dugMatting: Decomposed-Uncertainty-Guided Matting

Jiawei Wu [1] **Changqing Zhang** [2] **Zuoyong Li** [3] **Huazhu Fu** [4] **Xi Peng** [5] **Joey Tianyi Zhou** [4 6]

Abstract

Cutting out an object and estimating its opacity mask, known as image matting, is a key task in image and video editing. Due to the highly ill-posed issue, additional inputs, typically user-defined trimaps or scribbles, are usually needed to reduce the uncertainty. Although effective, it is either time consuming or only suitable for experienced users who know where to place the strokes. In this work, we propose a decomposed-uncertainty-guided matting (dugMatting) algorithm, which explores the explicitly decomposed uncertainties to efficiently and effectively improve the results. Basing on the characteristic of these uncertainties, the epistemic uncertainty is reduced in the process of guiding interaction (which introduces prior knowledge), while the aleatoric uncertainty is reduced in modeling data distribution (which introduces statistics for both data and possible noise). The proposed matting framework relieves the requirement for users to determine the interaction areas by using simple and efficient labeling. Extensively quantitative and qualitative results validate that the proposed method significantly improves the original matting algorithms in terms of both efficiency and efficacy.

[1]College of Mechanical and Electrical Engineering, Fujian Agriculture and Forestry University, Fuzhou, China [2]College of Intelligence and Computing, Tianjin University, Tianjin, China [3]Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou, China [4]Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore [5]College of Computer Science, Sichuan University, Chengdu, China [6]Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR), Singapore. Correspondence to: Changqing Zhang <zhangchangqing@tju.edu.cn>, Zuoyong Li <fzulzytdq@126.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

1. Introduction

Digital image matting is the estimation of the opacity of foreground or background from an image, which is one of the fundamental elements in many applications, e.g., compositing live-action and rendered elements together, and performing local color corrections. Specifically, given an image I, image matting can be regarded as a linear combination of foreground $F \in \mathbb{R}^{H\times W\times C}$ and background $B \in \mathbb{R}^{H\times W\times C}$ with the alpha matte $\mu \in [0,1]^{H\times W}$ as follows:

$$I_m = \mu_m F_m + (1-\mu_m)B_m,$$

where $m = (x, y)$ denotes the pixel position.

Since the estimation of μ without any extra information is a highly ill-posed problem, traditional algorithms (Levin et al., 2007; Chen et al., 2013; Lutz et al., 2018; Xu et al., 2017; Li & Lu, 2020; Liu et al., 2021b; Park et al., 2022) usually introduce a trimap to confine the solution space. The trimap separates a picture into two known foreground and background regions along with an unknown transition region. Hence, the matting task is simplified as the problem of estimating the opacity in the transition region. Based on this simplification, the recently proposed matteformer (Park et al., 2022) achieves the state-of-the-art performance. However, drawing a suitable trimap is still time-consuming and tedious. For some complex cases, it will even cost more than 10 minutes (Wei et al., 2021).

Recently, some trimap-free matting algorithms attempt to eliminate the model dependence on the prior labeling. However, the performance of trimap-free methods (Li et al., 2022; Ke et al., 2022; Qin et al., 2020; Chen et al., 2018; Li et al., 2021) still lags far behind the trimap-based methods. The inherent reason is that these models cannot determine which foreground target should be extracted without the guidance of trimap. Therefore, existing trimap-free methods are only able to extract the class-specific objects (e.g., portrait, animal) or salient objects after training on large-scale matting data. Moreover, trimap-free methods is powerless when users want to choose a new category. To balance the efficiency and effectiveness, some novel interactive strategies have been introduced for matting. With user scribbles or clicks, interactive matting achieves similar performance to the trimap-based approaches in relatively low labeling cost (Wei et al., 2021). However, a promising outcome

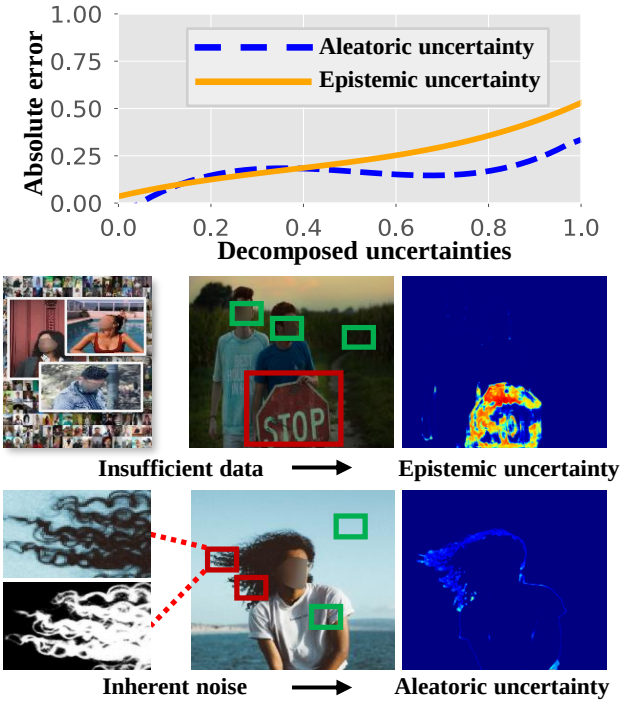


Figure 1. Motivation of the proposed dugMatting. The matting performance could be significantly improved by reducing the decomposed epistemic and aleatoric uncertainties (top row), where these uncertainties are ubiquitous in learning-based image matting (middle and bottom rows). The epistemic uncertainty cannot be reduced by model but can be reduced by user interaction, while the aleatoric uncertainty is difficult to be reduced by human but can be reduced by handling the data noise. Therefore, it is attractive to decompose these uncertainties and exploit them accordingly.

usually requires multiple interactions because the interactions heavily rely on user experience, leading to long-term interaction (the shortest click interaction method still takes about 20 seconds (Wei et al., 2021)). Besides, the matting performance may be unstable due to the ambiguity of the user interaction.

To relieve the restriction of user experience, we propose a decomposed-uncertainty-guided matting (dugMatting) algorithm, which elegantly exploits the decomposed epistemic and aleatoric uncertainties. As shown in Figure 1, the epistemic uncertainty basically results from insufficient training data while the aleatoric uncertainty often appears in transition regions due to the inherent noise. Based on the observation that the absolute error is highly correlated with the these uncertainties, a natural question is *can we effectively reduce the decomposed uncertainty*?

Contribution. Epistemic uncertainty is often due to a lack of training data and thus it is difficult to be reduced by models themselves, while it can be reduced by interaction. Aleatoric uncertainty refers to the uncertainty inherent in the observations, e.g., measurement noise or inaccurate labeling, which is more intricate but can be reduced by handling possible noise, e.g., using data augmentation (Ning et al., 2022; Sambyal et al., 2022). We propose a decomposed-uncertainty-guided matting framework, where the epistemic uncertainty (Kendall & Gal, 2017; Amini et al., 2020) is used to identify proposal regions for user interaction. Accordingly, users only need labeling these regions. To reduce the aleatoric uncertainty, a plug-and-play module based on the estimated data distribution is devised where the augmentation is realized. Specifically, we model the matting output as a Normal-Inverse-Gamma distribution, which hierarchically characterizes the uncertainties and accordingly promotes both regression accuracy and trustworthiness (Amini et al., 2020). Different from the standard setting, the Normal-Inverse-Gamma distribution depends on both the input image and interaction. Therefore, multiple interactions on an image yield multiple NIG distributions, where we introduce NIG summation (Ma et al., 2021; Qian, 2018) to combine these multiple NIG distributions improving the stability. The contributions of this work are summarized as follows:

- For the first time, we reveal the relationship between epistemic/aleatoric uncertainties and the matting error, and thus transform the matting promotion into the problem of epistemic/aleatoric uncertainties reduction.
- We propose a decomposed-uncertainty-guided matting algorithm, where the epistemic uncertainty is utilized to actively provide interaction proposals for users and the aleatoric uncertainty is used to guide the matte refinement in a plug-and-play module.
- We conduct extensive experiments on multiple real-world benchmarks, which demonstrate that the proposed method not only improves the performance of trimap-based matting, but also enables trimap-free matting to extract novel foreground.

