MAML-RAL: Learning Domain-Invariant HOI Rules for Real-Time Video Matting

Jiang Xin, *Graduate Student Member, IEEE*, Sheng Yue, Jinrui Zhang, *Student Member, IEEE*, Ju Ren, *Senior Member, IEEE*, Feng Qian, *Member, IEEE*, and Yaoxue Zhang, *Senior Member, IEEE*

***Abstract*— Real-time video matting is essential for applications like online video conferencing but faces challenges in human-object interaction (HOI) scenarios, known as the HOI-matting problem. This problem is challenging due to its open-recognition nature, where no dataset can cover the wide range of potential HOI cases, making it difficult for feature-learning-based methods to generalize effectively. To address this issue, we present an HOI-matting dataset and introduce a Model-Agnostic Meta-Learning-based rule-aware learning approach (MAML-RAL). MAML-RAL combines transfer learning and meta-learning to capture domain-invariant HOI rules, complemented by a fast local adaptation strategy to counter domain shifts and background interference. Our method achieves a mean intersection-over-union (mIoU) of 92.3%, outperforming current algorithms, with local adaptation further boosting performance to a remarkable mIoU of 95.84%.**

***Index Terms*— Video matting, HOI, meta-learning, domain adaptation, real-time.**

I. Introduction

REAL-TIME video matting has gained significant popularity across various applications, such as video conferencing, video calling, and live streaming, as it allows users to replace their personal background (e.g., from a bedroom) with virtual images/videos to safeguard their privacy. However, in practical scenarios where human-object interactions (HOI) occur, such as when users need to display an object on their hand while using a virtual background during a video conference, existing real-time video matting methods fail to produce satisfactory results. We refer to this matting task in HOI scenarios as real-time HOI matting. Real-time video HOI matting is still an open challenge problem.

Currently, conventional trimap-based methods [1], [2] have demonstrated strong performance on image HOI matting. However, due to their reliance on manual trimap generation for each image, they are unsuitable for real-time video matting. D-CL-RN [3] adopts a depth estimation-based approach, leveraging depth maps to infer foreground objects interacting with humans in videos. However, the computational complexity and time overhead of this approach limit its utility for real-time video streaming applications. Recent end-to-end algorithms for real-time video matting, such as MODNet [4], [5] and RVM [6], have achieved impressive performance for human body matting. Nevertheless, these methods are not well-suited for the HOI-Matting problem, as highlighted in Fig. 1 where objects interacting with humans are frequently misclassified as background elements. The cutting-edge BGM algorithm [7], [8] offers a promising solution for real-time video HOI matting. The method leverages both a background image and video frames to effectively separate foreground elements from the background, thus enabling accurate and efficient HOI matting. However, it is important to note that this approach requires a consistent background to be used during both the model training and inference phases. Any variations in the background, such as slight camera movements or changes in illumination, can result in the model malfunctioning and producing inaccurate matting results, as illustrated in Fig. 1. These limitations hinder the practicality of the BGM algorithm in real-world settings.

Real-time video human-object interaction (HOI) matting is widely recognized as an open-set recognition problem due to the inherent environmental and object variability in real-world applications [9]. Simultaneously achieving human and interacted object matting in real-time, without auxiliary input to the deep learning model, remains an open challenge. Firstly, **relying solely on feature representation-based learning methods for HOI matting is not reliable.** Existing HOI algorithms typically rely on feature representation learning to enable the model to learn features of the target and recognize them when presented with similar instances. However, in the context of HOI matting, the objects that humans interact with in real-world scenarios are often unpredictable and may not be seen or fully represented in the training set. Consequently, it is challenging to capture all relevant features of the objects using feature representation-based methods. Secondly, **the**

