Uncertainty-Guided Face Matting for Occlusion-Aware Face Transformation

Hyebin Cho
Korea Advanced Institute of Science & Technology
School of Electrical Engineering
Daejeon, Republic of Korea
hyebin.cho@kaist.ac.kr

Jaehyup Lee*
Kyungpook National University
School of Computer Science and Engineering
Daegu, Republic of Korea
jaehyuplee@knu.ac.kr

Figure 1: Application of face matting for occlusion-aware face filtering. (a) shows the original input image with partial occlusions. (b) presents the alpha matte predicted by our FaceMat framework, which separates occluding elements (e.g., hand, microphone) from the facial region. (c, d) compare visual filter results without and with alpha matte, respectively. Without matting (c), the filter is incorrectly applied to occluders, leading to unnatural results. In contrast, (d) shows alpha-guided compositing using the output from (b), where the filter is confined to the facial area. (e, f) show zoomed-in versions for clearer comparison.

Abstract

Face filters have become a key element of short-form video content, enabling a wide array of visual effects such as stylization and face swapping. However, their performance often degrades in the presence of occlusions, where objects like hands, hair, or accessories obscure the face. To address this limitation, we introduce the novel task of face matting, which estimates fine-grained alpha mattes to separate occluding elements from facial regions. We further present FaceMat, a trimap-free, uncertainty-aware framework that predicts high-quality alpha mattes under complex occlusions. Our approach leverages a two-stage training pipeline: a teacher model is trained to jointly estimate alpha mattes and per-pixel uncertainty using a negative log-likelihood (NLL) loss, and this uncertainty is then used to guide the student model through spatially adaptive knowledge distillation. This formulation enables the student to focus on ambiguous or occluded regions, improving generalization and preserving semantic consistency. Unlike previous approaches that rely on trimaps or segmentation masks, our framework requires no auxiliary inputs making it well-suited for real-time applications. In addition, we reformulate the matting objective by explicitly treating skin as foreground and occlusions as background, enabling clearer compositing strategies. To support this task, we newly constructed CelebAMat, a large-scale synthetic dataset specifically designed for occlusion-aware face matting. Extensive experiments show that FaceMat outperforms state-of-the-art methods across multiple benchmarks, enhancing the visual quality and robustness of face filters in real-world, unconstrained video scenarios. The source code and CelebAMat dataset are available at *https://github.com/hyebin-c/FaceMat.git*.

*Corresponding author

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755060

CCS Concepts

• **Computing methodologies** → **Video segmentation**.

Keywords

Image matting, Video matting, Face matting

1 Introduction

With the increasing popularity of short-form content on platforms like TikTok, Instagram, and YouTube Shorts, face filtering has emerged as a key feature for enhancing user engagement. These filters perform facial region detection and apply various visual effects such as stylization, overlays, and face swapping to enhance visual storytelling and provide personalized content.

However, as shown in Fig. 2, existing face filtering techniques often degrade under real-world conditions, particularly in the presence of occlusions caused by motion-blurred hands, accessories, or hair. Such occlusions often cause unnatural artifacts, as conventional segmentation methods relying on binary masks fail to capture fine-grained transparency. Consequently, they can not properly handle subtle transitions between foreground and background, resulting in degraded filter quality.

Image matting addresses this limitation by estimating a per-pixel alpha matte that models soft transitions between foreground and background. A pixel I_i in the image can be represented as:

$$I_i = \alpha_i F_i + (1 - \alpha_i) B_i, \quad \alpha \in [0, 1] \tag{1}$$

where F_i, B_i represent the foreground and background color components of the I_i, and α_i denotes the corresponding alpha value.

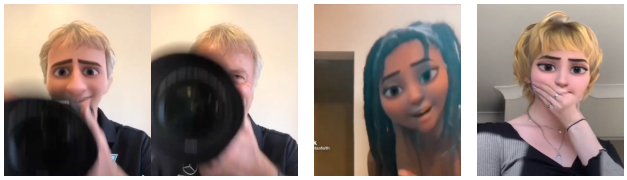
While image matting provides a precise representation, predicting alpha mattes from a single image remains an ill-posed problem. Most of the previous work relies heavily on auxiliary priors such as trimaps [21, 24, 35], binary masks [25, 39], or background images [19] to constrain the solution. However, such inputs are difficult to be obtained in real-time video settings, making them impractical for dynamic face filtering applications.

Recent advances in trimap-free matting aim to eliminate the dependence on auxiliary inputs by training data-driven models in specific domains such as human portraits [13, 20]. However, these models fail to operate reliably under challenging conditions such as occlusions, motion blur, and acquisition noise, which are common in short-form video content.

To address these challenges, we propose FaceMat, a robust and trimap-free face matting framework designed for occlusion-aware visual effects. At the core of our method is an uncertainty-aware knowledge distillation strategy. We first train a teacher model to jointly predict alpha mattes and per-pixel uncertainty using a negative log-likelihood (NLL) loss. The resulting uncertainty maps, which reflect the teacher model's confidence, are then used to adaptively adjust the distillation temperature across spatial locations. This allows the student model to focus its learning on ambiguous or occluded regions, resulting in improved generalization and robustness.

We further reinterpret the matting task by treating facial skin as background and occlusions such as hands or objects as foreground. Based on this definition, our proposed FaceMat operates in four stages: (1) Occlusion matting, where an alpha matte is predicted to isolate occluding elements from the facial region; (2) Face completion, where an inpainting module can optionally reconstruct occluded facial areas to obtain a clean face; (3) Face transformation, which applies visual effects such as swapping or stylization to the completed face; and (4) Compositing, where the transformed face is blended with the original occlusion using the predicted alpha matte to ensure visual consistency. This design enables natural and realistic face filtering under complex occlusions while preserving realistic appearances. Our contributions are summarized as follows:

- We define *face matting* as a new task that explicitly separates facial occlusions from skin regions, reformulating the foreground-background relationship.
- We introduce **FaceMat**, a trimap-free matting framework guided by per-pixel uncertainty to enable locally adaptive learning via knowledge distillation.
- We propose a multi-stage pipeline that integrates matting, inpainting, transformation, and compositing for high-quality filtering under occlusion.
- We construct a synthetic dataset, **CelebAMat**, designed for face matting and demonstrate that FaceMat achieves state-of-the-art results in challenging video matting applications with frequent facial occlusions.