2. Related Work

2.1. Image Matting

Image matting refers to extracting interesting foreground or background with fine details from an image, which can be divided into prior-based matting (Levin et al., 2007; Lutz et al., 2018; Xu et al., 2017; Yu et al., 2021c; Park et al., 2022) and prior-free matting (Li et al., 2022; Ke et al., 2022; Chen et al., 2018; Li et al., 2021; Qin et al., 2020). The prior-based matting methods require an additional prior for constraining the solution space. One typical trimap separates an image into foreground, background, and transition regions, where only the opacity of transition regions is unknown. Before the deep learning period, some well-established methods (Zheng & Kambhamettu, 2009; Chen et al., 2013; Levin et al., 2007; Grady et al., 2005; Chuang et al., 2001; Feng et al., 2016; He et al., 2011) solve the matting task based on

trimap prior. For example, the closed-form matting (Levin et al., 2007) derives a cost function based on local smoothing of foreground and background colors, and the globally optimal alpha matting is accordingly induced by solving a sparse system of linear equations. In the era of deep learning, data-driven methods have emerged in matting community, exhibiting much better performance than conventional methods. For example, deep image matting (DIM) (Xu et al., 2017) uses a convolutional network to refine the alpha matte predicted under the encoder-decoder framework, allowing for higher accuracy and sharper edges. A guided contextual attention block is designed in GCANet (Li & Lu, 2020) to integrate the alpha stream information and image information, and improve the details of matting as well. LPFNet (Liu et al., 2021b) models the long-range context features outside the reception fields to improve the alpha matte results. To relieve the load in manually constructing a trimap, the prior-free methods often divide the matting task into a triamp generation and a trimap-based matting subtasks (Li et al., 2022). However, these trimap-free methods fail to handle arbitrary foreground due to the model ambiguity without guidance.

2.2. Uncertainty Estimation

Uncertainty estimation in deep networks has attracted significant attention (Buisson et al., 2010; Gal & Ghahramani, 2016; Kendall & Gal, 2017; Amini et al., 2020; Sensoy et al., 2018; Angelopoulos et al., 2022; Zhou & Levine, 2021), especially when the systems are deployed in safety-critical tasks such as autonomous car control and medical diagnosis. Basically, uncertainty can be roughly divided into aleatoric uncertainty and epistemic uncertainty, in which aleatoric uncertainty captures noise inherent in the observations and epistemic uncertainty captures our ignorance about which model generated our collected data (Kendall & Gal, 2017). For modeling aleatoric uncertainty, the network often outputs a Gaussian distribution with a learnable variance. For modeling epistemic uncertainty, Bayesian-based methods (Weise & Woger, 1993; Maddox et al., 2019; Oakley & O'Hagan, 2002; Daxberger et al., 2021) form a predictive distribution by marginalizing the distribution over model parameters. To reduce the computation of Bayesian network, dropout or ensemble are used to approximate variational Bayesian inference (Buisson et al., 2010; Gal & Ghahramani, 2016), but these methods require multiple forwards. In contrast, some models directly predict the parameters of conjugate prior distribution on the predicted target distribution. Then, one forward pass can estimate the target and the associated uncertainty. Most of those models focus on classification and thus usually estimate the parameters of a Dirichlet distribution (Biloš et al., 2019; Sensoy et al., 2018; Charpentier et al., 2020; Stadler et al., 2021; Nandy et al., 2020). Since image matting is an intrinsically regression problem, we introduce a Normal-Inverse-Gamma (NIG) distribution (Kuleshov et al., 2018) to characterize the uncertainty.

3. Proposed Method

3.1. Preliminary of Evidence-based Uncertainty

We briefly introduce the regression under the evidence-based uncertainty estimation. Regression task can be solved from a maximum likelihood perspective with Gaussian distribution. Given the training data $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$, maximum likelihood estimation (MLE) is achieved by minimizing the negative log likelihood loss function

$$\mathcal{L}_i(\theta) = \frac{(y_i - \mu)^2}{2\sigma^2} + \log \sigma,$$

where θ denotes the parameters of matting network, μ and σ denote the mean and variance parameters of Gaussian distribution respectively, which are typically learned through deep neural networks. Existing matting networks target at learning the alpha matte (mean μ) only. When μ and σ are all learnable, the likelihood function successfully models the aleatoric uncertainty (variance), also known as the data uncertainty. However, epistemic uncertainty, also known as model uncertainty, often requires additional estimation based on the Bayesian framework, e.g., MC Dropout (Gal & Ghahramani, 2016) and ensemble (Buisson et al., 2010).

To jointly model aleatoric and epistemic uncertainties, the mean μ and variance σ^2 are assumed to be drawn from Gaussian and Inverse-Gamma distributions, respectively. Then the Normal Inverse-Gamma (NIG) distribution $\text{NIG}(\gamma, \omega, \alpha, \beta)$ can be considered as a higher-order conjugate prior of the Gaussian distribution

$$\begin{aligned} y_i \sim & \mathcal{N}(\mu, \sigma), \\ \mu \sim \mathcal{N}(\gamma, \sigma^2\omega^{-1}), & \qquad \sigma^2 \sim \Gamma^{-1}(\alpha, \beta), \end{aligned}$$

where $\Gamma(\cdot)$ denotes the gamma function. In this case, the distribution of y takes the form of a $\text{NIG}(\gamma, \omega, \alpha, \beta)$ distribution

$$p(\mu, \sigma|\gamma, \omega, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\sqrt{\omega}}{\sigma\sqrt{2\pi}} (\frac{1}{\sigma^2})^{\alpha+1} \exp -\frac{2\beta + \omega(\sigma - \mu)^2}{2\sigma^2},$$

where $\gamma \in R$, $\omega > 0$, $\alpha > 1$ and $\beta > 0$. The total evidence is the sum of all vitual-observations counts $2\omega + \alpha$. To solve the NIG distribution during training phase, the following loss (Amini et al., 2020) is induced to minimize the negative log likelihood

$$\begin{aligned} \mathcal{L}^{NLL}(\theta) = & \frac{1}{2}\log(\frac{\pi}{\omega}) - \alpha\log(\Omega) + \\ & (\alpha + \frac{1}{2})\log((y - \gamma)^2\omega + \Omega) + \log\Phi, \end{aligned} \tag{1}$$

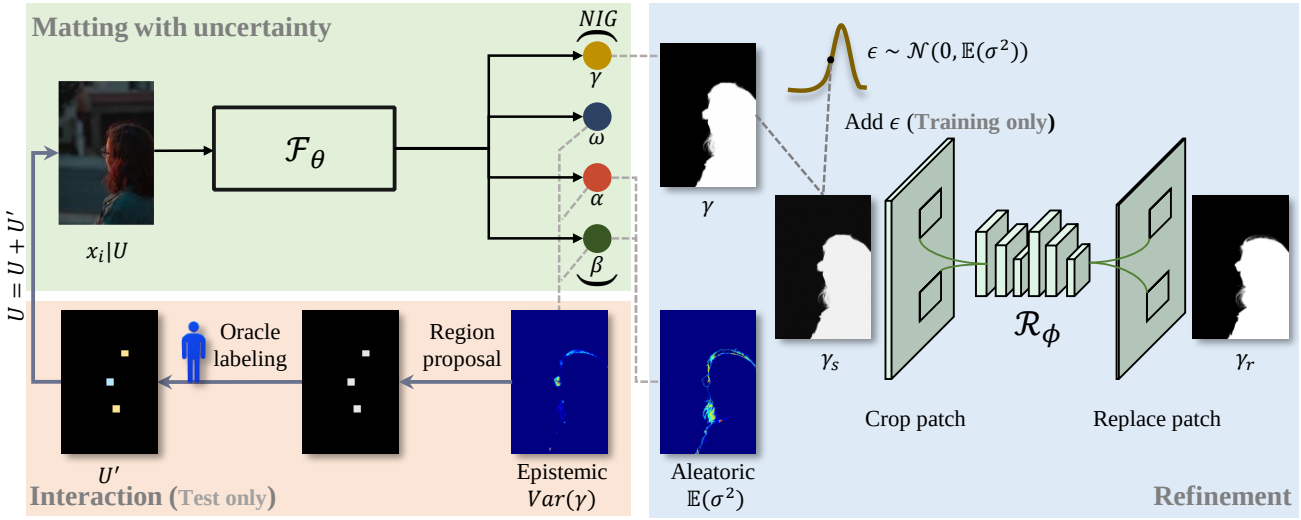


Figure 2. Illustration of the proposed decomposed-uncertainty-guided matting framework. The matting network fits a NIG distribution, proposing interactive regions for user based on epistemic uncertainty and detail regions for refined module based on aleatoric uncertainty.