Received 30 August 2024; revised 18 October 2024; accepted 18 November 2024. Date of publication 22 November 2024; date of current version 7 April 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62302260 and Grant 6240070645; in part by China Postdoctoral Science Foundation under Grant 2023M731956; in part by the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (CPSF) under Grant GZC20240832 and Grant GZB20230352; and in part by Tsinghua–Tencent Joint Laboratory for Internet Innovation Technology, funded by Tencent Technology (Shenzhen) Company Ltd., under Grant 20192911990. This article was recommended by Associate Editor B. Xiao. *(Corresponding authors: Sheng Yue; Jinrui Zhang.)*

Jiang Xin is with the School of Computer Science and Engineering, Central South University, Changsha, Hunan 410083, China (e-mail: xinjiang@csu.edu.cn).

Sheng Yue, Jinrui Zhang, Ju Ren, and Yaoxue Zhang are with the Department of Computer Science and Technology, BNRist, Tsinghua University, Beijing 100084, China (e-mail: shengyue@mail.tsinghua.edu.cn; jinruizhang@mail.tsinghua.edu.cn; renju@tsinghua.edu.cn; zhangyx@tsinghua.edu.cn).

Feng Qian is with the Ming Hsieh Department of Electrical and Computer Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089 USA (e-mail: fengqian@usc.edu).

Digital Object Identifier 10.1109/TCSVT.2024.3504838

Fig. 1. Examples of HOI-matting: Two recent state-of-the-art real-time video matting approaches, MODNet [4] and RVM [6], fail to recognize the objects interacting with humans. BGM [8] produces a poor alpha matte due to camera shake. In contrast, our approach produces an accurate alpha matte for both humans and objects.

performance of matting deteriorates in complex real-world environments. The deployment of an HOI-Matting model in real-world environments is characterized by uncertainty and unpredictability, leading to unanticipated challenges during the inference process. Specifically, background objects being erroneously identified as foreground elements can result in inaccurate or unpredictable matting outcomes, which is problematic.

To tackle the above challenges, we propose a rule-aware learning algorithm called MAML-RAL and a fast local adaptation strategy to improve the performance of the real-time video matting model in HOI scenarios: (1) To address the limitations of feature representation-based learning methods, we design MAML-RAL, a Model-Agnostic Meta-Learning based Rule-Aware Learning algorithm. MAML-RAL guides the model to focus on the learning of HOI rules during training, thereby transforming the open recognition problem into a limited recognition problem. Compared to the variable object features, the variation in HOI rules is limited. As a consequence, MAML-RAL could significantly enhance the model's performance and generalization ability in real-time video HOI matting. (2) To deal with the issue of model performance deterioration in complex real-world environments, we design a fast local adaptation framework that includes a semi-auto dataset collection module and an SGT training strategy. The augmented model derived from MAML-RAL model enjoys the ability of fast learning which ensures that the framework can fulfill the adaptation task in a limited time. The SGT strategy further reduces the requirements of computing resources for local adaptation and ensures the stability of convergence.

A. MAML-RAL

MAML-RAL initially trains a factory model based on a human-only video matting dataset, followed by the acquisition of domain-invariant HOI rules through the utilization of MAML in conjunction with continual learning and feature-aware constraints. MAML-RAL exhibits a key advantage in the realm of domain-invariant HOI rule learning by means of its effective supervision through the design of MAML tasks. Moreover, the integration of continual learning and feature-aware constraints guarantees the preservation of feature extraction capabilities within the MAML-based learning mechanism.

B. Fast Local Adaptation

Our approach commences with a MatteAnything-based data collection method, which is succeeded by a sampling and group-based training (SGT) strategy for local adaptation. The fast local adaptation can deal with challenging situations where the input video frames have complex backgrounds with high noise. We first use a sampling strategy to simplify the training process to reduce training time overhead, and use a group training strategy to reduce the impact of ground truth noise.