(a) Filter effect is lost (b) Hair occlusion artifact (c) Hand-mouth blending

Figure 2: Limitation of face manipulation under occlusion. The effectiveness of face-related applications may degrade significantly due to failures in face recognition or a lack of occlusion-aware design in manipulation techniques.

2 Related Work

2.1 Image and Video Matting

Image matting is an inherently ill-posed problem that often requires strong priors to produce reliable alpha mattes. One of the most commonly used priors is the trimap, which divides the image into foreground, background, and unknown regions in the corresponding context. By providing explicit spatial constraints, trimaps significantly enhance matting accuracy and stability. Consequently, a variety of trimap-based approaches [12, 22, 24, 35] have been proposed, leveraging deep neural networks to improve precision and generalization in image matting.

As previous research progressed from still images to videos, video matting emerged as a natural extension. However, collecting high-quality, per-frame alpha annotations for videos is expensive and time-consuming. Moreover, requiring users to provide auxiliary inputs (e.g., trimaps or masks) for every frame introduces substantial practical limitations. To mitigate these challenging issues, recent works [27, 30, 41] have proposed auxiliary-free or single-frame-conditioned video matting techniques. However, without any strong priors, these methods often suffer from degraded performance due to the ambiguity of the matting task.

To address these limitations, several methods [13, 20] have focused on human-centric scenarios, primarily targeting background removal in portrait images. These methods emphasize ease of use

and accurate boundary reconstruction, particularly for challenging regions such as hair.

While existing work has concentrated on background matting, we broaden the scope by targeting a wider range of occlusions, including hands, hair, transparent objects, and even semitransparent elements such as smoke or fire. These occlusions present unique challenges, such as intricate boundaries and the lack of clear trimap guidance. Moreover, since our proposed framework operates without any auxiliary inputs for test, conventional trimap- or mask-dependent approaches are inapplicable in real-world settings. In our trimap-free, real-world setting, the matting region must be inferred solely from the image content.

2.2 Face Occlusion Segmentation

Face occlusion segmentation is critical for a wide range of face-related tasks, such as face recognition [28, 31], face swapping [8, 23], and facial reconstruction [33, 38]. Occlusions caused by external objects, such as hands, accessories, or hair, can significantly degrade the performance of these applications, making robust occlusion handling essential for real-world deployment.

To address this issues, several datasets [2, 36] have been introduced, featuring occluded faces collected from in-the-wild scenarios. However, these datasets often suffer from key limitations, including limited scale, low resolution, and insufficient diversity in occlusion types. Furthermore, many lack high-quality segmentation annotations, or only provide coarse binary or categorical labels, which are inadequate for learning fine-grained occlusions.

Recent works [32, 37] have attempted to address these challenges. However, as illustrated in Figure 3, segmentation-based methods are fundamentally limited in their ability to model soft transitions and semi-transparent occlusions, such as motion blur or translucent objects. These cases introduce ambiguities that cannot be effectively resolved with discrete label maps. As a result, relying solely on segmentation often leads to hard boundaries and visible artifacts in downstream applications, especially in video-based scenarios.

To overcome these limitations, advocate for a matting-based formulation. Notably, unlike segmentation, matting predicts continuous alpha values, enabling smooth foreground-background transitions and more accurate integration of occluded elements. This is particularly beneficial for face filtering and editing tasks, where natural compositing is crucial.

In this work, we propose a novel framework for face occlusion matting that estimates high-quality alpha mattes for occluding regions. By replacing segmentation with matting, our method improves robustness to complex occlusion patterns and delivers more visually coherent results across various face-related applications.

3 Problem Formulation

3.1 Definition of Face Matting

Face matting is the task of accurately extracting the facial region while separating occlusions that appear in front of the face. Unlike general image matting, where the target object is often ambiguous and requires auxiliary inputs such as trimaps or binary masks, face matting benefits from a well-defined foreground. Although auxiliary inputs provide explicit guidance, they impose additional computation and user costs. Instead, auxiliary-free face matting

Figure 3: Comparison with FaceOcc [37]. (a) Input image. (b) FaceOcc fails to accurately separate hair from face due to hard binary masks. (c) Our method provides precise separation of facial and hair regions using alpha-based segmentation.

Figure 4: Face matting definition in our framework. (a) Input image with partial occlusions. (b) Predicted alpha matte separating face and occluders. (c, d) Extracted foreground and background for compositing and manipulation.

leverages domain-specific training to distinguish the foreground without requiring external input.

In face matting, the foreground is initially defined as the skin area, encompassing all facial components, while the background consists of the remaining regions outside the foreground. In real-world scenarios, occlusions exist in front of the facial skin from the camera's viewpoint, complicating the foreground-background distinction. Thus, while the skin is conceptually treated as the foreground, its physical placement in the scene contradicts this assumption. However, to maintain consistency with the conventional image matting formulation (1), we define the skin region as the foreground. Figure. 4 illustrates examples of input images, annotations, foreground, and background representations in the face matting task.

Performing face matting yields the alpha matte for the foreground, i.e. the skin, enabling flexible recomposition of either the skin or occlusions. In this study, occlusions are defined based on their feasibility in recomposing, specifically by determining which elements should be preserved after face-related tasks. Specifically, the skin region includes the face area with eyes, nose, and mouth, while the rest includes occlusions such as hair, ears, and hands. Non-body elements like heavy makeup are also considered occlusions, whereas purely transparent lenses without shadows or color are excluded.

