NutriDiff: Phenotype-Consistent Diffusion Augmentation for Malnutrition Screening

Hanqiu Yang[1,2], Chenming Zhang[3], Zhijie Shen[1,2], Xue Wang[3], Hongtao Bai[1,2], and Lili He[1,2]✉

[1] College of Computer Science and Technology, Jilin University, Changchun, China
[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China
[3] The Third Bethune Hospital of Jilin University, Changchun, China
✉ Corresponding author: helili@jlu.edu.cn

Abstract. Deep learning for facial-based elderly malnutrition screening shows preliminary feasibility, but clinical data scarcity limits diagnostic performance. Although diffusion models are widely used for medical image augmentation, their application to facial malnutrition screening is limited by insufficient control over malnutrition-related phenotypes. To address these limitations, we propose NutriDiff, which fine-tunes Stable Diffusion using view-class LoRA adapters and filters malnutrition-irrelevant synthetic samples through facial morphological feature extraction. In few-shot testing on 92 elderly subjects, NutriDiff achieves AUC 0.984 and MCC 0.854, outperforming baselines and offering a potential approach for elderly malnutrition screening.

Keywords: Diffusion Model· Facial Phenotype· Malnutrition Screening

1 Introduction

Malnutrition is a systemic nutritional and metabolic imbalance caused by multiple factors and is associated with prolonged hospital stays and increased mortality. It is particularly prevalent among elderly inpatients, with an incidence of 18.2%[16]. Early nutritional intervention can shorten hospital stays and reduce mortality[2]. A central pathological feature of malnutrition is the loss of muscle and adipose tissue, which manifests as surface depressions and related morphological changes[6].

Malnutrition is associated with characteristic facial morphological changes. Facial skeletal muscle thickness correlates positively with the appendicular skeletal muscle mass index (ASMI)[32, 7], suggesting value in nutritional assessment[8, 10]. Periorbital fat pads are routinely examined in nutrition-related physical assessments. Intermuscular adipose structures, such as buccal fat pads, are closely related to energy metabolism and nutritional status[15, 1], and show marked volume reduction with malnutrition[20]. Consequently, within the current diagnostic framework, phenotypic assessment in the Global Leadership Initiative on Malnutrition (GLIM) criteria[3] frequently relies on quantitative indicators

such as body mass index (BMI) and ASMI measured by specialized instruments, limiting routine follow-up feasibility in elderly populations. In contrast, the Subjective Global Assessment (SGA)[5] allows bedside evaluation based on visible morphological changes. However, results depend heavily on the evaluator's experience, reducing assessment consistency and reproducibility.

Deep learning models applied to facial images are being explored as tools for malnutrition screening. Given the association between facial phenotypes and nutritional status, Tay et al. demonstrated the feasibility of using facial morphology for nutritional assessment in elderly populations[26, 27]. Khan et al. employed multi-pose whole-body images to screen for malnutrition in children[13]. Wang et al. investigated 3D facial point clouds[29], although this approach depends on costly acquisition equipment. Wang et al. extracted periorbital features from 2D frontal images using large-scale annotated datasets[28], but the reliance on single-view information limits representational completeness. Given that many elderly individuals are bedridden or have limited mobility, these methods remain difficult to deploy as practical bedside screening tools due to equipment requirements and limited data availability.

Generative models have been widely adopted to address the scarcity of medical data[33]. Diffusion models can synthesize images with high visual fidelity. However, malnutrition screening relies on subtle morphological changes in specific regions, including the temporal, periorbital, malar, and jawline. Conditional generation methods based only on coarse class labels or text prompts often lack precise control over these local features[12, 25]. As a result, synthetic images may exhibit inconsistencies between phenotypic characteristics and assigned clinical labels[24]. Such mismatches introduce label noise and reduce the generalization performance of downstream discriminative models[30]. Consequently, constraining clinical phenotypic consistency in synthetic images and effectively filtering out those with feature deviation remain critical challenges in building reliable screening models[31].

