Robust Model-based Face Reconstruction through Weakly-Supervised Outlier Segmentation

Chunlu Li[1,2] Andreas Morel-Forster [2] Thomas Vetter [2] Bernhard Egger[3,*] Adam Kortylewski[4,5,*]
lcl@seu.edu.cn bernhard.egger@fau.de akortyle@mpi-inf.mpg.de
[1] School of Automation, Southeast University [2] Department of Mathematics and Computer Science, University of Basel
[3] Friedrich-Alexander-Universität Erlangen-Nürnberg [4]University of Freiburg [5]Max Planck Institute for Informatics

Abstract

In this work, we aim to enhance model-based face reconstruction by avoiding fitting the model to outliers, i.e. regions that cannot be well-expressed by the model such as occluders or make-up. The core challenge for localizing outliers is that they are highly variable and difficult to annotate. To overcome this challenging problem, we introduce a joint Face-autoencoder and outlier segmentation approach (FOCUS). In particular, we exploit the fact that the outliers cannot be fitted well by the face model and hence can be localized well given a high-quality model fitting. The main challenge is that the model fitting and the outlier segmentation are mutually dependent on each other, and need to be inferred jointly. We resolve this chicken-and-egg problem with an EM-type training strategy, where a face autoencoder is trained jointly with an outlier segmentation network. This leads to a synergistic effect, in which the segmentation network prevents the face encoder from fitting to the outliers, enhancing the reconstruction quality. The improved 3D face reconstruction, in turn, enables the segmentation network to better predict the outliers. To resolve the ambiguity between outliers and regions that are difficult to fit, such as eyebrows, we build a statistical prior from synthetic data that measures the systematic bias in model fitting. Experiments on the NoW testset demonstrate that FOCUS achieves SOTA 3D face reconstruction performance among all baselines trained without 3D annotation. Moreover, our results on CelebA-HQ and AR database show that the segmentation network can localize occluders accurately despite being trained without any segmentation annotation.

Figure 1. FOCUS conducts face reconstruction and outlier segmentation jointly under weak supervision. Top to bottom: target images, our reconstruction images, and estimated outlier masks.

1. Introduction

Monocular 3D face reconstruction aims at estimating the pose, shape, and albedo of a face, as well as the illumination conditions and camera parameters of the scene. Solving for all these factors from a single image is an ill-posed problem. Model-based face autoencoders [31] overcome this problem through fitting a 3D Morphable Model (3DMM) [1, 9] to a target image. The 3DMM provides prior knowledge about the face albedo and geometry such that 3D face reconstruction from a single image becomes feasible, enabling face autoencoders to set the current state-of-the-art in 3D face reconstruction [5]. The network architectures in the face autoencoders are devised to enable end-to-end reconstruction and to enhance the speed compared to optimization-based alternatives [19, 41], and sophisticated losses are designed to stabilize the training and to get better performance [5].

A major remaining challenge for face autoencoders is that their performance in in-the-wild environments is still limited by nuisance factors such as model outliers, extreme illumination, and poses. Among those nuisances, model outliers are ubiquitous and inherently difficult to handle because of their wide variety in shape, appearance, and loca-

* Denotes same contribution.
Codes available at: github.com/unibas-gravis/Occlusion-Robust-MoFA
C.Li is funded by the China Scholarship Council (CSC) from the Ministry of Education of P.R. China. B.Egger was supported by a Post-Doc Mobility Grant, Swiss National Science Foundation P400P2_191110. A.Kortylewski acknowledges support via his Emmy Noether Research Group funded by the German Science Foundation (DFG) under Grant No. 468670075. Sincere gratitude to Tatsuro Koizumi and William A. P. Smith who offered the MoFA re-implementation.

tion. The outliers are a combination of the occlusions that do not belong to the face and the mismatches which are the facial parts but cannot be depicted well by the face model, such as pigmentation and makeup on the texture and wrinkles on the shape. Fitting to the outliers often distorts the prediction (see Fig. 3) and fitting to the mismatches cannot improve the fitting further due to the limitation of the model. Therefore we propose to only fit the inliers, i.e. the target with the outliers excluded.

To prevent distortion caused by model outliers, existing methods often follow a bottom-up approach. For example, a multi-view shape consistency loss is used as prior to regularize the shape variation of the same face in different images [5, 10, 33], or the face symmetry is used to detect occluders [34]. Training the face encoder with dense landmark supervision also imposes strong regularization [37, 42], while pairs of realistic images and meshes are costly to acquire. Most existing methods apply face [27] or skin [5] segmentation models and subsequently exclude the non-facial regions during reconstruction. These segmentation methods operate in a supervised manner, which suffers from the high cost and efforts for acquiring a great variety of occlusion annotations from in-the-wild images.

In this work, we introduce an approach to handle outliers for model-based face reconstruction that is highly robust, without requiring any annotations for skin, occlusions, or dense landmarks. In particular, we propose to train a FaceautOenCoder and oUtlier Segmentation network, abbreviated as FOCUS, in a cooperative manner. To train the segmentation network in an unsupervised manner, we exploit the fact that the outliers cannot be well-expressed by the face model to guide the decision-making process of an outlier segmentation network. Specifically, the discrepancy between the target image and the rendered face image (Fig. 1 1st and 2nd rows) are evaluated by several losses that can serve as a supervision signal by preserving the similarities among the target image, the reconstructed image, and the reconstructed image under the estimated outlier mask.

