Emotionally Aware High-Resolution Image Generation Using ViT-based cVAEs

(April 2025)

Lana Almousa Hadeel Alshibani Daniyah AlTuwaijri Noura Alotaibi Shahad Alobaid

Computer Science Department, AI Program

Princess Norah bint Abdulrahman University, Kingdom of Saudi Arabia

***Abstract*— *This paper presents a novel multimodal framework for generating high-resolution images that are emotionally conditioned based on textual cues. The approach integrates a Vision Transformer (ViT) decoder within a Conditional Variational Autoencoder (cVAE) architecture, utilizing a CNN-based encoder and a high-resolution CNN head to enhance image quality. The model is trained using the Emotion Recognition Dataset, a publicly available collection of over 38,000 grayscale facial images labeled with one of seven core emotions: Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise. Each image is 48×48 pixels and captured under controlled conditions, making the dataset ideal for supervised learning in facial emotion detection. Evaluation using Fréchet Inception Distance (FID), Inception Score (IS), and emotion classification accuracy demonstrates that the proposed model outperforms traditional cVAE approaches in both visual quality and emotional consistency.*

INTRODUCTION

The generation of emotionally expressive facial images presents unique challenges at the intersection of generative modeling and affective computing. While conditional Variational Autoencoders (cVAEs) provide a principled framework for controlled image generation through their structured latent space, they often struggle to produce outputs with the fine details necessary for convincing emotional expressions. This limitation stems from their inherent trade-off between reconstruction quality and latent space regularization, typically resulting in overly smoothed facial features that lack emotional nuance.

Vision Transformers (ViTs) have emerged as a promising alternative, demonstrating superior capability in capturing global facial context through self-attention mechanisms. However, their patch-based processing and lack of inherent spatial biases make it difficult to maintain the precise local relationships that define subtle emotional cues. When applied independently, both architectures face fundamental limitations in balancing emotional fidelity with image quality.

This work investigates a synergistic approach that combines the strengths of cVAEs and ViTs for emotion-conditioned generation. By integrating cVAEs' structured latent representations with ViTs' powerful global attention mechanisms, we aim to develop a framework that better preserves both the macroscopic facial structure and microscopic emotional details. The hybrid architecture seeks to address key challenges in affective computing applications, where the ability to generate authentic, nuanced emotional expressions is crucial - from virtual avatars in mental health therapy to emotionally responsive human-computer interfaces.

Our research focuses on systematically evaluating this combined approach, examining how the complementary strengths of these architectures can be leveraged while mitigating their individual weaknesses. The findings contribute to ongoing efforts in developing more sophisticated emotion-aware generation systems, with implications for fields ranging from psychological research to interactive media design

I. Literature Review

A. Limitations of Vision Transformers (ViTs)

Vision Transformers (ViTs) have demonstrated strong performance in image classification, but they often struggle with local texture representation due to the lack of inherent

inductive biases, such as translation equivariance and locality. This shortcoming can reduce the model's ability to generate fine-grained textures, especially in emotionally detailed visual tasks like art synthesis [5]. Another key limitation of ViTs is their difficulty maintaining spatial coherence in structured image generation tasks. Without convolutional priors, ViTs can produce spatially distorted outputs, particularly when data is limited or lacks structure [6]. ViTs require large amounts of labeled data to train effectively and are highly data-hungry due to the absence of strong priors. This can be problematic in emotionally annotated datasets, which are often small or imbalanced.

B. Limitations of Variational Autoencoders (VAEs)

Variational Autoencoders (VAEs) are prone to producing blurry and low-resolution images due to the bottleneck in the latent space [2]. This compression often loses high-frequency visual information critical to emotion perception.

VAEs also face difficulty learning disentangled representations of emotion-related features, leading to poor emotional expressiveness in outputs [2]. VAE outputs can exhibit limited diversity due to collapsed variance in the latent space, especially in emotional or artistic domains.

