

Masked-to-Unmasked Face Translation: A Review of Datasets, Approaches, and Metrics

Yazid Safiny^{1,*}, Siham Ahmam¹, Nidal Lamghari¹, Abdelghani Ghazdali²

¹IDIA Laboratory, National School of Applied Sciences Khouribga, Sultan Moulay Slimane University, Morocco

²LaMDa Laboratory, National School of Applied Sciences Khouribga, Sultan Moulay Slimane University, Morocco

Abstract The emergence of face masks is causing major issues for computer vision systems as they prevent access to crucial parts of faces needed for applications like face recognition and expression analysis. To solve this issue, face mask removal which is a process that involves removing face masks and generating hidden facial details has become a popular topic of investigation. This paper provides an overview of various solutions proposed for face mask removal, paying particular attention to deep generative models, mainly based on GANs. Face mask removal methods can be grouped into three categories: one-stage networks, two-stage networks, and image inpainting techniques. Advantages, limitations, and architectural characteristics of each category will be analyzed comparatively. In addition, this paper looks at frequently utilized data sources and analyzes problems caused by synthetic mask generation. Metrics for evaluating face mask removal systems and losses used in them will be assessed in relation to realism and consistency of generated faces. Moreover, issues regarding benchmarking, reproducibility, and ethics of this area of research will be addressed. Future directions of the field will be outlined as well.

Keywords Masked Face Recognition, Face Mask Removal, Face Unmasking, Face de-occlusion, Masked-to-Unmasked Face Translation

DOI: 10.19139/soic-2310-5070-4153

1. Introduction

The widespread use of face masks in day-to-day life due to a variety of health and safety reasons has resulted in several challenges for computer vision systems. Use cases including face recognition [1], facial expression analysis [2], and lip reading [3] are particularly impacted due to the occlusion of key facial features. To overcome these limitations associated with face masks, a growing amount of research is focused on face mask removal. The main idea is to reconstruct face regions that are occluded by a mask. The problem is inherently difficult because the generated facial structure and texture must be realistic and consistent with the non-occluded face regions with respect to identity, head pose variations, and facial expressions.

In this review, the term face mask removal is used to describe the process of reconstruction of the occluded part of the face caused by the use of face masks through computational means. This problem is alternatively called face unmasking or face de-occlusion in the scientific literature; however, the two terms are sometimes used interchangeably. It should be noted that face de-occlusion is a much wider computer vision problem since it includes reconstruction of the occluded face regardless of what object occludes it, while face mask removal involves only reconstruction of the face occluded by face masks. Since this review focuses exclusively on mask-induced occlusions, the term face mask removal is used consistently throughout the remainder of the paper, except when referring to the terminology adopted in the cited literature.

*Correspondence to: Yazid Safiny (Email: yazid.safiny@usms.ac.ma). IDIA Laboratory, National School of Applied Sciences Khouribga, Sultan Moulay Slimane University, Morocco.

The recent developments in deep learning techniques, especially generative models, have been the key motivators for the existing face mask removal techniques. Most face mask removal techniques use large-scale datasets with paired images of faces with or without masks to establish a mapping from face-masked to unmasked facial areas. Evaluation of the face mask removal techniques is usually performed using quantitative image quality assessment metrics, as well as human perception studies to evaluate the realism of the generated face images.

In view of the rapid development of face mask removal research, several survey papers have addressed masked face analysis, primarily focusing on masked face recognition, detection, and general occlusion handling. However, current surveys place more importance on architecture descriptions and dataset analysis without sufficient critical comparison of face mask removal approaches and lack of evaluation reliability, dataset bias, and practicality discussions.

On the other hand, this study presents a critical analysis of face mask removal methods. More precisely, it offers:

- A unified taxonomy of face mask removal approaches by categorizing current techniques into one-stage, two-stage, and inpainting approaches to facilitate systematic comparisons among them in terms of architectural development.
- A critical comparative analysis of the existing techniques in light of their strengths, weaknesses, identity preservation capability, computational complexity, robustness, and practical deployment considerations rather than just analyzing individual architecture.
- A comprehensive analysis of optimization and evaluation approaches in terms of the most common loss functions, evaluation measures, and interaction among optimization goals within the context of face mask removal.
- An analysis of the datasets employed in face mask removal, focusing on the synthetic to real world domain gap, demographical diversity, data quality, dataset reality and the effects of all the above factors on generalization and benchmarking.
- An overview of the current research challenges and future directions including challenges with reproducibility, standardized benchmarking, diffusion models, transformers, and ethical challenges.

2. Related Work

Given that face masks have become very common nowadays, there are numerous reviews and surveys regarding the effects of wearing face masks on facial analysis, where there is an added focus on the area of masked face recognition. These surveys have considered issues like those resulting from partial occlusion, lack of datasets, and adjustment of face recognition algorithms to masked faces. For example, [4] includes a survey of different methods of masked face recognition ranging from popular deep learning models to synthetic masks generation techniques, datasets used, and issues of partial face recognition. [5] has provided a survey of deep learning based masked face recognition methods classified based on architecture type (CNN, GAN, Transformer) and face mask removal methods too. It also points out some problems, such as real-world occlusion, cross-domain generalization, and datasets with incorrectly worn masks. [6] is a survey paper that describes the advances in masked face recognition with emphasis on deep learning techniques, including CNN and Siamese networks. It describes the basic problems in facial variations caused by illumination, pose, type of masks, and partial occlusion, data augmentation, feature extraction, datasets and measures, and the applications in security, healthcare, and trends in the development of more efficient algorithms and performance in real-world scenarios. [7] is a unified review paper that includes masked face recognition, face mask detection, and face mask removal. It also includes detailed comparisons of the benchmark datasets, measures, and outstanding challenges and future research trends. [8] provides a focused review of face mask removal methods. It summarizes the main datasets, evaluation metrics, and loss functions used in face unmasking, while also discussing current challenges and future research directions, including dataset limitations, evaluation inconsistencies, and the gap between synthetic and real-world conditions.

Face mask removal is also closely related to several neighboring research areas that have significantly influenced its development. General image inpainting [9] and related image restoration techniques [10] provides many of the reconstruction strategies used to recover missing facial regions, including encoder-decoder networks [11], partial

convolutions [12], attention mechanisms [13, 14], transformer-based architectures [15, 16, 17], and more recently diffusion models [18, 19, 20]. Likewise, work on occlusion removal tackles the general problem of restoring image content after removing occluding objects in a manner that is visually consistent [21], providing algorithms that can be used to solve problems of facial occlusions. In parallel, face manipulation techniques such as face editing [22], face completion [23], face restoration [24], and face hallucination [25] have also seen the development of identity-specific loss functions, perceptual optimization, adversarial training, and facial semantics. While these areas have many overlapping techniques of image generation, there are several challenges faced by face mask removal tasks that do not apply to them as reconstructed images need to maintain both visual plausibility and biometric identity of the original face. Therefore, generic image generation techniques cannot be straightforwardly applied without addressing the identity preservation requirement.

The previous surveys are mostly devoted to the research area of masked face recognition and the task of face mask removal is seen as a part of the whole pipeline of the recognition system. Some surveys cover the topic of face unmasking and face reconstruction techniques, yet, the information provided is descriptive without a comprehensive analysis of reconstruction approaches, optimization techniques, and benchmarking. Moreover, existing surveys usually lack the discussion about the impact of loss function design, comparison of advantages and drawbacks of existing datasets, evolution of the reconstruction approaches and assessment of the methods in terms of their complexity, robustness, deployment feasibility and typical errors. Table 1 summarizes the scope of the existing surveys and highlights the aspects covered by each review.

Table 1. Comparison of existing surveys related to masked face recognition and face mask removal.

Feature	Alzu'bi et al. [4]	Saoud et al. [5]	Abdul Amir et al. [6]	Mahmoud et al. [7]	Safiny et al. [8]	Proposed Review
Primary focus	MFR	MFR	MFR	FMD, FMR, and MFR	FMR	FMR
Depth of FMR coverage	Limited	Limited	Limited	Moderate	Comprehensive	Comprehensive
Critical analysis of FMR methods	✗	✗	✗	Limited	Limited	Comprehensive
FMR methods summarized	✓	✓	✗	✓	✓	✓
Dataset overview	✓	✓	✓	✓	✓	✓
Comparative dataset analysis	Limited	Limited	Limited	✓	Limited	✓
Evaluation metrics	✓	✓	✓	✓	✓	✓
Loss function analysis	✗	✗	✗	✗	✓	✓
Interaction between loss components	✗	✗	✗	✗	✗	✓
Critical comparison of reconstruction methods	✗	✗	✗	✓	Limited	✓
Unified analytical framework	✗	✗	✗	✗	✗	✓
Standardized benchmarking recommendations	✗	✗	✗	✗	✗	✓
Identity preservation analysis	✗	✗	✗	✗	Limited	✓
Computational complexity analysis	Limited	✗	✗	Limited	✗	✓
Ethical, legal, and societal implications	✗	✗	✗	✗	✗	✓

MFR: Masked Face Recognition; FMD: Face Mask Detection; FMR: Face Mask Removal

From Table 1 it can be observed that most previous surveys focus on the masked face recognition problem, whereas the current survey focuses only on face mask removal. In addition to providing a comprehensive review of the techniques used, the current survey also introduces a classification scheme that classifies existing face mask removal techniques as one-stage, two-stage, and inpainting networks with a view to studying the evolution and shortcomings of these techniques. Furthermore, the current survey provides a multidimensional comparison by analyzing existing techniques taking into account factors such as identity preservation, robustness, computational cost, the requirement of training data, and applicability of the technique.

3. Training Objective and Evaluation Metrics

3.1. Evaluation Metrics

Face mask removal methods are commonly evaluated using pixel-level fidelity metrics, perceptual similarity measures, identity preservation metrics, and subjective human assessment.

The Structural Similarity Index (SSIM) quantifies the similarity between two images based on luminance, structure and contrast. It is appropriate for measuring the quality of reconstructed faces.

$$SSIM(x, y) = (l(x, y))^\alpha \cdot (c(x, y))^\beta \cdot (s(x, y))^\gamma \quad (1)$$

$l(x, y)$, $c(x, y)$, and $s(x, y)$ denote the luminance, contrast, and structural functions, respectively, where x , and y are the original (ground-truth) and the reconstructed (generated) images, respectively. The parameters $\alpha > 0$, $\beta > 0$, and $\gamma > 0$ control the relative importance of each component in the overall measure.

Multi-Scale Structural Similarity (MS-SSIM) [26] is the extension of SSIM that measures the structural similarity between the two images at different scales. MS-SSIM measures the luminance, contrast, and structural similarity at different scales by repeatedly downsampling the image. This is because MS-SSIM is capable of capturing the fine and coarse structural information in the image.

$$MS-SSIM(x, y) = [l_M(x, y)]^{\alpha_M} \cdot \prod_{j=1}^M [c_j(x, y)]^{\beta_j} \cdot [s_j(x, y)]^{\gamma_j} \quad (2)$$

Where c_j and s_j are contrast and structure comparison function at scale j . l_M is the luminance comparison function at the coarsest scale M . α_M , β_j , and γ_j are weights controlling the importance of each component. x and y represent the generated, and the target images, respectively.

The Peak Signal-to-Noise Ratio (PSNR) measures pixel-wise fidelity between the generated and target images, with higher values indicating improved reconstruction quality.

$$PSNR = 10 \cdot \log_{10} \left(\frac{MAX_i^2}{MSE} \right) \quad (3)$$

MAX_i denotes the maximum possible pixel value in the image, while MSE represents the Mean Squared Error. With:

$$MSE = \frac{1}{N} \sum_{i=1}^N (x_i - y_i)^2 \quad (4)$$

Where: N is the total number of pixels, x_i , and y_i correspond to the pixel values of the original and reconstructed images, respectively. N is the total number of pixels.

The Fréchet Inception Distance (FID) [27] evaluates the realism of generated images by measuring the discrepancy between feature distributions of real and synthesized samples extracted using a pre-trained Inception network [28]. Lower FID values indicate higher image quality and greater similarity to real data.

$$d^2((m, C), (m_w, C_w)) = \|m - m_w\|_2^2 + \text{Tr} \left(C + C_w - 2(CC_w)^{1/2} \right) \quad (5)$$

C , and C_w are the covariance matrices of the real and generated image distributions, respectively. m , and m_w indicate the mean vectors of the real and generated image distributions, respectively. And finally, Tr represents the trace of the matrix. $d^2(\cdot, \cdot)$ denotes the squared Fréchet distance between the feature distributions.

Perceptual similarity between reconstructed and ground-truth faces can be further evaluated using the Learned Perceptual Image Patch Similarity (LPIPS) [29] metric. Unlike pixel-based measures such as PSNR and SSIM, LPIPS computes the distance between deep feature representations extracted from a pretrained convolutional neural network, and has been shown to correlate well with human perceptual judgments. As a result, LPIPS is particularly effective at capturing perceptual artifacts and texture inconsistencies that may not be reflected in pixel-wise fidelity

metrics.

