Matting Moments: A Unified Data-Driven Matting Engine for Mobile AIGC in Photo Gallery

Yanhao Zhang[1] , **Fanyi Wang**[1] , **Weixuan Sun**[1,2] , **Jingwen Su**[1] , **Peng Liu**[1] , **Yaqian Li**[1] , **Xinjie Feng**[1] , **Zhengxia Zou**[3]
[1]OPPO Research Institute
[2]Australian National University
[3]Beihang University
{zhangyanhao, wangfanyi}@oppo.com

Abstract

Image matting is a fundamental technique in visual understanding and has become one of the most significant capabilities in mobile phones. Despite the development of mobile storage and computing power, achieving diverse mobile Artificial Intelligence Generated Content (AIGC) applications remains a great challenge. To address this issue, we present an innovative demonstration of an automatic system called "Matting Moments" that enables automatic image editing based on matting models in different scenarios. Coupled with accurate and refined matting subjects, our system provides visual element editing abilities and backend services for distribution and recommendation that respond to emotional expressions. Our system comprises three components: 1) photo content structuring, 2) data-driven matting engine, and 3) AIGC functions for generation, which automatically achieve diverse photo beautification in the gallery. This system offers a unified framework that guides consumers to obtain intelligent recommendations with beautifully generated contents, helping them enjoy the moments and memories of their present life.

1 Introduction

Precise matting of salient subjects in an image becomes pivotal in phone applications. And matting results can facilitate secondary creation including intelligent editing and Artificial Intelligence Generated Content (AIGC) [Du *et al.*, 2023]. As a key capability of visual systems, matting substantially impacts the editing efficiency and user experience of mobile phone products. To realize diverse AIGC applications on mobile terminals, matting serves as the foundation for translating technical capabilities into commercial value [Sun *et al.*, 2022; Chen *et al.*, 2022; Hu *et al.*, 2023; Li *et al.*, 2022]. Integrating the visual interface enables users to select and extract desired elements by holding, copying and dragging them. This expedites and simplifies editing or generation.

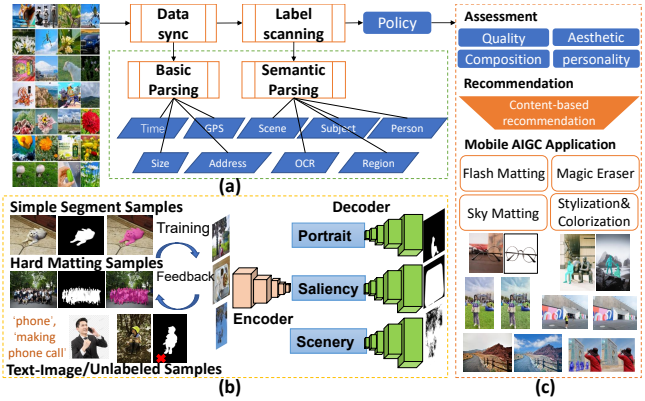

Figure 1: Overview of our unified data-driven system. (a) Photo content structuring. (b) Data-driven matting engine. (c) AIGC functions for generation.

Within the broader context of AIGC, matting functions and capabilities appear to be in high demand among users. They can be extended in several ways, such as through scene parsing and portrait editing. Combining matting abilities with super-resolution and stylization models could enable diverse intelligent creations. To maximize the effectiveness of matting functions, it would be useful to automatically identify matting subjects based on visual and contextual content. This could enable recommending relevant generated galleries to users. This would not only help users showcase the beauty of life's moments, but also improve the user experience by recommending recent and relevant content.

Advances in AI have enabled tremendous potential for automatic generation and creation. A wide range of innovative applications have emerged in this field, particularly AIGC applications on mobile devices. Matting technology can help users quickly and accurately isolate parts of images for editing, such as people, objects, and scenes. This makes it easier for users to perform various editing operations including beautification, retouching, and compositing of specific objects, thereby enhancing entertainment and visual effects. However, automatic recommendation of AIGC in conjunction with matting engines remain underdeveloped. This could improve the overall user experience and promote the development of more efficient and effective creation and generation applications on mobile devices.

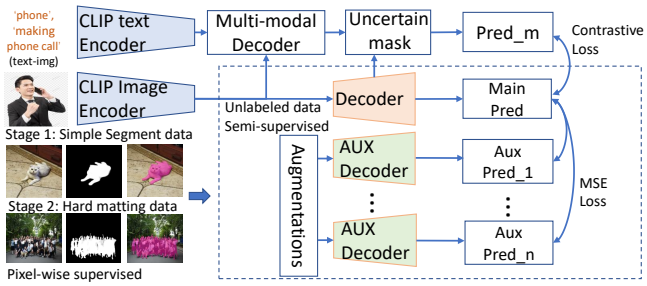

Figure 2: Data-driven matting engine including data and network structure.