3.2 Dataset Generation

To construct a dataset tailored for face matting, both a high-quality face dataset and diverse occlusion sources are essential. We synthesized training samples by compositing occlusions onto face images

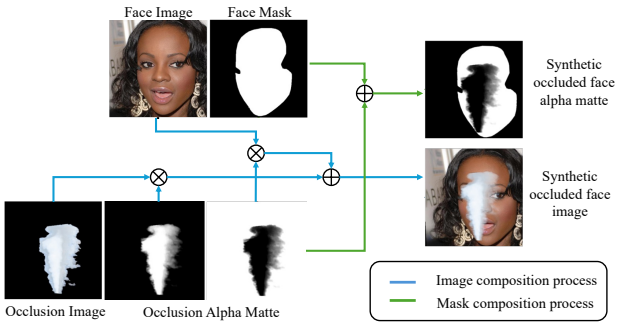


Figure 5: Overview of face matting data generation. A clean face image and an occlusion image are composited to create a synthetic occluded face. The corresponding soft alpha matte is generated by blending the face mask and occlusion mask, enabling pixel-level soft mask ground truth for training.

under various conditions. This process results in CelebAMat, a new dataset designed to support realistic and occlusion-aware face matting. We newly introduce and release CelebAMat as a novel benchmark dataset for evaluating face matting models under diverse and challenging occlusion scenarios. We use CelebAMat both as the training dataset and as a benchmark for evaluating face matting performance under diverse occlusion conditions.

The CelebAMask-HQ dataset [15] provides high-resolution face images with segmentation masks for facial attributes, which we employ for face matting. To ensure occlusion-free face images, samples with visible occlusions were removed from both training and testing sets, using the refined annotations and partitioning provided by [32]. This results in a dataset of 24,602 training images and 716 test images, matching the original quantity. However, their annotations lack the precision required for matting applications and often fail to accurately capture fine-grained facial details such as glasses and facial hair.

For occlusion modeling, we utilize several dataset from the image matting datasets: SIMD [29], AM2k [16], hand segmentation dataset: HIU-data [40], and Describe Textures Dataset (DTD) [7] for occlusion diversity. To enhance the realism of the dataset, we applied gaussian blur to the boundary regions of the HIU mask.

Fig 5 illustrates the overall data generation process. For each dataset, occlusions are sampled strictly from the original training split during training, and from the test split during evaluation. During training, occlusion instances are inserted with random variations in size, orientation, and position to increase robustness. In contrast, during evaluation, we adopt a fixed set of occlusion configurations to form a consistent benchmark setting.

4 Methodology

An overview of our proposed training framework **FaceMat**, is illustrated in Fig. 6. Conventional matting methods rely heavily on trimaps as spatial priors and loss weighting, with evaluation restricted to ambiguous regions. This often overlooks the semantic consistency of the full facial structure.

To address this limitation, we propose a novel two-stage training strategy: 1) **Boundary-aware Learning with Uncertainty Estimation**: In the first stage, a teacher network is jointly trained to predict both the alpha matte and associated uncertainty map. The teacher model is trained using trimaps to focus on the precise separation of facial boundaries, like other conventional matting pipeline [3, 4, 11, 17, 18, 24–26, 39]. 2) **Uncertainty-Guided Knowledge Distillation (UGKD)**: In the second stage, the estimated uncertainty map is utilized to drive a locally adaptive distillation strategy that guides the corresponding student model.

Uncertainty estimation has been increasingly adopted in deep learning to improve model robustness and interpretability [9, 10, 14, 34]. It has proven effective in various computer vision tasks, including segmentation [1], image classification [5], and object detection [6], where it is primarily used to quantify prediction confidence and refine model outputs. Unlike prior approaches that primarily utilize uncertainty as a measure of confidence, we propose a novel use of uncertainty as a supervisory signal. Specifically, we leverage the estimated uncertainty to guide the student model's learning in a fine-grained and spatially adaptive manner, which is crucial for modeling the soft transitions and boundary ambiguities inherent in image matting.

While the previous matting models that apply uniform supervision across the entire image rely heavily on handcrafted trimaps, our framework adaptively focuses on boundary-critical regions where prediction uncertainty is highest.

This mechanism encourages the student model to concentrate on boundary-critical regions, where alpha estimation exhibits the highest degree of uncertainty and thus requires more fine-grained supervision. As a result, the model not only learns to preserve comprehensive semantic representations across the entire image but also retains high fidelity in capturing intricate boundary details.

4.1 Boundary-aware Learning with Uncertainty Estimation

The matting task inherently involves a high degree of uncertainty, as it requires the prediction of alpha values that lie between the foreground and background from a visually mixed composition. This blending is influenced not only by color information but also by factors such as lighting direction, sensor characteristics, and scene dynamics. Compared to other vision tasks, matting exhibits greater ambiguity, particularly near object boundaries. Since these uncertain regions vary depending on local image conditions, a modeling approach that captures heteroscedasticity is essential for robust performance.

In our framework, the model is trained to jointly predict the alpha matte and a corresponding pixel-wise uncertainty map. This is achieved via an objective based on the NLL loss, which directly encourages the network to estimate a per-pixel variance conditioned on local image features. Consequently, the model implicitly learns to represent heteroscedastic aleatoric uncertainty, enabling it to allocate greater attention to ambiguous regions while maintaining robustness in more confident areas.

Furthermore, we apply the NLL formulation not only to the uncertainty map but also to the predicted alpha matte itself. This contributes to the spatial smoothness of the alpha map, yielding more natural transitions between regions. Additionally, it acts as an implicit regularizer for the uncertainty estimation, mitigating overconfidence and promoting stable learning dynamics.