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (\hat{y}_{hw}^l - \hat{y}_{0,hw}^l)\|_2^2 \quad (6)$$

Where x and x_0 represent the generated and the target images, $\hat{y}^l$ and $\hat{y}_0^l$ are the unit-normalized feature vectors for a layer l . w_l is the learned channel-wise weight vector, $\odot$ represents the element-wise multiplication. H_l and W_l are the spatial resolution of layer l . Lower values of $d(x, x_0)$ indicate higher perceptual similarity.

The Mean Opinion Score (MOS) is a subjective metric that aggregates human ratings of generated outputs, thereby reflecting perceptual qualities that may not be captured by automated evaluation methods.

$$MOS = \frac{1}{N} \sum_{i=1}^N S_i \quad (7)$$

N denotes the number of evaluators, while S_i corresponds to the score assigned by the i -th evaluator. Typically, the ratings range from 1 to 5, where 1 indicates poor reconstruction quality and 5 denotes excellent reconstruction quality.

Identity preservation is usually measured via pretrained face recognition algorithms, whereby the comparison between the reconstructed image and the actual image is done via embedding similarity through cosine similarity or Euclidean distance. Popular algorithms used for generating the embedding include ArcFace [30] and FaceNet [31], whose ability to generate an embedding makes them more reliable than pixel-wise similarity comparison. Apart from distance in embeddings, the verification score or accuracy of classification as the same identity under a threshold is also reported in certain studies.

However, unlike metrics such as PSNR or SSIM, there is no standardized protocol for identity evaluation in face mask removal. Different studies adopt different pretrained recognition models, similarity thresholds, and evaluation datasets, which makes direct quantitative comparison of identity preservation across methods difficult. As a consequence, consistency of identity is measured by how much improvement is achieved in the same experiment instead of being an absolute measure from one paper to another.

While both PSNR and SSIM have been used extensively in relation to face mask removal due to their computational simplicity and clearness of interpretation, both measures are concerned with checking accuracy and structure similarity at the pixel level. In this case, however, the issue emerges in that such measures would favor an overly smoothed image despite its numerical similarity but lack the necessary attributes needed to identify the subject. This limitation is particularly important in face mask removal applications, where the objective extends beyond visual plausibility to include accurate identity preservation.

Metrics like FID and LPIPS give additional information due to their ability to measure the visual realism and consistency. Indeed, it is known that LPIPS demonstrates higher correlation with human perception compared to pixel-level metrics, which makes it an efficient tool for texture comparison and detection of any artifacts. On the other hand, FID metric evaluates similarities between distributions of generated and real pictures and can be used to understand the level of realism in the reconstructed image. However, neither FID nor LPIPS directly measures whether the reconstructed face preserves the identity of the original subject.

While subjective metrics like MOS do attempt to mitigate this issue by relying on human evaluation, there is still the problem of consistency since such evaluation cannot be easily controlled for evaluator bias and inconsistencies in experimental design. In addition, data used in research from other authors cannot always be compared due to the inconsistency of datasets and preprocessing methods used.

Collectively, all these shortcomings reveal that the existing metrics offer only a partial evaluation of face mask removal systems. Future evaluation protocols could be strengthened through the development of identity-aware composite metrics specifically designed for face mask removal. Existing evaluation metrics assess complementary but isolated aspects of reconstruction quality. However, none of the metrics is able to cover all the objectives needed for successful face mask removal. In fact, a comprehensive identity-aware metric may integrate structural similarity, visual similarity, and biometric consistency through face recognition embeddings in a single score. Also, more importance can be placed on discriminative parts of the face including eyes, nose, and mouth, instead of equal importance being given to all image pixels while calculating the metrics. Currently, there is no universal identity-aware metric, but its development is an interesting direction in the future face mask removal research.

3.2. Loss Function Components

While evaluation metrics are used to assess visual quality, loss functions are designed to guide the training process toward generating realistic, identity-preserving unmasked faces. In practice, multiple loss terms are combined to jointly enforce pixel-level accuracy, structural consistency, perceptual similarity, and adversarial realism.

The L1 loss computes the pixel-wise absolute difference between the target image x and the generated image $\hat{x}$, promoting accurate low-level reconstruction.

$$\mathcal{L}_{L1} = \|x - \hat{x}\|_1 \quad (8)$$

The L2 loss measures the squared difference at the pixel level between the reference image x and the reconstructed image $\hat{x}$.

$$\mathcal{L}_{L2} = \|x - \hat{x}\|_2^2 \quad (9)$$

Derived from the SSIM metric, the SSIM loss assesses the perceptual resemblance between x and $\hat{x}$ by comparing their luminance, contrast, and structural information.

$$\mathcal{L}_{SSIM} = 1 - SSIM(x, \hat{x}) \quad (10)$$

The perceptual loss relies on feature representations extracted from a pre-trained network, commonly VGG [32], instead of directly comparing pixel values, encouraging $\hat{x}$ to preserve the perceptually relevant characteristics of x .

$$\mathcal{L}_{perc} = \sum_i \|\phi_i(x) - \phi_i(\hat{x})\|_2^2 \quad (11)$$

ϕ_i denotes the feature maps extracted from the i^{th} layer of the network.

The identity loss encourages the generated image $\hat{x}$ to preserve the high-level identity characteristics of the ground truth image x . It is commonly defined as the L1 or L2 distance between deep feature embeddings extracted using a pre-trained network.

$$\mathcal{L}_{id} = \|\psi(x) - \psi(\hat{x})\|_2^2 \quad (12)$$

$\psi(\cdot)$ denotes the feature extraction function of the pre-trained network.

The adversarial loss, originating from GAN theory, encourages the generator G to synthesize outputs that are indistinguishable from real samples by attempting to deceive the discriminator D .

$$\mathcal{L}_{Adv} = \mathbb{E} [\log (1 - D(G(x_m)))] \quad (13)$$

x_m represents the masked input.

In practice, these loss terms are often combined into a single hybrid objective, for example:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{L1} + \lambda_2 \mathcal{L}_{SSIM} + \lambda_3 \mathcal{L}_{perc} + \lambda_4 \mathcal{L}_{id} + \lambda_5 \mathcal{L}_{Adv} \quad (14)$$

Where $\mathcal{L}$ is the overall training objective, and $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ are the parameters to balance the effect of each loss function. Such formulations are widely used in face mask removal models to jointly enforce reconstruction accuracy while preserving visual realism.

3.3. Interaction Between Loss Function Components

Current face masks removal approaches do not often involve just one objective function for optimization, because each loss function is used for optimizing one aspect of the reconstruction task. State-of-the-art approaches, therefore, tend to use hybrid loss functions where complementary objectives are optimized together in order to achieve a proper trade-off between various elements of the task. The performance of such approaches thus depends both on the choice of loss functions and their mutual interaction during optimization.

Pixel-based loss functions such as L1 and L2 constitute an optimal criterion for optimization because they minimize the gap between the generated and the ground-truth images. These kinds of loss functions preserve the

overall structure of the face but create a very smooth face image owing to the fact that it averages several possible faces leading to the elimination of high frequency details in the face image. Combining the SSIM loss with pixel-wise loss functions resolves this problem because they provide a mechanism to enforce structure consistency and preserve the geometry of faces, edges, and local textures while still retaining stable reconstruction ability.

The advantage of perceptual loss is that it supplements the pixel-wise loss function through a comparison between the feature representation of the images obtained from deep learning models instead of individual pixels. Thus, the inclusion of perceptual loss in addition to L1/L2 or SSIM loss leads to the generation of images with more sharp textures and semantically sensible faces, which are otherwise difficult to achieve by pixel-wise loss functions. Nevertheless, perceptual loss alone cannot ensure the generated face conforms to the real face distribution.

The use of adversarial loss is useful for further improving the quality of the reconstruction process, since the generated faces will be similar enough to the real examples, and thus, there will be better texture, more natural look of the face. Nevertheless, adversarial optimization introduces additional challenges, including training instability and the possibility of hallucinating visually plausible facial features that do not belong to the original identity. For this reason, identity-preserving losses are commonly incorporated to constrain the optimization within the embedding space of pretrained face recognition networks. Identity loss counteracts the identity drift that may result from perceptual and adversarial objectives, ensuring that improvements in visual realism do not come at the expense of preserving discriminative facial characteristics.

Comparative ablation studies in related face restoration and inpainting literature generally confirm that no single loss dominates, instead, performance improvements arise from carefully weighted combinations, where hyperparameters (loss weights) control the trade-off between fidelity and realism. However, optimal configurations are highly dataset- and architecture-dependent, and most works rely on empirical tuning rather than a universal loss formulation.

4. Datasets

Most datasets used for training face mask removal deep learning models are not public. A face mask removal dataset should have image pairs with and without face masks, and since it's expensive to have real world face mask removal datasets, most approaches use only synthetic face masks to create and train the mask removal models. The addition of face masks can be manually, or automatically. MaskTheFace [33] is an open-source tool which detects facial landmarks in a face using dlib [34], and based on key features, adds a face mask that fits the face. The tool supports different mask variations and accommodates various face angles. It can be used to generate masked face datasets from existing face recognition datasets.

Several datasets have been employed in the literature for face mask removal. In many cases, datasets originally designed for tasks such as face recognition or face generation are reused by adding synthetic masks to construct paired training data. One of the most commonly used datasets is CelebA [35], which contains over 200,000 face images and is frequently adapted by overlaying artificial masks. A higher-quality variant, CelebA-HQ [36], provides approximately 30,000 high-resolution images and is widely used for face inpainting and mask removal tasks due to its improved visual fidelity.

Another commonly used dataset is LFW (Labeled Faces in the Wild) [37], which, despite its relatively smaller size, offers significant variability in pose, lighting, background, and age. It contains 13,233 images from 5,749 individuals and is often augmented with synthetic masks for this task. FFHQ (Flickr-Faces-HQ) [38] is also extensively used, as it provides 70,000 high-quality and diverse facial images covering variations in age, ethnicity, and appearance, making it well-suited for mask augmentation and inpainting.

In general, any sufficiently large and diverse face dataset can be adapted for face mask removal, provided it includes a wide range of identities and variations in pose, expression, and illumination. For instance, VGGFace2 [39] is another suitable dataset, consisting of 3.31 million images from 9,131 identities, with 80 to 800 images per individual.

The Masked and Unmasked Face Dataset [40] consists of paired images of faces with and without real face masks. It is organized into two subsets: one containing paired images with background information, and another

focusing exclusively on the cropped facial region without background. Each subset includes 330 image pairs, where both the masked and unmasked versions are captured using real face masks. To the best of our knowledge, this is the first dataset to provide paired real masked and unmasked facial images. However, its relatively limited size may restrict its effectiveness for training deep learning models, particularly in terms of generalization to unseen identities and diverse mask types.

Face Mask Lite Dataset [41] contains 10,000 image pairs, the dataset contains synthetic masks and faces that are generated using Style GAN 2 [42]. Unmasking Face Mask [43] is a dataset with two sets: training set with 9954 image pairs, and testing set with 100 image pairs. Unmask-My-Celeb-A-HQ [44] comprises 30,000 image pairs with a size of 1024×1024 sourced from CelebA-HQ [36] dataset, divided in two sets: training with 29,000 image pairs, and test set with 1000 image pairs.

Table 2 summarizes the datasets commonly used for the face mask removal task.

Table 2. Comparison of datasets for face mask removal

<i>Ref</i>	<i>Dataset</i>	<i>Images</i>
[35]	CelebA	200,000
[36]	CelebA-HQ	30,000
[37]	LFW	13,233
[38]	Flickr-Faces-HQ	70,000
[39]	VGGFace2	3,310,000
[40]	A Masked and Unmasked Face Dataset	330 pairs
[41]	Face Mask Lite Dataset (Kaggle)	10,000 pairs
[43]	Unmasking Face Mask (Kaggle)	9954 pairs
[44]	Unmask-My-Celeb-A-HQ (Kaggle)	30,000 pairs

Figure 1 presents image samples from five different datasets. The first three datasets were originally introduced for a different task (CelebA [35], and CelebA-HQ [36], as well as FFHQ [38] that we added synthetic masks to it using MaskTheFace [33]) but has been repurposed for face mask removal, whereas the last two datasets (A Masked and Unmasked Face Dataset [40], and Unmask-My-Celeb-A-HQ [44]) was specifically designed for the face mask removal task.

Although the aforementioned analysis highlighted the main datasets applied in the existing literature, a further comparative analysis is necessary to realize how their properties affect the performance of face mask removal algorithms. The next subsection analyzes these datasets from different practical perspectives.

Comparative Analysis of Existing Datasets

Differences can be seen in the current face mask removal datasets based on their variety, realism, quality of annotations, and their appropriateness for various kinds of studies. Most publicly available datasets were originally developed for face recognition, face attribute analysis, or face generation rather than face mask removal. Consequently, researchers typically construct face mask removal datasets by synthetically adding masks to existing face datasets. Although this approach is capable of generating large paired datasets, it also brings some limitations which should be noticed when designing and assessing face mask removal methods.