In this demonstration, we present a photo gallery application called "Matting Moments". It is a unified system based on diverse matting models and AIGC functions, allowing the consumers to obtain intelligent recommendations with beautifully generated content tailored to their photos. Specifically, our system comprises three modules: 1) photo content structuring, 2) data-driven matting engine, and 3) AIGC functions. Unlike previous works, we focus on editing photos that conform to the moment generation, helping users respond to sharing demands quickly and emotionally.

2 System Framework

We present an overview of our system framework in Figure 1, where users' photo gallery is used as input, and AIGC-generated contents are integrated as output. The workflow involves several steps: Firstly, the photo content structuring module extracts basic and semantic information from each image. Then, the data-driven matting engine guides users to highlight and matte diverse subjects. Finally, the AIGC functions module collects and presents beautiful moments. We will introduce each component in detail below.

2.1 Photo Content Structuring

Photo content structuring aims to parse photos in a gallery using both basic and semantic information. The proposed process requires foundational information about a photo, as shown in Figure 1. Data synchronization is used to obtain file path, size, time, and GPS. Additionally, each photo is fed into a 1000-class tagging model and an OCR model to categorize scenes, objects, and OCR texts through scanning. After categorization, images are scanned for content relevance and aesthetic quality. Quality assessment is conducted based on factors such as sizes, formats, composition, and aesthetic filters. These assessments determine the generation strategy based on the content and scene of each photo.

2.2 Data-driven Matting Engine

We introduce our data collection and training pipeline for the basic matting engine. Due to the diverse types of samples, we employ multiple training strategies [Liang *et al.*, 2022] to incorporate various available data into our training loop. First, we conduct supervised training with pixel-wise labeled samples. Then, we incorporate semi-supervised techniques cited in [Bao *et al.*, 2022; Lv *et al.*, 2022; Piao *et al.*,] and multi-modal learning capabilities to utilize unlabeled samples and text-image samples. As a result, our model can be continuously updated as new data arrives.

Pixel-Wise Supervised Training

Encoder-decoder structure is adopted for the matting model which is well-suited for dense prediction tasks such as U2Net [Qin *et al.*, 2020]. With the labeled samples, we propose a two-step training protocol. **a)** First stage: we train the model on a large dataset containing 160k simple samples to achieve accurate segmentation results. The large number of simple examples facilitates training and improves the model's robustness. **b)** Second stage: we fine-tune the model on a smaller set of 20k hard samples to refine the matting results. This stage focuses on training the model to handle complex scenes and ambiguous boundaries using the robust visual features learned in the first stage. The resulting model has a high capacity for generalization and can handle challenging corner cases.

Semi-Supervised and Multi-Modal Learning

We extend our framework to incorporate these expanding data into our training. Notably, the vast majority of these data do not include pixel-wise annotations and are typically labeled only with text-based tags or left unlabeled [Shin *et al.*, 2022; Cong *et al.*, 2022]. We aim to increase the trained model capacity to process these continuously arriving samples as in Figure 2, incorporating semi-supervised learning and multi-modal processing capabilities into our matting model.

Unlabeled data: We construct our semi-supervised framework following [Ouali *et al.*, 2020]. In addition to the encoder and the main decoder, multiple auxiliary decoders are considered and their inputs are perturbed feature maps of encoder in Figure 2. For an unlabeled sample, we generate a prediction from the main decoder as the pseudo target, and then apply various augmentations to each auxiliary decoder to obtain auxiliary predictions. The consistency between the auxiliary predictions and the pseudo-target is enforced. For training stability, we have regularized the mean squared error (MSE), and the loss is not back-propagated through the main decoder. The shared representation of encoder is improved through the use of a perturbed auxiliary decoder and the consistencies derived from unlabeled samples. Only the main decoder is employed during inference.

Text-image data: We also collect a substantial number of text-image samples. These texts can be easily extracted from manual annotations, web sources, and image caption models available on the market. Thus, we enable our framework further to process text-image samples. With a pretrained multi-modal model [Radford *et al.*, 2021], we propose multi-modal consistency learning and text-aware contrastive learning. For a text-image pair, the prediction is initially made using the image through the main decoder. The text feature intuitively complements the visual feature with semantically significant concepts. Then, we regularize the distance between pure-image prediction and text-image prediction to encourage the model to predict semantically significant salient objects [Liu *et al.*, 2022; Zhang *et al.*, 2022]. In addition, we propose a contrastive learning scheme to better couple image-text relationships. We consider the image-text feature to be the anchor, and encourage the pure-image feature to be adjacent to