Extensive experimental results demonstrate that our approach significantly improves the performance of HOI-matting. Specifically, the MAML-RAL algorithm achieves an mIoU of 92.3%, surpassing existing real-time video matting algorithms on the HOI-matting dataset. Furthermore, the MAML-RAL algorithm outperforms the HOI fine-tuning method, achieving a 2% improvement in mIoU. Additionally, after local adaptation, the algorithm's performance further improves, with the mIoU reaching 95.84%.

Our contribution can be summarized as follows:

- We build two distinct HOI-Matting datasets, comprising an image HOI-Matting dataset with 200 images and a video matting dataset with 112 videos. In total, these datasets contain 312 different objects that interact with humans.
- We propose **MAML-RAL**, a novel algorithm for learning domain-invariant HOI rules, improving the matting model's performance in various HOI scenarios. Additionally, we design an offline fast local adaptation strategy, which includes semi-automatic data generation and the sampling and group-based fast learning strategy for some complex scenarios.
- We conduct a comprehensive evaluation of the effectiveness of our proposed MAML-RAL algorithm and local adaptation strategy for real-time HOI-Matting tasks. Our results demonstrate superior performance compared to state-of-the-art real-time video matting models and HOI fine-tuning methods.

II. Related Works

A. Matting

Matting is the process of extracting a precise alpha matte that separates foreground objects from the background in an image or video frame. In general, the matting problem is formulated as Eq. 1, where I is an image or video frame, and α, F, and B are unknown variables. Specifically, α is a continuous value ranging from 0 to 1, where a value closer to 1 indicates that the pixel is more likely to be a part of the foreground. The goal of matting is to generate an accurate alpha matte.

$$I = \alpha F + (1 - \alpha) B \tag{1}$$

Traditional tripmap-based methods [1], [2] require a trimap along with the input image to generate the alpha matte. In a trimap, the area with a value of 0 is the background, the area with a value of 1 is the foreground, and the area with a value of 0.5 is an indeterminate area which is usually the transition area between foreground and background. Recently, the deep learning-based matting methods [10], [11], [12], [13] have received widespread attention. DIM [14] was proposed to use CNN for trimap-based image matting and shared a large

image matting dataset. Many subsequent works [15], [16], [17] have continued this scheme based on deep learning and trimap. However, obtaining trimap is a tedious task that usually requires manual annotation or automatic generation through complex algorithms [18], [19]. Therefore, this trimap-based approach is suitable for offline scenarios where human intervention can be easily performed, and is not suitable for real-time video matting.

Recently, trimap-propagation-based methods [20], [21], [22], [23], [24] are proposed to achieve video matting by propagating the tripmap of the first frame backward through spatio-temporal alignment and aggregation so that only one tripmap is needed for video matting. Although trimap-propagation-based methods have the potential to solve HOI-matting problems, the accumulation of trimap errors during matting leads to a continuous decline in matting quality. In offline video matting tasks, we can improve matting quality by updating the trimap at regular frame intervals, as with FTP-VM [22]. However, in real-time applications, we do not have the opportunity to update the trimap. Therefore, trimap-propagation-based methods are not suitable for real-time video HOI-matting tasks.

Considering the challenges faced by trimap-based methods in real-time video matting, several efficient trimap-free video matting algorithms have been proposed in recent years. BGM [7], [8] algorithms feed the video frames and the background image to the deep learning model and treats the background image as a green screen to realize video matting. BGM can achieve real-time performance and provide more reliable matting results. This kind of background-based algorithm must require the background to remain unchanged in the matting process, otherwise the algorithm fails to work. As a result, this kind of model is extremely limited in realistic applications. In recent years, several works [4], [6], [25], [26], [27], [28] have implemented fully end-to-end deep video human matting algorithms without any additional input. These algorithms only need to input the video frames to distinguish the human from the background, and the matting quality can reach hairline level. However, these methods only consider matting for humans and do not optimize for human-object interaction scenarios.

