VICL-CLIP: Enhancing Face Mask Detection in Context with Multimodal Foundation Models

Xinyi Gao[1[0000−0001−7727−9087]], Yanbin Liu[1[0000−0003−4724−8065]], Minh Nguyen[1[0000−0002−2757−8350]], and Wei Qi Yan[1[0000−0002−7443−3285]]

Auyckland University of Technology, Auckland 1010, NZ
{xinyi.gao,yanbin.liu,minh.nguyen,weiqi.yan}@aut.ac.nz

Abstract. Face mask detection has become crucial for public health and safety, especially during the COVID-19 pandemic. The existing methods, relying on large datasets of labeled human faces, pose privacy concerns and may not achieve high accuracy in diverse environments. In this paper, we present an innovative approach namely **VICL-CLIP**, which incorporates the Visual In-Context Learning (V-ICL) paradigm into the CLIP model to enhance face mask detection. By leveraging standardized cartoon images as learning context, our method addresses privacy issues while it also significantly improves detection accuracy. Specifically, we design effective multimodal prompts for in-context learning. Cartoon images with and without masks are proposed as the image prompts, while their corresponding text prompts are curated as the positive and negative contexts for the CLIP model. In this way, the model is able to be refined to generalize the capability from abstract representations to real human faces, through the inherent visual-text linkage. Our extensive experiments were conducted based on an real-world COVID Face Mask Detection Dataset. Our VICL-CLIP model achieves an excellent detection accuracy of 97%, outperforming all conventional methods and other state-of-the-art models. Moreover, this work underscores the potential of integrating the V-ICL learning paradigm into powerful vision-language foundation models to improve the mask detection accuracy while preserving privacy.

Keywords: Multimodal learning models · Visual in-context learning · Face mask detection

1 Introduction

Recent advances in the domain of Visual Language Models (VLMs) have led to significant advances in the field of machine learning [23]. The Contrastive Language-Image Pre-training (CLIP) method [15] has attracted much attention owning to its strong ability to understand and generate multimodal data. CLIP leverages a dual encoder architecture to improve the performance of the proposed method. Despite the recent progress, a great number of particular applications in a constrained scenarios, such as face mask detection, still pose significant challenges in terms of privacy and the need for high detection accuracy.

Face mask detection has been an essential task in a variety of public health and safety situations, especially in the context of a global health crisis such as the COVID-19 pandemic. Conventional face mask detection techniques often rely on large datasets of real human faces [6], which can raise privacy issues and ethical concerns. Furthermore, such techniques can struggle to achieve high accuracy when applied to diverse and unconstrained environments.

In order to solve those problems, we present an innovative method, namely **VICL-CLIP**, which incorporates the novel idea of Visual In-Context Learning (V-ICL) [24] into the powerful CLIP model. It can successfully improve the precision of face mask detection while simultaneously resolves the privacy issue. Specifically, our method relies on cartoon images of characters wearing and not wearing masks as visual contextual inputs. The introduction of cartoon images does not require any real human for training, thus effectively protecting personal privacy. To generalize the model capability from cartoon images to real faces, we further leverage the characteristics of zero-shot learning by designing positive and negative textual prompts aligned with the cartoon images. These aligned textual prompts explore the multimodal connection between image and text information, significantly facilitating the model performance by simultaneously leveraging the visual and textual contexts.

Visual In-Context Learning [24,25,10] is an emerging technique that improves the performance of large foundation models by incorporating relevant contextual information. For VLMs, this includes providing additional visual or textual cues to help the model understand the application context and process the target data. After applying visual in-context learning to the VLMs (e.g., CLIP [15]), our method is capable of bridging the gap between abstract representations and real-world applications. This enables our model to perform accurate mask detection without directly using sensitive personal data. Furthermore, by leveraging the vision-language model, our method also benefits from the zero-shot learning capability of VLMs, which enables the model to apply knowledge from unseen contexts to new datasets.