The training process follows the core idea of the Expectation-Maximization (EM) algorithm, by alternating between training the face autoencoder given the currently estimated segmentation mask, and subsequently training the segmentation network based on the current face reconstruction. The EM-like training strategy resolves the chicken-and-egg problem that the outlier segmentation and model fitting are dependent on each other. This leads to a synergistic effect, in which the outlier segmentation first guides the face autoencoder to fit image regions that are easy to classify as face regions. The improved face fitting, in turn, enables the segmentation model to refine its prediction.

We define in-domain misfits as errors in regions, where a fixed model can explain but constantly not fit well, which are observed in the eyebrows and the lip region. We assume that such misfits result from the deficiencies of the fitting pipeline. Model-based face autoencoders use image-level losses only, which are highly non-convex and suffer from local optima. Consequently, it is difficult to converge to the globally optimal solution. In this work, we propose to measure and adjust the in-domain misfits with a statistical prior. Our misfit prior learns from synthetic data at which regions these systematic errors occur on average. Subsequently, the learnt prior can be used to counteract these errors for predictions on real data, especially when our FOCUS structure excludes the outliers. Building the prior requires only data generated by a linear face model without any enhancement and no further improvement in landmark detection.

We demonstrate the effectiveness of our proposed pipeline by conducting experiments on the NoW testset [29], where we achieve state-of-the-art performance among model-based 3D face methods without 3D supervision. Remarkably, experiments on the CelebA-HQ dataset [20] and the AR database [22] validate that our method is able to predict accurate occlusion masks without requiring any supervision during training.

In summary, we make the following contributions:

1. We introduce an approach for model-based 3D face reconstruction that is highly robust, without requiring any human skin or occlusion annotation.

2. We propose to compensate for the misfits with an in-domain statistical misfit prior, which is easy to implement and benefits the reconstruction.

3. Our model achieves SOTA performance at self-supervised 3D face reconstruction and provides accurate estimates of the facial occlusion masks on in-the-wild images.

2. Related Work

Model-based face autoencoders [31] solve the 3D face reconstruction by fitting a face model to the target image with an encoder and a decoder containing the 3DMM and a renderer. Typically, the encoder first estimates parameters from a target image, including the shape, texture, and pose of the target, and the illumination and camera settings from the scene. Then the renderer synthesizes a 2D image using the estimated parameters with an illumination model and a projection function. The face is reconstructed by retrieving the parameters which result in a synthesized image most similar to the target image. The 3DMM [1] plays a paramount role in the face autoencoders, because it parameterizes the latent distribution space of faces, and therefore can connect the encoder with the renderer and enable end-to-end training. The model-based face autoencoders have been proven effective in improving the reconstruction.

They simplify the optimization step and enhance the reconstruction speed [32], improve the details of shape and texture [10, 12, 24, 34, 35], and can also reconstruct more discriminative features [5, 13].

Despite the advantages of face autoencoders, their performance is still limited by outliers. To solve the distortions caused by outliers, some early methods [25] resort to robust fitting losses, but they are not robust to illumination variations and appearance variations in eye and mouth regions. In recent years, shape consistency losses have been used as prior to constrain the face shape across images of the same subject [5, 10, 29, 33]. The variation of identity features of the 3D shape is restricted so that the fitting remains robust even in unconstrained environments. However, such methods usually need identity labels and do not promise robust texture reconstruction. Besides, many methods conduct face segmentation before reconstruction to lead the model to avoid fitting outliers. For example, a random forest detector for hair is proposed [23] to avoid fitting the hair region, and a semantic segmentation network is trained to better locate the face region [27]. A skin detector is employed to impose different weights on the pixels during reconstruction to guide the network to put more attention on the skin-colored regions [5]. However, skin-colored occlusion can not be distinguished correctly and the detector is sensitive to illumination. Yildirim *et al*. propose to explicitly model the 3D shapes and cast shadows of certain types of occlusions, in order to decompose the target into face regions and occlusions to exclude occlusions during fitting [39]. However, the types of occlusions are limited. Generally, these off-the-shelf segmentation models require labelled data for training. Although synthesized images can be used for training, there is a domain gap between the real images and the synthesized ones [18]. Unlike these methods, we merge the segmentation procedure into a model-based face autoencoder, which exploits the face model prior, and consequently does not require additional supervision.

The most relevant method to ours is proposed by Egger *et al*. [8]. They jointly adapt a face model to a target image and segment the target image into face, beard, and occlusion, and the segmentation models are trained with an EM-like algorithm, where different models for beard, foreground, and background are optimized in alternating steps. Compared to our method, their model can only be inferred instance-wise, and the inferred likelihood models for one instance cannot be used for another, while ours follows the learning paradigm, which empowers the segmentation model with greater generalization ability and speeds up our inference procedure. Besides, building the beard model in [8] requires annotations for the beard, while our model is completely label-free. Additionally, their beard model and background likelihood model are statistical models based on simple color histogram, but our model is guided by perceptual-level losses measuring fitting quality, which are intuitive and much easier to implement. De Smet *et al*. also propose to conduct face model fitting and occlusion segmentation jointly [3], but they estimate the occlusions based on an appearance distribution model, which is sensitive to illumination variation and many other subtle changes in appearance. Maninchedda *et al*. propose to solve face reconstruction and segmentation in a joint manner [21], but depth maps are required as supervision. In comparison, our FOCUS model learns from only weak supervision.The face autoencoder also enables us to adapt the face model more efficiently. In addition, we integrate perceptual losses which enable the segmentation network to reason over semantic features instead of only over independent pixels, which increases the robustness to illumination and other factors.