To address these challenges, we propose NutriDiff, a multi-view diffusion augmentation framework constrained by clinical phenotype consistency, as illustrated in Fig. 1. First, we introduce low-rank adaptation (LoRA) [9] into the text encoder and the attention layers of the denoising network in Stable Diffusion (SD) [23], and generate candidate multi-view images conditioned on the view and the nutritional state. Next, we localize facial keypoints in the synthesized images and extract phenotypic features from the temporal, periorbital, malar, and jawline. Using the feature distribution of real images from the same class as a reference, we remove images with phenotypic deviations measured by the Mahalanobis distance[19]. Finally, the downstream screening model is trained jointly on real and synthetic data.

The contributions of this work are as follows. 1) We introduce NutriDiff, a multi-view diffusion augmentation framework with clinical phenotype consistency constraints, to synthesize fine-grained facial phenotypes of malnutrition in elderly populations. 2) We further develop a malnutrition screening pipeline based on NutriDiff and show that anatomically guided feature filtering improves

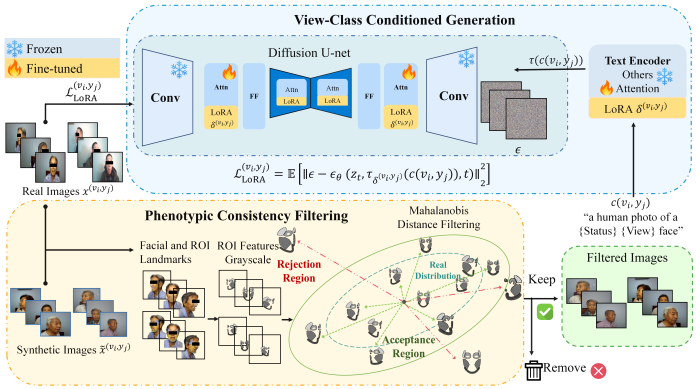

Fig. 1. Overview of the NutriDiff framework.

the generalization of downstream classifiers. 3) Experiments on clinical datasets show that, compared with an existing screening method [28] and a few-shot augmentation strategy [14], NutriDiff produces more realistic phenotypes and improves downstream screening performance.

2 Method

2.1 The NutriDiff Framework

View-Class Conditioned Generation. Given the limited sample size of malnutrition datasets in elderly populations, fully fine-tuning all parameters of large pretrained diffusion models is computationally demanding and susceptible to overfitting. To capture view-dependent phenotypic differences accurately under a parameter-efficient setting, inspired by DataDream[14], we propose a fine-grained view-class conditional Generation adaptation strategy.

Specifically, we adopt Stable Diffusion[23] V2.1 as the backbone and introduce LoRA [9] for parameter-efficient adaptation. We combine the view label $v_i \in \{F, L, R\}$, corresponding to frontal, left 45°, and right 45° facial images, with the nutritional status label $y_j \in \{N, M\}$, denoting normal nutrition and malnutrition, to construct independent training subsets $\mathcal{D}_{v_i,y_j}$.

For each view-class pair (v_i, y_j), we insert an independent LoRA adapter $\delta^{(v_i,y_j)}$ into the projection matrices (W_q, W_k, W_v, W_o) of all attention layers in both the text encoder and the U-Net. The backbone parameters θ are kept frozen, and only the LoRA parameters are optimized. For an attention layer with linear transformation $h = Wx$, we parameterize the weight as $W = W_0 + \Delta W$ with a low-rank update $\Delta W = BA$, yielding

$$h = (W_0 + \Delta W)x = W_0 x + BAx, \tag{1}$$

where W_0 denotes the frozen pre-trained weight, x and h denote the input and output features, and $\Delta W = BA$ corresponds to the trainable increment δ. We initialize B to zero so that $\Delta W = 0$ at training start.

During training, for a sample $x \in \mathcal{D}_{v_i,y_j}$, we encode it into the latent space using the pretrained VAE encoder E_ϕ and obtain $z_0 = E_\phi(x)$. We then apply the forward diffusion process, $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where t is a randomly sampled time step and $\epsilon \sim \mathcal{N}(0, I)$ denotes Gaussian noise.