Our method, VICL-CLIP, extends the capabilities of CLIP model to the specific task of mask detection. By leveraging the robust text and image encoders of CLIP, we are able to distinguish between masked and unmasked faces in a zero-shot setting. *Firstly*, we introduce a novel prompt generation technique using ChatGPT to create descriptive prompts for both masked and unmasked faces. These prompts are used to guide the text encoder in the CLIP model, enhancing its ability to differentiate between the two classes. *Secondly*, we develop a contrastive learning framework that effectively maximizes the similarity between matched pairs of text and image representations, e.g., a prompt description of a masked face and an image of a masked face, while minimizing the similarity between mismatched pairs. This approach ensures that the model learns to accurately associate the correct textual and visual features. Furthermore, we apply L2 regularization to the combined feature vectors. This step is crucial for enhancing model generalization and preventing overfitting, thereby ensuring robust performance when the model is exposed to unseen data. *Finally*, we con-

duct extensive evaluations using the COVID Face Mask Detection Dataset [7]. These evaluations demonstrate the effectiveness and efficiency of our proposed approach, highlighting its potential for real-world applications in mask detection during the pandemic.

To evaluate the effectiveness of the proposed method, we conduct extensive experiments on the COVID Face Mask Detection Dataset, which comprises diverse real human face images. The dataset includes various configurations of masked and unmasked faces to simulate real-world conditions. Our method demonstrated excellent performance in face mask detection, achieving an overall accuracy of 97%. This marks a substantial enhancement over state-of-the-art large visual and visual-language models, underscoring the potential of V-ICL in improving model performance through zero-shot learning techniques [10]. Moreover, we provide a feasible solution for applications where user privacy is crucial by using cartoon characters as context inputs.

The remainder of this paper is organized as follows. Section 2 reviews related work in mask detection, multimodal methods and in-context learning. In Section 3, we detail the methodology of our approach, including prompt generation, text encoder, image encoder, contrastive learning and mask detection. In Section 4, we describe the implementation details of the training and evaluation stages. In Section 5, we present our experimental results, comparing the performance of various face mask detection models and our enhanced VICL-CLIP method. Finally, Section 6 concludes the paper with a summary of our key contributions, and discusses future directions.

2 Related work

Face Mask Detection Face mask detection has gained prominence due to the COVID-19 pandemic, necessitating reliable systems to ensure public health and safety [5]. Traditional face mask detection methods rely heavily on deep learning models trained on large datasets of labeled images. These methods typically employ convolutional neural networks (CNNs) [19] for feature extraction [13], followed by classifiers to determine whether a face is masked or unmasked. Despite of the effectiveness, traditional methods face several limitations, such as requiring extensive labeled datasets and potential privacy issues. DINO (DIstillation with NO labels) employs a self-supervised learning method using vision transformers [18] to capture and distill essential visual features without labeled data [2]. This method is currently a commonly used zero-shot object detection method.

Multimodal Learning Models (MLLMs) MLLMs are designed to process and integrate information from multiple modalities, typically visual and textual, to perform a variety of tasks [22]. Early models such as Visual Semantic Embedding (VSE) [3] and Neural Image Caption (NIC) [4] laid the groundwork by learning joint representations of images and their textual descriptions. Recent advances have introduced models such as CLIP and BLIP (Bootstrapped Language Image Pre-training) [11]. CLIP model utilize large-scale datasets of

unstructured web images paired with text to learn more generalizable representations, significantly improving zero-shot learning tasks. BLIP focuses on enhancing the relationship between visual and textual data through a bootstrapped training approach.

Contrastive Language-Image Pre-training (CLIP) CLIP is a novel MLLMs approach. It leverages contrastive pre-training on diverse datasets of image-text pairs [17], providing robust zero-shot learning capabilities. The model consists of two encoders: One for images and one for text. Both encoders are trained to map their respective inputs into a shared multimodal embedding space. During inference, CLIP can match images with their corresponding textual descriptions without explicit training on the target task. This enables effective zero-shot classification.

Zero-Shot Learning Zero-shot learning (ZSL) is a challenging machine learning paradigm [16]. Its goal is to recognize previously unseen objects or concepts. Traditional models require extensive datasets for each expected category. In contrast, ZSL uses supplementary data, such as semantic attributes, to infer the properties of unseen categories. CLIP's ability to embed images and text into a shared space naturally extends this concept. It enables effective knowledge transfer from seen to unseen categories.

In-Context Learning In-Context Learning (ICL) is an emerging paradigm [20]. It enhances a model's learning ability by integrating relevant contextual information. This technique has shown promise in various natural language processing (NLP) tasks [8]. In these tasks, providing additional context significantly improves performance. In the visual domain, V-ICL extends this concept by incorporating visual context to aid in understanding and processing target images. Our approach involves using both positive and negative cartoon images and corresponding prompts as visual context. This enhances the model's ability to learn distinguishing features related to masked and unmasked faces.