II. Related Work

A. Conditional Variational Autoencoders (cVAEs)

VAE struggles with generating samples with specific properties [8]. In order to generate images with a restriction, the Autoencoder needs to learn how to decode variables when its given a hint, that's how cVAEs work. A one-hot encoded digit label conditions the decoder, guiding its learning of the decoding process [8].

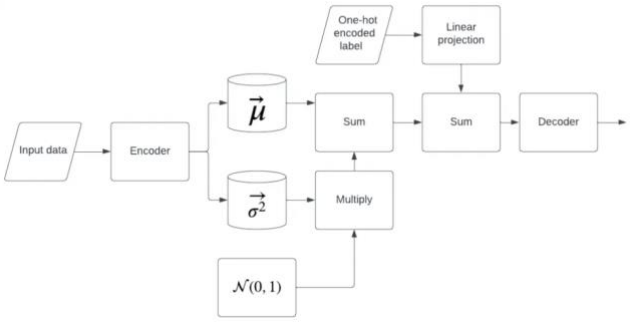

Fig. 1. The structure of cVAEs [8].

cVAEs have been widely used for image generation tasks due to their ability to model complex data distributions. However, they often suffer from generating blurry images and lack the capacity to capture fine-grained details, primarily due to limitations in their latent space representations [2].

In addition, designate one author as the "corresponding author." This is the author to whom proofs of the paper will be sent. Proofs are sent to the corresponding author only.

B. Vision Transformers (ViTs)

Vision Transformer (ViT) is a deep learning model that extends the transformer architecture, initially successful in NLP, to computer vision. Instead of using convolutional layers like CNNs, ViT splits the images into fixed-size patches, which are flattened and linearly projected into patch embeddings. Positional encodings are added to retain spatial information, and the sequence (along with a special [CLS] token) is fed into a transformer encoder. The encoder uses multi-head self-attention to model relationships between patches and a feed-forward network to refine features, with residual connections and layer normalization for stability. The [CLS] token, which aggregates global image information, is finally passed through an MLP head for classification [9].

This method adapts transformer architectures—originally designed for NLP—to vision tasks by processing image patches as tokenized inputs. By doing so, it effectively captures long-range dependencies, often outperforming traditional CNNs in tasks like image classification, object detection, and segmentation.

ViTs have revolutionized image processing by modeling images as sequences of patches, enabling the capture of long-range dependencies. Despite their success, ViTs may struggle with local texture representation, which is vital for detailed image synthesis [9].

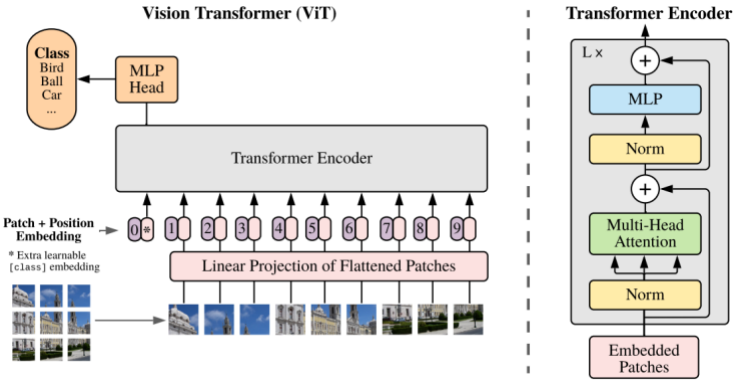

Fig. 2. The architecture of ViT [7].

C. Emotional Image Content Generation with Text-to-Image Diffusion Models

EmoGen is a framework that enables emotional control in text-to-image diffusion models by incorporating an emotion condition vector and an emotion classifier to guide the generation process [10]. It conditions the model to produce images that not only match the input text but also convey a desired emotional tone. A new benchmark dataset, EmoSet, is proposed to evaluate emotional alignment in generated images. Results show that EmoGen significantly improves emotional expressiveness while maintaining semantic consistency with the textual input [10].