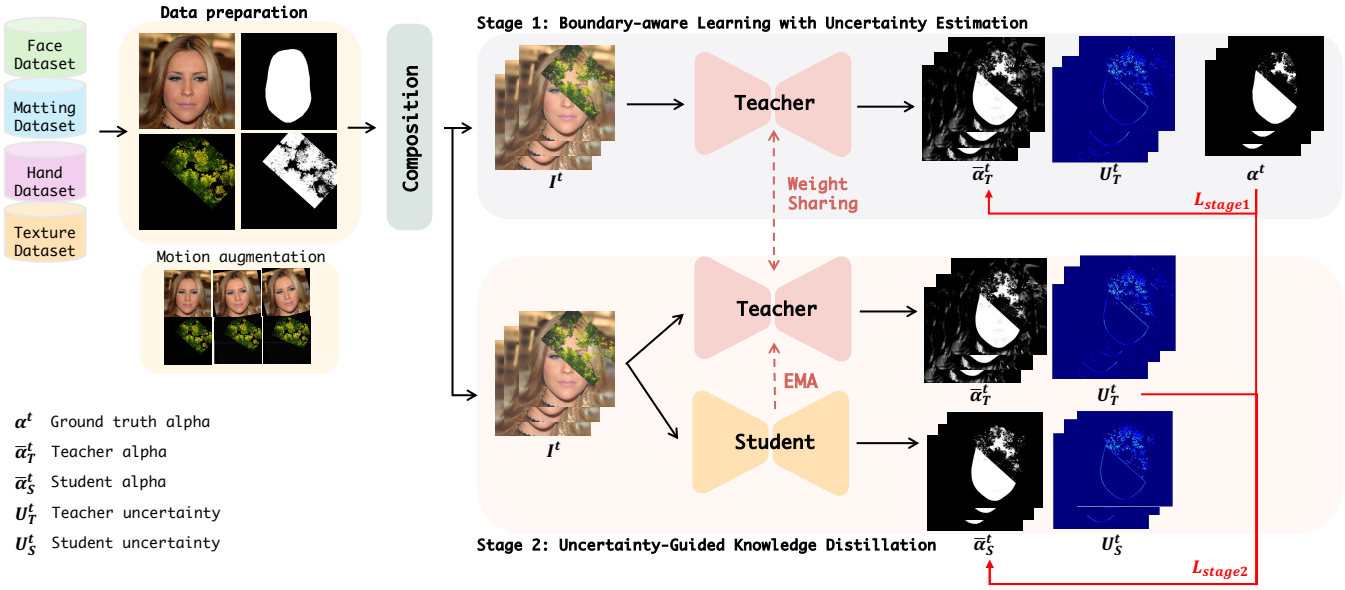


Figure 6: Overview of the full FaceMat pipeline. Multiple datasets, including face, hand, texture, and matting datasets, are combined through motion-aware composition to generate occlusion-rich training data. The framework then proceeds in two stages: Stage 1 trains a teacher model with boundary-aware learning and uncertainty estimation using trimaps; Stage 2 distills this knowledge to a student model via uncertainty-guided supervision, enabling trimap-free face matting under occlusions.

$$\mathcal{L}^{\beta}_{\text{NLL}-u} = \frac{1}{2}\left(\frac{(u-\mu_u)^2}{\sigma_u^2} + \log\sigma_u^2\right)\cdot(\sigma_u^2)^{\beta} \tag{2}$$

$$\mathcal{L}^{\beta}_{\text{NLL}-\alpha} = \frac{1}{2}\left(\frac{(\alpha-\mu_\alpha)^2}{\sigma_\alpha^2} + \log\sigma_\alpha^2\right)\cdot(\sigma_\alpha^2)^{\beta} \tag{3}$$

In the first stage, we aim to capture fine-grained boundary details of the alpha matte. To this end, we utilize trimaps generated from the ground truth alpha to spatially constrain the loss computation. We adopt RVM [20] as our baseline model, which is specifically designed for video matting, and apply all losses across the full temporal range of frames, i.e., for all $t \in [1, T]$.

Following RVM, we employ the L1 regression loss $\mathcal{L}_{\text{L1}}$, the pyramid Laplacian loss $\mathcal{L}_{\text{lap}}$, and the temporal consistency loss $\mathcal{L}_{\text{tc}}$. In addition, we incorporate negative log-likelihood (NLL) losses for both the predicted uncertainty map $\mathcal{L}^{\beta}_{\text{NLL}-u}$ and the alpha matte $\mathcal{L}^{\beta}_{\text{NLL}-\alpha}$. Note that all losses except for the NLL losses are masked by the unknown region in the trimap to emphasize learning in the ambiguous regions. The final objective function for the teacher model is defined as:

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{L1}} + \mathcal{L}_{\text{lap}} + \mathcal{L}_{\text{tc}} + \mathcal{L}^{\beta}_{\text{NLL}-u} + \mathcal{L}^{\beta}_{\text{NLL}-\alpha}.$$

4.2 Uncertainty-guided Knowledge Distillation

Conventional image matting models often rely on auxiliary inputs such as trimaps to guide the model by providing coarse annotations of foreground, background, and unknown regions. These auxiliary cues are commonly used not only to inform the model duroing inference but also to constrain the loss computation by masking the supervision to ambiguous regions. In contrast, our approach is trimap-free, meaning that no such auxiliary hints are available during training of inference. Consequently, the model must learn semantic priors and structural details directly from the data. This can introduce a key challenge: the inherent trade-off between capturing high-level semantics and preserving fine-grained boundary details. Effectively balancing this trade-off is crucial for generating high-quality alpha mattes in unconstrained settings.

To address this, we propose an uncertainty-guided knowledge distillation (UGKD) framework that adaptively modulates the supervision strength based on the estimated teacher model's confidence, or uncertainty. Specifically, since we have access to both the ground truth and the teacher's predictions, we leverage the teacher's uncertainty map to identify regions where the predictions are ambiguous or less reliable.

To enforce stronger learning in these regions, we introduce an uncertainty-weighted $L1$ regression loss between the ground truth and the student model's predicted alpha matte. This encourages the student model to focus more on semantically complex or visually uncertain areas during training. The objective of stage 2 is defined as $L_{\text{stage2}} = L_{l1}^{\text{soft}} + L_{\text{lap}}$. Here, L_{l1}^{soft} is the uncertainty-weighted $L1$ regression loss between the predicted alpha matte and the ground truth:

$$L_{l1}^{\text{soft}} = \left\| w_{\text{unc}} \odot (\alpha - \alpha_{\text{gt}}) \right\|_1 \tag{4}$$

where α denotes the predicted alpha matte by the student model, $\odot$ is element-wise multiplication, α_{gt} is the ground truth, and w_{unc} is the uncertainty-based spatial weight map derived from the teacher model's uncertainty.