Diversity is one of the key features of a dataset as it determines how well a model can generalize on unseen data. With respect to the diversity of identities, pose, illumination, image quality, and facial appearances, VGGFace2 [39] stands out as it includes the maximum number of identities among the analyzed datasets. Likewise, FFHQ [38] has a great diversity of age, hairstyles, facial accessories, backgrounds, along with high-resolution facial images. The datasets CelebA [35] and CelebA-HQ [36] also have a large number of identities and variation in facial expressions and poses, but demographic diversity seems less as compared to FFHQ [38] and VGGFace2 [39]. Contrarily, specialized face mask removal datasets such as Masked and Unmasked Face Dataset [40] contain considerably less diverse identities.

Also, there is an issue with dataset realism. The majority of datasets used for training, including CelebA [35], CelebA-HQ [36], FFHQ [38], LFW [37], and VGGFace2 [39], do not include masked faces in their native form. On the contrary, synthetic masks are put on them with the help of automatic mask generators, such as MaskTheFace

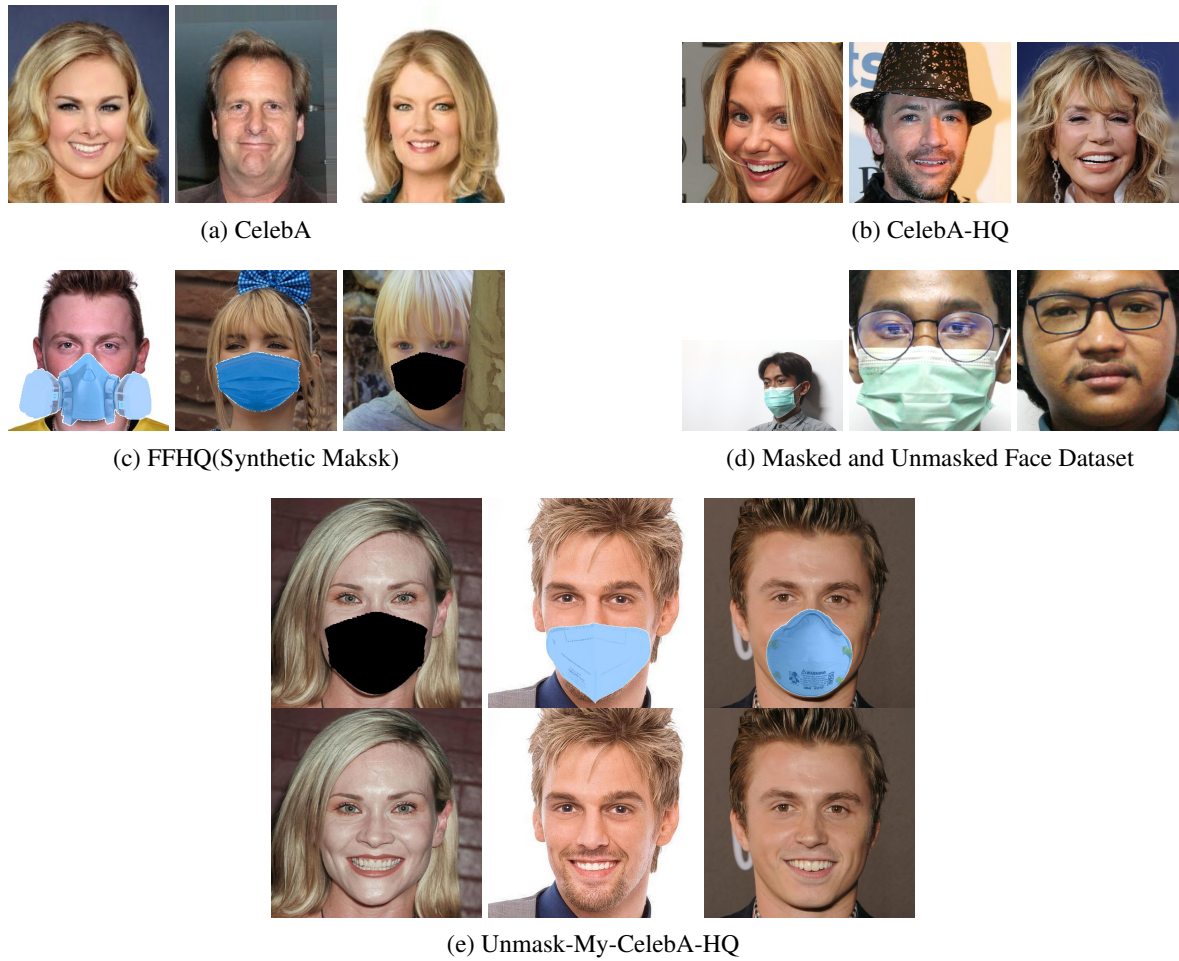


Figure 1. Representative examples of the principal datasets used in face mask removal research. (a) CelebA, a large-scale facial attribute dataset commonly used as the source of unmasked face images. (b) CelebA-HQ, a high-resolution version of CelebA that provides improved image quality for image generation and reconstruction tasks. (c) Flickr-Faces-HQ (FFHQ) with synthetically generated face masks illustrating the artificial masking strategy widely adopted for supervised training. (d) Masked and Unmasked Face Dataset, containing paired real masked and unmasked face images that reduce the domain gap between synthetic and real-world data. (e) Unmask-My-CelebA-HQ dataset, which provides paired masked and unmasked high-resolution facial images for face mask removal. The figure highlights the diversity of datasets used in the literature and illustrates the distinction between synthetic and real masked-face datasets.

[33]. This way, researchers can produce large amounts of paired datasets with a low amount of effort spent, however, even if synthetic masks imitate real ones quite closely, they cannot replicate such features of real masks as natural folds, distortions, shadowing, and small differences in masks' location. Thus, the model trained only on synthetic data will be hard to adapt to work with real world examples. Among the reviewed datasets, the Masked and Unmasked Face Dataset [40] is one of the few publicly available datasets containing paired images captured using real face masks, making it more representative of practical deployment scenarios despite its relatively very limited size.

Moreover, the annotated quality of the aforementioned datasets is also different. The datasets initially constructed for face recognition like CelebA [35], LFW [37], and VGGFace2 [39] contain reliable identity information but need further processing to obtain the pairs of masked-unmasked faces. The face mask removal datasets, such as Face Mask Lite [41], Unmasking Face Mask [43], Unmask-My-CelebA-HQ [44], and Masked and Unmasked Face Dataset [40], already provide the pair of data needed for supervised learning. It should be noted that most

datasets available lack rich annotations, such as mask segmentation map, mask attribute annotation, or annotations of occluded face regions.

Based on the demographic criteria, by visually inspecting the analyzed datasets, it can be concluded that there is under-representation of some facial features. Although FFHQ [38] and VGGFace2 [39] seem to represent diverse faces in terms of age, facial appearance, apparent ethnicity, and facial appearance in contrast to the other datasets, elderly people, individuals with facial hair, and certain skin colors are relatively less represented in several datasets. Moreover, celebrity-based face datasets mainly comprise faces of adults rather than children and senior citizens. This kind of limitation in demographics can impact the ability of the trained model to generalize well with the wide range of potential users. From an ethical perspective, improving demographic representation is important not only for achieving better generalization but also for reducing the risk of systematically lower reconstruction quality for underrepresented populations.

Mask variation is another important consideration. Since most face mask removal datasets are generated by adding synthetic masks to existing face datasets, the variety of mask types depends primarily on the mask generation algorithm rather than the original dataset itself. Although mask generation tools can produce several mask styles and viewing angles, many publicly available paired datasets rely predominantly on surgical masks with relatively limited variation in shape, texture, and wearing style. Real paired datasets better capture natural mask deformation and positioning but currently provide limited diversity in mask types because of their relatively small scale.

Environments also differ from one dataset to another. Unconstrained environments have been used in LFW [37] and VGGFace2 [39], which have contributed to high variations in background, lighting, camera angle, pose of the face, and image quality. CelebA [35] also has a variation in both background and poses, though it has lower resolution images compared to CelebA-HQ [36] and FFHQ [38]. High-resolution image datasets like CelebA-HQ [36] and FFHQ [38] contain clean faces because of lower background interferences, hence their suitability in image restoration and generative reconstruction. In contrast, dedicated real masked paired datasets are generally collected under more controlled conditions, resulting in less environmental variability.

Finally, the appropriateness of each of these datasets would be determined by the research goal. VGGFace2 [39], CelebA [35], and FFHQ [38] datasets are most appropriate for the training of deep learning algorithms due to their large amount of visual variability. The high-resolution datasets such CelebA-HQ [36] are ideal for image restoration techniques. LFW [37] remains useful for evaluating robustness under unconstrained imaging conditions. Conversely, the Masked and Unmasked Face Dataset [40] is more appropriate for evaluating reconstruction performance under realistic masking conditions than for training large-scale deep learning models because of its limited number of image pairs.

Overall, the analyzed datasets clearly show that there is a tradeoff between the scale of the dataset and realism of the face mask removal problem. On one hand, large-scale datasets offer great diversity and are perfect for training deep neural networks, however, all of those datasets use only artificial masks. On the other hand, the use of real pairs of masked and unmasked images makes datasets more realistic, however, they are relatively small, lack demographic diversity and variation of masks. In addition to enhancing scale and realism, future datasets should be created with respect to ethical practices in data gathering such as consent, licensing agreements, ensuring privacy, and balance in demographics. These aspects are crucial not only on the grounds of ethics but also because of their ability to reduce demographic bias and enhance the generalization capability of the face mask removal models. Therefore, future creation of datasets should be based on the creation of larger and diverse real-world datasets.

5. Methods

Face mask removal techniques target the restoration of the hidden parts of faces when the individuals are wearing face masks. It could assist non-masked face recognition algorithms to perform better and even help verify identities and surveillance applications when face masks obscure the essential facial attributes, thereby limiting the effectiveness of standard face recognition systems. However, such a task remains difficult as there is an increased need to create realistic and consistent identity in regions which have no prior information in them. In addition to

varying in terms of modeling the reconstruction technique, previous methods vary widely in terms of architecture used. To make a coherent classification, three categories have been proposed in this study based on the nature of the reconstruction process and architecture. Even though other classifications can be applied to face mask removal techniques, the current approach was preferred since it focuses on the actual reconstruction technique used.

- **One-stage networks:** In this category, the direct mapping from masked to unmasked faces is learned through an end-to-end process employing one network only. Despite their simple design and effective inference process, such approaches frequently encounter problems related to the precise reconstruction of complex facial structures owing to the lack of guidance along the intermediate stages.
- **Two-stage networks:** The key idea behind these approaches is the decomposition of the problem into two stages: mask localization or segmentation and subsequent reconstruction stage. Such decomposition facilitates greater structural guidance, resulting in improved identity retention but adds model complexity to the training stage.
- **Inpainting networks:** While not being specific to face mask removal, inpainting approaches tackle the problem of missing facial areas. They utilize context information obtained from available regions, yielding excellent local consistency. However, such approaches fail in retaining global identity information when large areas of faces are occluded.

Alternative taxonomies are also possible. Methods may be categorized according to supervision strategy (paired versus unpaired learning), backbone architecture (GAN-based, transformer-based, or diffusion-based models), or optimization objectives such as adversarial, perceptual, and identity-preserving losses. However, these dimensions often overlap within the same method, whereas the proposed pipeline-oriented taxonomy provides a clearer and more interpretable comparison framework for reviewing the evolution of face mask removal systems.

Figure 2 shows the general training process that many face mask removal approaches based on GANs use. Although individual implementations differ, most follow the same overall framework.

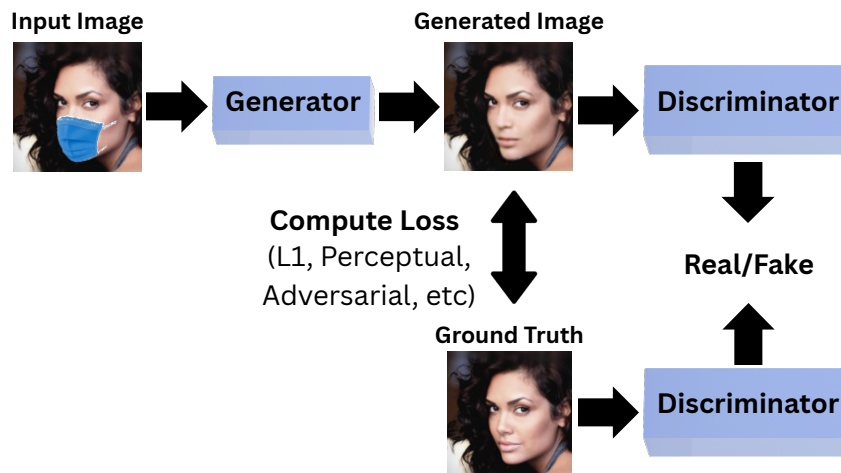


Figure 2. Generalized pipeline for face mask removal based on GAN. A face image with a mask is taken as an input of the generator, and the occluded part of the face is reconstructed to get an image of an unmasked face. In training, the generated image is compared to the ground truth unmasked image via the reconstruction loss and perceptual loss, while the discriminator learns to recognize the difference between generated and real images via adversarial learning.

