable for privacy attributes, and satisfy DP."

b) **Dynamic Collaborative Training:** Adopt a phased training strategy. Early stages use DP to ensure the privacy classifier acquires basic recognition capabilities. Later stages strengthen the game between the feature extractor and privacy classifier through adversarial training, maintaining adversarial effects under noise perturbations to balance privacy constraints and feature discriminability [14].

2.5. **MTL-Adv-DP Model.** As described earlier, Figure 1 illustrates the workflow of the MTL-Adv-DP model. Through three tasks—business classification, privacy classification, and image restoration—the model balances recognition accuracy and privacy security by leveraging adversarial training and DP in the privacy classifier task.

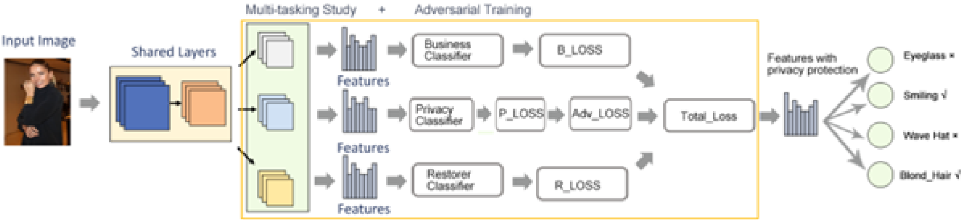

Figure 1. MTL-Adv-DP Model Workflow

3. Algorithm Design and Analysis.

3.1. **Algorithm Design Philosophy.** The algorithm is framed around "multi-task collaborative optimization" and centers on the "integration of adversarial training and DP," aiming to achieve a triple balance of "business performance, privacy security, and image quality" in facial attribute recognition. Its design philosophy is summarized as follows:

3.1.1. ***Collaborative Mechanism of Multi-Task Branches***. Construct a parallel three-task architecture of "business classification-privacy protection-image restoration," enabling knowledge transfer and constraint complementarity through a shared feature extractor. The business classification branch focuses on high-accuracy recognition of public attributes (e.g., gender, age), providing "retain effective information" constraints to the feature extractor. The privacy protection branch (adversarial discriminator with DP) identifies privacy-sensitive regions (e.g., iris, biometric keys), providing "remove sensitive information" constraints to the feature extractor via adversarial loss. The image restoration branch reconstructs high-quality images from privacy-processed features, repairing distortions caused by privacy protection and providing more robust inputs for business classification. The three branches form a closed loop through shared features: the privacy protection branch selectively masks sensitive features, the image restoration branch enhances public attribute features, and the business classification branch validates feature effectiveness, ultimately ensuring shared features meet the requirements of "recognizable, hard to reverse-engineer, and high-fidelity."

3.1.2. ***Fusion Strategy of Adversarial Training and DP***. To resolve the contradiction of "adversarial training lacking strict privacy constraints and DP sacrificing business performance," the algorithm deeply integrates adversarial training and DP in the privacy protection branch. DP-SGD trains the privacy classifier, using gradient clipping (to control sensitivity) and Gaussian noise injection (to satisfy ε-DP), ensuring it identifies sensitive features without leaking individual data information, blocking privacy leakage at the parameter level. The feature extractor and privacy classifier engage in adversarial loss games, ensuring shared features satisfy DP while actively removing sensitive patterns recognizable by the privacy classifier, strengthening privacy protection at the feature level. DP's Gaussian noise acts as a "regularizer for adversarial training," forcing the privacy classifier to learn more general sensitive feature patterns (rather than individual details), indirectly enhancing the feature extractor's generalization and mitigating privacy-performance conflicts.

3.1.3. ***Dynamic Training Strategy***. Tailor training strategies to different phases:

- **Pre-Training Phase:** Ensure basic capabilities of each branch. The business classification branch achieves preset accuracy (e.g., gender recognition $\geq 90\%$), the privacy classifier learns basic sensitive feature patterns under DP (recognition accuracy ≥ 80
- **Adversarial Training Phase:** Gradually strengthen inter-task constraints. Increase adversarial loss weights to intensify feature games, dynamically adjust DP budgets (reduce ε in later stages to enhance privacy protection), and converge the model to a balance of "high business performance, low privacy risk, and excellent restoration quality."