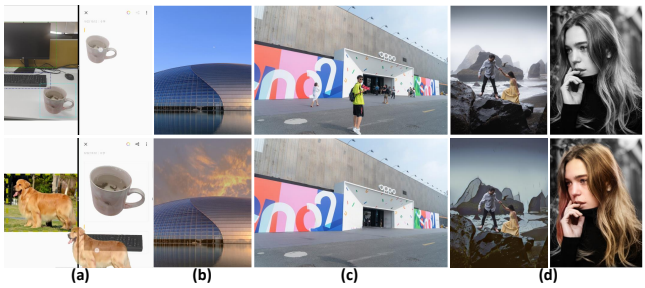

Figure 3: Functions of AIGC examples based on four matting schemes.

MSE	portrait	multi-person	pets	accessories	cards	delicacy	mean
Pixel_Sup	0.0141	0.0658	0.0123	0.0318	0.0179	0.0416	0.0323
+Semi_Sup	0.0114	0.0591	0.0123	0.0608	0.0243	0.0380	0.0297
+Mlt_Sup	0.0088	0.0576	0.0070	0.0329	0.0138	0.0342	**0.0261**
BIoU	portrait	multi-person	pets	accessories	cards	delicacy	mean
Pixel_Sup	0.8429	0.7339	0.8587	0.7732	0.8645	0.8118	0.8127
+Semi_Sup	0.8570	0.7528	0.8556	0.7225	0.8446	0.8263	0.8232
+Mlt_Sup	0.8743	0.7640	0.8814	0.7436	0.8872	0.8350	**0.8403**
MaxF	portrait	multi-person	pets	accessories	cards	delicacy	mean
Pixel_Sup	0.9440	0.9170	0.9615	0.9604	0.9734	0.9289	0.9396
+Semi_Sup	0.9529	0.9225	0.9514	0.9406	0.9587	0.9389	0.9437
+Mlt_Sup	0.9624	0.9271	0.9734	0.9519	0.9777	0.9412	**0.9525**
Conn	portrait	multi-person	pets	accessories	cards	delicacy	mean
Pixel_Sup	0.0166	0.0695	0.0151	0.0362	0.0201	0.0447	0.0353
+Semi_Sup	0.0139	0.0628	0.0151	0.0644	0.0265	0.0411	0.0326
+Mlt_Sup	0.0124	0.0644	0.0101	0.0519	0.0185	0.0397	**0.0309**

Table 1: The comparison of matting results between related learning components in matting engine.

it. Then, we sample some negative texts in order to generate negative image-text features, which are then pushed away from the anchor. Thus, the image encoder is encouraged to explore semantically significant entities.

Hierarchical Training Strategy

Given the diverse data sources, we integrate different training strategies to actively update the model with incoming data. For the images with pixel-wise labels, we categorize them into two groups by difficulties. A two-stage supervised training strategy is adopted. A group of simple samples ensures the quality of the features and the capacity to generalize, and the other group of fewer challenging examples handles corner cases and ambiguous boundaries. Such method ensures an optimal utilization of the fully labeled data. In addition, a semi-supervised method is used to include unlabeled data. We incorporate the multi-decoder consistency technique [Ouali *et al.*, 2020]. In particular, multiple feature perturbations are incorporated into the decoders, and the image encoder is encouraged to provide enhanced image representations. Finally, we propose the multi-modal learning method to leverage the image-text samples. The multi-modal consistency learning and text-aware contrastive learning enable our model to be aware of the semantically prominent objects in the images.

In summary, our matting engine training pipeline incorporates labeled, unlabeled, and text-image samples. Specially, we design the basic encoder with three specific decoders for saliency, portrait and scenery category to improve the accuracy, as shown in Figure 1(b).

2.3 Functions of AIGC Generation

The recommended content is generated by functions of AIGC as "Matting moments" from the perspective of time and space according to the category. To guarantee quality and computation efficiency of AIGC, our system presents a compact content selection and beautification functions with four matting schemes in Figure 3. **a) Flash matting**: to discover dominant subjects with easily copy and drag operations, flash matting is designed to make users easily edit the subjects. **b) Sky matting**: to achieve content beautification for foggy or smog weather, we apply blending and recoloring after sky matting to beautify the sky region, which greatly increases quality of scenery. **c) Magic eraser**: To meet with the demand of removing pedestrians in the scenery, we use the inpainting algorithm [Zeng *et al.*, 2021] on the accurately extracted masks to achieve eraser capability. **d) Stylization and colorization**: With prominent subject of portrait mask, we apply lightweight GAN [Chiu and Gurari, 2022] to human-preserving background stylizing and achieve portrait-only recoloring. With content selection and automatic beautification, the system presents the beauty of life moments with matting engine.