The in-domain misfits indicate systematic uncertainty and deficiencies in the fitting pipeline. Instead of improving each part of the pipeline individually, we propose to build a statistical prior that measures and adjusts the bias introduced by the pipeline.Our solution does not require any further improvements on the system such as landmark detection, model-landmark correspondence, or more supervision.

3. Approach

In this section, we introduce a neural network-based pipeline, FOCUS, that conducts 3D face reconstruction and outlier segmentation jointly. We first discuss our proposed pipeline architecture (Sec. 3.1) and then the EM-type training without any supervision regarding the outliers (Sec. 3.2). In Sec. 3.3 we show the unsupervised EM initialization. Finally, we show how to compensate for the systematic in-domain misfits with a statistical prior (Sec. 3.4)

3.1. Network Architecture

Our goal is to robustly reconstruct the 3D face from a single target image I_T with outliers, even severe occlusion. To solve this challenging problem, we integrate a model-based face autoencoder, R, with a segmentation network, S, and create synergy between them, as demonstrated in Fig. 2. For face reconstruction, the segmentation mask cuts the estimated outliers out during fitting, improving reconstruction robustness. For segmentation, the reconstructed result provides a reference, enhancing the segmentation accuracy. In this section, we explain how the two networks are connected together and how they benefit each other.

The model-based face autoencoder, R, is expected to reconstruct the complete face appearance from the visible face regions in the target image, I_T. It consists of an encoder and a computer graphics renderer as its decoder. The encoder estimates the latent parameters $\theta = [\alpha, \gamma, \phi, c] \in \mathbb{R}^{257}$, i.e. the 3D shape $\alpha \in \mathbb{R}^{144}$ and texture $\gamma \in \mathbb{R}^{80}$ of a 3DMM, as well as the illumination $\phi \in \mathbb{R}^{27}$ and camera parameters $c \in \mathbb{R}^{6}$ of the scene. Given the latent parameters,

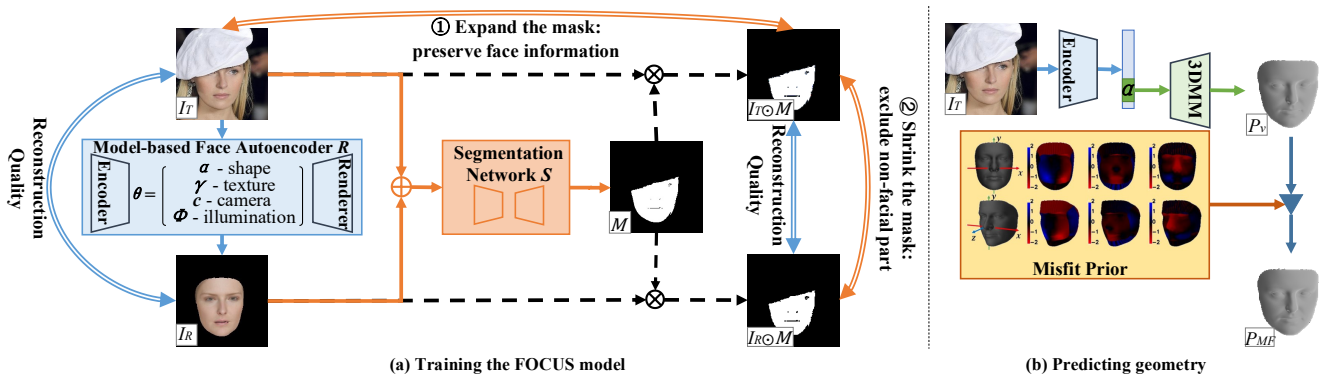


Figure 2. Overview of our method. The solid single lines show the forward path. (a) Given a target image I_T, the reconstruction network, R, estimates the latent parameters and subsequently renders an image I_R, containing only the face. Then, I_T and I_R are stacked and fed into the segmentation network, S, which predicts the mask M. The dashed lines show that M is used to mask out the estimated outliers in I_T and I_R to get assembly outlier-free images, namely $I_T \odot M$ and $I_R \odot M$. The double-lined arrows indicate the compared image pairs in the losses for S (orange) and losses for R (blue), as stated in Sec. 3.2. By training alternatively the two networks and exploiting the synergy between the segmentation and the reconstruction tasks, the proposed FOCUS pipeline is capable of both reconstructing faces even under severe occlusions robustly and conducting face segmentation. (b) Predicting the face geometry P_{mf} requires a single forward. The ▼ is a simple subtraction operation as introduced in Sec. 3.4.

Figure 3. In the presence of outliers, FOCUS reconstructs faces more faithfully than previous model-based face autoencoders. The images from top to bottom are: target images, results of the MoFA network [31], and our results.

the decoder renders a predicted face image $I_R = \mathbb{R}(\theta)$.

Standard face autoencoders [31] fit the face model parameters, regardless of whether the underlying pixels depict a face or occlusion. Consequently, the face model is distorted by the outliers, as shown in the second row in Fig. 3, it is obvious that the illumination, appearance, and shape are estimated incorrectly. To resolve this fundamental problem of face autoencoders, we introduce an unsupervised segmentation network, whose output can be used to mask the outliers out during model fitting and therefore make the autoencoder robust to outliers.