With the parameters frozen, we optimize the adapter weights $\delta^{(v_i,y_j)}$ by minimizing the conditional latent diffusion loss

$$\mathcal{L}_{\text{LoRA}}^{(v_i,y_j)} = \mathbb{E}_{x\in\mathcal{D}_{v_i,y_j},t,\epsilon}\left[\left\|\epsilon - \epsilon_{\theta,\delta^{(v_i,y_j)}}\big(z_t, \tau_{\delta^{(v_i,y_j)}}(c(v_i,y_j)), t\big)\right\|_2^2\right] \tag{2}$$

where $\epsilon_{\theta,\delta^{(v_i,y_j)}}$ denotes the U-Net denoiser with the LoRA adapters inserted, $c(v_i, y_j)$ is the template prompt constructed from the view and nutritional label, and $\tau(\cdot)$ represents the pretrained text encoder. $\tau_{\delta^{(v_i,y_j)}}$ denotes the same encoder with the corresponding LoRA adapter inserted.

During inference, we load the LoRA adapter corresponding to a specific (v_i, y_j) pair and combine it with a textual prompt, for example, "a human photo of a malnourished left three-quarter face", to guide sampling. The model generates a high-fidelity synthetic dataset that matches the target clinical distribution, denoted as $\widetilde{\mathcal{D}}_{v_i,y_j} = \{\tilde{x}\}$.

Phenotypic Consistency Filtering. Given that diffusion models may generate samples that do not fully reflect expected clinical phenotypes, NutriDiff applies a clinical phenotype consistency filtering strategy. We define four facial regions of interest (ROI) associated with nutritional status, denoted as $r_k \in \{\textit{temporal}, \textit{periorbital}, \textit{malar}, \textit{jawline}\}$, as shown in Fig. 2(left). For a facial image x, we apply MediaPipe Face Mesh [18] to extract facial landmarks $\Omega(x)$, and construct a binary mask $M_{r_k}(x)$ for each ROI r_k. The *temporal*, *periorbital*, and *malar* regions are defined by convex hulls of their corresponding keypoints, whereas the *jawline* region is specified by a closed polygon along the mandibular contour.

To mitigate illumination and exposure differences across subjects, we introduce a full-face skin mask $M_{\text{skin}}(x)$ as a normalization reference. This mask is generated from facial contour keypoints with non-skin areas, such as the eyes and lips, excluded. The image x is converted to a grayscale image $g(x)$. Let $\mathcal{P}$ denote the set of image pixels. We define the masked mean $\mu(h; M) = \frac{\sum_{\mathcal{P}} M(p)\, h(p)}{\sum_{\mathcal{P}} M(p)}$, and the masked variance $\sigma^2(h; M) = \mu\big((h - \mu(h; M))^2; M\big)$, where $M \in \{0, 1\}$. Samples are discarded if keypoint detection fails or if $\sum_\Omega M = 0$.

Based on these operators, we construct a four-dimensional phenotype descriptor $\mathbf{s}(x) = [s_1(x), \ldots, s_4(x)]^\top$. For the temporal and periorbital ($k \in \{1, 2\}$), soft tissue loss leads to relative darkening. We therefore define

$$s_k(x) = \mu(g(x); M_{r_k}(x)) \big/ \mu(g(x); M_{\text{skin}}(x)). \tag{3}$$

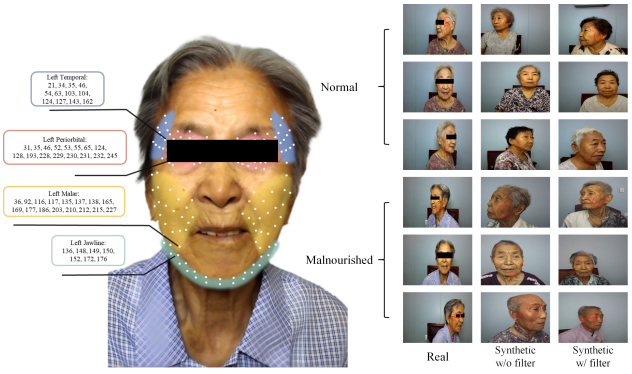

Fig. 2. Left: Facial ROIs defined by landmarks, with corresponding landmark indices. Four ROIs are delineated: temporal, periorbital, malar, and jawline. Right: Multi-view facial images from the Normal (top) and Malnourished (bottom) classes. Each column shows real images and synthetic images with and without filtering. Circled regions indicate inconsistent features.