Existing face mask removal methods not only vary in their reconstruction pipelines but also in their architectural design and optimization strategies. Depending on the approach, models incorporate different architectural components, such as encoder-decoder structures, residual learning, attention mechanisms, gated convolutions, and

facial priors, while their training objectives combine various reconstruction, perceptual, adversarial, and identity-preserving losses. These factors play a significant role in setting the balance between reconstruction fidelity, identity preservation, algorithmic complexity, and robustness against adverse environmental effects.

In the following subsections, we examine representative methods from each of these categories along with an analysis of their respective architectures and evolutionary developments.

5.1. One-Stage Networks

One-stage face mask removal approaches formulate the reconstruction task as a direct end-to-end mapping from masked to unmasked faces within a single network. In this setting, the model is expected to detect the occluded parts as well as generate the missing facial elements without depending on any form of explicit supervision like segmentation. The current literature suggests that most approaches involve an encoder-decoder architecture along with residual blocks, attention layers, and adversarial loss functions. Although this approach makes the framework structurally simple and easy to implement, it requires a lot of effort from the model to reconstruct the facial elements.

The autoencoder based methods like [45] demonstrate that a simple reconstruction strategy may do moderately well when ideal circumstances exist. However, the autoencoder method is typically trained on very few samples of synthetic images, implying that the autoencoder will not be able to generalize well enough to real-world masks because of the variation in appearance.

In contrast with other one-stage approaches developed subsequently, techniques using autoencoders do not utilize explicit knowledge about structure and semantics of images during the process of reconstruction; therefore, they are less computationally complicated, but their main disadvantage is the inability to reconstruct faces with real-world texture. This problem encouraged further research on architectures able to use more contextually and identity-related data.

To ensure better reconstruction results, some techniques rely on models with increased expressiveness. For instance, one approach presented by [46] relies on adding residual blocks along with attention layers to improve feature extraction and localization of relevant parts of an image. The same can be observed in approaches such as [47], which use attributes to include additional semantic information and thereby help reconstruct the occluded areas. Such techniques tend to perform well according to certain metrics like PSNR and SSIM, implying that the structural details have been preserved. Nonetheless, these metrics may not reflect the level of identity preservation of the face.

Even though residual connection, attention mechanism, and semantic attributes all enhance the reconstruction accuracy, they solve distinct problems of the face mask removal challenge. While residual connections make it easy for features to propagate while maintaining the spatial structure, the attention mechanism enables the network to concentrate on the occluded part of the face rather than the whole picture. Attribute-guided methods further constrain the generated appearance by introducing high-level semantic information. Consequently, improvements reported by different studies cannot be attributed solely to deeper architectures but also to the type of prior information incorporated during reconstruction.

GAN-based single-stage approaches further enhance the visual output using adversarial learning. In the case of the approach based on pix2pix [48], as used in [49], for example, special loss functions are employed that take into consideration facial landmarks and thus result in a better reconstruction of the masked parts. SAM C-GAN, proposed in [50], uses spatial attention along with the conditional GAN setting to synthesize natural looking unmasked faces. However, such techniques yield better quality textures in comparison to the autoencoder-based ones but suffer from increased instability during training.

Compared with reconstruction-based networks, GAN-based approaches generally produce more realistic facial textures because adversarial learning encourages the generated images to resemble the distribution of real faces rather than simply minimizing pixel differences. However, this improvement comes at the cost of more difficult optimization, increased sensitivity to hyperparameter selection, and the possibility of generating visually plausible but identity-inconsistent facial features.

Overall, one-stage methods demonstrate the efficiency of end-to-end training in the face mask removal task, however, their performance is highly dependent on the specifics of datasets and losses used. Even though methods

utilizing adversarial training and attention help achieve more realistic results visually, the absence of intermediate structural cues makes them highly vulnerable to noise. Thus, there exists an inherent contradiction in simplifying the network architecture while maintaining high reconstruction quality.

Compared to two-stage approaches, the one-stage framework provides a simpler and computationally cheaper pipeline but does not have any structure awareness and hence does not help maintain identity in cases where there are severe occlusions. In comparison to inpainting-based solutions, which use contextual information about the region, the one-stage approach is more dependent on the learning of the model itself. Table 3 lists discussed methods in the one-stage approaches of face mask removal.

Table 3. Summary of representative one-stage face mask removal methods and their reported evaluation results

<i>Ref</i>	<i>Method</i>	<i>Description</i>	<i>Dataset</i>	<i>Performance</i>
[45]	Autoencoders	Used autoencoders in order to remove the hidden area behind the mask and generate it.	Face Mask Lite Dataset [41]	91.00% (Accuracy)
[46]	A modified version of U-Net	An encoder-decoder structure, with residual blocks and attention units similar to U-Net [11] to unmask the face.	Synthesized CelebA dataset	0.9400 (SSIM) and 33.38 (PSNR)
[47]	Face inpainting via attributes	A dual-pipeline network to reconstruct the occluded areas of the face using attention mechanism and attributes relevant to the masked regions for accurate generation	Synthesized CelebA dataset	0.9180 (SSIM), 30.20 (PSNR), and 3.70 (FID)
[49]	A GAN-based method	Unmask faces using GAN [51] with a custom loss function focusing on facial landmarks and pixel errors in masked regions	A dataset of synthesized masked faces	0.9260 (SSIM) and 27.78 (PSNR) on 300 epochs
[50]	SAM C-GAN	A CGAN [52] with spatial attention module to remove the face mask from the input image and generate that area	A custom dataset from [53]	0.9123 (SSIM) and 30.98 (PSNR)

Reported results are reproduced from the original papers and may not be directly comparable due to differences in datasets, preprocessing pipelines, masking strategies, and evaluation protocols.

Evolution and Critical Analysis The development of one-stage face mask removal techniques is characterized by a continuous shift from simplistic approaches to more complex image reconstruction systems, which feature gradually increasing structural and semantic cues. The earliest systems used simple encoder-decoder autoencoders that learn how to perform the transformation from masked to unmasked faces. Despite showing the viability of face reconstruction through an end-to-end pipeline, such approaches yielded relatively smooth textures and could not reconstruct the occluded face area in detail. As research advanced, residual blocks, skip connections, attention, semantics, landmarks were added to help address this issue. Unlike making the model increasingly complex, these design elements were aimed at reducing the ambiguity related to the face reconstruction of the occluded area by adding more context to the process.

Although the reviewed methods belong to the same one-stage category, they adopt different forms of guidance during reconstruction. Residual connections improve feature propagation, attention mechanisms emphasize informative regions, landmark constraints preserve facial geometry, and semantic attributes constrain the generated appearance. Rather than representing competing alternatives, these mechanisms should be viewed as complementary components that address different sources of ambiguity in the reconstruction process. Consequently, recent methods increasingly combine multiple strategies within a single framework instead of relying on a single architectural innovation.

The development of objectives is another trend that can be observed. The previous models used the loss functions such as L1 and SSIM, which favor similarity between the pixels but generate extremely smooth images because there could be a number of faces matching the same masked image. The current approaches combine the reconstruction losses with the adversarial loss function or landmark-guided loss function in order to get a higher perceptual consistency. Although the current objective functions generate more realistic images, they complicate the training process and make the balancing of all the objectives difficult. The diversity of the loss functions used in various works makes the comparison hard.

The other limitation is related to the evaluation process. It is hard to test the robustness of single-stage approaches on real-life masked faces because they are usually tested and trained on synthetic datasets. Furthermore, reported performance is commonly based on PSNR and SSIM, whereas identity preservation and downstream face recognition performance remain less consistently evaluated.

Overall, developments in one-stage methods indicate that there is a shift from basic encoder-decoder networks to more guided reconstruction by means of attention-based approaches, semantic priors, and hybrid losses. Even though these improvements are helping to achieve better results in reconstruction, there is still a need for standard procedures and realistic benchmarks for fair comparison.

5.2. Two-Stage Networks

Two-stage face mask removal techniques split the whole process of reconstruction into several tasks, usually consisting of the segmentation task and the face reconstruction task. In this way, ambiguity involved in the direct reconstruction of large areas becomes minimized, and guidance about the process is explicitly included. The improvement achieved by such a procedure is especially valuable when there are heavy and abnormal occlusions. As a result, two-stage methods aim to improve reconstruction fidelity and identity preservation compared to single-stage approaches. Compared with one-stage reconstruction, decomposing the problem into multiple stages reduces the ambiguity associated with reconstructing large occluded facial regions because each stage solves a more constrained task.

Existing two-stage methods can be broadly grouped into three main design strategies. The most common paradigm is segmentation-guided reconstruction, where a dedicated network first predicts a binary mask indicating the occluded region, which is then used to guide a generative model for reconstructing the missing facial content [54, 55, 56]. The explicit localization can help guide the generation process to relevant areas and increase structural consistency. The second kind of methods involves the use of multi-tasking or feature-based approaches, where the tasks of emotion detection or feature extraction are employed to assist in adding semantics during the reconstruction process [57]. Finally, the latest methodologies involve the use of structure-assisted approaches for reconstructing the image through the inclusion of prior knowledge such as 3D face models or super-resolution systems [58, 59]. Such examples show that while the segmentation-reconstruction framework has dominated two-stage methods, other possible decomposition strategies can yield desirable results as well. Although these strategies all follow a multi-stage paradigm, they differ primarily in the type of guidance used during reconstruction. Segmentation-guided methods provide explicit spatial localization of the occluded region, multitask approaches exploit complementary semantic information such as facial expressions, whereas structure-assisted methods introduce geometric or resolution priors. Consequently, their differences lie not only in network architecture but also in the form of prior information available to reduce uncertainty during facial reconstruction.

According to the results from [54] and [56], an efficient localization of the mask is crucial for the successful performance of such methods. The ablation experiments carried out by the authors prove that the absence of the segmentation part or the poor localization lead to an inefficient mask removal and facial reconstruction.

While two-stage approaches differ from their one-stage counterparts due to the explicit presence of intermediate representations that enhance accuracy by reducing uncertainty during reconstruction, they help improve the accuracy of occlusion localization and identity maintenance, especially when there is complex masking. However, such an approach entails high architectural and training complexities. Moreover, multi-stage networks rely heavily on the first stage. If the first stage fails to generate an accurate representation of the input image, this can lead to poor reconstruction results at subsequent stages.

In addition, while inpainting-based techniques rely heavily on context from nearby regions, two-stage approaches utilize intermediate results to guide the process, providing some sort of explicit structural or spatial guidance. As such, this approach is more robust for dealing with large occlusions because simply using context might not be enough. On the downside, the use of intermediate guidance means that such approaches are less flexible and might perform poorly if the guidance is inaccurate.

Even with these advancements, however, there are some drawbacks still. The success of the reconstruction process is highly dependent on the fidelity of the intermediate representation, thus any errors arising in the previous stages will pass on to the next phase. Currently, all existing techniques are mostly tested on artificial masks for faces and not on natural ones, which makes it hard to measure their resilience under natural circumstances. In addition, differences in architectures and evaluation protocols complicate direct comparison across studies.

Overall, two-stage networks present a more systematic way of handling face mask removal with the use of intermediate supervision, which results in better face reconstruction and identity preservation. However, this is

accompanied by higher levels of complexity, sensitivity to intermediate errors, and generalization problems. Table 4 summarizes representative two-stage methods.

Table 4. Summary of representative two-stage face mask removal methods and their reported evaluation results

<i>Ref</i>	<i>Method</i>	<i>Description</i>	<i>Dataset</i>	<i>Performance</i>
[54]	GAN	Two stages were used: one to segment the mask, the other to generate an unmasked version of the face using a generator and two discriminators.	Synthesized CelebA dataset	0.8640 (SSIM), 26.19 (PSNR), and 3.54 (FID)
[55]	M-CSAM	Remove the face mask and generate that part with two stages: one for segmentation and the other having two discriminators and one generator with gated convolutions and attention mechanism along with dilated residual blocks.	Masked-Faces generated from CelebA dataset	0.9511 (SSIM), 30.38 (PSNR)
[57]	CCGAN	Combine CNN to classify facial emotions of masked faces, and CGAN for the reconstruction of the masked area of the face.	UCFCrime [60] and Youtube	90.00% (Accuracy) in anomaly detection
[56]	GANMasker	A two-stage network for face mask removal using GAN with an encoder-decoder, and an additional Murf mechanism to refine the generated output	Synthesized CelebA dataset	0.9500 (SSIM), 30.96 (PSNR), and 16.34 (FID)
[58]	JDSR-GAN	A network with two modules: a denoising module that generate an unmasked version of the input, and a super-resolution module to create a high-resolution version of the unmasked face.	Synthesized CelebA dataset	0.8500 (SSIM), 29.18 (PSNR), and 5.74 (FID)
[59]	Face mask removal based on 3D priors	Combined 3D face reconstruction and inpainting techniques to output a reconstructed the unmasked version of masked face	Synthesized CelebA-MaskHQ and FFHQ datasets	0.9120 (SSIM), 27.23 (PSNR), and 9.74 (FID)

Reported results are reproduced from the original papers and may not be directly comparable due to differences in datasets, preprocessing pipelines, masking strategies, and evaluation protocols.