The segmentation network, S, takes the stacked target image I_T and the synthesized image I_R as input and predicts a binary mask, $M = S(I_T, I_R)$, to describe whether a pixel depicts the face (1) or outliers (0). Since I_R contains the estimated intact face, it provides the segmentation network with prior knowledge and helps the estimation.

The face autoencoder and the segmentation network are coupled together during training to induce a synergistic effect which makes the segmentation more accurate and reconstruction more robust under outliers, as shown in the last row in Fig. 3. Sec. 3.2 describes how the pipeline can be trained end-to-end, despite the entanglement between the two networks, and the high-level losses that relieve our pipeline of any occlusion or skin annotation.

3.2. EM-type Training

Due to the mutual dependencies between the face autoencoder and the segmentation network, we conduct an Expectation-Maximization (EM) like strategy, where we train the two networks in an alternating manner. This enables a stable convergence of the model training process. Similar to other EM-type training strategies, our training process starts from a rough initialization of the model parameters which is obtained in an unsupervised manner (as described in Sec. 3.3). We then optimize the two networks in an alternating manner, as described in the following.

Training the segmentation network. When training the segmentation network, the parameters of the face autoencoder are fixed and only the segmentation network is optimized. Instead of hunting for labelled data, we propose four losses enforcing intrinsic similarities among the images. Each loss works to either include pixels indicating face or the opposite. Since the proposed losses have overlapped or opposite functions to each other, only by reaching a balance among these losses can the network yield good segmentation results. The losses work either on the perceptual level or the pixel level, to fully exploit the visual clues. The perceptual-level losses compare the intermediate features of

two images extracted by a pretrained face recognition model F. We use the cosine distance, $cos(X,Y) = 1 - \frac{X \cdot Y}{\|X\|\|Y\|}$, to compute the distance between the features. Perceptual losses are common for training face autoencoders, which encourage encoding facial details that are important for face recognition [10]. We found that the perceptual losses also benefit segmentation (see Sec. 4.4).

The proposed losses are as follows:

$$L_{nbr} = \sum_{x \in \Omega} \left\| \min_{\forall x' \in \mathcal{N}(x)} \|I_T(x) - I_R(x')\| \right\|_2^2 \quad (1)$$

$$L_{dist} = cos(F(I_T \odot M), F(I_R \odot M)) \quad (2)$$

$$L_{area} = -S_M / S_R \quad (3)$$

$$L_{presv} = cos(F(I_T \odot M), F(I_T)) \quad (4)$$

The pixel-level neighbor loss in Eq. (1), L_{nbr}, compares a pixel, $I_T(x)$, at location x on the target image, with the pixels on the rendered image in the neighboring region, $\mathcal{N}(x)$ of this pixel, so that this loss is stable even if there are small misalignments. Note that it only accounts within the face region, Ω, predicted by the segmentation network. A higher neighbor loss at x indicates that this pixel is not fitted well and is more likely to be an outlier.

Similarly, L_{dist} in Eq. (2) is introduced to compare the target and reconstructed face at perceptual level. Eqs. (1) and (2) aim at shrinking the mask on the outliers, where the pixel-level and perceptual differences are large. Without any other constraints, the segmentation network would output an all-zero mask to make them both 0. On the contrary, once there is a force to encourage the network to preserve some image parts, parts with smaller losses are more likely to be preserved, which in fact are the ones well-explained by the face model and is much more likely to depict face.

Therefore, Eqs. (3) and (4) are proposed to counterwork Eqs. (1) and (2). Eq. (3) is an area loss, L_{area} that enlarges the ratio between the number of estimated facial pixels, S_M, and the number of pixels in the rendered face region, S_R. It prevents the segmentation network from discarding too many pixels. L_{presv} (Eq. (4)), ensures that the perceptual face features remain similar after the outliers in the target image are masked out and encourages the model to preserve as much of the visible face region as possible. Likewise, the network would keep the most-likely face region to decrease Eqs. (3) and (4) in the presence of Eqs. (1) and (2).

We use an additional regularization term, $L_{bin} = -\sum_x (M(x) - 0.5)^2$, to encourage the face mask M to be binary. The total loss for training the segmentation network is: $L_S = \eta_1 L_{nbr} + \eta_2 L_{dist} + \eta_3 L_{area} + \eta_4 L_{presv} + \eta_5 L_{bin}$, with $\eta_1 = 15$, $\eta_2 = 3$, $\eta_3 = 0.5$, and $\eta_4 = 2.5$, and $\eta_5 = 10$. Analysis of the influence of the hyper-parameters is provided in the supplementary material.

During training, the segmentation network is guided seeking a balance between discarding pixels that cannot be explained well by the face model and preserving pixels that are important to retain the perceptual representations of the target and rendered face images. Therefore no supervision for skin or occlusions is required.

Training the face autoencoder. In the second step, we continue to optimize the parameters of the face autoencoder, while keeping the segmentation network fixed. The losses for training the face autoencoder include:

$$L_{pixel} = \left\| (I_T - I_R) \odot M \right\|_2^2 \quad (5)$$

$$L_{per} = cos(F(I_T), F(I_R)) \quad (6)$$

$$L_{lm} = \left\| lm_T - lm_R \right\|_2^2 \quad (7)$$

Above are two reconstruction losses: L_{pixel} (Eq. (5)) at the image level and L_{per} (Eq. (6)) at the perceptual level, and a landmark loss (Eq. (7)) used to estimate the pose, where lm_T and lm_R stand for the 2D landmark coordinates on I_T and I_R, respectively [5]. We set the weights for the landmarks on the nose ridge and inner lip as 20, and the rest as 1. A regularization term is also required for the 3DMM: $L_{reg} = \left\| \theta \right\|_2^2$. To sum up, the loss for training the face autoencoder can be represented as:

$$L_R = \lambda_1 L_{pixel} + \lambda_2 L_{per} + \lambda_3 L_{lm} + \lambda_4 L_{reg} \quad (8)$$

, where $\lambda_1 = 0.5$, $\lambda_2 = 0.25$, $\lambda_3 = 5e-4$, and $\lambda_4 = 0.1$.