In addition, the whole system provides real-time mobile online services for a large number of photos. We develop a heterogeneous computing inference engine with hardware and software adaptation for real-time matting, inpainting and stylization for mobile AIGC application.

3 Experiments

We investigate the contributions of several learning components in "Data-driven Matting Engine" and implement alternative models of learning components. The experiments are conducted on the dataset containing 3K images of six categories including portrait, multi-person, pets, accessories, cards, and delicacy. MSE, Maxf, BoundaryIoU (BIoU), Connectivity (Conn) are adopted as evaluation metric. The contributed components in "Matting Engine" are composed of visual modality with pixel-wise supervised training (Pixel_Sup), semi-supervised training (Semi_Sup), and multiple modalities with text-image (Mlt_Sup). We observe that Mlt_Sup achieves the continuous better results along all the categories in Table 1, which demonstrates the superior performance for data-driven strategy of accurate matting.

4 Conclusion

In this paper, we propose a unified data-driven matting engine for exact subject matting for mobile AIGC application. The system included three aspects: 1) photo content structuring based on semantic parsing, 2) data-driven matting engine, 3) functions of AIGC generation. Our system achieves favourable performance in mobile AIGC products, which provides satisfying user experience.

References

[Bao *et al.*, 2022] Yunqing Bao, Hang Dai, and Abdulmotaleb Elsaddik. Semi-supervised cross-modal salient object detection with u-structure networks. *arXiv preprint arXiv:2208.04361*, 2022.

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

[Chiu and Gurari, 2022] Tai-Yin Chiu and Danna Gurari. Pca-based knowledge distillation towards lightweight and content-style balanced photorealistic style transfer models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7834–7843. IEEE, 2022.

[Cong *et al.*, 2022] Runmin Cong, Qi Qin, Chen Zhang, Qiuping Jiang, Shiqi Wang, Yao Zhao, and Sam Kwong. A weakly supervised learning framework for salient object detection via hybrid labels. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(2):534–548, 2022.

[Du *et al.*, 2023] Hongyang Du, Zonghang Li, Dusit Niyato, Jiawen Kang, Zehui Xiong, Dong In Kim, et al. Enabling ai-generated content (aigc) services in wireless edge networks. *arXiv preprint arXiv:2301.03220*, 2023.

[Hu *et al.*, 2023] Liangpeng Hu, Yating Kong, Jide Li, and Xiaoqiang Li. Effective local-global transformer for natural image matting. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[Li *et al.*, 2022] Jiachen Li, Vidit Goel, Marianna Ohanyan, Shant Navasardyan, Yunchao Wei, and Humphrey Shi. Vmformer: End-to-end video matting with transformer. *arXiv preprint arXiv:2208.12801*, 2022.

[Liang *et al.*, 2022] Yanhua Liang, Guihe Qin, Minghui Sun, Jun Qin, Jie Yan, and Zhonghan Zhang. Multi-modal interactive attention and dual progressive decoding network for rgb-d/t salient object detection. *Neurocomputing*, 490:132–145, 2022.

[Liu *et al.*, 2022] Jiang-Jiang Liu, Qibin Hou, Zhi-Ang Liu, and Ming-Ming Cheng. Poolnet+: Exploring the potential of pooling for salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):887–904, 2022.

[Lv *et al.*, 2022] Yunqiu Lv, Bowen Liu, Jing Zhang, Yuchao Dai, Aixuan Li, and Tong Zhang. Semi-supervised active salient object detection. *Pattern Recognition*, 123:108364, 2022.

[Ouali *et al.*, 2020] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 12671–12681, 2020.

[Piao *et al.*,] Yongri Piao, Chenyang Lu, Miao Zhang, and Huchuan Lu. Semi-supervised video salient object detection based on uncertainty-guided pseudo labels. In *Advances in Neural Information Processing Systems*.

[Qin *et al.*, 2020] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R. Zaïane, and Martin Jägersand. U^2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognit.*, 106:107404, 2020.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021.

[Shin *et al.*, 2022] Gyungin Shin, Samuel Albanie, and Weidi Xie. Unsupervised salient object detection with spectral cluster voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3971–3980, 2022.

[Sun *et al.*, 2022] Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. Human instance matting via mutual guidance and multi-instance refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2647–2656, 2022.

[Zeng *et al.*, 2021] Yu Zeng, Zhe Lin, Huchuan Lu, and Vishal M. Patel. Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 14144–14153. IEEE, 2021.

[Zhang *et al.*, 2022] Jin Zhang, Yanjiao Shi, Qing Zhang, Liu Cui, Ying Chen, and Yugen Yi. Attention guided contextual feature fusion network for salient object detection. *Image and Vision Computing*, 117:104337, 2022.