Evolution and Critical Analysis Two-stage face mask removal techniques have also been witnessing a gradual change from basic segmentation-based reconstruction to more advanced systems with higher-level guidance. In earlier works, researchers mainly broke down the task into two consecutive steps of localization of the mask followed by generation of the masked faces using GAN. The breakdown clearly showed that the identification of the occluded area prior to generation was significantly helpful in reducing ambiguity when reconstructing large masked areas. With the development of this research line, various mechanisms like attention, gated convolutions, residual connections, multiple discriminators, and feature fusion techniques were used to make the generated facial textures more realistic without compromising the facial structure. Some recent approaches to two-stage face mask removal systems used auxiliary tasks like super-resolution, facial expression recognition, and 3D facial reconstruction.

Although the reviewed methods all belong to the two-stage category, they employ different forms of intermediate supervision. Segmentation-guided approaches explicitly localize the occluded region before reconstruction, multitask methods introduce semantic information through auxiliary learning objectives, whereas structure-assisted approaches exploit geometric priors or super-resolution to improve facial consistency. Rather than representing competing solutions, these guidance mechanisms address different limitations of the reconstruction process. Consequently, recent architectures increasingly integrate multiple complementary components instead of relying on a single source of intermediate information.

Yet another trend that can be observed is the development of the objective functions used. The older techniques would typically rely on the combination of reconstruction and adversarial losses. The newer techniques would also include perceptual, style, identity, landmark, and feature consistency losses, which would ensure preservation of facial features and identity of the face being reconstructed. While such a combination of objective functions leads to improved realism of the resulting images, it also increases the complexity of the optimization task and makes it more difficult to balance between conflicting objectives.

Intermediate supervision is one of the main features of two-stage algorithms. The results of ablation experiments performed in both [54] and [56] clearly show that without the segmentation step, the quality of reconstruction drops considerably, resulting in partial visibility of the mask in the resulting image. This suggests that the success of the reconstruction phase depends heavily on the quality of the intermediate representation. It means that there is an

inherent trade-off in two-stage approaches: whereas the guidance makes the reconstruction process less ambiguous, any errors of the first stage carry over to future stages.

Although two-stage methods generally outperform one-stage architectures in reconstruction quality, their reported performance should be interpreted carefully because evaluation protocols remain highly heterogeneous. Most studies employ different datasets, masking strategies, and combinations of image-quality metrics, while relatively few assess downstream face recognition performance or identity preservation using standardized protocols.

In general, the development of two-stage approaches shows that intermediate supervision can significantly enhance facial reconstruction by decreasing ambiguities in the generation process. In future work, the robustness of intermediate representation needs to be enhanced, and benchmarks and evaluation standards have to be created for comparing different approaches.

5.3. *Inpainting Networks*

As opposed to techniques that have been particularly developed for face mask removal, inpainting techniques provide solutions to the more general problem of image reconstruction where there is an occlusion. With respect to face mask removal, these approaches regard the occluded face area as a missing portion which should be reconstructed through information that is available within the visible portions of the face without taking the mask into consideration. In contrast to face mask removal models that are specifically created for this task, the inpainting techniques perceive face masks as just one type of occlusion within the domain of image completion. Therefore, face mask removal in the case of the inpainting is mainly carried out through context-based reasoning, and not based on the explicit facial modeling or recognition., making these methods naturally applicable to a broader range of occlusions.

Existing approaches for inpainting can be generally described in terms of their generative modeling approach, which employs GANs alongside the enforcement of global and local consistency [61, 23]. One typical strategy entails the deployment of encoders-decoders as the generator, inferring the missing parts based on the information from the contextual parts, while the global and local discriminators guarantee consistency on the various spatial levels. Some models also include semantic regularization to maintain structural consistency, thereby making sure that the reconstructed areas are consistent with other features of the face like eyes, nose, and mouth. While both representations make use of adversarial learning, the difference lies in how they exploit the context information. The earlier methods like [23] make use of traditional encoder-decoder generators to determine the missing part from the surrounding facial context. On the other hand, the recent methods include architectural features like partial convolutions, squeeze and excitation blocks, and localization of the objects [61]. This aids in propagating the features and distinguishing between the valid and occluded regions in the image. These additions demonstrate a gradual transition from generic image completion toward more context-aware facial reconstruction.

As compared with one-stage face mask removal techniques, inpainting techniques are less dependent on learned mappings and instead use contextual information to fill in the gaps. As a result, these techniques have higher adaptability and can handle other types of occlusions. Yet, such flexibility is achieved at the expense of the lack of any explicit modeling of the area that needs to be unmasked, making these techniques unable to recover the identity of the face especially in cases where most of the face is hidden.

The difference between two-stage methods becomes much more noticeable when compared to inpainting. Two-stage methods utilize some intermediate representations like segmentation masks or priors that help to reconstruct the object and provide more precise localization. Inpainting algorithms, on the other hand, do not rely on any prior knowledge and depend completely on context. Although this approach makes inpainting much easier to implement, in practice, it results in lower accuracy in the case of heavy occlusion because context alone is insufficient to restore identity-consistent elements.

As opposed to methods designed specifically for face mask removal tasks, the aim of inpainting is not to recreate the original concealed appearance of the face; rather, it is to produce a plausible restoration which matches the rest of the environment. Hence, although an inpainted face might look real, it might not contain the biometric information about the individual. This is the reason why realistic images do not necessarily lead to better recognition results.

Although flexible, existing inpainting solutions suffer from some key shortcomings. Existing techniques are usually trained on artificial occlusions that can be created either by the means of simple geometric masks or by overlaying faces with objects. Such training data can be a poor match for real-world face masks since occluded areas usually vary in shape and size. Moreover, the larger the area of occlusion is, the worse the results are obtained since the amount of context becomes insufficient for generating correct facial parts. Without any facial structures or identities being provided, the generated areas could look realistic but semantically wrong.

Overall, inpainting networks offer a general and flexible framework for handling occlusions, making them applicable beyond face mask removal. But they lack modeling and structure, which makes them inefficient in terms of recovering identity-important information in cases of extensive or complicated occlusions. All these considerations demonstrate the difference between a generic and special approach to restoring occluded areas of an image. Representative techniques that use inpainting principles are listed in Table 5.

Table 5. Summary of representative inpainting-based face mask removal methods and their reported evaluation results

<i>Ref</i>	<i>Method</i>	<i>Description</i>	<i>Dataset</i>	<i>Performance</i>
[61]	GAN for face de-occlusion	Use GAN with two discriminators and one generator to remove specific objects from the input image	Synthesized CelebA and CelebA-HQ datasets	0.9120 (SSIM) and 29.96 (PSNR) as the average on multiple objects
[23]	Generative Face Completion	Used a GAN architecture with one generator and two discriminators to reconstruct the missing parts of the input image	Synthesized CelebA dataset	0.8410 (SSIM) and 20.20 (PSNR)

Reported results are reproduced from the original papers and may not be directly comparable due to differences in datasets, preprocessing pipelines, masking strategies, and evaluation protocols.

Evolution and Critical Analysis The progress made in the development of inpainting-based methods is associated with the general trend in image completion algorithms rather than the design of specific architectures for masked face reconstruction. The earliest algorithms posed the problem as image completion and reconstructed the missing portion using only the context around the object through encoder-decoder GANs architecture. Even though these models managed to prove the power of adversarial training in generating convincing facial textures, they failed to recover subtle facial details due to their purely context-driven nature.

With further advancements made within the research domain, additional architectural components were designed in order to assist the model in enhancing contextual reasoning for the process of image completion. The utilization of partial convolutions [12] assisted the model in distinguishing between valid and occluded pixels during the feature extraction step, whereas the incorporation of squeeze and excitation blocks [13] helped in representing the channel-based features. In addition to this, an object localization component was also included to assist the model in localizing the occluded region before the process of reconstruction.

Another trend that needs to be noted is the emergence of optimization approaches. The first approaches were based on reconstruction losses and adversarial losses, which were meant to ensure visual authenticity of the resulting images. Nevertheless, recent approaches include perceptual losses and structural similarity losses in order to ensure texture authenticity and maintain facial appearance. On the other hand, unlike approaches specifically designed for the task of face mask removal, not much importance is given to ensuring preservation of biometric identity while optimizing images.

One of the common constraints among the existing methods for image inpainting is that the consistency of context does not guarantee consistency of identity. Since the occluded facial region is reconstructed solely from visible context, multiple plausible facial appearances may satisfy the reconstruction objective. Consequently, generated faces may appear visually realistic while differing from the true hidden facial characteristics of the original subject. This limitation becomes increasingly pronounced as the size of the occluded region increases because less contextual information is available to guide the reconstruction.

Even though inpainting techniques result in plausible reconstructions, evaluation is performed using only image quality evaluation measures that do not account for identity preservation. Moreover, there are not many approaches to face mask removal using inpainting techniques, hence the difficulty of comparing it with reconstruction approaches.

Overall, the development of inpainting methods proves to be effective in the reconstruction of faces through the use of context-based image completion. However, achieving identity-preserving facial feature restoration poses an important task, implying that future systems of face masks removal could greatly benefit from the combination of the two techniques.

5.4. *Emerging Generative Paradigms*

In recent times, diffusion models [18] and transformers [15] have shown exceptional results in image generation and restoration tasks [62, 16]. In diffusion models, the images are produced using iterative denoising, often producing higher perceptual quality than GAN-based methods for unconditional image generation [63]. Transformers use self-attention mechanisms to model long-range dependencies, which might help in reconstructing occluded parts of a face.

In comparison with GANs, diffusion models and transformers provide unique benefits for face reconstruction. The diffusion model gradually improves the quality of the reconstructed face in multiple denoising steps, which has already proven its efficiency in generating high-quality textures with less visual artifacts in various image generation tasks. The transformer model takes advantage of the self-attention mechanism for encoding long-range dependencies; hence, data from visible parts of the face can be used for reconstructing heavily occluded parts of the image. In turn, GAN-based solutions require only one pass of the image through the network to produce realistic reconstructions.

However, the application of diffusion and transformer-based approaches to face mask removal remains limited in the current literature [64, 20]. One key reason is the scarcity of real large-scale paired masked-unmasked face datasets, although diffusion models can sometimes work with smaller datasets via fine-tuning. Another important factor is the high computational complexity and slower inference of diffusion models due to their iterative sampling procedures. In addition, transformer models also demand heavy pre-training and large computational costs. Additionally, GANs can seamlessly incorporate perception and identity-preservation losses, which are extensively applied in face restoration applications. Thus, almost all current face mask removal models still use GAN-based architecture due to their training efficiency and reasonable reconstruction quality even with little data. Nonetheless, diffusion and transformer models have shown promise for future research in enhancing face reconstruction and preserving identities in scenarios with increasing datasets and improved sampling approaches.

Another critical factor to be considered is identity preservation. Even though the diffused-based models and transformer-based models have shown great results regarding visual reconstruction in general image restoration applications [65, 66, 67], it still remains unclear whether the improved results are being reflected on identity preservation during mask removal of the face. Currently, most research works only focus on assessing the performance in terms of image reconstruction based on image similarity and perceptual measures, while the evaluation of face recognition has been less studied. Therefore, there is currently insufficient evidence to conclude that these emerging paradigms outperform conventional GAN-based approaches for identity-preserving face mask removal.

Overall, GAN-based approaches are still the most commonly used due to the presence of the appropriate compromise among such components as reconstruction quality, computational efficiency, and the possibility of introducing identity-preserving criteria. However, it is necessary to note that both diffusion models and transformers will become an important subject of future research in the field under consideration since bigger real-world datasets, more efficient diffusion sampling techniques, and more lightweight transformer architectures continue to appear. It should be mentioned that few works have been done in order to compare them with GAN-based approaches directly.

5.5. *Summary of Face Mask Removal Approaches*

The face mask removal techniques can be classified into one-stage techniques, two-stage techniques, and techniques based on image inpainting. All three methods make use of deep generative models like autoencoders and GANs, however, the effectiveness of these techniques is contingent upon how well they perform localization and reconstruction of the occluded portions of the image.

The one-stage methods approach face mask removal directly from the input image to the output reconstructed face as an end-to-end solution. This is ideal for practical implementations that demand real-time performance or have strict constraints in resources availability. The primary disadvantage of these methods lies in the possibility of achieving a poor-quality reconstruction due to the lack of intermediate instruction that could result in less structure or identity consistency in complex occlusions cases. In contrast, two-stage approaches divide the task into two sequential steps. The major advantage of the proposed approach is the guidance by the predicted mask, which makes the reconstruction focused on the necessary occluded portions of the face.