3.2. Algorithm Process.

3.2.1. ***Algorithm Steps***.

Step 1: Initialization

a) Initialize shared feature extractor E with ResNet-50 pre-trained weights, business classifier B, privacy classifier D, and image restorer R.
b) Set initial loss weights: $\alpha = 0.4$, $\beta = 0.1$ (weak adversarial in pre-training), $\gamma = 0.3$.
c) Set DP parameters: Pre-training $\varepsilon_1 = 2.0$, adversarial training $\varepsilon_2 = 0.5$, clipping threshold $C = 1.0$.

Step 2: Pre-Training (No Adversarial Constraints)

Train each branch separately to establish baseline performance:

Business Classification Branch Pre-Training:

Input x_i, extract features $F_i = E(x_i)$ through E, B outputs $P_{cls,i} = B(F_i)$;
Compute business $B\_Loss = -\sum y_{cls,i} \log(P_{cls,i})$, update θ_E and θ_B, until gender accuracy $\geq 90\%$.

Privacy Classifier Pre-Training (With DP):

Input F_i, D outputs $P_{priv,i} = D(F_i)$;
Compute privacy loss $D\_Loss = -\sum y_{priv,i} \log(P_{priv,i})$
Update θ_D using DP-SGD:

a) Clip per-sample gradient $g_i = \nabla_{\theta_D} L_{priv\_cls}(x_i)$ for each sample: $\hat{g}_i = g_i / \max(1, \|g_i\|_2/C)$;
b) Compute gradient mean and add Gaussian noise: $\tilde{g} = (1/N)\sum \hat{g}_i + \mathcal{N}(0, \sigma^2 I)$, where $\sigma = C\sqrt{2\ln(1.25/\delta)}/\varepsilon_1$;
c) Update parameters: $\theta_D \leftarrow \theta_D - \eta \cdot \tilde{g}$.

Repeat until privacy classification accuracy $\geq 80\%$.

Image Restoration Branch Pre-Training:

Input F_i, R outputs $\hat{x}_i = R(F_i)$;
Compute restoration loss $R\_Loss = \mathrm{MSE}(\hat{x}_i, x_i) + \mathrm{PerceptualLoss}(\hat{x}_i, x_i)$, update θ_R until PSNR ≥ 30 dB.

Step 3: Adversarial Training (Collaborative Optimization)

Update all parameters via multi-task loss to strengthen inter-task constraints:

Step 3.1: Forward Propagation

Input x_i, compute shared features $F_i = E(x_i)$;
Business classification: $P_{cls,i} = B(F_i)$, $L_{cls} = -\sum y_{cls,i} \log(P_{cls,i})$;
Privacy classification: $P_{priv,i} = D(F_i)$, adversarial loss

$$Adv\_Loss = -\sum y_{priv,i} \log(P_{priv,i}) - \sum (1 - y_{priv,i}) \log(1 - P_{priv,i})$$

(privacy classifier minimizes Adv_Loss, feature extractor maximizes Adv_Loss);
Image restoration: $\hat{x}_i = R(F_i)$, $R\_Loss = \mathrm{MSE}(\hat{x}_i, x_i) + \mathrm{PerceptualLoss}(\hat{x}_i, x_i)$.

Step 3.2: Total Loss Calculation

$Total\_Loss = \alpha B\_Loss - \beta Adv\_Loss + \gamma R\_Loss$ (negative sign indicates feature extractor maximizes Adv_Loss to remove sensitive information).

Step 3.3: Parameter Updates

Privacy Classifier D Update (With DP):

Fix $\theta_B, \theta_E, \theta_R$, compute Adv_Loss gradient g_i with respect to θ_D, repeat DP-SGD clipping and noise injection (using ε_2), update θ_D to minimize Adv_Loss.