3 Methodology

Previous methods for face mask detection relies on a large dataset to train the deep neural networks. For example, MaskedFace-Net [1] includes 133,783 images for correctly or incorrectly worn mask detection. Yu et al. [21] curated a dataset of 10,855 images for face mask wearing detection. The reliance on such large datasets has two challenges: (1) There is potential privacy concerns about revealing the person identities in the training set, (2) The performance will drop significantly when applied to unseen application scenarios.

To address the above challenges, we propose an innovative method called VICL-CLIP, as shown in Fig. 1. The introduction of visual in-context learning with cartoon images tackles the first challenge, while the adoption of large vision-language models with visual-textual prompts address the second challenge.

The overall network architecture comprises two primary components: The text encoder and the image encoder. The text encoder processes positive and negative prompts, while the image encoder processes cartoon images representing

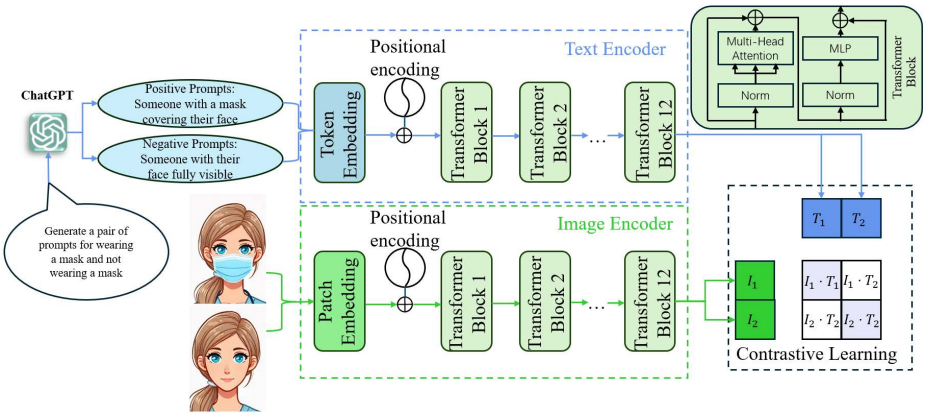

Fig. 1. The structure of our VICL-CLIP method. We design effective visual-textual prompts to finetune the large multimodal model such as CLIP. Specifically, cartoon images free from privacy concerns are adopted with aligned textual prompts to serve as the positive and negative multimodal prompt pairs. For inference, relevant questions are asked to match the positive (mask) or negative (no mask) prompts.

masked and unmasked faces. These encoded representations are utilized in a contrastive learning framework to facilitate effective mask detection.

3.1 Prompt Generation

The first step in our framework involves designing effective visual-textual prompts for the concrete face mask detection task. For *visual prompts*, we leverage the cartoon images depicting characters both wearing and not wearing masks to serve as the context to finetune the model for mask detection. However, cartoon images have a domain gap with realistic human faces. Therefore, we fix this gap by paring those cartoon images with meaningful and corresponding *textual prompts*.

To generate the prompts used in our model, we utilized the ChatGPT language model. We input the phrase "*Generate a pair of prompts for wearing a mask and not wearing a mask*" into ChatGPT, which generated the following descriptive prompts:

- **Positive Prompt:** "Someone with a mask covering their face"
- **Negative Prompt:** "Someone with their face fully visible"

These prompts are employed to guide the text encoder in distinguishing between masked and unmasked faces.

We pair the cartoon character wearing a mask with the prompt "*Someone with a mask covering their face*", whereas the (cartoon, text) pair serves as a positive example. In contrast, we generate the negative example by paring the

cartoon without a mask with the prompt "*Someone with their face fully visible*". By leveraging the proposed multimodal prompts (visual-textual) to finetune the model, the linkage between image and text can be explored to mitigate the gap between cartoon and real images.

3.2 Text Encoder

The text encoder processes the generated prompts. The prompts are firstly tokenized into word tokens, which are then embedded into high-dimensional vectors using a token embedding layer. To incorporate positional information, positional encoding is applied to these token embeddings. The resulting embeddings are fed into a transformer block comprising 12 layers. Each layer of the transformer block includes the following sub-layers:

1. **Layer Normalization:** Normalizes the input to stabilize and accelerate training.
2. **Multi-Head Attention:** Applies attention mechanisms to capture relationships between different tokens.
3. **Layer Normalization:** Another normalization step post-attention.
4. **Multi-Layer Perceptron(MLP):** Applies a feed-forward network to transform the input representations.