The inpainting techniques employ a more generic approach where the occluded area is considered as missing information that must be recovered using the contextual information from the surroundings. The advantage here is increased flexibility and the ability to deal with different kinds of occlusions other than the face masks. On the other hand, the lack of explicit consideration of masks or any kind of structural assumptions means that they cannot effectively preserve the identity when large areas of the face are occluded.

The prevalent utilization of synthetic data for mask images is another typical challenge across all classes. Most times, this entails overlaying synthetic masks on existing facial datasets. The implication of this is that there will exist a domain gap, as the masks encountered during training may differ significantly from real-world masks in terms of shape, appearance, and occlusion patterns. It might be highly misleading to assess the efficacy of such algorithms using metrics like PSNR and SSIM since they do not take into account identity but rather focus on image similarity.

Another key limitation that should be mentioned refers to the comparability of the quantitative metrics reported by various works. Currently, existing face mask removal techniques are tested on different datasets, with varying approaches to mask generation, preprocessing, resolution of images, training/testing splitting criteria, and overall evaluation metrics. Furthermore, some research works use custom-built datasets based on CelebA [35], or CelebA-HQ [36] applying different masking and reconstruction techniques. Thus, metrics like PSNR and SSIM may only be considered performance indicators and cannot be used to conclusively prove that one method is better than another, due to the lack of common benchmarking criteria and testing conditions. Therefore, comparisons presented in Tables 3, 4, and 5 are intended to provide a qualitative overview of architectural trends, datasets, and evaluation practices rather than a standardized benchmark ranking.

Although these limitations reduce the reliability of direct comparisons across studies, the reviewed approaches still demonstrate the capability of modern generative architectures to reconstruct semantically consistent facial regions from partially occluded inputs. The widespread use of synthetic masks is primarily motivated by the scarcity of large-scale real masked-face datasets containing paired masked and unmasked images, whose collection is difficult and resource-intensive. Therefore, many existing methods should currently be interpreted as proof-of-concept systems rather than fully validated real-world solutions. It is reasonable to expect that these approaches could achieve improved robustness and stronger real-world generalization if trained and evaluated on realistic large-scale datasets containing real masked-face images.

Another issue is that of balancing the realism of visual reconstruction and biometric consistency. Several algorithms have been successful in providing highly realistic face reconstructions, however, they have not been able to ensure biometric consistency for face regions in which some portion is obscured or hidden. Consequently, it becomes difficult to associate the reconstructed image with the correct person due to identity drift. The lack of standardized evaluation protocols for identity preservation further complicates fair comparison across methods. However, the increasing trend towards identity-preserving reconstruction also brings out the two-sided nature of the removal of face masks technology. Although better identity preservation will increase the effectiveness of applications like biometric recognition and security and forensics, it may also increase the danger of misuse for the purpose of identifying people in an unwanted manner or violating their anonymity provided by wearing a face mask if used irresponsibly. Therefore, future methodological developments should consider responsible deployment alongside improvements in reconstruction accuracy and identity preservation.

Future studies can consider working towards minimizing the domain gap problem through the creation of datasets containing genuine examples of masked-unmasked image pairs. Identity-aware losses and consistent evaluation criteria could be considered when trying to increase the performance of face reconstruction models for identification purposes. From a practical perspective, model pruning techniques and lightweight networks are

important because of the necessity to deploy the algorithms in a real-time environment. Lastly, self-supervised [68], and semi-supervised [69] learning frameworks could potentially minimize the reliance on synthesized data.

5.6. Recommendations for Standardized Benchmarking

Lack of a standard benchmarking approach makes it impossible to compare the effectiveness of different methods for face mask removal, due to differences in the used dataset, mask generation, preprocessing, train-test splits, and evaluation measures in previous works. To improve reproducibility and enable meaningful comparisons, future research should adopt a common benchmarking protocol.

- **Standardized datasets:** It is essential that benchmarks comprise synthetic as well as real masked face datasets. Though synthetic datasets like CelebA-HQ and FFHQ still remain relevant tools in creating experiments in controlled settings, real datasets must be employed, involving paired images of masked and unmasked faces to validate generalization ability in realistic settings. The future datasets intended to be used as benchmarks must comprise a wide variety of identities, ages, ethnicities, lighting conditions, head poses, facial expressions, and types of masks. Publicly available data sets must be favored whenever possible over private datasets.
- **Unified train-validation-test splits:** Benchmark datasets should include a set of fixed identity-separated train, validation, and test sets. It is crucial to prevent identity leakage by avoiding having an individual appear in both training and test sets. Having official partitions will guarantee that all performance differences arise from the model architecture.
- **Common preprocessing procedures:** A consistent process should be maintained in the pre-processing pipeline adopted by the researchers which includes facial detection, landmark alignment, image cropping, image resolution, color normalization, and mask creation. It is essential that researchers maintain the details of their pre-processing pipeline since this will prevent any performance discrepancies that may arise due to inconsistencies in the data.
- **Unified evaluation protocol:** The performance should be assessed with the help of a combination of complementary metrics rather than assessing performance only with the help of PSNR and SSIM. Perceptual quality, pixel-wise fidelity, and subjective assessment should all be provided in addition to identity preservation metrics. The benchmarking report should also consist of computation efficiency statistics such as inference time, number of parameters, and GPU memory usage.
- **Standardized identity evaluation:** The process of identity preservation assessment must be based on a certain protocol in order to allow an equal comparison between studies. First, reconstructed and ground-truth facial images should undergo the same preprocessing pipeline, including face detection, landmark alignment, cropping, and normalization. Identity embeddings should then be extracted using a common pretrained face recognition network, with identical feature extraction settings. Identity preservation should be quantified using cosine similarity between embeddings and verification accuracy under a decision threshold. Receiver operating characteristic (ROC) curves and the associated area under the ROC curves (AUCs) should also be included in order to assess the performance of verification at different operating points. Identity Leakage is supposed to be avoided by performing any kind of evaluation using exclusively disjoint from the training process datasets. In conclusion, all experiments have to specify the face recognition algorithm being used, preprocessing steps applied, method of choosing the threshold value, and the datasets used for evaluation.
- **Cross-dataset evaluation:** In addition to evaluating the model using the training dataset, the models need to be evaluated using other datasets collected under different circumstances. This type of evaluation will give a more realistic measure of robustness and generalization in view of the existing dependence on synthetic mask creation.

Based on these suggestions, future research related to face mask removal technology needs to adopt a more comprehensive benchmarking approach that is capable of evaluating the reconstruction accuracy, perceptual reality, preservation of the identity, computational efficiency, and generalization performance over multiple face datasets at once. In contrast to the traditional one-dimensional approach of performance evaluation based on a single performance measure, this new approach should include both objective measures and standardized evaluation methodology involving identity-disjoint data partitions, unified preprocessing, and evaluation on both synthetic and realistic face-masked datasets. In addition, detailed information about the implementation and hyperparameters of the pre-trained model would facilitate reproduction of results and unbiased comparison between different methods. Such a benchmarking framework would provide a more reliable basis for assessing future advances in face mask removal than the heterogeneous evaluation practices currently used in the literature.

6. Multidimensional Comparative Analysis of Face Mask Removal Methods

Based on the limitations identified throughout this review, we propose a multidimensional evaluation methodology consisting of five complementary assessment dimensions: identity preservation, robustness, computational complexity, training data requirements, and deployment readiness. These dimensions are used in the following subsections to analyze existing methods and are intended as a standardized evaluation protocol for future face mask removal research.

6.1. Identity Preservation

Identity preservation can be considered the primary objective of face mask removal since the reconstructed faces are often required for further uses such as face recognition, biometric verification, surveillance, and forensic analysis. The visually realistic reconstruction does not mean much if it changes the specific features of the identity.

The two-stage approaches tend to show the most promise in the domain of identity preservation. These techniques minimize the ambiguity of generating the missing face region by estimating some intermediate information prior to reconstruction; examples of this include estimating masks, facial geometry, semantics, etc. The additional structural guidance helps constrain the reconstruction process and decreases the likelihood of identity drift.

One-stage methods reconstruct unmasked faces directly through a single end-to-end network without explicit intermediate supervision. Their simpler design enables efficient inference and often produces visually convincing results. However, because the network must simultaneously localize the occlusion and reconstruct missing facial content, identity-specific characteristics may be modified, particularly when large facial regions are occluded.

Inpainting methods depend on the use of contextual information from nearby areas to reconstruct missing regions as opposed to facial priors. Even though such methods yield locally consistent and plausible reconstructions, they do not guarantee consistency in terms of identity since the larger the missing region becomes, the more the possibilities of different reconstructions there are.

Overall, current evidence suggests that architectural guidance plays an important role in preserving facial identity. However, identity preservation is influenced not only by network architecture but also by the optimization objectives adopted during training. Recent methods increasingly incorporate identity-preserving losses, perceptual losses, feature-matching losses, or facial landmark constraints to encourage reconstructed faces to remain consistent with the original subject while maintaining perceptual realism. In particular, identity-preserving losses computed using pretrained face recognition networks directly constrain biometric feature representations, whereas perceptual losses encourage preservation of high-level semantic features that are often correlated with facial identity. However, these objectives are typically complementary to architectural advancements, in that they reduce identity drift during reconstruction. However, a direct comparison is hard to make since few researchers have considered assessing identity preservation through established face recognition or benchmark standards.

6.2. *Robustness to Real-World Conditions*

Generalization to real-world masked faces remains one of the most significant unresolved challenges in face mask removal. Most published methods are trained and evaluated using synthetically generated masks that exhibit limited variability in shape, position, texture, and illumination. Consequently, the results achieved do not necessarily represent actual performance.

While one-stage approaches are effective for training datasets having a variety of synthetic masks, their robustness decreases when there is a large variation in poses, illuminations, or unusual placement of masks due to the fact that reconstruction is based only on learned features.

Two-stage techniques prove more robust as they take into account the occluded part before reconstructing the image. Guidance at an intermediate stage helps recover the geometry of the face in case there is significant occlusion of the face by masks. Nevertheless, their effectiveness heavily relies on the correctness of the segmentation or guidance stage.

Inpainting-based methods are inherently flexible because they treat masks as a generic image completion problem and can therefore be applied to a wider variety of occlusions. Nevertheless, reconstruction quality deteriorates rapidly when contextual information becomes insufficient, particularly for large facial masks or heavily occluded lower facial regions.

Despite recent progress, relatively few studies perform cross-dataset evaluation using real masked-face images. Consequently, the robustness of current methods under unconstrained real-world conditions remains largely unverified.

6.3. *Computational Complexity*

Computational efficiency is another important factor to consider especially for those applications that require real-time computation like surveillance applications, mobile computing, and embedded biometrics platforms.

Among the three categories, one-stage methods generally exhibit the lowest computational complexity because reconstruction is performed in a single forward pass. Their simpler pipeline typically results in lower inference latency, reduced memory consumption, and easier deployment on resource-constrained devices.

Two-stage techniques need to run multiple processing phases in sequence, such as mask localization and facial reconstruction. While this separation enhances image reconstruction and identity recognition performance, it comes at the expense of higher costs and increased latency over one-stage techniques.

The computational complexity of inpainting algorithms depends mainly on the architecture used for image reconstruction. Traditional convolutional inpainting models possess computational complexities similar to those of other CNNs, while newer transformer and diffusion image completion models have much higher computational needs because of their self-attention mechanisms or iterative noise removal processes.

Therefore, there is an existing compromise between reconstruction accuracy and computation efficiency, wherein more complex models usually entail higher computational costs.

6.4. *Training Data Requirements*

Deep learning based face mask removal relies heavily on the availability of appropriate training data. Supervised learning models rely on pairs of masked and unmasked faces, which are rare to obtain because collecting such facial images before and after masks is costly and time-consuming.

In order to address this shortcoming, most of the current methods use a strategy of creating synthetic masked face datasets by placing synthetic masks on top of public datasets of faces, for example, CelebA or FFHQ. While this method allows training a model in a supervised way using a large dataset, the synthetic masks do not capture all the variations of real-world masks.

Other than the difficulties of employing artificial masks in generating such models, there is also an influence brought about by the demographic composition of the training datasets for the construction of such models. In that most of these models are built through the usage of images of the face, any imbalance brought about by certain demographic differences in terms of age, skin color, facial features, and the conditions under which images are captured might lead to poor performances in reconstruction of the faces. Consequently, future datasets should

emphasize demographic diversity alongside realistic masked-unmasked image pairs to improve both generalization and fairness of reconstructed facial representations.

Furthermore, there are techniques that need auxiliary supervision information in the form of landmarks on faces, semantic segmentation maps, identity labels, or three-dimensional face priors. While this extra information helps to enhance reconstruction accuracy, it simultaneously complicates data preparation and annotation.

6.5. Deployment Readiness

While several techniques for removing masks from faces show promise from a quantitative perspective, only a few seem suitable for implementation into practical biometric systems. While most experiments performed by researchers on the subject are carried out on well-captured images of faces taken in highly-controlled lab settings, practical implementations require much more flexibility in terms of acquisition environment.