B. HOI Detection

HOI detection aims to not only detect humans and objects, but also understand the interaction relations between humans and objects. There are already a few approaches [29], [30] focusing on this problem to detect HOI with two-stage approaches. They first detect all the objects in the image and then identify the relationships between human and objects. Recently, some works [31], [32] have proposed one-stage HOI detection methods, with a special focus on object detection. However, the existing HOI detection methods cannot be integrated to the video matting task with strict real-time requirements. For example, D-CL-RN [3] implemented HOI matting by estimating the depth information of the image. RIM [33] achieves object extraction in images based on text prompts using a diffusion model. However, both D-CL-RN and RIM focus more on the quality of matting rather than the real-time performance of the algorithms. Therefore, these two methods are more suitable for offline video matting rather than real-time video matting.

TABLE I
A COMPARISON OF THE PROPOSED WORK WITH EXISTING WORK

Method	Auxiliary-free	Real-time	HOI-supported
D-CL-RN [45]	N	N	Y
OTVM [21]	N	N	Y
HSTSG [46]	N	N	Y
SparseMatt [47]	Y	N	N
RIM [33]	Y	N	Y
FactorMatte [48]	N	N	Y
VIM [49]	N	N	Y
BGM [8]	Y	Y	N
MODNet [5]	Y	Y	N
RVM [6]	Y	Y	N
VideoMatt [50]	Y	Y	N
FTP-VM [22]	N	Y	Y
VMFormer [27]	Y	Y	N
AdaM [26]	Y	Y	N
LPM [28]	Y	Y	N
Ours	Y	Y	Y

C. Model-Agnostic Meta-Learning

Model-Agnostic Meta-Learning (MAML) [34] is a famous meta-learning algorithm, it can learn a meta-model that can quickly obtain a model adapted to the local data through one or a few iterative training rounds. MAML has been widely used in few-shot learning [35], [36] and fast adption [37], [38], [39]. Recent works [40], [41] show that meta-learning has the potential to learn from highly diverse domains. However, the existing domain generalization based on meta-learning [42], [43], [44] is mainly for classification problems and does not handle unknown categories well. We are the first to address the real-time video HOI-Matting problem using the meta-learning techniques.

Table I compares recent advanced video matting algorithms from three aspects: the requirement for auxiliary input (e.g., trimap, mask, background, etc.), support for real-time inference, and support for HOI-matting. It is worth noting that we will focus only on the performance of the real-time video matting algorithms in the experiment section. Our proposed method supports real-time HOI-matting by learning HOI rules and an on-demand local adaptation, addressing the issue that existing real-time video matting methods do not adequately support HOI-matting.

III. DATASETS

This study utilizes four datasets: VideoMatte240K, Meta Dataset, IBHOI, and VBHOI. While VideoMatte240K is an open-source dataset, the other three were specifically collected for this research due to the lack of publicly available datasets suitable for evaluating HOI-matting algorithms. VideoMatte240K and the Meta Dataset are used for model training, whereas IBHOI and VBHOI are exclusively for performance evaluation. The Meta Dataset, IBHOI, and VBHOI primarily feature human-object interaction (HOI) scenarios.

To create these HOI scenes, we selected commonly encountered objects and realistic video conferencing environments, such as living rooms, bedrooms, meeting rooms, and offices.

A. VideoMatte240K

VideoMatte240K is a publicly available dataset for video matting, comprising 484 high-resolution green screen videos and generating a total of 240,709 unique frames of alpha mattes [8]. The videos are split into 479 for training and 5 for testing. The training set is used to build the human-only video matting model, and the testing set is used to evaluate its performance on non-HOI video matting tasks.

B. Meta Dataset

We collect a dataset consisting of 10 HOI green-screen videos (named *Meta Dataset*). Each video contains scenes of human-object interactions, where all objects have varying morphological characteristics. The dataset is used for training the MAML-RAL algorithm and carrying out the HOI-finetuning procedure. Notably, the number of videos thereof is deemed sufficient for demonstrating the performance of MAML-RAL as we can generate 90 task samples by applying permutation and combination.