The output of the text encoder generates two representations: T_1 for the positive prompt and T_2 for the negative prompt.

3.3 Image Encoder

The image encoder processes visual input in the form of cartoon images, one depicting a masked face and the other an unmasked face. These two cartoon images were generated through the GPT-4 language model. The images are first divided into fixed-size patches, each of which is embedded into a vector representation through a patch embedding layer. Positional encoding is applied to these patch embeddings to retain spatial information. The encoded patches are then processed through a series of 12 transformer blocks, identical in structure to those used in the text encoder.

The output of this image encoder produces two representations: I_1 for the masked image and I_2 for the unmasked image.

3.4 Contrastive Learning

Contrastive learning forms the core of our method, aiming to align corresponding text and image representations while pushing apart non-corresponding pairs. The objective is to maximize the similarity between T_1 and I_1 (masked pairs) and T_2 and I_2 (unmasked pairs), while minimizing the similarity between mismatched pairs such as T_1 and I_2. The contrastive loss function is formulated as follows:

$$\mathcal{L} = -\log \frac{\exp(\text{sim}(T_1, I_1)/\tau)}{\sum_{i=1}^{2}\sum_{j=1}^{2}\exp(\text{sim}(T_i, I_j)/\tau)}, \tag{1}$$

where $\text{sim}(a, b)$ represents the cosine similarity between vectors a and b, and τ is a temperature parameter adopted from the CLIP model to scale the logits.

Additionally, we apply L2 regularization to the tensor outputs of both the text and image encoders to enhance stability and generalization of the embeddings. This is achieved by scaling the tensors to mitigate overfitting. This ensures that the model generalizes well to unseen data.

3.5 Mask Detection

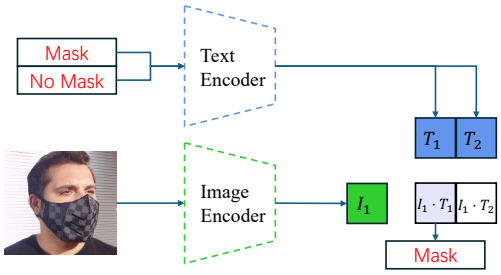

Fig. 2. The process of mask detection: The "mask" and "no mask" prompts are input into the text encoder, while the image to be detected is input into the image encoder. The similarity between the encoded representations is calculated through contrastive learning to determine whether the person in the image is wearing a mask.

Fig. 2 shows the process of mask detection. In real-world mask detection, the trained text and image encoders are utilized. Given an input image, its representation is obtained using the image encoder. This representation is then compared with the representations of the "mask" and "no mask" prompts through the contrastive learning framework. By assessing the similarity scores, we determine whether the individual in the image is wearing a mask.

By employing the new model, we enhance the accuracy of face mask detection while simultaneously addressing privacy concerns. We take use of cartoon images and corresponding prompts as context pairs. Our methodology demonstrates the potential of combining advanced machine learning techniques to improve model performance in a privacy-preserving manner. Future work will explore additional applications and further refinements to our approach.

4 Training and Evaluation Details

To evaluate the effectiveness of our approach for detecting face masks, we carried out comprehensive experiments by using the COVID Face Mask Detection

Dataset [7]. In this section, we provide a detailed explanation of the dataset, the preprocessing steps, the model training procedures, the evaluation metrics, and the baseline comparisons.

4.1 Dataset Description

The COVID Face Mask Detection Dataset contains real human face images divided into masked and unmasked classes [7]. The dataset is divided into training, validation, and test sets. Fig. 3 shows some examples from the dataset. This split simulates real-world conditions and ensures our model evaluation is robust. For our experiments, we specifically used the test dataset to validate the performance of our VICL-CLIP model.

Fig. 3. Sample images from the COVID Face Mask Detection Dataset.