To emphasize regions where the teacher model is less confident, we define the uncertainty-based weighting map w_{unc} as a linear

Table 1: Quatitative comparison of matting methods on the CelebAMat benchmark under various occlusion types. We report MSE and SAD scores (lower is better) across four test sets: SIMD, AM2k, HIU, and Random (Rand). This table serves as a baseline reference for evaluating performance under divese occlusion. Best and second-best results are marked in bold and underlined, respectively.

Model Configuration			Matting Test Dataset				Segmentation Test Dataset			
Network	Auxiliary Input	Encoder Type	SIMD		AM2k		HIU		Rand	
			MSE(↓)	SAD(↓)	MSE(↓)	SAD(↓)	MSE(↓)	SAD(↓)	MSE(↓)	SAD(↓)
UNet	–	ResNet18	0.0693	31.2129	0.0454	17.1462	0.0578	20.2689	0.0651	25.5367
Aematter[21]	Trimap	Transformer	0.0942	37.6412	0.0472	16.6684	0.0719	24.0475	0.0484	17.2676
MGMatting[39]	Mask	ResNet34	0.0645	29.3725	0.0360	14.2273	0.0469	17.2082	0.0319	13.1362
MODNet[13]	–	MobileNetV2	0.0457	23.1507	0.0311	11.2826	0.0266	**9.6049**	0.0350	12.3250
RVM[20]	–	MobileNetV3	**0.0301**	**20.0812**	**0.0105**	**5.5623**	**0.0199**	10.1634	**0.0123**	**6.2677**

Table 2: Quantitative comparison across different training settings on CelebAMat. Trimap(✓) indicates that a trimap is used as part of the loss function. Each metric is reported as mean ± standard deviation (std). Best and second-best results are marked in bold and UNDERLINED, respectively. The lowest std for each metric is shown in red text.

Setting	Trimap	MSE(↓)	SAD(↓)	Grad(↓)	Conn(↓)	IoU(↑)	Accuracy(↑)
Stage 1: Comparison of NLL-based Multi-Task Learning Variants							
RVM[20]	✓	0.0202	10.2900	0.0385	7.0865	0.8249	0.9528
Stage1 (NLL)	✓	**0.0169 ± 0.00077**	9.25 ± 0.3400	0.0437 ± 0.0014	13.3785 ± 2.0844	0.6017 ± 0.0998	0.8180 ± 0.0797
Stage 2: Comparison of Uncertainty-guided Knowledge Distillation							
Stage1 Extended	✓	0.0183 ± 0.00203	9.82 ± 1.8800	0.0437± 0.0006	12.6937 ± 1.620	0.6333 ± 0.1340	0.8424 ± 0.0891
Stage1 Extended		0.0236 ± 0.00041	16.38 ± 8.9800	**0.0379 ± 0.0018**	7.6577 ± 1.5681	**0.8579 ± 0.0036**	**0.9602 ± 0.0017**
Stage2 (UGKD)		0.0182 ± 0.00028	**8.43 ± 0.1600**	0.0424 ± 0.0004	**6.5764 ± 0.4824**	0.8408 ± 0.0181	0.9529 ± 0.0067

function of the predicted variance $\sigma_u^{\text{teacher}}$:

$$w_{\text{unc}} = w_1 + w_2 \cdot \sigma_u^{\text{teacher}}$$

where $w_1 = 2$ and $w_2 = 2$. This formulation ensures that higher uncertainty leads to greater loss contribution, thereby encouraging the student model to prioritize learning from difficult of ambiguous regions. To improve training stability, the teacher model is updated throughout the training process using an Exponential Moving Average (EMA) of the student model parameters.

5 Experiments

5.1 Benchmarking

To benchmark our proposed face matting dataset, CelebAMat, we evaluate five representative matting models: U-Net with a ResNet-18 encoder pretrained on ImageNet; AEMatter [21], a transformer-based model that utilizes trimaps as auxiliary input; MGMatting [39], a mask-guided approach with a ResNet-34 backbone; and two trimap-free video matting models, MODNet[13] and RVM [20].

Although methods with auxiliary inputs are generally expected to generalize well, we observe a significant performance degradation when applied to CelebAMat without additional adaptation. This phenomenon reflects the auxiliary input inconsistency problem, where a mismatch between training and inference distribution of auxiliary inputs undermines generalization. To ensure fair and consistent evaluation, we trained all models on CelebAMat using the same training settings.

Quantitative results are reported in Table 1. Among all evaluated models, RVM demonstrates the best overall performance across diverse occlusion scenarios. Accordingly, we adopt RVM as the baseline in our framework and extend it with our proposed uncertainty-guided distillation strategy to further enhance matting performance under various occlusions.

5.2 Implementation Details

For the occlusion segmentation datasets, HIU [40] and DTD [7], which are significantly larger than occlusion matting datasets such as SIMD [29] and AM2k [16], we randomly selected 200 samples from each dataset for training to ensure a balanced comparison. As our method only utilize still image datasets, we synthesize motion across adjacent frames by applying affine transformations, resizing, horizontal flipping, color jittering, and random pauses to occluding objects, following the approach of RVM [20].

We evaluate our model on the CelebAMat test dataset using four standard matting metrics: Mean Squared Error (MSE), Sum of Absolute Differences (SAD), Gradient error (Grad), and Connectivity error (Conn), all computed within the unknown region of the trimap. Additionally to measure the model's ability to capture semantic object understanding at the image level, we report binary segmentation metrics: Intersection-over-Union (IoU) and pixel-wise

Table 3: Quantitative comparison on RealOcc [32] in terms of IoU, accuracy, and recall between RVM and our method.

Setting	IoU	Accuracy	Recall
RVM[20]	0.4099	0.8432	0.4876
Ours	0.7121	0.9197	0.9084

accuracy. To further evaluate robustness in real-world scenarios, we conduct additional experiments on the RealOcc [32] which is face occlusion segmentation dataset, reporting IoU, accuracy, and recall as complementary evaluation metrics.