For the malar ($k = 3$), subcutaneous fat loss alters surface texture. We quantify this change by the variance of the Laplacian response within the ROI,

$$s_3(x) = \sigma^2(\Delta g(x); M_{r_3}(x)). \tag{4}$$

For the jawline ($k = 4$), fat loss sharpens the mandibular contour, measured by the mean gradient magnitude within the region,

$$s_4(x) = \mu(\|\nabla g(x)\|_2; M_{r_4}(x)). \tag{5}$$

To quantify the overall deviation of a synthetic image in the ROIs phenotype space, we adopt the Mahalanobis distance[19]. For each (v_i, y_j), we estimate the mean vector $\boldsymbol{\mu}^{(v_i,y_j)}$ and covariance matrix $\boldsymbol{\Sigma}^{(v_i,y_j)}$ from the real sample set $\mathcal{D}_{v_i,y_j}$. For a candidate $\tilde{x} \in \widetilde{\mathcal{D}}_{v_i,y_j}$, the deviation is

$$d_M^2(\tilde{x}; v_i, y_j) = \big(\mathbf{s}(\tilde{x}) - \boldsymbol{\mu}^{(v_i,y_j)}\big)^\top \big(\boldsymbol{\Sigma}^{(v_i,y_j)}\big)^{-1} \big(\mathbf{s}(\tilde{x}) - \boldsymbol{\mu}^{(v_i,y_j)}\big). \tag{6}$$

An acceptance threshold $\mathcal{T}^{(v_i,y_j)}$ is derived from the real samples' Mahalanobis distance distribution. A candidate is retained only if $d_M^2(\tilde{x}; v_i, y_j) \leq \mathcal{T}^{(v_i,y_j)}$, yielding the filtered synthetic set $\widehat{\mathcal{D}}_{v_i,y_j}$.

2.2 Downstream Malnutrition Screening

Based on the real dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ and the synthetic dataset $\widehat{\mathcal{D}} = \{(\tilde{x}_j, y_j)\}_{j=1}^m$ generated and filtered by NutriDiff, we train a binary malnutrition screening model f_{cls}. We adopt CLIP ViT-B/16[22] as the backbone and apply

LoRA to both the image encoder and the text encoder. The model is trained with a weighted cross-entropy objective:

$$\mathcal{L}_{\text{cls}} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\Big[\mathcal{L}_{\text{CE}}\big(f_{\text{cls}}(x), y\big)\Big] + \gamma\, \mathbb{E}_{(\tilde{x},y)\sim\widehat{\mathcal{D}}}\Big[\mathcal{L}_{\text{CE}}\big(f_{\text{cls}}(\tilde{x}), y\big)\Big], \tag{7}$$

where $\gamma \geq 0$ controls the contribution of synthetic samples during training.

During inference, we aggregate predictions from multiple views of the same subject and compute a subject-level malnutrition risk score by averaging the predicted probabilities across the three views.

3 Experiments

3.1 Datasets and Implementation Details

Data Collection. As no publicly available dataset targets malnutrition in individuals aged 65 years or older, we construct a private cohort to evaluate the proposed model. We recruit patients visiting the Department of Clinical Nutrition between January and August 2025. Inclusion criteria are age ≥ 65 years and the ability to complete body composition analysis. Exclusion criteria include facial structural changes due to surgery, facial abnormalities related to disease or medication, and other conditions deemed unsuitable by the investigators. The study was approved by the Ethics Committee of The Third Bethune Hospital of Jilin University (No. 20251219004), and all participants provided written informed consent.

Body composition is measured using the InBody S10, from which BMI and ASMI are derived. Facial images are captured with the Gemini Pro device. Participants remove facial occlusions before acquisition, and three views per subject are obtained: one frontal and two lateral at 45°. Data annotation is performed independently by three attending nutrition physicians using the SGA. When evaluations agree, the label is assigned directly; otherwise, consensus is reached through discussion with reference to BMI and ASMI under the GLIM framework. Due to limited samples at each severity level, all malnutrition grades are merged into one class, resulting in a binary classification of normal and malnourished.

In total, 92 subjects are included, yielding 276 facial images. Among them, 45 are diagnosed with malnutrition, and 47 are nutritionally normal. Eighteen malnourished patients, able to cooperate with image acquisition, and 20 normal subjects constitute the training set. Twenty-seven malnourished bedridden patients or those receiving enteral nutrition, together with 27 normal subjects, form the test set to evaluate performance under complex clinical conditions. **Implementation Details.** All experiments run on a single RTX5090 GPU with CUDA 12.8. The implementation uses PyTorch, and the diffusion module is based on SD v2.1. For each view-class pair, images are generated at 512×512 resolution with batch size 8. The LoRA rank is 8. We train with AdamW [17] at learning rate 1×10^{-4} for 240 epochs. Sampling uses 50 diffusion steps with classifier-free guidance scale 3.5. The number of generated candidate images equals $\beta = 8$ times the number of real images.