D. Global Context Vision Transformers

The Global Context Vision Transformer (GC ViT) presents a hierarchical vision transformer architecture that efficiently models both local and long-range spatial dependencies through a novel global self-attention mechanism [11]. Unlike standard vision transformers that suffer from quadratic complexity or local-window approaches like Swin Transformer that limit cross-window interactions, GC ViT introduces shared global query tokens that interact with local key-value pairs, enabling efficient global context integration without expensive operations.

The architecture employs modified Fused-MBConv blocks for downsampling, injecting CNN-like inductive biases while maintaining transformer benefits. This design achieves state-of-the-art performance across image classification (85.7% Top 1 on ImageNet-1K), object detection (58.3% AP on COCO), and semantic segmentation (49.2 mIoU on ADE20K), outperforming ConvNeXt, Swin Transformer, and MaxViT counterparts [11]. By unifying global context modeling with local attention in a parameter-efficient manner, GC ViT advances the Pareto frontier for vision transformers, particularly in high-resolution scenarios where long-range dependencies are critical.

III. Methodology

A. Dataset: Emotion Recognition Dataset

The Emotion Recognition Dataset utilized in this project comprises a publicly accessible collection of over 38,000 facial images, each meticulously labeled with one of seven fundamental emotions: Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise. Each image consists of a 48x48 pixel grayscale representation of a face, captured under controlled environmental conditions. This controlled acquisition process makes the dataset particularly well-suited for training machine learning models specifically designed for facial emotion detection. The dataset exhibits a balanced distribution across the various emotion classes, ensuring that no single emotion unduly dominates the training process. This balanced nature provides a robust foundation for supervised learning tasks, allowing models to learn to accurately distinguish between the different emotional expressions. The Emotion Recognition Dataset is widely adopted in both research and educational projects, serving as a valuable resource for building and evaluating deep learning models aimed at recognizing emotions from facial expressions. Its widespread use underscores its importance in advancing the field of affective computing and enabling the development of more emotionally intelligent artificial intelligence systems. *The full dataset (Appendix 1) confirms these findings.*

"Happy"

"Angry"

Fig. 3. Two samples from "The Emotion Recognition" dataset

B. Dataset Preparation and Preprocessing

To enable emotion-based image generation, we developed a specialized image dataset organized into labeled directories that correspond to seven primary emotions. Each subdirectory housed images that exemplified a particular emotional category. A custom PyTorch Dataset class was created to streamline data loading and labeling, ensuring compatibility with standard image formats such as .jpg, .png, and .bmp. This class utilized LabelEncoder from scikit-learn to transform emotion labels into integer class representations, thus standardizing the input for supervised learning.

During the initialization phase, the dataset loader systematically navigated the root directory, excluding non-image files and invalid entries. Any defective or unreadable images were automatically bypassed during iteration, thereby enhancing the robustness of the process. The dataset was organized to maintain class balance whenever feasible and included images taken under diverse lighting and compositional conditions.

To ensure visual consistency and meet the model's input requirements, all images were converted to RGB format and standardized through a series of PyTorch transforms, which included resizing, normalization, and conversion to tensors. This preprocessing pipeline guaranteed that the model received clean, uniformly processed inputs that retained emotion-relevant facial features, which is essential for generating semantically meaningful outputs in emotion-based synthesis tasks.

C. Model Architecture

The proposed model integrates a conditional Variational Autoencoder (cVAE) framework with a Vision Transformer (ViT)-based decoder, combining probabilistic latent-space modeling with the expressive power of transformer architectures for high-fidelity generation, outlined in Table 1:

Table 1

Component	Description
Encoder (CNN-based)	A CNN that extracts hierarchical features from the input image associated with an emotion, compressing it into a lower-dimensional latent representation.
Latent representation (z)	A compressed, low-dimensional vector sampled from the distribution N(μ, σ^2). It captures the essential features while discarding redundant details.
Decoder (ViT-based)	A ViT that reconstructs the image from the latent representation. It processes images using self-attention mechanisms, capturing long-range dependencies.
High-resolution Head (CNN-based)	A module that refines the output of the ViT decoder using CNN layers to enhance spatial details and produce a high-resolution image.