3.3. Unsupervised Initialization

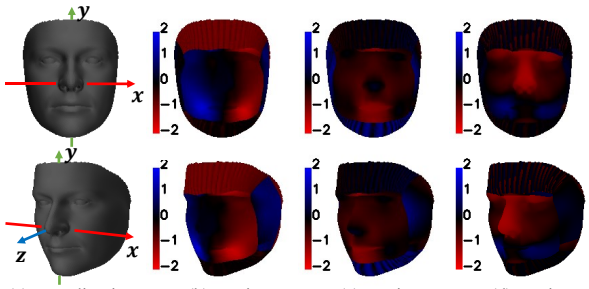
As every other EM-type training strategy, our training needs to be roughly initialized. To achieve unsupervised initialization, we generate preliminary masks using an outlier robust loss [8]:

$$log(P_{face}(x)) = -\frac{1}{2\sigma^2}(I_T(x) - I_R(x))^2 + N_c \quad (9)$$

$$M_{pre}(x) = \begin{cases} 1 & \text{if } (I_T(x) - I_R(x))^2 < \xi \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

We assume that the pixel-wise error at pixel x in the face regions follows a zero-mean Gaussian distribution. Therefore, we can express the log-likelihood that a pixel belongs to the face regions as $log(P_{face})$ (Eq. (9)), where σ and N_c are constant. We also assume that the values of the non-face pixels follow a uniform distribution, i.e., $log(P_{non-face})$ is a constant. Finally, a pixel at position x is classified as face or non-face by comparing the log-likelihoods. This reduces to thresholding of the reconstruction error with a constant parameter ξ (Eq. (10)). When ξ increases, the initialized masks allow the pixels on the target image to have a larger difference from the reconstructed pixels and encourage the reconstruction network to fit these pixels. Empirically, we found that $\xi = 0.17$ leads to a good enough initialization.

To initialize the face autoencoder, the preliminary mask, $M_{pre}(x)$, is obtained in the forward pass using Eq. (10), after the reconstructed face is rendered. Then $M_{pre}(x)$ is used



(a) Coordination (b) x axis (c) y axis (d) z axis

Figure 4. Visualization of the misfit prior. We provide two views of the coordinate system in (a) and the misfits along x, y, and z axis in (b), (c), and (d), respectively.

to mask out the roughly-estimated outliers as in Eq. (5), preventing the face autoencoder from fitting to any possible outliers. Subsequently, the segmentation network is pre-trained using these preliminary masks as ground truths.

3.4. Solving misfits

The misfits in image regions that the model can explain yet not fitted well indicate systematic errors in the fitting pipeline. We propose an in-domain Misfit Prior (abbreviated as MP), E_{MP}, to measure and adjust such misfits.

To build the in-domain prior, we first synthesize images using the face model where theoretically every facial part should be fitted well. Hence, any systematic error is due to systematic deficiencies. We draw random vectors θ_i in the face model latent space to generate ground truth (GT) T_{vi} geometry and texture. Then the renderer in the face autoencoder is employed to render target images, T_i, $i = 1, ..., N$.

A face autoencoder, R_{syn}, with the same structure as R is then trained on the synthesized images using losses Eq. (8), except that no segmentation mask is required in pixel-wise loss, resulting in: $L_{pixel} = \left\| I_T - I_R \right\|_2^2$, since there are no outliers on the synthesized images.

We built the statistical prior as the average error of the vertex-wise reconstruction deviation from the predicted geometry $P_{vi} \in \mathbb{R}^{px3}$ to the GT geometry T_{vi}, where p is the number of vertices:

$$E_{MP} = \frac{1}{N} \sum_{i \in [1,N]} (P_{vi} - T_{vi}) \tag{11}$$

This prior visualized in Fig. 4 shows per-vertex bias introduced by the fitting pipeline. After inference, it could be used to adjust the in-domain misfits and the corrected prediction is $P_{mfi} = P_{vi} - E_{MP}$.

4. Experiments

In this section, results of systematic experiments show that our weakly-supervised method reaches the state-of-the-art face shape reconstruction accuracy and competitive occlusion segmentation results compared to the state-of-the-art self-supervised pipelines and methods that use full supervision in terms of skin or occlusion labels. Our ablation study shows the effectiveness of the segmentation network, our proposed losses, and the misfit prior. More detailed analysis can be found in the supplementary material.

4.1. Experiment setting

Our face encoder shares the structure of the ResNet 50 [16] and uses the Basel Face Model (BFM) 2017 [14] as the 3D face model, with the differentiable renderer proposed in [17]. The segmentation network follows the UNet architecture [26]. Our FOCUS pipeline is trained on the CelebA-HQ trainset [20], following their protocol. Facial landmarks are detected using the method of [2], and images are pre-processed in the same way as [5]. The perceptual features are extracted by the pre-trained ArcFace [4]. More details can be found in the supplementary materials. Note that there is no fine-tuning with 3D data on any of the testsets in our experiments.