where $\Omega = 2\beta(1+\omega)$ and $\Phi = \left(\frac{\Gamma(\alpha)}{\Gamma(\alpha+\frac{1}{2})}\right)$. To further constrain the incorrect evidence, a regularizer is introduced in the total loss

$$\mathcal{L}_{NIG}(\theta) = \mathcal{L}^{NLL}(\theta) + \lambda \mathcal{L}^{R}(\theta),$$

where $\mathcal{L}^R(\theta) = |y_i - \gamma| \cdot (2\omega + \alpha)$ is the penalty for incorrect evidence, and the coefficient $\lambda > 0$ balances these two loss terms.

3.2. Integrating Uncertainty into Matting

Image matting can be considered as a regression task, where the output is the alpha matte $\mu \in [0, 1]$ conditioning on the user map U

$$\mu = \mathcal{F}_\theta(x_i|U),$$

where $\mathcal{F}_\theta$ denotes the matting network, and the user map U is empty for trimap-free matting. In order to characterize uncertainty for existing matting networks, we propose to replace the deterministic output with a NIG distribution following Section 3.1

$$NIG(\gamma, \omega, \alpha, \beta) = \mathcal{F}_\theta(x_i|U),$$

where $\gamma \in [0, 1]$, $\omega > 0$, $\alpha > 1$, and $\beta > 0$. Specifically, we first extend the last layer of matting network to output $\gamma, \omega, \alpha, \beta$ by four independent linear layers with shared features as shown in Figure 2. Then, we apply activation functions $sigmoid, softplus, softplus + 1, softplus$ for $\gamma, \omega, \alpha, \beta$ to ensure the proper ranges. Although simple, the modification well fits most existing matting networks. Accordingly, the aleatoric and epistemic uncertainties are obtained as

$$\underbrace{\mathbb{E}[\sigma^2] = \frac{\beta}{\alpha - 1}}_{aleatoric}, \quad \underbrace{Var[\gamma] = \frac{\beta}{\omega(\alpha - 1)}}_{epistemic}.$$

Algorithm 1 Uncertainty-Guided Interaction.

Input: Epistemic uncertainty u_{epis}, predicted matte γ, input image x, threshold t, patch number K, and selection number N.
Initialization: Initialize the user map U.
Divide u_{epis} into $K \times K$ patches.
Compute the patch-level uncertainty $u_p \in \mathbb{R}_+^{K \times K}$.
$\mathcal{P} \leftarrow$ Top N uncertainty patches from $\{u_p | u_p > t\}$.
for p *in* $\mathcal{P}$ **do**
  Calculate the index I of p in input image x.
  Users select a label from foreground, background, or transition for $x[I]$.
  Update user map U.
end
Output: User map U.

3.3. Epistemic Uncertainty-based Interaction

Traditional interaction (Wei et al., 2021; Ding et al., 2022) implicitly contains two steps: users first empirically locate the interacted regions and then conduct interactive operation by trimap, scribble, or click. In our method, we estimate the epistemic uncertainty to automatically determine the interaction regions and then users could only select labels (foreground, background, or transition) for them as shown in Figure 3. This novel interaction avoids the time-consuming region searching. Specifically, we divide the epistemic un-

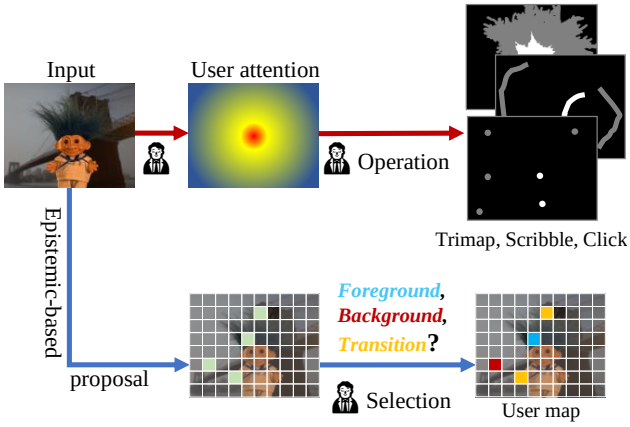


Figure 3. The proposed interaction allows the users to focus on selection.

certainty map into $K \times K$ patches, where the patch-level epistemic uncertainty is the average on all pixels in each patch. Then, the proposal patch set for interaction is constructed satisfying two conditions: top N patch-level epistemic uncertainty and greater than a threshold t. Finally, users select a label for each proposal patch. The simplified interactive process is summarized in Algorithm 1.

To incorporate the results in previous interactions for stabilization, a direct way is to integrate the corresponding NIG distributions into a uniform one. A natural way is using the following additive way

$$NIG(\gamma, \omega, \alpha, \beta) = \frac{1}{M} \sum_{m=1}^{M} NIG(\gamma_m, \omega_m, \alpha_m, \beta_m),$$

where M denotes the number of interactions. Although simple in form, unfortunately, it is intractable to infer the parameters for the fused NIG distribution since there is no closed-form solution. Therefore, inspired by multi-modal learning (Ma et al., 2021) and multi-source learning (Qian, 2018), we employ the simple NIG summation operation to approximately solve this problem

$$\begin{aligned} NIG(\gamma, \omega, \alpha, \beta) \triangleq\ & NIG(\gamma_1, \omega_1, \alpha_1, \beta_1) \\ & \oplus NIG(\gamma_2, \omega_2, \alpha_2, \beta_2) \\ & \oplus \cdots \\ & \oplus NIG(\gamma_M, \omega_M, \alpha_M, \beta_M), \end{aligned} \tag{2}$$

where M denotes the number of interaction, $\oplus$ denotes the summation operation of two NIG distributions as follows,

$$\oplus \begin{cases} \gamma = (\omega_1 + \omega_2)^{-1}(\omega_1\gamma_1 + \omega_2\gamma_2), \\ \omega = \omega_1 + \omega_2, \\ \alpha = \alpha_1 + \alpha_2 + \dfrac{1}{2}, \\ \beta = \beta_1 + \beta_2 + \dfrac{1}{2}\omega_1(\gamma_1 - \gamma)^2 + \dfrac{1}{2}\omega_2(\gamma_2 - \gamma)^2. \end{cases}$$

The NIG summation can reasonably make use of predictions with different qualities. Specifically, the parameter ω indicates the confidence of a NIG distribution for the mean γ. If one matte is more confident with its prediction, then it will contribute more to the final prediction. Moreover, β directly reflects both aleatoric uncertainty and epistemic uncertainty which consists of two parts, i.e., the sum of β_1 and β_2 from multiple mattes and the variance between the final prediction and that of every single matte.