C. IBHOI

We constructed an HOI image matting dataset consisting of 200 images with manually processed alpha mattes, named *IBHOI*. The images were sourced from publicly available websites, featuring diverse objects in the IBHOI dataset. Photoshop software was then used to generate the alpha mattes for the humans and interacting objects in the images. The entire dataset is used to evaluate the efficacy of the matting models.

D. VBHOI

The VBHOI dataset contains a total of 112 videos. Some of these videos were obtained from free online sources, while others were collected by us. We used the chroma-key function in Adobe After Effects to manually generate alpha mattes for all the videos. The dataset is designed with three primary objectives: 1) to evaluate the robustness of matting models in the context of HOI, 2) to assess the impact of background objects on HOI matting, and 3) to evaluate the performance of matting models in real-world HOI scenarios. To achieve these goals, we created four types of videos in VBHOI: 1) videos containing HOI content on a solid-color background, 2) HOI videos with interfering objects added to a solid-color background, 3) HOI videos recorded in real-world settings, such as living rooms, conference rooms, and laboratories, and 4) other HOI videos obtained from websites. We utilize all the videos in VBHOI to evaluate the performance of matting models.

IV. Method

The framework of our proposed method is depicted in Fig. 2. We begin by learning an augmented model through the application of the MAML-RAL algorithm to the video matting model. In this paper, we refer to the model trained on human-only videos as the *factory model* and the model trained for real-time HOI matting as the *target model*. MAML-RAL first uses human-only video data to train the factory model, which performs well in human matting. This training process follows the traditional deep learning scheme. Next, HOI video data is used to build meta tasks. These meta tasks, along with the MAML training mechanism and the feature-aware and regularization-based constraints, are then integrated to guide the target model in learning domain-invariant HOI rules. The final model obtained from MAML-RAL is referred to as the augmented model. MAML-RAL encourages the model to learn domain-invariant rules, such as human-object connections or depth information, by designing meta-tasks. Additionally, the MAML-RAL-enhanced model gains the capability for rapid adaptation.

In the deployment environment of user devices, we further design a pipeline for local domain adaptation to address challenging situations. We develop a convenient local data collection method by combining the MatteAnything [51] model with manual frame selection. The processed data is then used to perform adaptation training on the augmented model derived from MAML-RAL. Finally, we design a sampling and group-based training (SGT) strategy to accelerate the training process and mitigate the influence of noisy labels. In the following sections, we describe the key components in detail.

A. Continual Learnning Constraint

We first design a regularization-based continual learning constraint to avoid the decline of the model's ability to extract human features. Regularization-based continual learning constraint is used to constrain the parameters of the target model close to the parameters of the factory model while training with MAML mechanism which is described later. Accordingly, the updated parameters θ of the target model are constrained by the parameters θ_f of the factory model. In this paper, we use a $l2$ regularizer, shown in Eq. 2, to constrain θ of the target model. For more complex constraints, please refer to [52].

$$\mathcal{L}^{reg} = \left\| \theta - \theta_f \right\|_2 \tag{2}$$

B. Feature-Aware Constraint

In the matting problems including image matting and video matting, the recognition of details is of crucial importance, such as the recognition of the transitional area between the human body and the background. To this end, we need to protect the fine-grained features of human body from being forgotten while training the model with the MAML mechanism. As a result, we design a feature-aware constraint to ensure the model to preserve the ability of the model to extract rich fine-grained human body features.

We denote the feature-aware constraint as $\mathcal{L}^{feat}$ which is achieved by Eq. 3, where $h^k_{\theta_f}$ represents the k-th hidden features output by the factory model and k is from 1 to K. h^k_{θ} represents the kth hidden features output by the target model.