In the filtering stage, MediaPipe Face Mesh extracts 468 facial landmarks. The covariance matrix is estimated separately for each view-class combination, and a shrinkage estimator improves robustness under small-sample conditions. The acceptance threshold is the 95th percentile of the Mahalanobis distance distribution computed from real images. Candidate images satisfying the threshold are ranked in ascending order of d_M^2, and the top-K images are retained, where $K = \lceil \beta \cdot n_{\text{real}} \rceil$. We first generate a larger candidate pool and then apply phenotype filtering. After filtering, 912 synthetic images are retained.

The downstream screening model uses CLIP ViT-B/16 as the backbone with an input resolution of 224×224. Training uses AdamW with an initial learning rate of 1×10^{-5} and weight decay of 1×10^{-4}. The batch size is 32, and the model is trained for 40 epochs. A cosine learning rate schedule with 4 warmup epochs is applied. When real and synthetic samples are jointly used, the contribution weight of synthetic data is $\gamma = 0.3$. The dataset is split at the subject level to prevent identity overlap between training and testing. During inference, prediction probabilities across views are averaged to obtain subject-level results.

3.2 Comparison with Existing Works

Comparison with Prior Work. To assess downstream classification performance, we compare NutriDiff with prior methods. As shown in Table 1, NutriDiff outperforms existing approaches across all metrics. The current state-of-the-art (SOTA) study [28] relies on single-view features and conventional machine learning models. Such approaches have limited representation capacity under the inherent class imbalance of medical datasets, yielding an MCC of 0.193. Incorporating a few-shot CLIP baseline markedly improves classification performance, and NutriDiff further builds on it, increasing the MCC from 0.759 to 0.854. These results indicate that multi-view facial images provide complementary, informative representations and support targeted data augmentation to mitigate data scarcity and severe class imbalance in malnutrition screening.

Analysis of Diffusion-Based Baselines. We evaluate synthesis quality and downstream diagnostic performance. NutriDiff achieves the best FID and AFFID. Although DataDream reports a much lower FID than SDXL (121.2 vs. 204.3), their downstream classification metrics remain comparable, suggesting that simply enlarging the synthetic data scale offers limited benefit for few-shot medical image classification. By introducing clinical phenotype consistency filtering, NutriDiff preserves fine-grained pathological characteristics in generated samples and improves downstream screening performance, raising the F1 score and MCC to 0.929 and 0.854, respectively.

3.3 Ablation Study

Contribution of Synthetic Data Augmentation. To assess the impact of synthetic data across architectures, we perform subject-level 5-fold cross-validation on the training set. Table 2 shows that generative augmentation improves feature representations consistently across classifier backbones. However,

Table 1. Comparison with prior work and diffusion-based augmentation baselines. FID and AFFID [11] (Fréchet distance in ArcFace [4] space) evaluate synthesis quality. Downstream performance is reported as accuracy (Acc), area under the ROC curve (AUC), F1-score (F1), and Matthews correlation coefficient (MCC).

Method	Training Data	Synthesis Quality		Downstream Performance			
		FID ↓	AFFID ↓	Acc ↑	AUC ↑	F1 ↑	MCC ↑
Wang et al. [28]	Real	–	–	0.593	0.706	0.656	0.193
Baseline (CLIP)	Real	–	–	0.877	0.936	0.884	0.759
SD2.1 [23]	Real + Synth	272.3	1.201	0.889	0.981	0.897	0.787
SDXL [21]	Real + Synth	204.3	1.173	0.907	0.981	0.912	0.820
DataDream [14]	Real + Synth	121.2	0.820	0.907	0.978	0.912	0.820
NutriDiff (Ours)	**Real + Synth**	**120.0**	**0.782**	**0.926**	**0.984**	**0.929**	**0.854**

Table 2. Ablation study on backbone. Results are reported as mean ± std (%) over subject-level 5-fold cross-validation on the training set. "w/ filter" and "w/o filter" denote training with and without the phenotypic consistency filtering step, respectively.