3.4. Aleatoric Uncertainty-based Refinement

Some methods (Sambyal et al., 2022; Ning et al., 2022) constrain invariant predictions for the simulated inherent noise by data augmentation, i.e., enhancing the robustness by explicitly modeling noise to reduce the aleatoric uncertainty. Given the noise $\epsilon \sim \mathcal{N}(0, \varepsilon)$ for input χ, a simple way to reduce aleatoric uncertainty is constraining consistent prediction for samples from $\mathcal{N}(\chi, \varepsilon)$ (Sambyal et al., 2022). However, characterizing the noise of the data requires additional self-supervised training, such as image reconstruction. To simplify the training steps, we propose a plug-and-play module $\mathcal{R}_\phi$ to reduce aleatoric uncertainty and also to refine the matting details. Instead of modeling the noise during the input, we directly model the output noise in terms of the aleatoric uncertainty $\mathbb{E}(\sigma^2)$. In other words, we regard the matting output γ as χ, and the noise $\epsilon \sim \mathcal{N}(0, \mathbb{E}(\sigma^2))$, and then, we attempt to keep the consistent prediction for data sampling from $\mathcal{N}(\gamma, \mathbb{E}(\sigma^2))$. Furthermore, we use the variance $Var(\sigma^2)$ to filter out the regions whose aleatoric uncertainty $\mathbb{E}(\sigma^2)$ may be inaccurate. The $Var(\sigma^2)$ (Cook, 2008) is defined as

$$Var[\sigma^2] = \frac{\beta^2}{(\alpha - 1)^2(\alpha - 2)},$$

where $\alpha > 2$.

The objective of the refinement module is to restore high-aleatoric uncertainty matting details without redundant calculation, thus it only concentrates on local patch refinement. We first obtain the coarse matte γ_s, sampling once from $\mathcal{N}(\gamma, \mathbb{E}(\sigma^2))$ due to small variance as shown in Figure 2. Then, we use OTSU (Otsu, 1979) and $Var[\sigma^2]$ to adaptively select the pixels of reliable high aleatoric uncertainty. Finally, the 32×32 patches centered on the selected pixels corresponding to γ_s are fed into our refinement module $\mathcal{R}_\phi$, and the obtained predictions replace the coarse matte corresponding position to obtain the refined alpha matte γ_r.

3.5. Optimization

We train the proposed dugMatting in two stages to enhance the stability, i.e., optimizing $\mathcal{F}_\theta$ to empower the epistemic uncertainty-based interaction and $\mathcal{R}_\phi$ to refine details.

For the optimization of $\mathcal{F}_\theta$, intuitively, we need to simu-

Table 1. Comparison results on the benchmarks P3M-500-P (Li et al., 2022) and P3M-500-NP (Li et al., 2022). ‡, † denote predictions without and with user map, respectively. For all metrics, the smaller value indicates the better performance.

	P3M-500-P							P3M-500-NP						
Method	SAD	MSE	MAD	Grad	SAD_{bf}	SAD_t	Conn	SAD	MSE	MAD	Grad	SAD_{bf}	SAD_t	Conn
SHM (Chen et al., 2018)	26.84	1.26	1.65	20.18	16.90	9.94	23.30	30.20	1.46	1.93	20.31	17.99	12.21	26.06
U^2Net (Qin et al., 2020)	73.48	1.99	4.51	33.06	48.54	26.91	53.81	70.67	1.89	4.51	34.89	42.75	27.91	53.29
MODNet (Ke et al., 2022)	23.86	1.11	1.46	23.74	16.40	7.46	21.02	25.39	1.20	1.61	21.15	17.41	7.98	22.22
GFM (Li et al., 2022)	12.90	0.58	0.79	14.61	5.98	6.93	11.33	17.01	0.85	1.09	14.54	8.84	8.17	14.86
P3MNet (Li et al., 2021)	12.73	0.56	0.78	13.89	5.95	6.78	11.14	16.49	0.80	1.05	12.75	8.97	7.54	14.35
SHM (dugMatting) ‡	21.43	1.26	1.51	17.82	11.57	10.07	19.32	39.67	1.66	2.43	17.23	28.27	11.40	33.88
U^2Net (dugMatting) ‡	60.21	1.76	4.24	28.74	31.66	28.55	47.34	82.67	2.29	5.07	31.65	51.29	31.38	60.12
MODNet (dugMatting) ‡	18.15	0.72	1.04	15.57	9.59	8.55	16.75	35.66	1.49	2.07	16.04	24.26	11.40	32.83
GFM (dugMatting) ‡	9.25	0.40	0.63	13.79	3.18	6.71	9.29	19.01	0.86	1.14	14.14	9.27	9.73	16.45
P3MNet (dugMatting) ‡	10.08	0.46	0.69	14.61	4.03	7.01	10.30	16.12	0.66	0.94	14.15	6.92	9.18	13.81
△ Average gain ‡	-2.13	-0.17	-0.21	0.13	-6.01	1.86	-1.54	14.31	0.15	0.29	1.02	6.67	5.02	7.86
SHM (dugMatting) †	13.87	0.36	0.85	15.21	4.83	9.04	11.42	18.22	0.51	1.11	13.99	6.57	11.65	15.16
U^2Net (dugMatting) †	35.23	1.35	2.16	19.32	22.78	12.45	32.76	39.86	1.67	2.44	20.12	21.26	18.60	33.91
MODNet (dugMatting) †	9.62	0.29	0.55	12.88	2.63	6.98	9.05	11.08	0.33	0.64	11.75	3.19	7.88	10.99
GFM (dugMatting) †	7.90	0.23	0.46	12.31	1.29	6.60	6.59	9.55	0.28	0.55	11.01	2.12	7.42	7.95
P3MNet (dugMatting) †	7.72	0.22	0.45	12.56	1.01	6.71	6.42	8.79	0.24	0.51	11.08	1.34	7.47	7.23
△ Average gain †	**-15.09**	**-0.61**	**-0.94**	**-6.64**	**-12.24**	**-3.24**	**-10.87**	**-14.45**	**-0.63**	**-0.98**	**-7.13**	**-12.29**	**-2.15**	**-11.10**

late and supervise different predictions in real interaction process, including the initial prediction, the prediction after interaction, and the fused prediction. To simplify the training process, we analyze the purpose of supervision in the three predictions. The supervision of the initial prediction aims to train the network to conduct matting without user map. The supervision of the prediction after interaction aims to relate the network predictions to interaction, which assumes the user map is generated according to epistemic uncertainty. The supervision of the fused prediction aims to stabilize the the fusion result. Based on above analysis, we can jointly supervise the initial prediction and the prediction after interaction by generating random user map U including the empty case. The details of user map can be found in Appendix B.1. The supervision of the fused prediction can be removed because the fusion strategy in Equation (2) is exactly for stabilization. Therefore, the simplified supervision is similar to the previous interactive matting methods (Wei et al., 2021), which only needs to pass through the model once in each iteration. The loss of the first stage can be expressed as

$$\mathcal{L}_{stage1} = \mathcal{L}_{NIG}(\gamma, \omega, \alpha, \beta; \theta) + \mathcal{L}_M(\gamma; \theta_\gamma),$$

where minimizing $\mathcal{L}_{NIG}$ optimizes the parameters of NIG distribution to replace the regression loss (e.g., l_1 loss or l_2 loss) in common matting methods, and $\mathcal{L}_M$ denotes the additional terms (e.g., Laplacian loss (Li et al., 2022)) about matte in the original matting methods.

For the optimization of $\mathcal{R}_\phi$, we first freeze the parameters θ of $\mathcal{F}_\theta$. Then, we can obtain the γ_s and k patches of interest $\{\gamma_s^p\}^k$ according to Section 3.4. The l_1 distance with the ground truth y in matte and gradient map are used for supervision. The loss of the second stage is

$$\mathcal{L}_{stage2} = ||y - \gamma_r||_1 + ||\nabla y - \nabla \gamma_r||_1,$$

where $\gamma_r = \mathcal{R}_\phi(\gamma_s^k, \gamma_s), \gamma_s = (\gamma + \epsilon), \epsilon \sim \mathcal{N}(0, \mathbb{E}(\sigma^2))$.