Baselines. We compare our method with two SOTA self-supervised model-based face autoencoders, i.e. the MoFA [31] and the Deep3D [5]. Additionally, to achieve a fair comparison between our proposed weakly-supervised

Figure 5. Qualitative comparison on the reconstruction and segmentation results of the Deep3D [5] network (the 2nd and 3rd rows), the MoFA network [31] (the 4th row), and our FOCUS (the last two rows) on occluded faces from the CelebA-HQ testset (the first 8 columns) and the AR database (the last 2 columns). Note that all the masks are binarized.

Table 1. RMSE on the CelebA-HQ testsets and the AR testset.

Testset	MoFA [31]	Backbone-Supervised	Backbone-cutmix	Backbone-cutout	Deep3D [5]	FOCUS (ours)
CelebA-Unoccluded	8.77 ± 0.40	8.71 ± 0.38	8.75 ± 0.39	8.72 ± 0.40	8.49 ± 0.39	$\mathbf{8.38 \pm 0.42}$
CelebA-Occluded	9.20 ± 0.45	9.01 ± 0.45	9.04 ± 0.44	8.99 ± 0.45	8.79 ± 0.45	$\mathbf{8.71 \pm 0.48}$
CelebA-Overall	8.99 ± 0.47	8.86 ± 0.44	8.90 ± 0.44	8.85 ± 0.45	8.64 ± 0.44	$\mathbf{8.55 \pm 0.48}$
AR-Overall	9.53 ± 0.33	9.34 ± 0.33	9.33 ± 0.32	9.28 ± 0.31	9.11 ± 0.37	$\mathbf{8.93 \pm 0.35}$

method and fully-supervised counterparts, we train our backbone reconstruction network in supervised settings using the GT masks provided by the CelebA-HQ database to exclude occlusions during training. The GT masks of the CelebA-HQ database are the merge of their manually labelled masks for skin, hair, accessories, and so on. Two data augmentation methods for occlusion handling, i.e. the cutmix [40] and cutout [6], are also implemented to enhance the performance of the supervised pipelines. We refer to the three baselines as Backbone-Supervised, Backbone-cutmix, and Backbone-cutout, respectively. Note that we abbreviate a model with misfit prior as '-MP' for simplicity.

Databases We evaluate the shape reconstruction accuracy on the NoW database [28]. The publicly-available CelebA-HQ testset [20] and the AR database [22] are used for validating the effectiveness of fitting and face segmentation. For the AR dataset, we take as GT masks the 120 manually-segmented masks in [8] that are publicly shared by the authors. The standard deviation is provided after $'\pm'$.

4.2. Reconstruction Quality

Fig. 5 shows results of the Deep3D network, the MoFA network, and our method for qualitative comparison. Note that all the masks are binarized by rounding the pixels. The segmentation masks provided by the Deep3D result from a skin detector which assumes that skin color follows the simple multivariate Gaussian distribution. It shows that in our segmentation results, some small occlusions and skin-colored occlusions are better detected. Furthermore, our segmentation is more robust to illumination variations. It can also be observed from the reconstructed images that the illumination and texture of the faces are better estimated. Visually speaking, our method reaches competitive fitting results and improved outlier segmentation. More quantitative results are provided in the supplementary materials.

Image fitting accuracy shows how much the fitting results get misled by outliers. We evaluate the Root Mean Square Error (RMSE) between the input image and the reconstructed image inside visible face regions, with provided GT segmentation masks. We compare different methods on the AR database, CelebA-HQ testset, and two randomly-selected occluded (750 images) and unoccluded subsets (558 images), referred to as 'CelebA-Overall', 'CelebA-Occluded', and 'CelebA-Unoccluded', respectively. As shown in Tab. 1, our fitting accuracy is competitive even to the fully supervised counterpart with data augmentation.

Table 2. Reconstruction error (mm) on the NoW testset [28].

Method	median	mean	std
MICA [42]	**0.90**	**1.11**	**0.92**
Wood *et al.* [37]	1.02	1.28	1.08
DECA [10]	1.09	1.38	1.18
RingNet [28]	1.21	1.53	1.31
Deep3D [5]	1.23	1.54	1.29
3DDFA V2 [15]	1.23	1.57	1.39
Dib *et al.* [7]	1.26	1.57	1.31
SynergyNet [38]	1.27	1.59	1.31
MGCNet [30]	1.31	1.87	2.63
PRNet [11]	1.50	1.98	1.88
3DMM-CNN [36]	1.84	2.33	2.05
FOCUS (ours)	1.04	1.30	1.10
FOCUS-MP (ours)	1.02	1.28	1.09

Table 3. Reconstruction error (mm) on the non-occluded and occluded data in the NoW validation subset.