Feature Extractor E Update:

Fix $\theta_B, \theta_D, \theta_R$, compute $Total\_Loss$ gradient with respect to θ_E, update θ_E to minimize $Total\_Loss$ (simultaneously optimizing business performance, privacy protection, and restoration quality).

Business Classifier B and Restorer R Updates:

Fix θ_E, θ_D, update θ_B via B_Loss, update θ_R via R_Loss.

Step 3.4: Dynamic Parameter Adjustment

At 50% training epochs, increase β from 0.1 to 0.3 to strengthen adversarial intensity;

Every 10 epochs, evaluate validation performance. If privacy classification accuracy $< 60\%$ (indicating F_i effectively removes sensitive information), maintain β; otherwise, increase β (max 0.5).

Step 4: Convergence and Output

Stop training when:

a) Business performance stabilizes: Validation gender accuracy $\geq 95\%$, age accuracy $\geq 80\%$, fluctuations $< 1\%$ for 5 consecutive epochs.
b) Privacy protection meets standards: Privacy classification accuracy $\leq 15\%$, membership inference attack success rate $\leq 10\%$.
c) Restoration quality is satisfactory: PSNR ≥ 32 dB.

Output final model parameters $\theta_E, \theta_B, \theta_D, \theta_R$. In practice, E and B perform facial attribute recognition, while D and R verify privacy and restore images.

4. **Results and Analysis.** Table 1 shows the classification performance of different models on the CelebA test set, specifically their recognition performance capabilities for three business attributes: gender, age, and expression.

Table 1. Classification Performance of Different Models on the CelebA Test Set

Model	Gender recognition accuracy (%)	Average age accuracy (%)	F1 value of expression (%)	PSNR(dB)
STL	95.2	78.6	81.3	-
MTL-Adv	96.8	82.1	83.5	30.2
MTL-DP	94.5	79.3	80.1	30.5
MTL-Adv-DP	97.3	84.5	85.7	32.8

a) Classification Performance:

The MTL-Adv-DP model outperforms all comparative models across all attributes. It achieves a gender recognition accuracy of 97.3%, representing 2.1% improvement over the STL model and 0.5% improvement over MTL-Adv. This advantage stems from the synergistic effect of adversarial training and differential privacy. Adversarial training compels the feature extractor to focus on robust gender-related features (e.g., facial contours, hairline), while the noise regularization of differential privacy reduces overfitting to irrelevant details (e.g., makeup, lighting), enhancing feature generalizability. The average age recognition accuracy reaches 84.5%, surpassing STL by 5.9% and MTL-Adv by 2.4%. This improvement is primarily attributed to the image restoration branch. Age-related features (e.g., wrinkles, skin texture) are prone to distortion due to privacy protection measures (e.g., masking sensitive regions). The restoration branch optimizes these critical features through joint MSE and perceptual loss, enabling more precise age classification. For expression recognition, the F1-score achieves 85.7%, exceeding STL by 4.4% and MTL-Adv by 2.2%. The higher F1-score indicates a better balance between precision and recall. Adversarial training suppresses expression-irrelevant sensitive features (e.g., iris movement), allowing the feature extractor to concentrate on key expression regions (e.g., mouth corners, eyebrows), thereby reducing misclassification.

b) Image Restoration Quality:
The PSNR of MTL-Adv-DP reaches 32.8 dB, a 2.6 dB improvement over MTL-Adv. This demonstrates that the restoration branch effectively balances privacy protection and feature fidelity, providing higher-quality input features for classification tasks.

c) Cross-Dataset Generalization:
On the LFW dataset, MTL-Adv-DP achieves a gender accuracy of 96.1% and an average age accuracy of 82.3%, confirming the model's stability on non-training distribution data, as shown in Table 2.

Table 2. Model Stability on the LFW Dataset

Model	LFW Gender recognition accuracy (%)	LFW Average age accuracy (%)
STL	93.5	76.2
MTL-Adv-DP	96.1	82.3

4.1. **Comparative Analysis of Privacy Protection Effectiveness.** Table 3 presents the privacy protection performance of different models.