When it comes to deployment, one-stage approaches appear to be the most suitable ones at present due to their relatively low computational complexity and effective inference. On the other hand, two-stage approaches usually exhibit superior performance in terms of reconstruction quality and identity preservation, although they require more computation and are more sensitive to prediction errors at the intermediate stages. Inpainting approaches demonstrate high adaptability in restoration, although they may only be applied in offline image editing rather than biometric applications.

In spite of considerable progress made by academic researchers, the practical use of techniques to remove face masks is still extremely rare. As opposed to the reconstruction of occluded facial areas, most commercial systems for face recognition have concentrated on the development of recognition algorithms that would work with masked faces. This tendency is caused by the ambiguity of the synthesized facial image and the potential inconsistencies in personal identification that could result from its usage.

Despite this fact, there can be some possible areas of application for this method in forensics, crime investigation, and photo restoration, where the face masks reconstructed can give an analyst more information. Nevertheless, in all these cases there will be the need for validation, testing of the approach, and understanding of its limitations. This difference between theory and practice lies not only in the accuracy of face reconstruction but also in robustness, reliability, efficiency, and some other aspects.

Apart from the accuracy of reconstruction, there is a need for lightweight architectures, evaluation on unconstrained datasets, estimation of uncertainty in the reconstructed facial areas, and privacy-sensitive deployment procedures.

6.6. Taxonomy of Common Failure Modes

While there have been significant advancements made in face mask removal, the literature contains some repeated problems that occur regardless of the architectural design adopted. Although the problem appears in different forms in single-stage, two-stage, and inpainting-based architectures, the main source of these problems lies in one fundamental issue: it is necessary to restore facial regions that are fully occluded and therefore lack any observable data. Contrary to being unique problems to specific methodologies, these recurring problems are failure modes that affect the applicability of the current face mask removal approaches in practice. According to the literature examined, the primary failure modes may be divided into six categories: identity inconsistency, reconstruction ambiguity under severe occlusion, visual artifacts and texture degradation, limited generalization to real-world conditions, error propagation in multi-stage architectures, and evaluation-related shortcomings. The identification of the reasons for these failure modes provides insight into why current methods often perform well on benchmark datasets while remaining insufficiently robust for practical deployment.

Identity preservation remains among the most consistent problems that arise in all techniques for face mask removal. Because facial masks cover regions on the face that can be used to discriminate between people such as the nose, mouth, and chin, the reconstruction task becomes an under-constrained problem. As a result, there could exist several possible appearances for the same face covered by the mask, thus leading generative models to make assumptions about the missing facial features instead of preserving the actual identity of the person. Even though some state-of-the-art techniques utilize identity-preserving, perceptual, or landmark loss functions to solve the problem, the reconstructed faces could still be different from the original one. This limitation becomes particularly

problematic when face mask removal is intended to support downstream applications such as face recognition or biometric authentication.

The performance of reconstruction typically gets worse when the amount of facial occlusion is greater. If just a small part of the face is occluded, then the surrounding information is enough to guide the reconstruction process. In case of a large mask, a lot of structure and semantics information will be lost, making it difficult for the model to make accurate judgments about the occluded area of the face since there is not enough visual information to infer that area. This ambiguity affects all categories of face mask removal methods but is particularly pronounced in context-driven inpainting approaches, where reconstruction relies almost exclusively on surrounding pixels.

Yet another typical issue is associated with visual quality degradation, which includes blurring of textures, inconsistencies in the appearance of the skin, and artifacts resulting from image reconstruction in the vicinity of the mask. Initially, encoder-decoder models often suffered from overly smooth textures of faces since they were based on pixel-wise reconstruction losses, like L1 or SSIM. Even though the introduction of adversarial and perceptual losses significantly increased realism of the output images, some artifacts might emerge while blending the reconstructed areas with the visible part of the face. In two-stage architectures, inaccurate mask localization further increases the likelihood of residual mask traces or discontinuities at reconstruction boundaries, demonstrating the strong dependence of image quality on accurate intermediate representations.

Most existing face mask removal methods are trained and evaluated using synthetically generated masked-face datasets. Although synthetic data facilitate supervised learning with paired images, they do not fully capture the variability of real-world masks, illumination conditions, facial poses, and demographic characteristics. Consequently, models that perform well on benchmark datasets may exhibit substantially reduced performance when applied to real masked-face images. This synthetic-to-real domain gap remains a major obstacle to practical deployment.

While one-stage techniques do not possess this drawback, two-stage techniques create an additional potential for failure in the process of sequential processing. Two-stage methods require performing some intermediary tasks that help to achieve the result, such as mask segmentation, facial landmarks detection, or structural guidance prior to reconstruction. The use of intermediate supervision helps to enhance the quality of reconstruction since the latter becomes less ambiguous; however, the errors at the previous stage will always influence the next stage. This illustrates the inherent trade-off between stronger structural guidance and increased architectural dependency.

Another less direct but equally significant issue relates to the methodologies applied in evaluating the models. In most cases, the results are expressed in terms of reconstruction metrics that measure the similarity on the level of pixels. However, this approach does not reveal much about how well identity is preserved and how effective is the recognition process. Therefore, even though the reconstructed images may be quite convincing, this does not guarantee their correctness since the reconstructed images may actually belong to a different person. Moreover, the non-existence of standardized dataset, training, and evaluation methodologies makes a fair comparison among works difficult, and it is impossible to tell whether any gains in performance have resulted from better architectures or from experimental setup.

6.7. A Unified Analytical Framework for Face Mask Removal

As seen from the comparison above, there is no paradigm that always outperforms other techniques on any criterion. One-stage models usually allow reducing computational cost and inference speed, while two-stage approaches typically ensure better identity preservation due to their structural constraints, and inpainting-based models ensure more flexibility to handle different types of occlusions. From this we can conclude that the effectiveness of a face mask removal algorithm is not determined by its architecture alone, but rather by the nature of the reconstruction problem itself. Based on this synthesis of the existing literature, we propose a unified analytical framework that relates occlusion characteristics, reconstruction difficulty, and application-specific identity preservation requirements. Together, these interacting factors provide a structured basis for understanding existing methods, selecting appropriate reconstruction strategies, and identifying future research directions.

The suggested framework starts with analyzing the occlusion parameters that define the amount and quality of information on the face for reconstruction. The severity of occlusion depends on a number of parameters such as the area size and shape, geometric configuration of the mask, material and texture, and its positioning. All of these

factors define the amount of visible facial data to be used for reconstruction. Regular and small occlusions ensure enough contextual and identifying data for successful reconstruction while large and irregular occlusions cover a significant portion of discriminative facial features thus making reconstruction more uncertain.

The second component of the framework is the reconstruction difficulty, which captures the uncertainty associated with reconstructing the missing facial information. In the suggested framework, the reconstruction difficulty is not considered a separate attribute but is directly a result of the occlusion attributes. With the decreasing availability of facial information, the process of reconstruction becomes increasingly dependent on the learned prior information as opposed to the facial information available from the visible face. While partial occlusions can be recovered by exploiting the local contextual information, in case of extensive occlusions, the model has to recover the facial geometry and appearance, which are no longer available in the input image. This implies that the process of reconstruction involves more uncertainty as the model becomes increasingly dependent on the learned prior information, and thus it is highly likely for the identity to be drifted.

The third component relates to identity preservation and depends on the purpose of the reconstructed face rather than the procedure itself. In case of image enhancement, photography, or restoration of images, the primary attention is paid to the perceptual accuracy of the reconstruction, therefore, allowing some degree of deviation from the true identity. However, in case of biometric identification, surveillance, and forensic investigation, the necessity arises for the reconstructed face to be very close to the true one because any modifications of discriminant features will lead to decreasing the accuracy of the recognition procedure. Consequently, identity preservation should be regarded as an application-dependent design constraint that determines the acceptable level of reconstruction uncertainty.

Such interactions of the above-listed components constitute a foundation for choosing a suitable reconstruction approach. Lightweight one-stage networks would be sufficient when there is a relatively small amount of occlusion and reconstruction complexity is low, while the use of two-stage networks can be considered as the most suitable choice when there is greater difficulty in the reconstruction process and identity preservation becomes an essential requirement due to the fact that such network structure makes reconstruction less ambiguous and improves biometric consistency. Such an inpainting approach would still be appropriate for irregular or non-mask occlusions due to the fact that it is more flexible, but its dependence on surrounding context can affect identity preservation when facial occlusion is severe. Accordingly, the most appropriate reconstruction strategy depends not on architectural sophistication alone but on the combined influence of occlusion characteristics, reconstruction difficulty, and identity preservation requirements.

Overall, the proposed unified analytical framework can be considered complementary to the architectural taxonomy introduced above since it moves the attention from the way the face is reconstructed to the right moment of applying the certain approach to reconstruction. Instead of considering the one-stage, two-stage, and inpainting approaches to facial reconstruction as rivals in terms of the architectural design, the framework can consider them as different approaches which are applicable depending on the relationship between the nature of occlusion, complexity of reconstruction, and specific needs for preserving personal identity.

7. Ethical, Legal, and Societal Implications

Face mask removal and face recognition technology pose many ethical, legal, and social issues that go far beyond the technological framework of the process and the mathematical metrics associated with it. This type of biometric technology uses information that is very personal to the individual and is connected directly to his or her identity and privacy. In this way, the large-scale use of such technology for surveillance, identification, and security purposes poses many problems regarding individual privacy if the technology is not used with proper consent and control. The individual has no input in the collection, storage, and analysis of his or her face images and this leads to an abuse of power and the right to assemble freely and express opinions without feeling like they are under constant surveillance [70].

From the legal standpoint, it may be argued that the biometric data falls into the category of personal information, which is highly secured by different data protection acts like the GDPR (General Data Protection Regulation) in

Europe. However, such regulations does not have global reach. For example, there is a high degree of uncertainty surrounding the creation and application of technologies associated with facial recognition due to regulations that differ in the international community when it comes to the use of such technologies in non-governmental contexts [71].

Algorithmic fairness and bias are two other major ethical considerations. Research indicates that facial recognition algorithms as well as face-unmasking algorithms can display some form of bias associated with demographic factors [72]. This is another question of equity and fairness related to the data used to train such algorithms and how they are deployed.

There have been a number of instances from the real world that highlight the societal implications of facial analysis techniques. According to research carried out at the United States National Institute of Standards and Technology (NIST), there have been cases where many of the commercial facial recognition systems demonstrated much higher percentages of false positives for women, older people, and individuals with darker skin tones compared to white male subjects [73]. Such demographic disparities may lead to disproportionate misidentification and discriminatory outcomes in security and law enforcement applications.

Misuse issues have also become a concern in the context of surveillance. Facial recognition systems have been implemented in public surveillance systems to such an extent that there were concerns over the surveillance of masses, tracking of individuals without their knowledge, and abuse of individual rights [74]. Moreover, face reconstruction and unmasking techniques can work against the very purpose of using facial covers to protect individual privacy, anonymity, or even health. Face mask removal techniques, for example, that are able to reconstruct faces based on incomplete data can be misused in social networking photos as well [75].

The other major issue that needs to be considered is that of biased datasets and consent. Face datasets that are used to train recognition and generative models are frequently obtained from the internet from freely available pictures, where there is no explicit informed consent taken from the people who were captured in the picture [74]. In addition, if the training datasets do not include certain demographics, it will have a detrimental effect on their generalization.

From the above examples, it is clear that there is more to the risk factors associated with face mask removal besides mere technical performance. Hence, there is need for future research to concentrate on areas such as evaluation criteria, bias mitigation techniques, privacy-preserving machine learning, and ethical and legal considerations.

Societally, the acceptance of both face mask removal and face recognition technology can possibly affect society's attitudes towards surveillance and trust. While face recognition technology may be used for safety or during a public health emergency, the technology can also be misused as part of surveillance, misidentification, and racial profiling. It is important to make sure that society is not being exploited by technology.

Researchers working on face mask removal techniques must not only concentrate on improving the generation fidelity and identification results but must also look at the ethical, privacy, and legal aspects of the techniques they are working on. The inclusion of law, ethics, and social sciences would be beneficial, as it would ensure that the future research on face mask removal techniques is aligned with societal norms and expectations.

8. Conclusion and Future Directions

Face mask removal is an emerging research field that attracted much attention due to the difficulties encountered in dealing with face masks that cover some of the most important facial areas and complicate face recognition and expression analysis. as well as other computer vision tasks. Most of the existing works on face mask removal have been performed through GAN-based architectures. Although recent progress has shown that automated face mask removal is feasible, the problem is still not completely solved.

One of the limitations that can be identified among existing studies is the insufficiency of available data. Despite the applicability of the method of synthetic face masking in datasets to train deep networks, it does not capture the complexity of the problem at hand in any form. In contrast, recent datasets, such as [40], have made substantial progress in this matter by providing paired real examples. However, the lack of data containing real masked pair

images is currently hindering the research into realistic face mask removal. Future face mask removal datasets should feature a diverse set of participants in demographic characteristics, mask variations, and environmental conditions. Moreover, all necessary provisions should be included regarding the consent of participants and usage agreement of the data.