4. Experiments

4.1. Experimental Setup

Dataset. We conduct extensive experiments on standard natural matting dataset Composition-1k (Xu et al., 2017) and the real-world portrait dataset P3M-10K (Li et al., 2021). Composition-1k (Xu et al., 2017) contains 43,100 synthetic images for training and 1000 synthetic images for testing. P3M-10K (Li et al., 2021) consists of $10,000$ anonymized high-resolution portrait images with face obfuscation, containing $9,421$ images for training and 500 images denoted as P3M-500-P for testing. Besides, for P3M-10K there are additional 500 public Internet images without face obfuscation to test the matting performance on regular portrait images, denoted as P3M-500-NP.

Implementation Details. For class-specific matting, we train all models with the same data augmentations setting for a fair comparison, including random horizontal flipping, random blurring, random sharpen, random shadow, and then random cropping to 512×512 in the end. All models are optimized using the Adam optimizer (Kingma & Ba, 2014), and the base learning rate is set to 1×10^{-3} with the cosine learning rate scheduler (He et al., 2019), 100 epochs iteration, and batch size of 16. For natural image matting, we use the standard setting as specified by MatteFormer (Park et al., 2022). Our implementation [1] is based on the open source framework Pytorch . All the experiments were run on two GeForce RTX 3090 GPUs.

Evaluation Metrics. For Composition-1k, we employ mul-

[1]Code is available at https://github.com/Fire-friend/dugMatting.

Table 2. Quantitative comparison results of natural matting on Composition-1K (Xu et al., 2017) benchmark.

Method	User Map	SAD (10^3) ↓	MAD↓	MSE (10^{-3}) ↓	Grad ↓	Conn↓
Learning Based Matting (Zheng & Kambhamettu, 2009)	Trimap	113.9	0.0501	48.0	91.6	122.2
Closed-Form Matting (Levin et al., 2007)	Trimap	168.1	0.0739	91.0	126.9	167.9
KNN Matting (Chen et al., 2013)	Trimap	175.4	0.0771	103.0	124.1	176.4
Deep Image Matting (Xu et al., 2017)	Trimap	50.4	0.0221	14.0	31.0	50.8
AlphaGan (Lutz et al., 2018)	Trimap	52.4	0.0231	30.0	38.0	-
IndexNet (Lu et al., 2019)	Trimap	45.8	0.0201	13.0	25.9	43.7
HAttMatting (Qiao et al., 2020)	Trimap	44.0	0.0193	7.0	29.3	46.4
AdaMatting (Cai et al., 2019)	Trimap	41.7	0.0183	10.0	16.8	-
sampleNet (Tang et al., 2019)	Trimap	40.4	0.0177	9.9	-	-
Fine-Grained Matting (Liu et al., 2021a)	Trimap	37.6	0.0165	9.0	18.3	35.4
Context-Aware Matting (Hou & Liu, 2019)	Trimap	35.8	0.0157	8.2	17.3	33.2
GCA Matting (Li & Lu, 2020)	Trimap	35.3	0.0155	9.1	16.9	32.5
HDMatt (Yu et al., 2021b)	Trimap	33.5	0.0147	7.3	14.5	29.9
MG Matting (Yu et al., 2021c)	Mask	31.5	0.0138	6.8	13.5	27.3
TIMNet (Liu et al., 2021c)	Trimap	29.1	0.0128	6.0	11.5	25.4
SIM (Sun et al., 2021)	Mask	28.0	0.0123	5.8	10.8	24.8
MatteFormer (Park et al., 2022)	Trimap	23.8	0.0104	4.0	8.7	18.9
MG Matting (dugMatting)	w/o	36.5	0.0161	8.5	17.8	33.6
MG Matting (dugMatting)	1-Selection	32.3	0.0142	7.1	14.2	28.6
MG Matting (dugMatting)	2-Selection	30.2	0.0132	6.4	11.8	26.1
MatteFormer (dugMatting)	w/o	34.1	0.0149	5.7	15.6	31.2
MatteFormer (dugMatting)	1-Selection	25.8	0.0112	4.3	9.7	22.3
MatteFormer (dugMatting)	2-Selection	23.4	0.0102	3.9	7.2	18.8

tiple quantitative metrics, i.e., sum of absolute differences (SAD), mean absolute difference (MAD), mean squared error (MSE), gradient (Grad), and connectivity (Conn). For P3M-10K, we also adopt the above metrics and report the additional SAD_{bf} and SAD_t to compute the SAD within the foreground-background regions and transition regions.

4.2. Quantitative Analysis

Class-specific Matting. To validate our methods on class-specific matting task, we compare our algorithm with state-of-the-art trimap-free methods (Chen et al., 2018; Qin et al., 2020; Ke et al., 2022; Li et al., 2022; 2021) on real-world portrait dataset (Li et al., 2021). As shown in Table 1, dugMatting without interaction outperforms the original trimap-free methods on P3M-500-P, demonstrating that the way of modeling uncertainty can improve the matting performance. In addition, dugMatting significantly improves performance when introducing once interaction, particularly by roughly 50% on P3M-500-NP, demonstrating that the interaction is still useful even when dealing with data from different domains.

Natural Image Matting. The natural image matting expects to extract the interesting foreground with the guidance of user interaction. We first investigate the natural matting methods (Zheng & Kambhamettu, 2009; Chen et al., 2013; Xu et al., 2017; Levin et al., 2007; Lutz et al., 2018; Lu et al., 2019; Qiao et al., 2020; Cai et al., 2019; Tang et al., 2019; Liu et al., 2021a; Hou & Liu, 2019; Li & Lu, 2020; Yu et al., 2021b;c) on Composition-1k (Xu et al., 2017). Then, we employ the effective MG Matting (Yu et al., 2021c) and MatteFormer (Park et al., 2022) as foundation models, integrating our method to validate the performance on natural matting task. Since the Composition-1k is a synthetic set, it allows for the extraction of target objects without any initial interaction. However, when dealing with arbitrary images in real-world, we suggest providing an initial user map through a single click and then utilizing our method for further interaction. The quantitative results are shown in Table 2. With only one or two interactions, our dugMatting outperforms advanced trimap-based matting algorithms. The reason is that reducing the decomposed uncertainties can accurately improve the matte. We also conduct experiments to compare the efficiency of existing interaction methods in Appendix C.1.

4.3. Qualitative Analysis

Visual Comparison with State-of-the-art Methods. In Figure 4, we visualize some results for intuitive comparison. Although dugMatting uses a weaker prior, the results is comparable to other trimap-based methods. In addition, benefiting from modeling data noise, dugMatting produces a matte that is more uniform and smooth. For instance, the ground truth of the second example has some local opacity mutations that do not occur in the real world, but dugMatting also achieves a smooth outcome.

Visualization of Step-by-step Results in dugMatting. Figure 5 visualizes the step-by-step results of our dugMatting. Our interaction can effectively improve the incorrect matting