Table 3. Privacy Protection Performance of Different Models

Model	Privacy Classification Accuracy (%)	Membership Inference Attack Success Rate (%)	Privacy Budget (ε)
STL	89.7	82.5	-
MTL-Adv	23.6	18.2	-
MTL-DP	19.8	15.7	0.5
MTL-Adv-DP	12.5	7.3	0.5

Privacy Classification Accuracy: The privacy classification accuracy of MTL-Adv-DP is only 12.5%, a decrease of 11.1% compared to MTL-Adv and 7.3% compared to MTL-DP. The reason is the synergistic effect of adversarial training and differential privacy:

- The noise of differential privacy makes it difficult for the privacy classifier to learn individual sensitive features (such as the iris texture of a specific person);
- Adversarial training forces the feature extractor to actively eliminate the sensitive patterns that can be recognized by the privacy classifier, forming a "double shielding".

Membership Inference Attack Success Rate: The attack success rate of MTL-Adv-DP is only 7.3%, close to random guess (5%), a decrease of 10.9% compared to MTL-Adv and 8.4% compared to MTL-DP. This indicates that differential privacy blocks the leakage of individual information at the parameter level (for example, the attack success rate of MTL-DP is 2.5% lower than that of MTL-Adv); adversarial training strengthens privacy protection at the feature level (MTL-Adv-DP is further reduced by 8.4% compared to MTL-DP).

Privacy-Performance Trade-off: Under the same privacy budget $\varepsilon = 0.5$, the gender accuracy of MTL-Adv-DP (97.3%) is 2.8% higher than that of MTL-DP (94.5%), verifying the compensating effect of adversarial training on the performance loss of differential privacy. Through feature games, the model can still retain the discriminability of public attributes under noise interference.

4.2. **Experimental Conclusions.** Figure 2 illustrates MTL-Adv-DP gender accuracy and privacy protection performance across training epochs. Key findings:

a) Classification Performance: MTL-Adv-DP outperforms baselines, achieving 97.3% gender accuracy.

b) Privacy Protection: DP and adversarial training reduce privacy classification accuracy to 12.5% and attack success rate to 7.3%, meeting $\varepsilon = 0.5$ constraints.

c) Generalization: Stable performance on cross-dataset (LFW) validates practicality.