Training Data	Acc (%) ↑	AUC (%) ↑	F1 (%) ↑	MCC (%) ↑
Backbone: CLIP ViT-B/16				
Real	80.12±13.01	89.17±16.56	77.87±15.35	64.23±24.96
Real + Synth (w/o filter)	89.64±16.39	93.75±13.98	88.57±18.63	80.16±32.25
Real + Synth (w/ filter)	**95.00±11.18**	**96.25±8.39**	**96.00±8.94**	**91.55±18.90**
Backbone: ResNet-50				
Real	74.64±18.92	84.17±13.31	65.78±36.97	54.13±32.55
Real + Synth (w/o filter)	77.50±8.34	74.58±11.45	73.78±10.23	62.96±11.32
Real + Synth (w/ filter)	**84.17±7.38**	**87.87±9.15**	**83.42±5.80**	**69.76±16.03**

appending unfiltered synthetic data can reduce stability, reflected by increased cross-fold variance (e.g., the standard deviation of MCC for CLIP increases from 24.96% to 32.25%) and a drop in ResNet AUC (from 84.17% to 74.58%).

Efficacy of Phenotypic Consistency Filtering. We evaluate the proposed filtering mechanism by comparing training with unfiltered and filtered synthetic data under two backbone architectures. Without filtering, AUC decreases for ResNet, suggesting that phenotypically mismatched synthetic samples degrade performance. After applying filtering, all four metrics improve for both backbones, and cross-fold standard deviations are lower than in the unfiltered setting. These results suggest that removing synthetic samples whose ROI descriptors deviate from the real training distribution yields more stable downstream classification. Representative synthetic images before and after filtering are shown in Fig. 2 (right), where circles indicate regions of phenotypic inconsistency in unfiltered samples.

4 Conclusion

In this paper, we present NutriDiff, a multi-view diffusion-based augmentation framework with clinical phenotype consistency constraints to synthesize fine-grained facial phenotypes of malnutrition in elderly populations. Experiments show that anatomically guided filtering reduces generative label noise and improves generalization and stability of the downstream malnutrition screening model. Although the 2D image-based design enables practical, low-cost bedside deployment, it remains sensitive to extreme poses where 2D facial landmark detection is unreliable. In addition, the limited size of our single-center cohort restricts the model to binary classification. Future work will incorporate 3D priors to better handle complex bedside poses and expand to large multi-center cohorts for multi-class severity grading to develop a more robust clinical screening tool.

Acknowledgments. This work was supported by the National College Students Innovation and Entrepreneurship Training Program (No. 202510183235) and the Undergraduate Research Training Program at Jilin University.

Disclosure of Interests. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Akşamoğlu, M., Muluk, N.B., Şahan, M.H.: Changes of the buccal fat pad volume according to the different age groups, gender, and body mass index: An evaluation with computed tomography. Journal of Computer Assisted Tomography **49**(1), 156–164 (2025)
2. Bellanti, F., Lo Buglio, A., Quiete, S., Vendemiale, G.: Malnutrition in hospitalized old patients: screening and diagnosis, clinical outcomes, and management. Nutrients **14**(4), 910 (2022)
3. Cederholm, T., Jensen, G.L., Correia, M.I.T., et al.: The glim consensus approach to diagnosis of malnutrition: a 5-year update. Clinical Nutrition **49**, 11–20 (2025)
4. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
5. Detsky, A.S., McLaughlin, J.R., Baker, J.P., et al.: What is subjective global assessment of nutritional status? Journal of parenteral and enteral nutrition **11**(1), 8–13 (1987)
6. Fischer, M., JeVenn, A., Hipskind, P.: Evaluation of muscle and fat loss as diagnostic criteria for malnutrition. Nutrition in Clinical Practice **30**(2), 239–248 (2015)
7. González-Fernández, M., Perez-Nogueras, J., Serrano-Oliver, A., et al.: Masseter muscle thickness measured by ultrasound as a possible link with sarcopenia, malnutrition and dependence in nursing homes. Diagnostics **11**(9), 1587 (2021)
8. Hasegawa, Y., Yoshida, M., Sato, A., et al.: Temporal muscle thickness as a new indicator of nutritional status in older individuals. Geriatrics & Gerontology International **19**(2), 135–140 (2019)