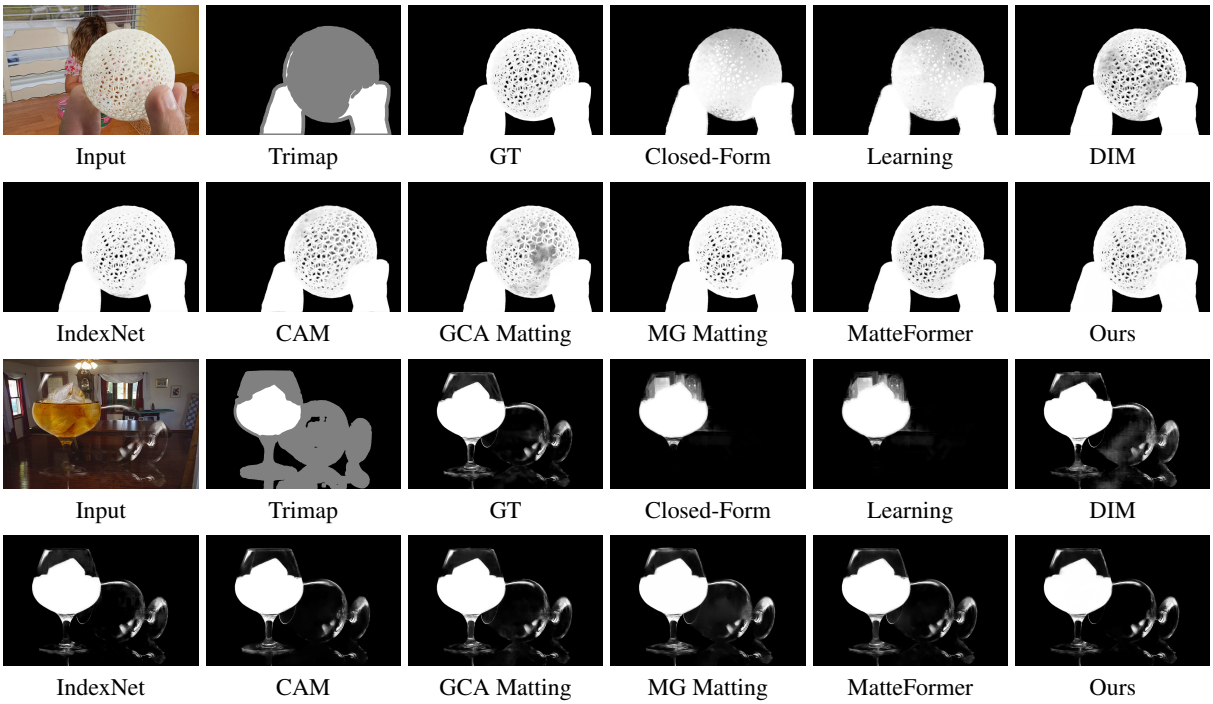


Figure 4. Qualitative examples on the Composition-1k (Xu et al., 2017) test set.

regions, and our refinement module can improve the details. The reason is that external knowledge by user interaction significantly reduces epistemic uncertainty, complementing the unlearned foreground and background patches. Meanwhile, our refinement of modeling high-frequency noise reduces the aleatoric uncertainty, enhancing the robustness in patches containing more details.

Uncertainty Evaluation. We evaluate the uncertainty from two aspects. The first one is to verify the region proposal of our interaction and refinement, while the second one is to validate the ability of uncertainty estimation which is detailed in Appendix C.2. As shown in Figure 6, there are much higher proportion of foreground and background regions with large epistemic uncertainty (a). Thus selecting patches with top K patch-levels epistemic uncertainty enables the user to concentrate on the annotation of foreground and background. We further evaluate the ROC curve between the regions obtained by two strategies and the real transition (b). Our refined aleatoric uncertainty-based algorithm significantly improves the AUC, demonstrating the refined aleatoric uncertainty can improve more details.

4.4. Ablation Study

In this subsection, we first investigate the proposed components and then independently analyze our plug-and-play module. Furthermore, we perform additional experiment to investigate the hyper-parameter of interaction numbers.

Table 3. Ablation study (SAD↓) of the NIG distribution and the proposed module on the P3M-500-P dataset.

Method	Original	w/ NIG	w/ NIG & Module
SHM (Chen et al., 2018)	26.84	24.65	**21.43**
U^2Net (Qin et al., 2020)	73.48	69.76	**60.21**
MODNet (Ke et al., 2022)	23.86	20.04	**18.15**
GFM (Li et al., 2022)	12.90	10.89	**9.25**
P3MNet (Li et al., 2021)	12.73	12.03	**10.38**

Table 4. Ablation study (SAD↓) on our refined module on the P3M-500-P dataset. Baseline uses the original trimap-free methods.

Method	Baseline (Ke et al., 2022)	Gaussian	Module (our)
SAD_f	3.69	3.36	**3.36**
SAD_b	6.46	6.55	**6.23**
SAD_t	9.88	8.75	**8.55**
Aleatoric	0.0021	0.0015	**0.0013**

The Effectiveness of Each Component. We first evaluate the uncertainty integration in matting, i.e., replacing the deterministic output with a Normal-Inverse-Gamma distribution, and then adding the proposed plug-and-play module. As shown in Table 3, both NIG distribution and our refinement module can improve the matting performance over original methods, demonstrating the efficacy of the key components in dugMatting.

The Effectiveness of Reducing Aleatoric Uncertainty.

Figure 5. Visualization of step-by-step results in dugMatting. From left to right are input image, initial prediction, epistemic uncertainty, user map, prediction after interaction, prediction after refinement, respectively.

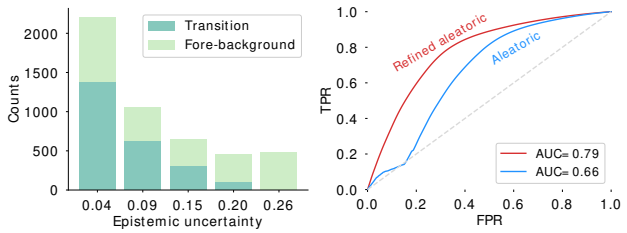


Figure 6. The correlation regions of decomposed uncertainties. The proportion of foreground and background regions is higher in high epistemic uncertainty. ROC of the obtained regions and the real transition regions, refined aleatoric uncertainty-based algorithm achieves better performance.

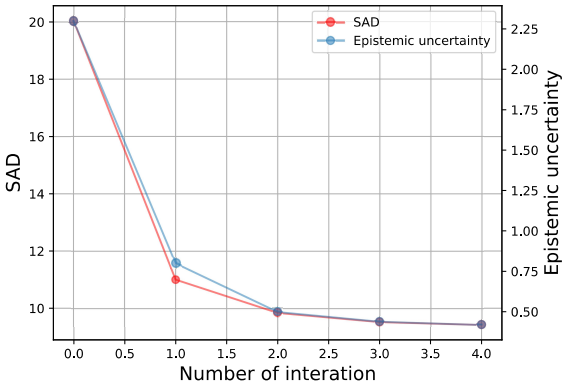


Figure 7. SAD and epistemic uncertainty at different number of interaction.

Following (Sambyal et al., 2022), we compare the augmentation of our module and a Gaussian noise. The variance of Gaussian noise is fixed, determined by the average aleatoric uncertainty of all pixels. The augmentation of our module also belongs to a Gaussian noise, but the variance is dynamic and determined by the aleatoric uncertainty of the current pixel. The result of reducing the aleatoric uncertainty is shown in Table 4. The proposed module achieves the best performance, significantly decreasing the aleatoric uncertainty and improving the performance.

The Hyper-parameter of Interaction Numbers. As shown in Figure 7, regardless of SAD or epistemic uncertainty, the most obvious improvement occurs in the first interaction, and the performance improvement is slight improved after the second interaction. Therefore, in order to balance the performance and interaction time, the interaction number is set as 1 unless otherwise specified.

5. Conclusion

In this paper, we propose a decomposed-uncertainty-guided matting (dugMatting) algorithm for both trimap-free and trimap-based matting. We first introduce epistemic uncertainty to actively propose interactive regions, which simplifies the search of difficult regions by user for trimap-based matting. Besides, we propose a plug-and-play module, which not only reduces the aleatoric uncertainty but also improves the matting details. This is exciting because it first explores different types of uncertainties in an explainable and elegant way in matting. Extensive experiments are conducted on natural matting and class-specific matting which validates that the existing matting methods equipped with dugMatting achieve superior performance than the original ones. It would be interesting to further explore the image structures (e.g., segments) for the goal of further computational efficiency and performance improvement. Another direction for further research is to apply the proposed dugMatting to other related domains such as interactive image segmentation.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (No. 2022YFC3302200), the National Natural Science Foundation of China (No. 61972187, 61976151), and the A*STAR Central Research Fund. The authors appreciate the comments from reviewers.

References

Amini, A., Schwarting, W., Soleimany, A., and Rus, D. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33:14927–14937, 2020.

Angelopoulos, A. N., Kohli, A. P., Bates, S., Jordan, M., Malik, J., Alshaabi, T., Upadhyayula, S., and Romano, Y. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, pp. 717–730. PMLR, 2022.