4.2 Preprocessing

In order to ensure the consistency and applicability of the image input CLIP model, we implemented a rigorous pre-processing process. Firstly, all images were resized to the standard size of 224×224 pixels to match the input requirements of the CLIP model. Next, pixel values were normalized to a mean of 0.5 and a standard deviation of 0.5 to ensure a standard distribution across all images. We fine-tuned the pre-trained CLIP model using selected context image pairs and their corresponding text prompts. The training process involved several key steps. We created: (1) Positive pairs consisting of cartoon images wearing masks paired with the prompt "*Someone with a mask covering their face*"; (2) Negative pairs consisted of cartoon images not wearing masks paired with the prompt

"*Someone with their face fully visible*". The context images and prompts were encoded using the CLIP model's image and text encoders, respectively. This process mapped the inputs into a shared embedding space. Finally, the encoded image and text features are contrastively learned.

The evaluation process uses the COVID Face Mask Detection Dataset test set. During inference, for each test image, we computed similarity scores between the image representation and the representations of the "mask" and "no mask" prompts. Based on these similarity scores, the image was classified as either "mask" or "no mask". The evaluation metrics were then calculated based on the classification results.

4.3 Evaluation Metrics

In order to evaluate the performance of our VICL-CLIP model, we adopted the following evaluation metrics: Accuracy, Precision, Recall, and F1 score.

Accuracy measures the proportion of correctly classified images out of the total number of images. It provides an overall measure of the model's performance. The formula for accuracy is given by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}\,, \tag{2}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

Precision is the ratio of true positive predictions to the sum of true positive and false positive predictions. It indicates the accuracy of the positive predictions made by the model. The formula for precision is:

$$\text{Precision} = \frac{TP}{TP + FP}\,. \tag{3}$$

Recall measures the ratio of true positive predictions to the sum of true positive and false negative predictions. It reflects the model's ability to capture all relevant instances in the dataset.

$$\text{Recall} = \frac{TP}{TP + FN}\,. \tag{4}$$

The F1 score is the mean of precision and recall, providing a balanced measure of the model's performance. It is particularly useful when dealing with imbalanced datasets. The formula for F1-Score is:

$$\text{F1}_{\text{Score}} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}\,. \tag{5}$$

These metrics collectively offer a comprehensive evaluation of the model's performance, ensuring it accurately distinguishes between masked and unmasked faces in various scenarios.

4.4 Comparison Methods

To demonstrate the effectiveness of our proposed methodology, we compared the performance of our VICL-CLIP model with the baseline CLIP [15], BLIP [11], and DINO [2] models. The baseline models took use of the standard pre-trained versions without additional fine-tuning. Our VICL-CLIP model, which incorporates visual in-context learning with context pairs, showed significant improvements over the baseline models. This comparison highlighted the enhancements achieved through our approach, particularly in terms of accuracy, precision, recall, and F1 score.

Our model was implemented by using PyTorch and the Hugging Face Transformers library. The steps included initializing the model and processor, preparing context images and prompts, encoding features, applying L2 regularization, and evaluating the model on the test dataset. All experiments are conducted on a Windows 11 machine with an NVIDIA RTX 4080 GPU and 48GB RAM.

5 Experimental Results

In this section, the experimental results are presented to elaborate on the enhancement in performances with our approach as proposed. The performance of the VICL-CLIP method measured by accuracy, precision, recall and F1 scores are compared with the strong baselines such as CLIP, BLIP and DINO. Furthermore, we analyze the results in depth and backed them up using graphs.

5.1 Performance Metrics

The main evaluation metrics for evaluating model performance are accuracy, precision, recall, and F1 score. These metrics provide a comprehensive understanding of the model's ability to accurately distinguish between faces with and without masks.

Table 1. Performance Comparison of Different Models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
BLIP	47.00	46.94	46.00	46.46
DINO	45.00	45.28	48.00	46.60
CLIP	56.00	**100.00**	12.00	21.43
VICL-CLIP	**97.00**	97.96	**96.00**	**96.97**

Table 1 provides a comparative analysis of the performance metrics of various models. The models evaluated include BLIP, DINO, CLIP, and our proposed VICL-CLIP. The performance metrics considered for comparison are Accuracy, Precision, Recall, and F1-Score, each expressed as a percentage.