5.3 Quantitative Comparison

Quantitative results for our FaceMat framework on the CelebAMat benchmark are summarized in Table 2.

In stage 1, we apply negative log-likelihood (NLL) loss independently to both the predicted uncertainty map and the alpha matte. This dual supervision leads to performance improvements by enabling the model not only to estimate accurate alpha values but also to self-assess the confidence of its predictions. Moreover, this setup promotes implicit multitask learning, which has regularizing effects and helps mitigate overfitting.

To extend the model's capacity beyond boundary refinement, we compare several training strategies that progressively build toward full-scene semantic understanding. In particular, we investigate: (1) extended training of the Stage 1 model, (2) training without using trimaps in the objective, and (3) full Stage 2 training with uncertainty-guided knowledge distillation (UGKD).

We observe that relying on trimaps in Stage 1 improves local boundary quality but limits global generalization, especially in semantic metrics such as IoU and pixel accuracy. Removing the trimap yields better semantic performance but often degrades accuracy in matting-specific metrics like MSE, SAD. In contrast, our full Stage 2 model with UGKD achieves a strong balance between the two, maintaining competitive boundary precision while improving global semantic consistency.

The results demonstrate that the proposed two-stage framework effectively preserves fine-grained boundaries while learning to reason over semantically meaningful structures. This outcomes is largely attributed to the use of pixel-wise uncertainty, which provides soft, spatially adaptive guidance, which cannot flexibly account for varying degrees of ambiguity across the image.

Furthermore, we also observe that UGKD reduces intra-model variance, called as epistemic uncertainty, leading to more stable and consistent alpha predictions. This is especially beneficial in high-frequency or semi-transparent regions, where even minor inconsistencies may result in perceptual artifacts.

To further assess the generalizability of our method, we also evaluate it on the RealOcc. The results are summarized in Table 3.

5.4 Qualitative Comparison

Figure 7 shows qualitative comparisons between our FaceMat and several state-of-the-art matting models on the CelebAMaat benchmark. Each example includes an occluded input image (a), followed by the predicted alpha mattes from previous methods (b-f) our result (g), and the teacher-predicted uncertainty map (i).

Table 4: Ablation study on occlusion dataset composition in ResNet18

Occlusion Train Dataset				Matting Test Dataset				Segmentation Test Dataset			
				SIMD		AM2K		HIU		Rand	
SIMD	HIU	Rand	AM2K	MSE (↓)	SAD (↓)	MSE (↓)	SAD (↓)	MSE (↓)	SAD (↓)	MSE (↓)	SAD (↓)
✓				0.1248	51.7981	0.1337	47.1108	0.1213	42.7249	0.1581	53.6633
✓	✓			0.0806	35.9724	0.0703	25.5113	0.0738	25.8287	0.0947	31.9772
✓	✓	✓		0.0871	40.2320	0.0820	31.1358	0.0793	30.0110	0.1015	37.0328
✓	✓	✓	✓	**0.0693**	**31.2129**	**0.0454**	**17.1462**	**0.0578**	**20.2689**	**0.0651**	**25.5367**

As highlighted in the yellow boxes, previous methods exhibit difficulties in accurately separating facial structures from occluding elements, especially in regions such as hair boundaries, ears. These models often produce over-smoothed or overly hard transitions, resulting in perceptual artifacts.

In contrast, our method (g) delivers more precise alpha mattes, preserving fine-grained facial boundaries while excluding occluders. Notably, the uncertainty map (i) reveals that the teacher model assigns higher uncertainty to ambiguous regions such as transition zones skin and hair. The spatial distribution aligns closely with the unknown regions of conventional trimap and demonstrates that the uncertainty map effectively serves as a soft attention prior durinig distillation.

These results highlights the benefit of our uncertainty-guided distillation strategy, which enables the student model to concentrate supervision on visually ambiguous areas, leading to more coherent and robust alpha matte predictions under real-world occlusion conditions.

As shown in Figure 7, the predicted uncertainty map exhibits high activation near facial boundaries and occlusion edges, which align well with the unknown region of the trimap. This observation supports the idea that the uncertainty map has been effectively trained to reflect inherently ambiguous regions, such as the transition zones between foreground and background.

Figure 8 illustrates qualitative comparison on RealOcc. Compared to CelebAMat, which belongs to the same domain as the training data, our method exhibits improved generalization to unseen domains. UGKD facilitates the learning of structural representations that are crucial for domain adaptation.

Table 5: Ablation study on fixed occlusion ratio in the training dataset for ResNet18

Fixed Occlusion Ratio	Matting Test Dataset				Segmentation Test Dataset			
	SIMD		AM2K		HIU		Rand	
	MSE (↓)	SAD (↓)	MSE (↓)	SAD (↓)	MSE (↓)	SAD (↓)	MSE (↓)	SAD (↓)
1	0.3258	99.3865	0.4201	112.3609	0.4445	117.2768	0.4297	117.7057
0.75	0.1343	52.8224	0.1478	49.9889	0.1210	41.7132	0.1740	56.4217
0.5	0.1232	49.1341	0.1151	39.7154	0.1139	38.2213	0.1314	44.1395
0.25	**0.0693**	**31.2129**	**0.0454**	**17.1462**	**0.0578**	**20.2689**	**0.0651**	**25.5367**

5.5 Ablation study and Analysis

To demonstrate the effectiveness of occlusion diversity and the impact of the occlusion ratio in the CelebAMat dataset, we conduct an ablation study using a simple U-Net architecture. As shown in Table 4, training with the full set of occlusion datasets improves performance, highlighting the benefit of incorporating diverse occlusion types. In addition, Table 5 indicates that an occlusion ratio of