Biloš, M., Charpentier, B., and Günnemann, S. Uncertainty on asynchronous time event prediction. *Advances in Neural Information Processing Systems*, 32, 2019.

Buisson, L., Thuiller, W., Casajus, N., Lek, S., and Grenouillet, G. Uncertainty in ensemble forecasting of species distribution. *Global Change Biology*, 16(4):1145–1157, 2010.

Cai, S., Zhang, X., Fan, H., Huang, H., Liu, J., Liu, J., Liu, J., Wang, J., and Sun, J. Disentangled image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8819–8828, 2019.

Charpentier, B., Zügner, D., and Günnemann, S. Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts. *Advances in Neural Information Processing Systems*, 33:1356–1367, 2020.

Chen, Q., Li, D., and Tang, C.-K. Knn matting. *IEEE transactions on pattern analysis and machine intelligence*, 35(9):2175–2188, 2013.

Chen, Q., Ge, T., Xu, Y., Zhang, Z., Yang, X., and Gai, K. Semantic human matting. In *Proceedings of the 26th ACM international conference on Multimedia*, pp. 618–626, 2018.

Chuang, Y.-Y., Curless, B., Salesin, D. H., and Szeliski, R. A bayesian approach to digital matting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pp. II–II. IEEE, 2001.

Cook, J. D. Inverse gamma distribution. *online: http://www. johndcook. com/inverse gamma. pdf, Tech. Rep*, 2008.

Daxberger, E., Nalisnick, E., Allingham, J. U., Antorán, J., and Hernández-Lobato, J. M. Bayesian deep learning via subnetwork inference. In *International Conference on Machine Learning*, pp. 2510–2521. PMLR, 2021.

Ding, H., Zhang, H., Liu, C., and Jiang, X. Deep interactive image matting with feature propagation. *IEEE Transactions on Image Processing*, 31:2421–2432, 2022.

Feng, X., Liang, X., and Zhang, Z. A cluster sampling method for image matting via sparse coding. In *European Conference on Computer Vision*, pp. 204–219. Springer, 2016.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Grady, L., Schiwietz, T., Aharon, S., and Westermann, R. Random walks for interactive alpha-matting. In *Proceedings of VIIP*, volume 2005, pp. 423–429, 2005.

He, K., Rhemann, C., Rother, C., Tang, X., and Sun, J. A global sampling method for alpha matting. In *CVPR 2011*, pp. 2049–2056. IEEE, 2011.

He, T., Zhang, Z., Zhang, H., Zhang, Z., Xie, J., and Li, M. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 558–567, 2019.

Hou, Q. and Liu, F. Context-aware image matting for simultaneous foreground and alpha estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4130–4139, 2019.

Ke, Z., Sun, J., Li, K., Yan, Q., and Lau, R. W. Modnet: Real-time trimap-free portrait matting via objective decomposition. AAAI, 2022.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pp. 2796–2804. PMLR, 2018.

Levin, A., Lischinski, D., and Weiss, Y. A closed-form solution to natural image matting. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):228–242, 2007.

Li, J., Ma, S., Zhang, J., and Tao, D. Privacy-preserving portrait matting. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 3501–3509, 2021.

Li, J., Zhang, J., Maybank, S. J., and Tao, D. Bridging composite and real: towards end-to-end deep image matting. *International Journal of Computer Vision*, pp. 1–21, 2022.

Li, Y. and Lu, H. Natural image matting via guided contextual attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11450–11457, 2020.

Liu, C., Ding, H., and Jiang, X. Towards enhancing fine-grained details for image matting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 385–393, 2021a.

Liu, Q., Xie, H., Zhang, S., Zhong, B., and Ji, R. Long-range feature propagating for natural image matting. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 526–534, 2021b.

Liu, Y., Xie, J., Shi, X., Qiao, Y., Huang, Y., Tang, Y., and Yang, X. Tripartite information mining and integration for image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7555–7564, 2021c.

Lu, H., Dai, Y., Shen, C., and Xu, S. Indices matter: Learning to index for deep image matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3266–3275, 2019.

Lutz, S., Amplianitis, K., and Smolic, A. Alphagan: Generative adversarial networks for natural image matting. *arXiv preprint arXiv:1807.10088*, 2018.

Ma, H., Han, Z., Zhang, C., Fu, H., Zhou, J. T., and Hu, Q. Trustworthy multimodal regression with mixture of normal-inverse gamma distributions. *Advances in Neural Information Processing Systems*, 34:6881–6893, 2021.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., and Wilson, A. G. A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Nandy, J., Hsu, W., and Lee, M. L. Towards maximizing the representation gap between in-domain & out-of-distribution examples. *Advances in Neural Information Processing Systems*, 33:9239–9250, 2020.

Ning, Q., Tang, J., Wu, F., Dong, W., Li, X., and Shi, G. Learning degradation uncertainty for unsupervised real-world image super-resolution. In Raedt, L. D. (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 1261–1267. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/176. URL https://doi.org/10.24963/ijcai.2022/176. Main Track.

Oakley, J. and O'Hagan, A. Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika*, 89(4):769–784, 2002.

Otsu, N. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

Park, G., Son, S., Yoo, J., Kim, S., and Kwak, N. Matteformer: Transformer-based image matting via prior-tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11696–11706, 2022.

Qian, H. Big data bayesian linear regression and variable selection by normal-inverse-gamma summation. *Bayesian Analysis*, 13(4):1011–1035, 2018.

Qiao, Y., Liu, Y., Yang, X., Zhou, D., Xu, M., Zhang, Q., and Wei, X. Attention-guided hierarchical structure aggregation for image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13676–13685, 2020.

Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O. R., and Jagersand, M. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404, 2020.

Sambyal, A. S., Krishnan, N. C., and Bathula, D. R. Towards reducing aleatoric uncertainty for medical imaging tasks. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–4. IEEE, 2022.

Sensoy, M., Kaplan, L., and Kandemir, M. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Stadler, M., Charpentier, B., Geisler, S., Zügner, D., and Günnemann, S. Graph posterior network: Bayesian predictive uncertainty for node classification. *Advances in Neural Information Processing Systems*, 34:18033–18048, 2021.

Sun, Y., Tang, C.-K., and Tai, Y.-W. Semantic image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11120–11129, 2021.

Tang, J., Aksoy, Y., Oztireli, C., Gross, M., and Aydin, T. O. Learning-based sampling for natural image matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3055–3063, 2019.

Wei, T., Chen, D., Zhou, W., Liao, J., Zhao, H., Zhang, W., and Yu, N. Improved image matting via real-time user clicks and uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15374–15383, 2021.

Weise, K. and Woger, W. A bayesian theory of measurement uncertainty. *Measurement Science and Technology*, 4(1): 1, 1993.

Xu, N., Price, B., Cohen, S., and Huang, T. Deep image matting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2970–2979, 2017.

Yu, C., Xiao, B., Gao, C., Yuan, L., Zhang, L., Sang, N., and Wang, J. Lite-hrnet: A lightweight high-resolution network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10440–10450, 2021a.

Yu, H., Xu, N., Huang, Z., Zhou, Y., and Shi, H. High-resolution deep image matting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 3217–3224, 2021b.

Yu, Q., Zhang, J., Zhang, H., Wang, Y., Lin, Z., Xu, N., Bai, Y., and Yuille, A. Mask guided matting via progressive refinement network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1154–1163, 2021c.

Zheng, Y. and Kambhamettu, C. Learning based digital matting. In *2009 IEEE 12th international conference on computer vision*, pp. 889–896. IEEE, 2009.

Zhou, A. and Levine, S. Amortized conditional normalized maximum likelihood: Reliable out of distribution uncertainty estimation. In *International Conference on Machine Learning*, pp. 12803–12812. PMLR, 2021.