The BLIP model achieved an Accuracy of 47.00%, with a Precision of 46.94%, a Recall of 46.00%, and an F1 score of 46.46%. This indicates a balanced performance across all metrics, but with relatively low overall effectiveness in distinguishing between masked and unmasked faces. The DINO model demonstrated slightly lower Accuracy at 45.00%, but it exhibited a marginally higher Precision of 45.28%, Recall of 48.00%, and F1 score of 46.60%. This suggests that while DINO had a higher recall rate, indicating it was more successful in identifying all relevant instances, it did not perform as well in terms of precision and overall accuracy. The CLIP model showed a significant disparity between its Precision and Recall metrics. It achieved a high Precision of 100.00%, but this was accompanied by a notably low Recall of 12.00%, resulting in an F1-Score of 21.43%. The overall Accuracy for CLIP was 56.00%. The high precision indicates that when CLIP predicted a masked face, it was almost always correct, but its low recall suggests it failed to identify a large number of masked faces correctly.

In contrast, our proposed VICL-CLIP model demonstrated superior performance across all metrics. It achieved an impressive Accuracy of 97.00%, with a Precision of 97.96%, a Recall of 96.00%, and an F1-Score of 96.97%. These results indicate that VICL-CLIP not only correctly identified masked and unmasked faces with high precision but also had a high recall rate, making it the most effective model among those evaluated. Overall, the experimental results highlight the robustness and effectiveness of the VICL-CLIP model in accurately distinguishing between masked and unmasked faces, outperforming BLIP, DINO, and CLIP by a substantial margin across all key performance metrics.

The confusion matrices shown in Fig. 4 for the BLIP, DINO, CLIP, and VICL-CLIP models illustrate their respective abilities to distinguish between masked and unmasked faces. The BLIP and DINO models exhibit balanced but moderate performance, with significant misclassifications in both categories. Specifically, the BLIP model correctly identified 46.00% of masked faces and 48.00% of unmasked faces, while misclassifying 54.00% of masked faces and 52.00% of unmasked faces. Similarly, the DINO model correctly identified 48.00% of masked faces and 42.00% of unmasked faces, with misclassification rates of 52.00% and 58.00%, respectively.

In contrast, the CLIP model, though accurately identifying 100.00% of masked faces, struggled significantly with unmasked faces, correctly identifying only 12.00% and misclassifying 88.00%. The VICL-CLIP model demonstrated superior performance, accurately identifying 98.00% of masked faces and 96.00% of unmasked faces, with minimal misclassifications of 2.00% and 4.00%, respectively. This indicates that the VICL-CLIP model effectively enhances face mask detection accuracy by leveraging visual in-context learning with context pairs.

5.2 Ablation Study of Context Pairs

The integration of different prompt pairs significantly influenced the performance of the VICL-CLIP model in face mask detection. As shown in Table 2, the use of the prompt pair "*Someone with a mask covering their face*" and "*Someone with their face fully visible*" yielded the highest performance metrics, with an accuracy

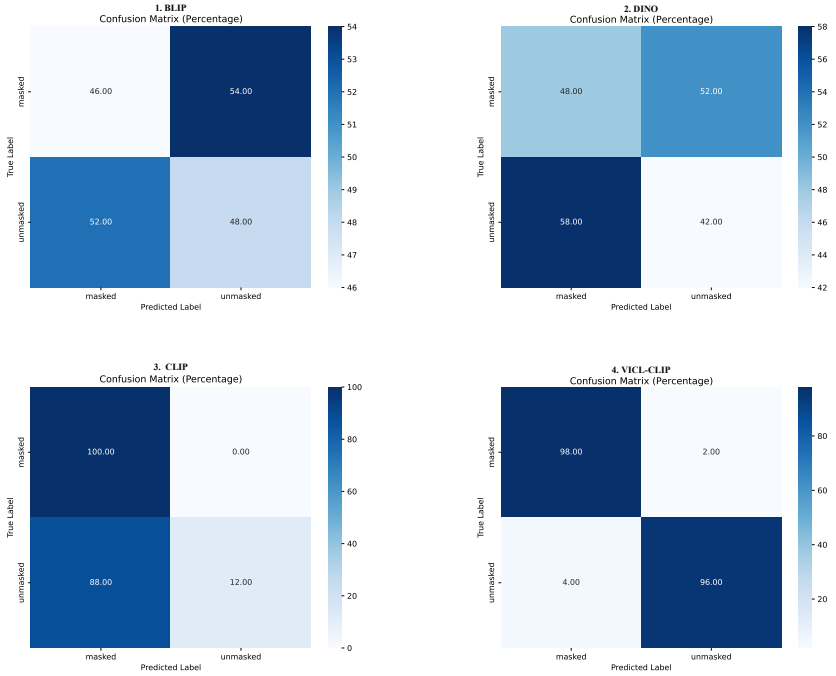

Fig. 4. The confusion matrix for BLIP, DINO, CLIP, and the proposed VICL-CLIP.