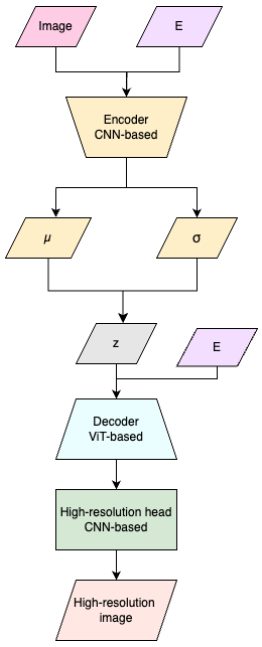

Fig. 3. cVAE-ViT model archeticture diagram.

The model begins by processing an input imagewith its emotion through a CNN-based encoder, which compresses it into a compact latent space representation, capturing essential features while reducing spatial dimensions. This latent representation then serves as input to a Vision Transformer (ViT)-based decoder, which reconstructs the image by leveraging self-attention to model long-range dependencies and global structure. The decoded features are further refined by a CNN-based high-resolution head, which upsamples and enhances fine details to produce the final high-resolution output image.

By combining CNN-based encoding (for efficient local feature extraction), a latent bottleneck (for compact representation), and ViT-based decoding (for global coherence), this hybrid architecture effectively bridges convolutional and transformer-based approaches for high-quality image generation or reconstruction.

D. Training Procedure

The model undergoes a training process designed to minimize a carefully constructed composite loss function. This loss function is composed of three components, each designed to optimize a specific aspect of the model's performance:

- **Reconstruction Loss**: Measures the difference between the generated image and the ground truth (e.g. MSE loss)
- **Structural Similarity Index Measure (SSIM)**: Evaluates the perceptual similarity between the generated and real images.
- **Perceptual Loss**: Assesses high-level feature differences using a pre-trained neural network (e.g. VGG)

Total Loss = recon_loss + 0.2 * perc_loss + 0.2 * ssim_loss

The full implementation code is provided in Appendix 2.

IV. Experiments and Results

A. Evaluation Metrics

The performance of the model is subjected to a thorough evaluation process, employing a carefully selected set of evaluation metrics. These metrics are chosen to provide a comprehensive and nuanced understanding of the model's strengths and weaknesses across a range of performance characteristics:

- **Fréchet Inception Distance (FID)**: Assesses the quality of generated images. The FID obtained was 232.70, reflecting a noticeable divergence between generated images and the real image distribution.
- **Inception Score (IS)**: Evaluates the diversity and quality of generated images. The model achieved an

IS of 1.47 ± 0.09, reflecting moderate image quality and diversity under the given emotional constraints.

B. Model Output Examples

To visually demonstrate the tangible impact of incorporating a resolution head into our model's architecture, we present a series of illustrative image examples. These examples serve to highlight the qualitative differences in performance achieved with and without the inclusion of the high-resolution head.