Another question to be discussed is that there is currently no single benchmarking technique employed to evaluate face mask removal algorithms. Many studies apply their own splitting approaches, preprocessing techniques, as well as combinations of metrics, this creates additional difficulties in comparing results across different articles. Hence, some of the progress reported might be attributable to differences in methodology and not necessarily in the architectures used.

Reproducibility is still a major concern for the current work done on face mask removal. While many papers detail the network design used and loss function used, reproducibility is difficult to achieve without open-source code, pre-trained models, and testing procedures. Some important considerations such as the settings of the optimization methods, learning rate schedule, data pre-processing pipeline, random seed generation, and hardware specifications are usually not reported. This results in the implemented code, despite being faithful to the published algorithm, performing differently from other independent implementations. To improve reproducibility and facilitate fair comparison across studies, future face mask removal research should adhere to a common set of reporting standards. At a minimum, published studies should provide (1) publicly available source code, (2) pretrained model weights, (3) complete documentation of synthetic mask generation procedures and data preprocessing pipelines, (4) training hyperparameters, optimizer settings, learning-rate schedules, random seeds, software versions, and hardware specifications, and (5) evaluation on standardized benchmark datasets using common identity-aware evaluation protocols. Adoption of these minimum reporting standards would substantially improve reproducibility, independent validation, and meaningful comparison between competing approaches.

Based on the literature overview provided in this paper, it can be argued that future development of the technology of removing masks from images should involve overcoming three main challenges, beginning with the most important one. In particular, realistic pairs of masked and unmasked images need to be created in order to reduce the gap between synthetic and real masked faces. It is unlikely that an increase in the volume of images in the dataset would help solve the problem without proper data diversity. Secondly, there should be a set of procedures for evaluation in the community where reconstruction quality is not the only thing to be considered, but also identity preservation and face recognition as well. Without common evaluation criteria, it would be hard to understand whether any improvements were achieved thanks to architecture or other factors. Third, future architectural progress should be concerned with generalization, not just reconstruction quality, because existing algorithms fail at preserving identity in challenging situations.

In future face mask removal systems, it is recommended that approaches shift away from just pixel similarity optimization to identity-based reconstruction. Future systems should not only optimize for PSNR and SSIM but should instead optimize on facial identity, semantics, and perceptual quality at the same time. A particularly promising direction is task-oriented face mask removal, in which reconstruction is optimized jointly with downstream objectives such as face recognition or expression analysis rather than treated as an isolated image generation problem. Such end-to-end optimization may reduce the discrepancy between visually plausible reconstructions and representations that are useful for biometric analysis.

Future improvements will most probably not rely on one particular architecture. Rather, hybrid models leveraging the strengths of efficient GAN reconstructions, transformer context modeling, diffusion, and identity-preserving tasks could provide a better trade-off between reconstruction performance and efficiency.

The effectiveness of face mask removal is ultimately measured based on how well it performs in further processes like face recognition, face expression detection, biometric authentication, and human-computer interaction.

As the field evolves, it is anticipated that the major difficulty would transition from enhancing the accuracy of the reconstruction process to creating models that can generalize well on unseen mask types, demographics, and imaging settings.

Finally, the issue of ethics should never be overlooked, especially in view of the possibility of de-anonymization. Although face mask removal can be used in the fields of security, healthcare, and human-computer interaction, it

is important that measures be put in place to ensure that it is not misused. In general, the future of this area will depend not only on innovation but also on ethics.

Acknowledgement

This work was supported by the CNRST under the PhD-Associate Scholarship (PASS) Program in the field of digital technology – Horizon 2027, Morocco National Strategy 2030.

Conflict of interest

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

1. Govind Jeevan, Geevar C. Zacharias, Madhu S. Nair, and Jeny Rajan. An empirical study of the impact of masks on face recognition. *Pattern Recognition*, 122:108308, 2022.
2. Michael Christian Leitner, Verena Meurer, Florian Hutzler, Sarah Schuster, and Stefan Hawelka. The effect of masks on the recognition of facial expressions: A true-to-life study on the perception of basic emotions. *Frontiers in psychology*, 13:933438, 2022.
3. Taiki Arakane, Chihiro Kai, and Takeshi Saitoh. Can you read lips with a masked face? In *2023 18th International Conference on Machine Vision and Applications (MVA)*, pages 1–5. IEEE, 2023.
4. Ahmad Alzu'bi, Firas Albalas, Tawfik AL-Hadhrani, Lojin Bani Younis, and Amjad Bashayreh. Masked face recognition using deep learning: A review. *Electronics*, 10(21), 2021.
5. Bilal Saoud, Abdul Hakim H. M. Mohamed, Ibraheem Shaya, Ayman A. El-Saleh, and Abdulaziz Alashbi. Review of masked face recognition based on deep learning. *Technologies*, 13(7), 2025.
6. Ali Haitham Abdul Amir and Zainab N Nemer. Inclusive review on advances in masked human face recognition technologies. *arXiv preprint arXiv:2508.00841*, 2025.
7. Mohamed Mahmoud, Mahmoud SalahEldin Kasem, and Hyun-Soo Kang. A comprehensive survey of masked faces: Recognition, detection, and unmasking. *Applied Sciences*, 14(19), 2024.
8. Yazid Safiny, Siham Ahmam, Nidal Lamghari, and Abdelghani Ghazdali. Face mask removal: Datasets, methods, and evaluation metrics. In Nouredine Aboutabit and Imad Hafidi, editors, *Advances in Machine Intelligence and Computer Science Applications*, pages 133–145, Cham, 2026. Springer Nature Switzerland.
9. Omar Elharrouss, Noor Almaadeed, Somaya Al-Maadeed, and Younes Akbari. Image inpainting: A review: O. elharrouss et al. *Neural Processing Letters*, 51(2):2007–2028, 2020.
10. Mehamdia Abd Elhamid, Issam AR Moghrabi, Basim A Hassan, and Nivine Guler. An improved hybrid nonlinear conjugate gradient method and application to image restoration problems. *Statistics, Optimization & Information Computing*, 16(1):1–14, 2026.
11. Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
12. Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *European conference on computer vision*, pages 89–105. Springer, 2018.
13. Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.
14. Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.
15. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
16. Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1833–1844, 2021.
17. Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
18. Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
19. Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in neural information processing systems*, 35:23593–23606, 2022.
20. Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022.

21. Jixin Zhao, Zhouxia Wang, Peiqing Yang, and Shangchen Zhou. Precise object and effect removal with adaptive target-aware attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19370–19379, 2026.
22. Mengting Wei, Tuomas Varanka, Yante Li, Xingxun Jiang, Huai-Qian Khor, and Guoying Zhao. Towards consistent and controllable image synthesis for face editing. *arXiv preprint arXiv:2502.02465*, 2025.
23. Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3911–3919, 2017.
24. Yu Wang, Gencheng Wang, Rong Chen, and Yuzhen Chen. An efficient diffusion for blind face restoration. In *Proceedings of the 2024 7th International Conference on Image and Graphics Processing, ICIGP '24*, page 340–347, New York, NY, USA, 2024. Association for Computing Machinery.
25. Klemen Grm, Walter J Scheirer, and Vitomir Štruc. Face hallucination using cascaded super-resolution and identity priors. *IEEE Transactions on Image Processing*, 29:2150–2165, 2019.
26. Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thirtieth-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003, volume 2, pages 1398–1402 Vol.2, 2003.
27. Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.
28. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
29. Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.
30. Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.
31. Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
32. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
33. Aqeel Anwar and Arijit Raychowdhury. Masked face recognition for secure authentication. *arXiv preprint arXiv:2008.11104*, 2020.
34. dlib. <http://dlib.net/>.
35. Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.
36. Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arxiv 2017. arXiv preprint arXiv:1710.10196*, pages 1–26, 2018.
37. Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition*, 2008.
38. Tero Karras. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2019.
39. Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.
40. Indah Siradjuddin, Muhammad Ilham Zakaria, and Mohammad Habibul Akhyar. A masked and unmasked face dataset, 2024.
41. Prasoon Kottarathil. Face mask lite dataset. <https://www.kaggle.com/prasoonkottarathil/face-mask-lite-dataset>, 2020.
42. Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.
43. Saddam Sinatrya Jalu Mukti. Unmasking face mask. <https://www.kaggle.com/datasets/saddamazzyzy/unmasking-face-mask?select=train>.
44. Satyam Dubey and Prathmesh Pawar. Unmask-my-celeb-a-hq. <https://www.kaggle.com/datasets/prosad123/unmask-my-celeb-a-hq>.
45. Apoorva Shete, Rachael Nihalaani, and Amit Hatekar. Approach to prediction of unmasked face from masked face using deep learning. *International Journal of Computer Applications*, 975:8887.
46. Md Imran Hosen and Md Baharul Islam. Masked face inpainting through residual attention unet. In *2022 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–5. IEEE, 2022.
47. Yefan Jiang, Fan Yang, Zhangxing Bian, Changsheng Lu, and Siyu Xia. Mask removal: Face inpainting via attributes. *Multimedia Tools and Applications*, 81(21):29785–29797, 2022.
48. Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
49. Hitoshi Yoshitani, Naoto Ienaga, and Maki Sugimoto. Gan-based face mask removal using facial landmarks and pixel errors in masked region. In *VISIGRAPP (5: VISAPP)*, pages 125–133, 2022.
50. Akhil Kumar, Manisha Kaushal, and Akashdeep Sharma. Sam c-gan: a method for removal of face masks from masked faces. *Signal, Image and Video Processing*, 17(7):3749–3757, 2023.
51. Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
52. Mehdi Mirza. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.
53. Rajesh Kumar. Bollywood celebrity faces. <https://www.kaggle.com/datasets/havingfun/100-bollywood-celebrity-faces>.
54. Nizam Ud Din, Kamran Javed, Seho Bae, and Juneho Yi. A novel gan-based network for unmasking of masked face. *IEEE Access*, 8:44276–44287, 2020.
55. Minmin Yang. Face mask removal with region-attentive face inpainting. *arXiv preprint arXiv:2409.06845*, 2024.

56. Mohamed Mahmoud and Hyun-Soo Kang. Ganmasker: A two-stage generative adversarial network for high-quality face mask removal. *Sensors*, 23(16), 2023.
57. Mohamed Elhoseny, Ahmed Hassan, Marwa H. Shehata, and Mohammed Kayed. Advanced deep learning for masked individual surveillance. *International Journal of Cognitive Computing in Engineering*, 5:406–415, 2024.
58. Guangwei Gao, Lei Tang, Fei Wu, Huimin Lu, and Jian Yang. Jdsr-gan: Constructing an efficient joint learning network for masked face super-resolution. *IEEE Transactions on Multimedia*, 25:1505–1512, 2023.
59. Xiangnan Yin, Di Huang, and Liming Chen. Non-deterministic face mask removal based on 3d priors. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2137–2141. IEEE, 2022.
60. Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488, 2018.
61. Nizam Ud Din, Kamran Javed, Seho Bae, and Juneho Yi. Effective removal of user-selected foreground object from facial images using a novel gan-based network. *IEEE Access*, 8:109648–109661, 2020.
62. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
63. Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
64. Jingxia Lu, Xianxu Hou, Hao Li, Zhibin Peng, Linlin Shen, and Lixin Fan. Maskdiffuse: Text-guided face mask removal based on diffusion models. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 435–446. Springer, 2023.
65. Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.
66. Zongsheng Yue and Chen Change Loy. Difface: Blind face restoration with diffused error contraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9991–10004, 2024.
67. Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5718–5729. IEEE, 2022.
68. Zohair Elmourabit and Oumayma Banouar. Gan based approaches for self-supervised segmentation: A comparative study. *Statistics, optimization & information computing*, 12(3):646–659, 2024.
69. Augustus Odena. Semi-supervised learning with generative adversarial networks. *arXiv preprint arXiv:1606.01583*, 2016.
70. IEEE Public Safety Technology. Ethical considerations in the use of facial recognition for public safety. Accessed: 2026-02-02.
71. Xukang Wang, Ying Cheng Wu, Mengjie Zhou, and Hongpeng Fu. Beyond surveillance: privacy, ethics, and regulations in face recognition technology. *Frontiers in big data*, 7:1337465, 2024.
72. Zhifang Sun and Zhe Liu. Ensuring privacy in face recognition: a survey on data generation, inference and storage. *Discover Applied Sciences*, 7(5):441, 2025.
73. P. Grother, M. Ngan, K. Hanaoka, National Institute of Standards, and Technology (U.S.). *Face Recognition Vendor Test (FVRT): Part 3, Demographic Effects*. NISTIR; NIST IR; NIST interagency report; NIST internal report. National Institute of Standards and Technology, 2019.
74. Lucas Introna and Helen Nissenbaum. Facial recognition technology: A survey of policy and implementation issues. 07 2009.
75. Guangcan Mai, Kai Cao, Pong C. Yuen, and Anil K. Jain. On the reconstruction of face images from deep face templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(5):1188–1202, 2019.