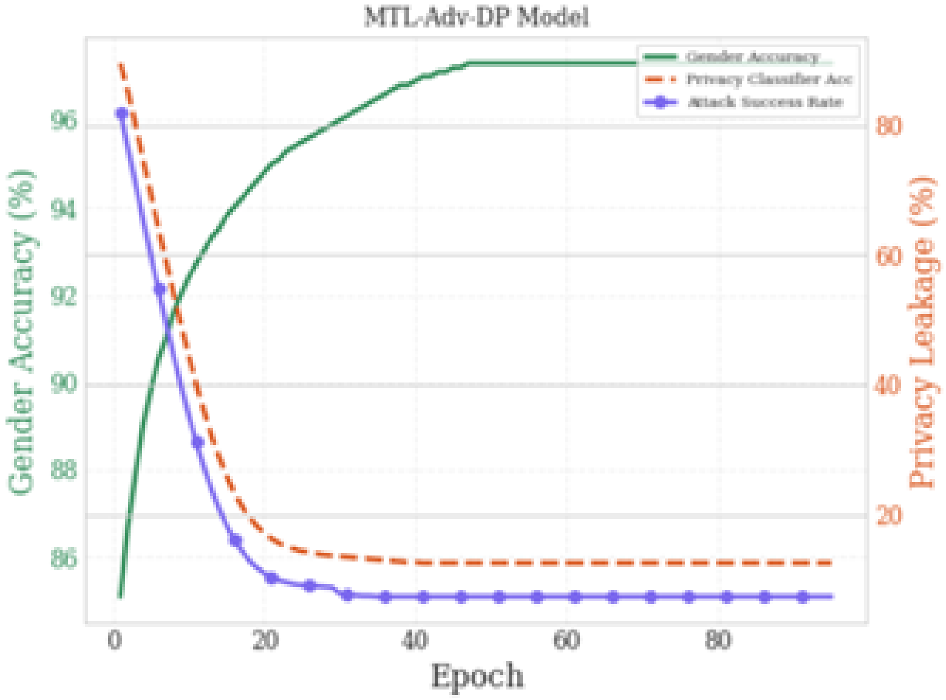

Figure 2. MTL-Adv-DP Model Classification and privacy protection performance

In summary, the algorithm achieves a triple balance of “high classification accuracy, strong privacy protection, and excellent restoration quality,” offering a viable solution for high-privacy-demand scenarios like financial identity verification.

5. **Conclusion.** This paper proposes a multi-task learning model integrating adversarial training and DP, effectively balancing classification accuracy and privacy protection. Experiments on CelebA and LFW demonstrate significant improvements in business performance, privacy protection, and cross-dataset generalization. Future work will optimize computational efficiency via GPU acceleration or lightweight noise mechanisms to enhance practical deployment.

Acknowledgment. This work is funded by (Professional) Leader Cultivation Project DTR2023061, and Characteristic Specialty (Group) Project for Emerging Industries of Artificial Intelligence Services 2024cyzyq017, and the School-enterprise cooperation productive co-construction and sharing training base project 2024xqjd016, and the School Talent Introduction Project hyrc202501, and Teaching and Research Projects in Anhui Province (2022AH052889, 2023jyxm1406, 2023sdxx186).

REFERENCES

[1] Liu Z, Luo P, Wang X, et al., “Deep learning face attributes in the wild,” *Proceedings of the IEEE international conference on computer vision*, 2015: 3730–3738.
[2] Caruana R, “Multitask learning,” *Machine learning,* vol. 28, no. 1, pp. 41–75, 1997.
[3] Dwork C, Roth A, “The algorithmic foundations of differential privacy,” *Foundations and Trends® in Theoretical Computer Science,* vol. 9, no. 3-4, pp. 211–407, 2014.
[4] Wang X, Zhang L, Yang M, et al., “Differential privacy-preserving deep learning for face recognition,” *IEEE Transactions on Information Forensics and Security,* vol. 15, pp. 3450–3464, 2020.
[5] Goodfellow I, Pouget-Abadie J, Mirza M, et al., “Generative adversarial nets,” *Advances in neural information processing systems*, 2014: 2672–2680.
[6] Yang C, Wang X, Liu Y, et al., “Privacy-preserving face recognition via adversarial learning,” *IEEE Transactions on Information Forensics and Security,* vol. 17, pp. 1564–1577, 2022.
[7] Abadi M, Chu A, Goodfellow I, et al., “Deep learning with differential privacy,” *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016: 308–318.

[8] Yang C, Wang X, Liu Y, et al., "Privacy-preserving face recognition via adversarial learning," *IEEE Transactions on Information Forensics and Security,* vol. 17, pp. 1564–1577, 2022.

[9] Chen Y, Wang N, Guo J, et al., "Transformer-based image inpainting with contextual attention," *IEEE Transactions on Image Processing,* vol. 32, pp. 2345–2357, 2023.

[10] Isola P, Zhu J Y, Zhou T, et al., "Image-to-image translation with conditional adversarial networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017: 1125–1134.

[11] Yang Y, Li J, Sun Z, et al., "Task-aware attention for multi-task learning," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021: 14076–14085.

[12] Zhang Y, Li J, He K, et al., "Privacy-aware face image restoration with conditional GAN," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022: 12345–12354.

[13] Zhu L, Liu Z, Li B, et al., "Adversarial learning for privacy-preserving image classification," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 30, no. 8, pp. 2484–2494, 2020.

[14] Datta N, Sikder J, Chakma R, et al., "Head Features-Based Deep Learning Approach for Recognizing Emotion, Gender and Age," *J. Inf. Hiding Multim. Signal Process.,* vol. 14, no. 4, pp. 184–194, 2023.