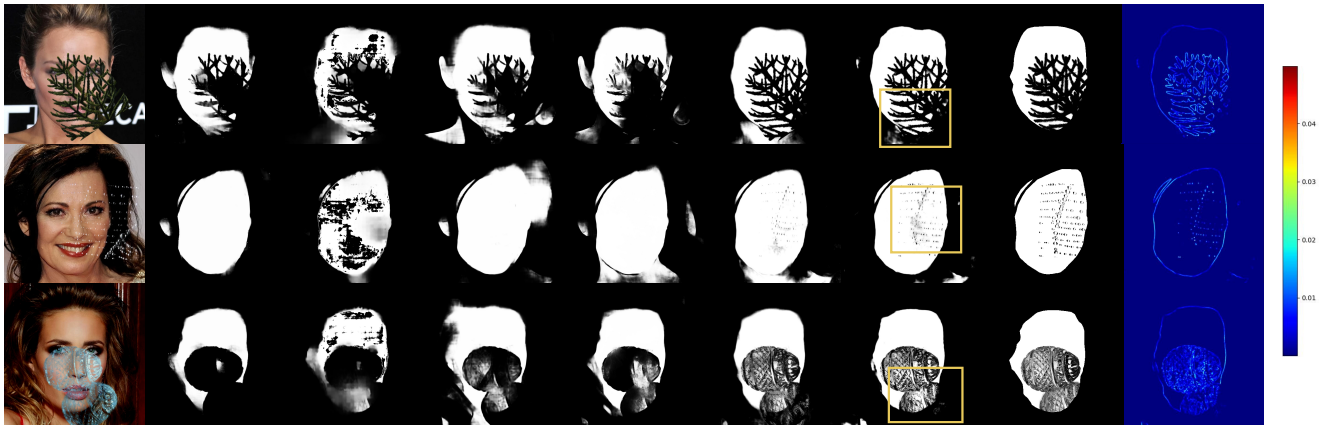


Figure 7: Qualitative comparison on CelebAMat benchmark. (a) shows the occluded input image. As highlighted in the yellow boxes, previous methods (b-f) struggle to preserve fine facial boundaries under occlusions. In contrast, our proposed method (g) predicts a sharper and more accurate alpha matte. (i) presents the uncertainty map estimated by the teacher model in Stage 1, which guides the student model to focus on ambiguous regions and enhances boundary accuracy during distillation.

Figure 8: Qualitative results on the RealOcc dataset. (a) shows an input image with natural occlusions. (b) illustrates a failure case where the baseline struggles with complex, in-the-wild appearance. In contrast, (c) successfully separates the face region from heavy occlusions.

Table 6: Ablation study on occlusion ratio scheduling patterns in the training dataset for RVM. Pattern 1 increases the ratio by 0.2 every 10 epochs: [0.1, 0.3, 0.5, 0.7, 0.9, 1.0]. Pattern 2 increases it by 0.1 every 10 epochs: [0.1, 0.2, 0.3, 0.4, 0.5, 0.6].

Occlusion pattern	Matting Test Dataset				Segmentation Test Dataset			
	SIMD		AM2K		HIU		Rand	
	MSE (↓)	SAD (↓)	MSE (↓)	SAD (↓)	MSE (↓)	SAD (↓)	MSE (↓)	SAD (↓)
fixed 0.25	**0.0301**	**20.0812**	**0.0105**	**5.5623**	**0.0199**	**10.1634**	**0.0123**	**6.2677**
pattern 1	0.0318	22.3034	0.0187	8.4664	0.0206	11.0830	0.0184	8.2502
pattern 2	0.0303	21.4591	0.0176	7.8937	0.0206	10.7454	0.0190	7.9013

0.25 yields the best results, enabling the model to better learn facial boundaries and components under partially occluded conditions.

Building on these findings, Table 6 presents an additional ablation study using the RVM framework to investigate scheduled occlusion ratio increases during training. Although various patterns were explored, the fixed 0.25 ratio consistently achieved the best performance. We hypothesize that maintaining a moderate level of occlusion encourages the model to focus on learning high-level semantic structures of the face, rather than overfitting to heavily occluded cases.

5.6 Real-world Application

To demonstrate the generality and practical potential of our occlusion-aware face matting model, we introduce an application for face filters under occluded conditions. As illustrated in Figure 1, our method enables stable face filter rendering even when key facial attributes are partially or fully occluded.

6 Conclusion

We introduce a novel task, face matting, aimed at enabling occlusion-aware face transformation in real-world scenarios. To support this task, we construct CelebAMat, a large-scale benchmark dataset synthesized from clean facial images and diverse occluders with realistic motion augmentations. Leveraging this dataset, we propose FaceMat, a two-stage learning framework that accurately predicts alpha mattes under occlusions by integrating boundary-aware learning and uncertainty-guided knowledge distillation.

Our extensive experiments demonstrate that our FaceMat not only preserves fine-grained facial boundaries but also improves robustness under challenging visual conditions, such as hands, accessories, or motion blur. We also validate its practical utility by applying it to downstream video tasks, where existing methods often fail due to occlusion artifacts or lack of semantic understanding.

Finally, as shown in *Supplementary material*, ensuring temporal consistency in facial inpainting remains an open challenge, pointing to a valuable direction for future work beyond alpha matte prediction.

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 12 (2017), 2481–2495. doi:10.1109/TPAMI.2016.2644615

[2] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. 2013. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE international conference on computer vision*. 1513–1520.

[3] Huanqia Cai, Fanglei Xue, Lele Xu, and Lili Guo. 2022. Transmatting: Enhancing transparent objects matting with transformers. In *European Conference on Computer Vision*. Springer, 253–269.

[4] Guowei Chen, Yi Liu, Jian Wang, Juncai Peng, Yuying Hao, Lutao Chu, Shiyu Tang, Zewu Wu, Zeyu Chen, Zhiliang Yu, et al. 2022. PP-matting: high-accuracy natural image matting. *arXiv preprint arXiv:2204.09433* (2022).

[5] Steve Chiu, Zhong Jin, and Yingjie Gu. 2015. Active learning combining uncertainty and diversity for multi-class image classification. *IET Computer Vision* 9 (06 2015), 400–407. doi:10.1049/iet-cvi.2014.0140

[6] Jiwoong Choi, Dayoung Chun, Hyun Kim, and Hyuk-Jae Lee. 2019. Gaussian YOLOv3: An Accurate and Fast Object Detector Using Localization Uncertainty for Autonomous Driving. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 502–511. https://api.semanticscholar.org/CorpusID:104292012

[7] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3606–3613.