	Unoccluded Subset			Occluded Subset		
Method	median	mean	std	median	mean	std
Deep3D [5]	1.33	1.67	1.41	1.40	1.73	1.41
DECA [10]	1.18	1.47	1.24	1.29	1.56	1.29
MoFA [31]	1.35	1.69	1.42	1.36	1.69	1.41
Backbone	1.21	1.46	1.18	1.33	1.59	1.27
Backbone-Supervised	**1.02**	1.25	1.04	**1.05**	**1.29**	**1.09**
Backbone-cutmix	1.05	1.28	1.04	1.08	1.33	1.11
Backbone-cutout	1.03	1.28	1.06	1.09	1.34	1.10
FOCUS (ours)	1.03	1.25	1.03	1.07	1.34	1.19
FOCUS-MP (ours)	**1.02**	**1.24**	**1.02**	1.08	1.34	1.20

Shape reconstruction accuracy is evaluated on the NoW Dataset. The cumulative errors on the testset in Tab. 2 indicate that FOCUS reaches the state-of-the-art among the pipelines without 3D supervision even with a considerably smaller training set and no constraints on identity consistency. Note that [37, 42] are trained with GT geometry of real images or synthesized images and we are comparable to the work of Wood *et al.* [37]. To further evaluate the robustness fairly, 62 pairs of images in the evaluation set are selected with comparable poses with or without occlusions in the publicly-available validation set to normalize pose variation. Tab. 3 shows that the shape reconstruction accuracy of our pipeline is barely affected by occlusions, and reaches a similar level as the fully-supervised pipelines. Please refer to the supplementary for a more detailed analysis.

4.3. Outlier Segmentation

In this section, the performance of the outlier segmentation is indicated with occlusion segmentation accuracy, since occlusions account for a large portion of the outliers

Table 4. Evaluation of occlusions segmentation accuracy on the AR testsets.

	Unoccluded				Glasses				Scarf			
Method	ACC	PPV	TPR	F1	ACC	PPV	TPR	F1	ACC	PPV	TPR	F1
Deep3D [5]	**0.88**	0.93	**0.94**	$\mathbf{0.93 \pm 0.04}$	**0.88**	0.92	**0.92**	$\mathbf{0.92 \pm 0.04}$	0.80	0.80	**0.93**	0.86 ± 0.05
Egger *et al*. [8]	-	-	-	0.90	-	-	-	0.87	-	-	-	0.86
FOCUS (ours)	**0.88**	**0.96**	0.91	$\mathbf{0.93 \pm 0.03}$	**0.88**	**0.98**	0.85	0.91 ± 0.04	**0.86**	**0.97**	0.81	$\mathbf{0.88 \pm 0.05}$

Table 5. Ablation study on the AR testsets and the NoW validation subset.

	AR-unoccluded				AR-glasses				AR-scarf				NoW Evaluation Set		
Method	ACC	PPV	TPR	F1	ACC	PPV	TPR	F1	ACC	PPV	TPR	F1	median	mean	std
Pretrained	0.75	0.95	0.77	0.85 ± 0.05	0.78	0.97	0.72	0.82 ± 0.05	0.70	0.89	0.62	0.73 ± 0.07	1.06	1.32	1.14
Baseline	0.81	**0.96**	0.83	0.89 ± 0.04	0.81	0.97	0.76	0.85 ± 0.05	0.79	0.96	0.71	0.82 ± 0.07	1.06	1.32	1.15
Neighbor	0.85	0.95	0.88	0.91 ± 0.04	0.84	0.95	0.81	0.87 ± 0.04	0.83	0.94	0.79	0.85 ± 0.06	1.06	1.32	1.15
Perceptual	**0.89**	**0.96**	**0.92**	$\mathbf{0.94 \pm 0.03}$	**0.89**	**0.98**	**0.87**	$\mathbf{0.92 \pm 0.04}$	**0.87**	**0.97**	**0.84**	$\mathbf{0.90 \pm 0.05}$	1.06	1.32	1.14
FOCUS	0.88	**0.96**	0.91	0.93 ± 0.03	0.88	**0.98**	0.85	0.91 ± 0.04	0.86	**0.97**	0.81	0.88 ± 0.05	1.05	1.31	1.14
FOCUS-MF	-	-	-	-	-	-	-	-	-	-	-	-	**1.03**	**1.29**	**1.12**

and their labels are feasible. Four indices are calculated inside the rendered regions, including accuracy (ACC), precision (Positive Predictive Value, PPV), recall rate (True Positive Rate, TPR), and F1 score (F1). We separate the AR dataset into three subsets, which include faces without occlusions (neutral), faces with glasses (glasses), and faces with scarves (scarf). According to Tab. 4, the masks predicted by our method show a higher accuracy, recall rate, and F1 score, and competitive precision compared to the skin detector used in [5] and the segmentation method proposed in [8]. The results validate the potential of the outlier segmentation network to locate occlusions.

4.4. Ablation Study

In this section, we first verify the utility of the segmentation network and the proposed neighbor loss, L_{nbr} (Eq. (1)), and the coupled perceptual losses, L_{dist} (Eq. (2)) and L_{presv} (Eq. (4)). We compare the segmentation performances of ablated pipelines on the AR testset, since it contains heavier occlusion, and test the shape reconstruction quality on the NoW evaluation subset. The pre-trained model as introduced in Sec. 3.3 is referred to as 'Pretrained'. We refer to the segmentation network trained without the neighbor loss or perceptual losses as 'Baseline' and Eq. (5) is used to compensate for the lack of such losses. The 'Neighbor' and 'Perceptual' pipelines refer to the proposed segmentation network trained only with L_{nbr} and only with two perceptual losses, L_{dist} and L_{presv}, respectively. The results in Tab. 5 verify the usefulness of the segmentation network, since their results excel the pretrained model in almost all the indices. Comparison among the results of the ablated pipelines shows that both losses contribute significantly to the segmentation accuracy, indicating that the segmentations are more semantic with the FOCUS model. The gain in the reconstruction accuracy also validates the usefulness of the FOCUS model. As for the misfit prior, results in Tabs. 2, 3 and 5 proves that it helps reduce the reconstruction error. Please refer to the supplementary material for more comparisons among the ablated pipelines.