of 97.00%, precision of 97.96%, recall of 96.00%, and F1-score of 96.97%. This indicates that more descriptive and contextually relevant prompt pairs substantially enhance the model's ability to accurately classify masked and unmasked faces. In contrast, the approach using a single prompt, where only one type of textual description (either masked or unmasked) is provided, resulted in notably lower performance. Additionally, approaches that employ either two positive prompts (both masked) or two negative prompts (both unmasked) also show poor performance. This highlights the effectiveness of using contrasting context pairs (both masked and unmasked descriptions) for improving the model's generalization and robustness. Our ablation study demonstrates that the optimal performance is achieved when contrasting prompts are used, as this setup facilitates a stronger alignment between the image and text encoders. By identifying this optimal prompt pair, we aim to reduce the sensitivity of the model to the choice of prompts, thereby ensuring that future applications of the model are not disproportionately influenced by prompt design.

5.3 Comparison with Other Methods

The performance comparison with other state-of-the-art models demonstrates the superiority of the VICL-CLIP model. Specifically, VICL-CLIP achieved an

Table 2. Impact of Positive and Negative Context.

Prompt Pair	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Single Prompt	50.00	0.00	0.00	0.00
"A person *wearing* a mask" and "Someone with a *mask covering their face*"	43.00	46.24	86.00	60.14
"A person *without* a mask" and "Someone with their *face fully visible*"	80.00	87.50	70.00	77.78
"A person *wearing* a mask" and "A person *without* a mask"	94.00	95.83	92.00	93.88
"A face *with* a mask" and "A face *without* a mask"	95.00	95.92	94.00	94.95
"Someone with a *mask covering their face*" and "Someone with their *face fully visible*"	**97.00**	**97.96**	**96.00**	**96.97**

accuracy of 97.00%, which is significantly higher than that of the other models. Additionally, VICL-CLIP reached a precision of 97.96% and a recall of 96.00%, outperforming the other methods. In terms of the F1 score, VICL-CLIP achieved 96.97%, indicating its strong balance between precision and recall.

In contrast, other models performed relatively less well. For example, SSDMNV2 [14] achieved an accuracy of 92.64%, with a precision and recall of 94.00% and 93.00%, respectively, while YOLOv3 [9] had an accuracy of 93.90%. Although SwinTransformer+YOLOv8 [5] achieved a precision of 96.10%, its recall was only 90.60%.

Table 3. The Comparison with Other Face Mask Detection Methods.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
SSDMNV2 [14]	92.64	94.00	93.00	93.00
YOLOv2+ResNet50 [12]	-	81.00	-	-
YOLOv3 [9]	93.90	-	-	-
SwinTransformer+YOLOv8 [5]	-	96.10	90.60	-
VICL-CLIP (Ours)	**97.00**	**97.96**	**96.00**	**96.97**

In conclusion, the results of our experiments clearly indicate that the integration of Visual In-Context Learning with CLIP, using carefully selected positive and negative context pairs, significantly enhances the accuracy and robustness of face mask detection. Our approach not only outperforms the baseline CLIP model but also demonstrates superior performance compared to other face mask detection methods.

6 Conclusion and Future Work

In this paper, we present VICL-CLIP, a novel approach to enhancing face mask detection accuracy by integrating Visual In-Context Learning with the Contrastive Language-Image Pre-training model. Our method leverages context pairs using cartoon images of characters with and without masks paired with corresponding prompts to provide clear distinctions between masked and unmasked faces, addressing privacy concerns while improving generalization capabilities. Extensive experiments on the COVID Face Mask Detection Dataset, particularly focusing on real human face images, demonstrate that VICL-CLIP significantly outperforms the baseline CLIP model, achieving an overall accuracy of 94% compared to 56.0%. Precision, recall, and F1-score metrics also showed marked improvements. The use of positive and negative context pairs is crucial for enhancing model performance, enabling better generalization to real-world scenarios. VICL-CLIP also surpasses other leading methods such as YOLOv8 and Swin Transformer-YOLOv8 in accuracy, precision, recall, and F1-score, underscoring its substantial potential in improving face mask detection accuracy and reliability compared to current advanced techniques.