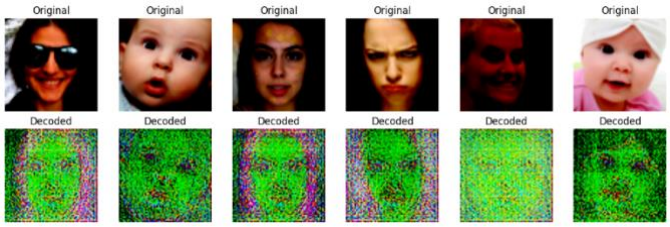

Fig. 4 Model output without high-resolution head

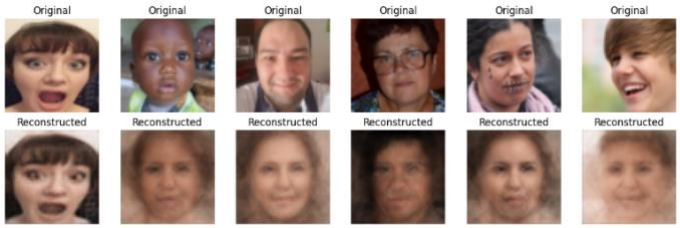

Fig.5 Model output with high-resolution head

C. Comparative Analysis

The marked contrast in reconstruction quality, as illustrated in Fig. 4 and Fig. 5, underscores the critical role of incorporating a resolution head within the model architecture. In the absence of a dedicated resolution head, the model's learned representation frequently proves insufficient to effectively capture the intricate, fine-grained details inherent in the original image. This limitation subsequently manifests as a reconstruction process that yields blurred or distorted outputs, lacking the sharpness and precision of the source material. Conversely, the strategic addition of a resolution head fosters the development of a more robust and comprehensive representation. This enhanced representation empowers the model to reconstruct images with substantially improved clarity and fidelity, enabling it to more accurately reproduce the nuances and subtleties of the original input.

While the inclusion of a resolution head significantly mitigates the issue of blurred or distorted reconstructions, it's important to acknowledge that the reconstructed images may still exhibit some residual imperfections, highlighting the ongoing challenges in achieving perfect image reconstruction.

V. DISCUSSION

A. Addressing Limitations of cVAEs and ViTs

Combining Vision Transformers (ViTs) with conditional Variational Autoencoders (cVAEs) offers benefits for global feature modeling but encounters challenges in generating high-resolution images. ViTs, which use patch-based attention mechanisms, are effective at capturing long-range dependencies within an image. However, this approach can lead to a loss of fine-grained spatial details during the latent compression stage. The cVAE's bottleneck further restricts the amount of information that can be effectively passed through the network, which makes it difficult to reconstruct high-frequency textures when upsampling the image. The decoder architectures typically used in cVAEs also contribute to blurring effects, as standard upscaling methods often struggle to synthesize sharp and coherent details from compressed representations.

While recent hybrid approaches have attempted to address these limitations by incorporating multi-scale feature fusion techniques, fundamental trade-offs remain between achieving high reconstruction fidelity and maintaining generative flexibility, especially when dealing with complex, high-resolution outputs.

B. *Ethical Considerations*

Employing the Emotion Recognition Dataset for training cVAE-ViT models necessitates careful consideration of several key ethical concerns. Privacy stands as a paramount consideration, given that facial data inherently involves sensitive biometric information. Consequently, it is imperative to verify the presence of proper consent mechanisms and robust anonymization techniques to safeguard individual privacy. Bias mitigation is also of critical importance, as the presence of imbalanced data within the dataset can lead to unfair or discriminatory model performance across different demographic groups. Furthermore, the potential for misuse of emotion recognition technology, such as its application in unauthorized surveillance systems, warrants the establishment of strict ethical guidelines to prevent harmful applications. Transparency with users regarding data usage practices and the provision of clear opt-out options are essential components of upholding trust and ensuring accountability in the development and deployment of these models.

Addressing these multifaceted ethical issues proactively ensures the responsible and ethical development of AI systems for emotion analysis, promoting fairness, privacy, and user well-being.

C. Future Work

While our hybrid cVAE-ViT model advances emotion-conditioned image generation, several promising directions

remain. First, we aim to enhance emotion fidelity by incorporating dynamic intensity control (e.g. scaling "joy" from subtle smiles to exuberant laughter) and cross-cultural emotion embeddings to address dataset biases. Second, architectural improvements like replacing Gaussian latent space with VQ-VAE or diffusion-based representations could better capture multimodal emotion-semantic relationships. Third, we will explore efficient ViT-CNN fusion via cross-attention to reduce computational overhead while preserving detail. Additionally, we propose developing a progressive training strategy where the model first learns low-resolution emotional semantics before gradually incorporating finer details, potentially improving both training stability and final output quality. Also, Extending the framework to video generation with temporal emotion consistency (e.g. mood transitions) and integrating user feedback loops for personalized emotional styling present exciting avenues. Finally, rigorous fairness evaluations will be conducted to mitigate potential biases in emotion portrayal across demographics.