A. Proof

The marginal likelihood of Normal-Inverse-Gamma distribution by Type-II maximum likelihood technique is defined by

$$
\begin{aligned}
p(y|\tau) &= \int_{\zeta} p(y|\zeta)p(\zeta|\tau)\mathrm{d}\zeta \\
&= \int_{\sigma^2}^{\infty}\int_{\mu=-\infty}^{\infty} p(y|\mu,\sigma^2)p(\mu,\sigma^2|\tau)\mathrm{d}\mu\mathrm{d}\sigma^2 \\
&= \int_{\sigma^2}^{\infty}\int_{\mu=-\infty}^{\infty} p(y|\mu,\sigma^2)p(\mu,\sigma^2|\gamma,\omega,\alpha,\beta)\mathrm{d}\mu\mathrm{d}\sigma^2 \\
&= \int_{\sigma^2}^{\infty}\int_{\mu=-\infty}^{\infty} \left[\sqrt{\frac{1}{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}\right]\left[\frac{\beta}{\omega\alpha}\frac{\sqrt{\omega}}{\sqrt{2\pi\sigma^2}}\left(\frac{1}{\sigma^2}\right)^{\alpha+1}\exp\left\{-\frac{2\beta+\omega(\gamma-\mu)}{2\sigma^2}\right\}\right]\mathrm{d}\mu\mathrm{d}\sigma^2 \\
&= \int_{\sigma^2}^{\infty}\frac{\beta^{\alpha}\sigma^{-3-2\alpha}}{\sqrt{2\pi}\sqrt{1+1/\omega}\Gamma(\alpha)}\exp\left\{-\frac{2\beta+\frac{\omega(y-\gamma)^2}{1+\omega}}{2\sigma^2}\right\}\mathrm{d}\sigma^2 \\
&= \frac{\Gamma(1/2+\alpha)}{\Gamma(\alpha)}\sqrt{\frac{\omega}{\pi}}(2\beta(1+\omega))^{\alpha}\left(\omega(y-\gamma)^2+2\beta(1+\omega)\right)^{-(\frac{1}{2}+\alpha)} \\
&= St\left(y;\gamma,\frac{\beta(1+\omega)}{\omega\alpha},2\alpha\right).
\end{aligned}
$$

Maximizing the likelihood as Equation (1) by using the standard parameterization for Student t distribution makes our model fit the data.

According to $\sigma^2 \sim \Gamma^{-1}(\alpha,\beta)$, the $Var(\sigma^2)$ is derived from

$$Var(\sigma^2) = \mathbb{E}((\sigma^2)^2) - \mathbb{E}((\sigma^2))^2,$$

where

$$
\begin{aligned}
\mathbb{E}((\sigma^2)^n) &= \frac{\beta}{\Gamma(\alpha)}\int_0^{\infty}\sigma^{n-2\alpha-2}\exp(-\beta/\sigma^2)\mathrm{d}\sigma^2 \\
&= \frac{\beta^{\alpha}}{\Gamma(\alpha)}\frac{\Gamma(\alpha-n)}{\beta^{\alpha-n}} \\
&= \frac{\beta^n\Gamma(\alpha-n)}{(\alpha-1)\cdots(\alpha-n)\Gamma(\alpha-n)} \\
&= \frac{\beta^n}{(\alpha-1)\cdots(\alpha-n)},
\end{aligned}
$$

For $\alpha > 1$, we have

$$\mathbb{E}(\sigma^2) = \frac{\beta}{\alpha-1},$$

and for $\alpha > 2$, we have

$$\mathbb{E}((\sigma^2)^2) = \frac{\beta^2}{(\alpha-1)(\alpha-2)}.$$

Accordingly, we can obtain the variance as

$$Var(\sigma^2) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}.$$

B. More Details

B.1. Details of User Map

For the construction of user map U, we randomly sample L patches with 15×15, where L is drawn from a geometric distribution with $p = \frac{1}{6}$. The user map $U \in [-1, 0, 0.5, 1]^{1\times H\times W}$ where foreground is 1, background is -1, transition is 0.5 and unknown is 0.

B.2. Details of Refinement Module

Since the refinement module aims to recover the high-frequency details, we use the Naive Lite-HRNet-18 (Yu et al., 2021a) and bilinear interpolation as the refinement module. The Naive Lite-HRNet-18 can efficiently preserve high-resolution features with only 0.7M parameters.

C. More Experiments

C.1. Resource Comparison of Major Interaction

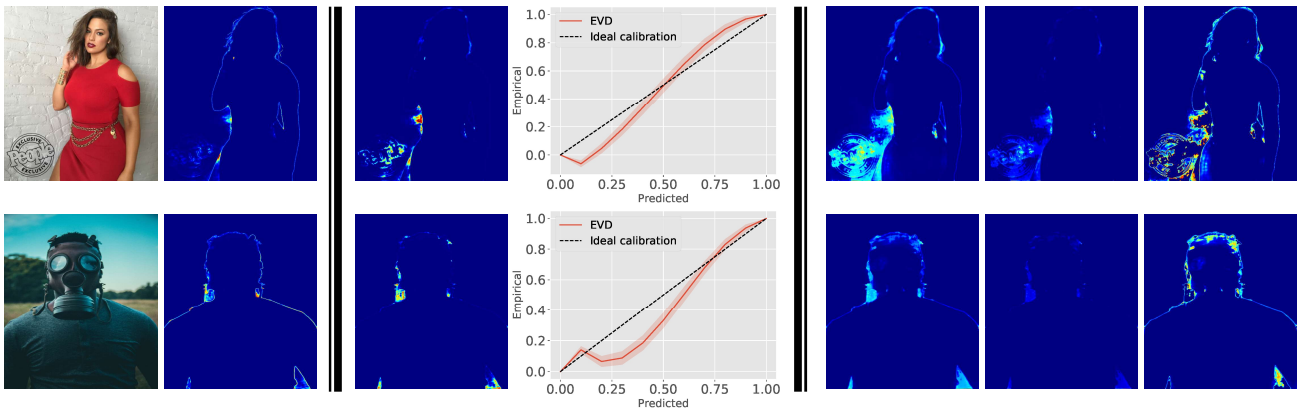
We also conduct a comparison experiment to explore the resource consumption of the major interaction methods. As shown in Table 5, the trimap, scribble, and click methods do not require extra parameters while they need to take times between 17 and 260 seconds. In contrast, our method only takes 8 seconds and requires almost no extra parameters. The reason is that our interaction method actively proposes the interaction area based on the epistemic uncertainty, allowing the user to focus on the annotation. It significantly enhances the interaction efficiency.

Table 5. Comparison results of resource consuming on 10 samples of the Conposition-1K (Xu et al., 2017) benchmark.

Interaction method	Times	Extra Parameters
Trimap	261s	-
Mask	234s	-
scribble	171s	-
Click	17s	-
Selection (ours)	**8s**	0.7M

C.2. Uncertainty Estimation

We evaluate the epistemic uncertainty and aleatoric uncertainty on unseen P3M-500-NP test dataset using MODNet. The input, absolute error, evaluation of epistemic uncertainty and aleatoric uncertainty are depicted in Figure 8. For the evaluation of epistemic uncertainty, we use calibration curves to evaluate the estimation. Calibration curves are computed according to (Kuleshov et al., 2018), and ideally follows $y = x$ to represent, for example, that a target falls in a 90% confidence interval approximately 90% of the time. It is observed that epistemic uncertainty matches error regions in most time. For the evaluation of aleatoric uncertainty, we can find that the aleatoric uncertainty is misestimated in some cases, and the variance of the aleatoric uncertainty can serve as an additional metric to identify these regions. Thus, it is appropriate for our strategy to utilize epistemic uncertainty to identify areas of user interaction and aleatoric uncertainty to guide the refinement of details.



(a) Input (b) Absolute error (c) Epistemic uncertainty (d) Calibration curves (e) Aleatoric uncertainty (f) Variance of aleatoric (g) Refined aleatoric

Figure 8. Uncertainty evaluation of MODNet. Epistemic uncertainty matches error regions in most time. Aleatoric uncertainty may capture erroneous transition regions, the variance of aleatoric uncertainty can help to more precisely indicate transition regions.