[8] Kaiwen Cui, Rongliang Wu, Fangneng Zhan, and Shijian Lu. 2023. Face Transformer: Towards High Fidelity and Accurate Face Swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 668–677.

[9] Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or Epistemic? Does It Matter? *Structural Safety* 31 (03 2009), 105–112. doi:10.1016/j.strusafe.2008.06.020

[10] Paul Goldberg, Christopher Williams, and Christopher Bishop. 1997. Regression with Input-dependent Noise: A Gaussian Process Treatment. In *Advances in Neural Information Processing Systems*, M. Jordan, M. Kearns, and S. Solla (Eds.), Vol. 10. MIT Press. https://proceedings.neurips.cc/paper_files/paper/1997/file/afe434653a898da20044041262b3ac74-Paper.pdf

[11] Xiaobin Hu, Xu Peng, Donghao Luo, Xiaozhong Ji, Jinlong Peng, Zhengkai Jiang, Jiangning Zhang, Taisong Jin, Chengjie Wang, and Rongrong Ji. 2024. DiffuMatting: Synthesizing Arbitrary Objects with Matting-Level Annotation. In *European Conference on Computer Vision*. Springer, 396–413.

[12] Yihan Hu, Yiheng Lin, Wei Wang, Yao Zhao, Yunchao Wei, and Humphrey Shi. 2024. Diffusion for natural image matting. In *European Conference on Computer Vision*. Springer, 181–199.

[13] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson WH Lau. 2022. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1140–1147.

[14] Alex Kendall and Yarin Gal. 2017. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf

[15] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. 2020. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5549–5558.

[16] Jizhizi Li, Jing Zhang, Stephen J Maybank, and Dacheng Tao. 2022. Bridging composite and real: towards end-to-end deep image matting. *International Journal of Computer Vision* 130, 2 (2022), 246–266.

[17] Jizhizi Li, Jing Zhang, and Dacheng Tao. 2021. Deep automatic natural image matting. *arXiv preprint arXiv:2107.07235* (2021).

[18] Yaoyi Li and Hongtao Lu. 2020. Natural image matting via guided contextual attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11450–11457.

[19] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2021. Real-time high-resolution background matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8762–8771.

[20] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. 2022. Robust high-resolution video matting with temporal guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 238–247.

[21] Qinglin Liu, Shengping Zhang, Quanling Meng, Ru Li, Bineng Zhong, and Liqiang Nie. 2023. Rethinking Context Aggregation in Natural Image Matting. *arXiv preprint arXiv:2304.01171* (2023).

[22] Sebastian Lutz, Konstantinos Amplianitis, and Aljosa Smolic. 2018. Alphagan: Generative adversarial networks for natural image matting. *arXiv preprint arXiv:1807.10088* (2018).

[23] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. 2018. On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 98–105.

[24] GyuTae Park, SungJoon Son, JaeYoung Yoo, SeHo Kim, and Nojun Kwak. 2022. Matteformer: Transformer-based image matting via prior-tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11696–11706.

[25] Kwanyong Park, Sanghyun Woo, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. 2023. Mask-guided Matting in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1992–2001.

[26] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. 2020. Attention-guided hierarchical structure aggregation for image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13676–13685.

[27] Hongje Seong, Seoung Wug Oh, Brian Price, Euntai Kim, and Joon-Young Lee. 2022. One-trimap video matting. In *European Conference on Computer Vision*. Springer, 430–448.

[28] Lingxue Song, Dihong Gong, Zhifeng Li, Changsong Liu, and Wei Liu. 2019. Occlusion robust face recognition based on mask learning with pairwise differential siamese network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 773–782.

[29] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. 2021. Semantic image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11120–11129.

[30] Yanan Sun, Guanzhi Wang, Qiao Gu, Chi-Keung Tang, and Yu-Wing Tai. 2021. Deep video matting via spatio-temporal alignment and aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6975–6984.

[31] Zhonglin Sun, Chen Feng, Ioannis Patras, and Georgios Tzimiropoulos. 2024. LAFS: Landmark-based Facial Self-supervised Learning for Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1639–1649.

[32] Kenny TR Voo, Liming Jiang, and Chen Change Loy. 2022. Delving into high-quality synthetic face occlusion segmentation datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4711–4720.

[33] Zidu Wang, Xiangyu Zhu, Tianshuo Zhang, Baiqin Wang, and Zhen Lei. 2024. 3D Face Reconstruction with the Geometric Guidance of Facial Part Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1672–1682.

[34] W. A. Wright. 1999. Bayesian approach to neural-network modeling with input uncertainty. *IEEE transactions on neural networks* 10 6 (1999), 1261–70. https://api.semanticscholar.org/CorpusID:28445688

[35] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. 2017. Deep image matting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2970–2979.

[36] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. 2016. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5525–5533.

[37] Xiangnan Yin and Liming Chen. 2022. FaceOcc: A diverse, high-quality face occlusion dataset for human face extraction. *arXiv preprint arXiv:2201.08425* (2022).

[38] Xiangnan Yin, Di Huang, Zehua Fu, Yunhong Wang, and Liming Chen. 2023. Segmentation-reconstruction-guided facial image de-occlusion. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–8.

[39] Qihang Yu, Jianming Zhang, He Zhang, Yilin Wang, Zhe Lin, Ning Xu, Yutong Bai, and Alan Yuille. 2021. Mask guided matting via progressive refinement network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1154–1163.

[40] Xiong Zhang, Hongsheng Huang, Jianchao Tan, Hongmin Xu, Cheng Yang, Guozhu Peng, Lei Wang, and Ji Liu. 2021. Hand image understanding via deep multi-task learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11281–11292.

[41] Yunke Zhang, Chi Wang, Miaomiao Cui, Peiran Ren, Xuansong Xie, Xian-Sheng Hua, Hujun Bao, Qixing Huang, and Weiwei Xu. 2021. Attention-guided temporally coherent video object matting. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5128–5137.