These steps would further bridge the affective gap in generative AI while enabling applications in therapy, entertainment, and human-computer interaction.

VI. Conclusion

This paper presents a new method for emotion-conditioned image generation, which combines a conditional variational autoencoder (cVAE) with a Vision Transformer (ViT)-based decoder. By using the Emotion Recognition Dataset, the hybrid cVAE-ViT model effectively merges the hierarchical feature extraction capabilities of CNNs with the global contextual understanding provided by ViTs. A high-resolution refinement module is also incorporated to improve the clarity of the generated images. Quantitative evaluations and qualitative comparisons show that the model is better at generating perceptually coherent and emotionally aligned facial expressions compared to traditional cVAEs. However, limitations in image resolution and ethical concerns, such as potential data bias and privacy issues, indicate areas where further improvements are needed. Future research will focus on exploring dynamic control of emotion intensity, developing advanced latent representations, and implementing bias mitigation techniques to improve both the model's performance and its fairness.

This research contributes to the advancement of emotion-aware generative AI and opens up possibilities for applications in areas such as affective computing, interactive media, and mental health technologies.

VII. Appendix

APPENDIX 1 - Data Description

The dataset used in this study is publicly available at:

- Emotion Recognition Dataset – Kaggle: https://www.kaggle.com/datasets/sujaykapadnis/emotion-recognition-dataset

This dataset contains facial images labeled with corresponding emotions such as Happy, Sad, Angry, and others. It was used to train and evaluate the emotion recognition model presented in this paper.

APPENDIX 2 - Code Availability

The complete implementation is provided through two Google Colab notebooks:

- Data Preprocessing and Model Training: https://colab.research.google.com/drive/1j24u76LhS3qmKZhC5PcJ2ZHgmbY_XJLq?usp=sharing

- Evaluation: https://colab.research.google.com/drive/11lNZ2TuRiS5KL-ldQFay2oymLr-vzNvH?usp=sharing

These notebooks contain the full code, including data loading, model training, and evaluation.

VIII. References

[1] Overcoming the Limitations of Vision Transformers in Local Texture Representation, PMC10830169, 2024.PMC

[2] Efrat Taig, "VAE: The Latent Bottleneck – Why Image Generation Processes Lose Fine Details," Medium, March 31, 2025.Medium

[3] Yang et al., "EmoGen: Emotional Image Content Generation with Text-to-Image Diffusion Models," CVPR 2024.

[4] Lei Cai, Hongyang Gao, Shuiwang Ji, "Multi-Stage Variational Auto-Encoders for Coarse-to-Fine Image Generation," arXiv:1705.07202, 2017.arxiv.org

[5] Y. Wu, J. Xu, and Y. Lin, “Overcoming the Limitations of Vision Transformers in Local Texture Representation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[6] A. Wang, T. Wang, and B. Xu, “Identifying Limitations of Vision Transformers in Structured Image Generation,” *CVPR Workshops*, 2022.

[7] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[8] S. Kovalchuk, "Implementing conditional variational auto-encoders (CVAE) from scratch," *Medium*, 2023. Meduim.

[9] "Vision Transformer (ViT) Architecture," *GeeksforGeeks*, 2025.GeeksforGeeks

[10] X. Yang, S. Wu, J. Zhao, S. Liu, Y. Liu, and Y. N. Wu, “EmoGen: Emotional Image Content Generation with Text-to-Image Diffusion Models,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2024.

[11] A. Hatamizadeh, H. Yin, G. Heinrich, J. Kautz, and P. Molchanov, "Global Context Vision Transformers," arXiv preprint arXiv:2206.09959, 2022.