Our future research work will explore this methodology in other domains, such as emotion recognition, gesture recognition, and object detection, where context pairs provide valuable distinctions.

References

1. Cabani, A., Hammoudi, K., Benhabiles, H., Melkemi, M.: Maskedface-net–a dataset of correctly/incorrectly masked face images in the context of COVID-19. Smart Health **19**, 100144 (2021)
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9650–9660 (2021)
3. Chen, J., Hu, H., Wu, H., Jiang, Y., Wang, C.: Learning the best pooling strategy for visual semantic embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15789–15798 (2021)
4. Chen, S., Song, Z., Haque, M., Liu, C., Yang, W.: Nicgslowdown: Evaluating the efficiency robustness of neural image caption generation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15365–15374 (2022)
5. Gao, X., Nguyen, M., Yan, W.Q.: Enhancement of human face mask detection performance by using ensemble learning models. In: Pacific-Rim Symposium on Image and Video Technology. pp. 124–137. Springer (2023)
6. Gao, X., Nguyen, M., Yan, W.Q.: A high-accuracy deformable model for human face mask detection. In: Pacific-Rim Symposium on Image and Video Technology. pp. 96–109. Springer Nature Singapore Singapore (2023)
7. Gedik, O., Demirhan, A.: Comparison of the effectiveness of deep learning methods for face mask detection. Traitement du Signal **38**(4) (2021)
8. Kang, Y., Cai, Z., Tan, C.W., Huang, Q., Liu, H.: Natural language processing (NLP) in management research: A literature review. Journal of Management Analytics **7**(2), 139–172 (2020)

9. Li, C., Wang, R., Li, J., Fei, L.: Face detection based on yolov3. In: Recent Trends in Intelligent Computing, Communication and Devices: Proceedings of ICCD 2018. pp. 277–284. Springer (2020)
10. Li, F., Jiang, Q., Zhang, H., Ren, T., Liu, S., Zou, X., Xu, H., Li, H., Yang, J., Li, C., et al.: Visual in-context prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12861–12871 (2024)
11. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
12. Loey, M., Manogaran, G., Taha, M.H.N., Khalifa, N.E.M.: Fighting against covid-19: A novel deep learning model based on yolo-v2 with resnet-50 for medical face mask detection. Sustainable Cities and Society **65**, 102600 (2021)
13. Mbunge, E., Simelane, S., Fashoto, S.G., Akinnuwesi, B., Metfula, A.S.: Application of deep learning and machine learning models to detect COVID-19 face masks-a review. Sustainable Operations and Computers **2**, 235–245 (2021)
14. Nagrath, P., Jain, R., Madan, A., Arora, R., Kataria, P., Hemanth, J.: Ssdmnv2: A real time dnn-based face mask detection system using single shot multibox detector and mobilenetv2. Sustainable Cities and Society **66**, 102692 (2021)
15. 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)
16. Sun, X., Gu, J., Sun, H.: Research progress of zero-shot learning. Applied Intelligence **51**, 3600–3614 (2021)
17. Tu, W., Deng, W., Gedeon, T.: A closer look at the robustness of contrastive language-image pre-training (CLIP). Advances in Neural Information Processing Systems **36** (2024)
18. Xiao, B., Nguyen, M., Yan, W.Q.: Fruit ripeness identification using transformers. Applied Intelligence **53**(19), 22488–22499 (2023)
19. Yan, W.Q.: Computational methods for deep learning: theory, algorithms, and implementations. Springer Nature (2023)
20. Ye, J., Wu, Z., Feng, J., Yu, T., Kong, L.: Compositional exemplars for in-context learning. In: International Conference on Machine Learning. pp. 39818–39833. PMLR (2023)
21. Yu, J., Zhang, W.: Face mask wearing detection algorithm based on improved YOLOv4. Sensors **21**(9), 3263 (2021)
22. Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.T., Sun, M., et al.: Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13807–13816 (2024)
23. Zhang, G., Zhang, Y., Zhang, K., Tresp, V.: Can vision-language models be a good guesser? exploring vlms for times and location reasoning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 636–645 (2024)
24. Zhang, J., Wang, B., Li, L., Nakashima, Y., Nagahara, H.: Instruct me more! random prompting for visual in-context learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2597–2606 (2024)
25. Zhang, Y., Zhou, K., Liu, Z.: What makes good examples for visual in-context learning? Advances in Neural Information Processing Systems **36** (2024)