4.5. Limitations

Despite the accurate reconstruction and segmentation proven in the experiments, there are several limitations. The main issue is that the potential of our outlier segmentation to occlusion segmentation is limited by the expressiveness of the face model. For further improvement, a model capable of depicting facial details, makeup, and beard is required. Additionally, although the misfit prior reduces the overall reconstruction error, it does not promise enhancement for every single prediction.

5. Conclusion

In this paper, we address two sources of errors of the model-based face auto-encoder pipelines: the outliers and the misfits. We have shown how to solve face reconstruction and outlier segmentation jointly in a weakly-supervised way, so as to enhance the robustness for model-based face autoencoders in unconstrained environments. We have also shown how to reduce misfits with a statistical prior. Comprehensive experiments have shown that our method reaches state-of-the-art reconstruction accuracy on the NoW testset among methods without 3D supervision and provides promising segmentation masks as well.

Theoretically, our method can be integrated with the existing face autoencoders and/or non-linear parametric face models to achieve better performance. More importantly, we believe that the fundamental concepts of our approach can go beyond the context of face reconstruction and will inspire future work, such as human body reconstruction, or object reconstruction, with a reliable generative model. We also expect that the masks will be useful for other tasks, *e.g.* image completion, recognition, or more. An analysis of the societal impact is provided in the supplementary materials.

References

[1] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9):1063–1074, 2003. 1, 2

[2] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 6

[3] M. De Smet, R. Fransens, and L. Van Gool. A generalized em approach for 3d model based face recognition under occlusions. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1423–1430, 2006. 3

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 6

[5] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2, 3, 5, 6, 7, 8

[6] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 7

[7] Abdallah Dib, Cedric Thebault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 7

[8] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision*, 126(12):1269–1287, 2018. 3, 5, 7, 8

[9] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020. 1

[10] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *arXiv preprint arXiv:2012.04012*, 2020. 2, 3, 5, 7

[11] Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *ECCV*, 2018. 7

[12] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[13] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8377–8386, 2018. 3

[14] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018. 6

[15] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3D dense face alignment. In *European Conference on Computer Vision (ECCV)*, pages 152–168, 2020. 7

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[17] Tatsuro Koizumi and William AP Smith. "look ma, no landmarks!"–unsupervised, model-based dense face alignment. In *European Conference on Computer Vision*, pages 690–706. Springer, 2020. 6

[18] Adam Kortylewski, Andreas Schneider, Thomas Gerig, Bernhard Egger, Andreas Morel-Forster, and Thomas Vetter. Training deep face recognition systems with synthetic data. *arXiv preprint arXiv:1802.05891*, 2018. 3

[19] Adam Kortylewski, Mario Wieser, Andreas Morel-Forster, Aleksander Wieczorek, Sonali Parbhoo, Volker Roth, and Thomas Vetter. Informed mcmc with bayesian neural networks for facial image analysis. *arXiv preprint arXiv:1811.07969*, 2018. 1

[20] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 2, 6, 7

[21] Fabio Maninchedda, Christian Häne, Bastien Jacquet, Amaël Delaunoy, and Marc Pollefeys. Semantic 3d reconstruction of heads. In *European conference on computer vision*, pages 667–683. Springer, 2016. 3

[22] A. Martinez and Robert Benavente. The ar face database. *Tech. Rep. 24 CVC Technical Report*, 01 1998. 2, 7

[23] Andreas Morel-Forster. *Generative shape and image analysis by combining Gaussian processes and MCMC sampling*. PhD thesis, University_of_Basel, 2016. 3

[24] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[25] Sami Romdhani and Thomas Vetter. Efficient, robust and accurate fitting of a 3d morphable model. In *ICCV*, volume 3, pages 59–66, 2003. 3

[26] 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. 6

[27] Shunsuke Saito, Tianye Li, and Hao Li. Real-time facial segmentation and performance capture from rgb input. In *European conference on computer vision*, pages 244–261. Springer, 2016. 2, 3

[28] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 7

[29] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019. 2, 3

[30] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3D face reconstruction by occlusion-aware multi-view geometry consistency. In *European Conference on Computer Vision (ECCV)*, volume 12360, pages 53–70, 2020. 7

[31] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017. 1, 2, 4, 6, 7

[32] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3

[33] Hitika Tiwari, Vinod K Kurmi, KS Venkatesh, and Yong-Sheng Chen. Occlusion resistant network for 3d face reconstruction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 813–822, 2022. 2, 3

[34] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3935–3944, 2018. 2, 3

[35] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1126–1135, 2019. 3

[36] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5163–5172, 2017. 7

[37] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 3d face reconstruction with dense landmarks. In *European Conference on Computer Vision*, pages 160–177. Springer, 2022. 2, 7

[38] Cho-Ying Wu, Qiangeng Xu, and Ulrich Neumann. Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry. In *2021 International Conference on 3D Vision (3DV)*, 2021. 7

[39] Ilker Yildirim, Michael Janner, Mario Belledonne, Christian Wallraven, Winrich Freiwald, and Josh Tenenbaum. Causal and compositional generative models in online perception. In *CogSci*, 2017. 3

[40] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. 7

[41] Xiangyu Zhu, Junjie Yan, Dong Yi, Zhen Lei, and Stan Z Li. Discriminative 3d morphable model fitting. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8. IEEE, 2015. 1

[42] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. *arXiv preprint arXiv:2204.06607*, 2022. 2, 7