9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
10. Hwang, Y., Lee, Y.H., Cho, D.H., et al.: Applicability of the masseter muscle as a nutritional biomarker. Medicine **99**(6), e19069 (2020)
11. Kabra, K., Balakrishnan, G.: F?d: On understanding the role of deep feature spaces on face generation evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 8327–8332 (2024)
12. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D.: Diffusion models in medical imaging: A comprehensive survey. Medical image analysis **88**, 102846 (2023)
13. Khan, M., Singh, R., Vatsa, M., Singh, K.: Domainadapt: Leveraging multitask learning and domain insights for children's nutritional status assessment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 606–616. Springer (2024)
14. Kim, J.M., Bader, J., Alaniz, S., Schmid, C., Akata, Z.: Datadream: Few-shot guided dataset generation. In: European Conference on Computer Vision. pp. 252–268. Springer (2024)
15. Kruglikov, I., Trujillo, O., Quick, K., et al.: The facial adipose tissue: a revision. Facial Plastic Surgery **32**(06), 671–682 (2016)
16. Liu, Y., Song, C., Wang, X., et al.: Prevalence of malnutrition among adult inpatients in china: a nationwide cross-sectional study. Science China Life Sciences **68**(5), 1487–1497 (2025)
17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
18. Lugaresi, C., Tang, J., Nash, H., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)
19. Mahalanobis, P.C.: On the generalized distance in statistics. Sankhyā: The Indian Journal of Statistics, Series A (2008-) **80**, S1–S7 (2018)
20. Miehle, K., Stumvoll, M., Fasshauer, M., Hierl, T.: Facial soft tissue volume decreases during metreleptin treatment in patients with partial and generalized lipodystrophy. Endocrine **58**(2), 262–266 (2017)
21. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
23. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
24. Seo, S., Lee, I.K., Kim, H.W., Min, J., Jung, C.H.: Diffusion-based user-guided data augmentation for coronary stenosis detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 149–159. Springer (2025)
25. Susladkar, O., Deshmukh, G., Tur, Y., Durak, G., Bagci, U.: Victr: Vital consistency transfer for pathology aware image synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22772–22782 (2025)
26. Tay, W., Quek, R., Kaur, B., Lim, J., Henry, C.J.: Use of facial morphology to determine nutritional status in older adults: opportunities and challenges. JMIR Public Health and Surveillance **8**(7), e33478 (2022)

27. Tay, W.L.W., Quek, R.Y.C., Lim, J., Kaur, B., Ponnalagu, S., Toh, D.W.K., Leow, M.K.S., Henry, C.J.: Real-time and digital remote nutritional assessment framework with the use of smartphone-enabled facial morphometrics and machine learning—a proof of concept. European Journal of Clinical Nutrition **80**(1), 104–112 (2026)
28. Wang, J., He, C., Long, Z.: Establishing a machine learning model for predicting nutritional risk through facial feature recognition. Frontiers in Nutrition **10**, 1219193 (2023)
29. Wang, X., Liu, Y., Rong, Z., Wang, W., Han, M., Chen, M., Fu, J., Chong, Y., Long, X., Tang, Y., et al.: Development and evaluation of a deep learning framework for the diagnosis of malnutrition using a 3d facial points cloud: A cross-sectional study. Journal of Parenteral and Enteral Nutrition **48**(5), 554–561 (2024)
30. Wei, Q., Feng, L., Sun, H., Wang, R., Guo, C., Yin, Y.: Fine-grained classification with noisy labels. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11651–11660 (2023)
31. Xie, J., Zhang, Z., Weng, Z., Zhu, Y., Luo, G.: Meddiff-ft: Data-efficient diffusion model fine-tuning with structural guidance for controllable medical image synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 306–316. Springer (2025)
32. Yilmaz, E., Arsava, E.M., Topcuoglu, M.A.: Temporalis muscle thickness: Ultrasound measurement, clinical significance and correlation with sarcopenic indices. Clinical Nutrition **54**, 140–152 (2025)
33. Zhang, L., Jindal, B., Alaa, A., et al.: Generative ai enables medical image segmentation in ultra low-data regimes. Nature Communications **16**(1), 6486 (2025)