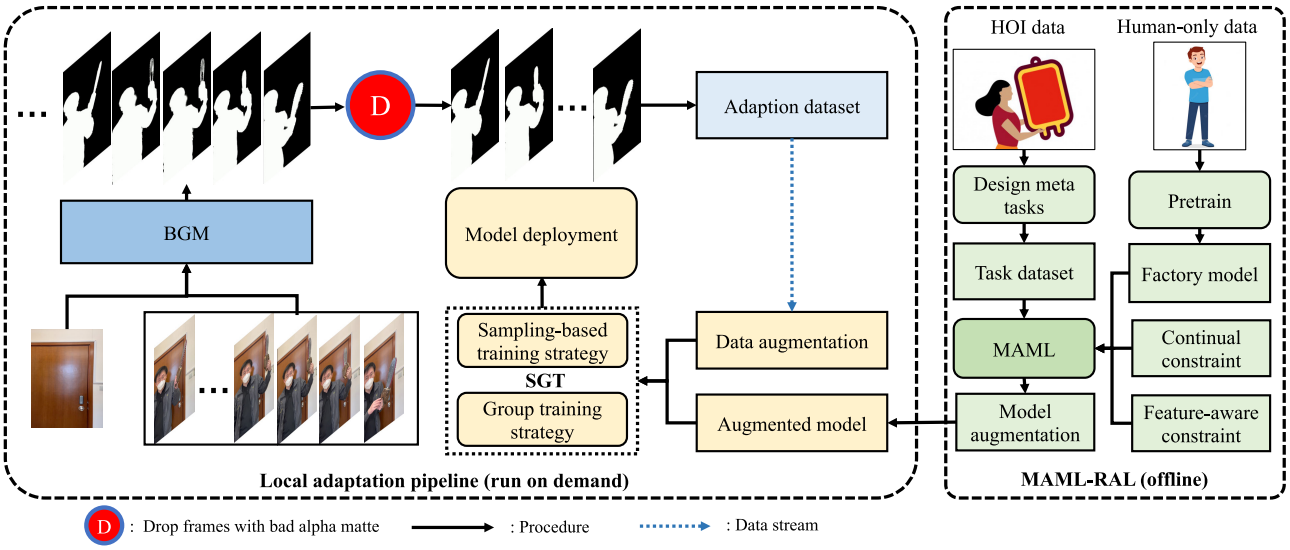


Fig. 2. Architecture of MAML-RAL and fast local adaptation.

$\mathcal{H}$ is a mapping function that is used to map all the channels of hidden features to one channel. Specifically, we first get K hidden features from the factory model. We then compute the K hidden features of the target model corresponding to the factory model. Then we apply an average hashing function to each of $h_{\theta_f}^k$ and concatenate them to generate a feature embedding. The same operation is also applied to h_{θ^k}. Finally we compute the $l2$ distance between $\mathcal{H}(h_{\theta_f}^k)$ and $\mathcal{H}(h_\theta^k)$ to get the final feature-aware constraint $\mathcal{L}^{feat}$.

$$\mathcal{L}^{feat} = \frac{1}{K}\sum_{k=1}^{K}\left\|\mathcal{H}\left(h_{\theta_f}^k\right) - \mathcal{H}\left(h_\theta^k\right)\right\|_2 \tag{3}$$

C. The Loss Function for HOI Matting

We first adopt a traditional video matting loss function following the SOTA methods, which is formulated by Eq. 4 - 9 where α_t^* means the ground truth of alpha matte. $\mathcal{L}_{l1}^{\alpha}$ means the $L1$ loss between the predicting α_t and the ground truth α_t^*. $\mathcal{L}_{lap}^{\alpha}$ means the pyramid Laplacian loss. A coherence loss $\mathcal{L}_{tc}^{\alpha}$ is used in the loss function to capture the temporal information since we use a sequence model which takes a frame sequence as the input. The specific meanings of Eq. 4 - 8 are referenced in [6].

$$\mathcal{L}_{l1}^{\alpha} = \left\|\alpha_t - \alpha_t^*\right\|_1 \tag{4}$$

$$\mathcal{L}_{lap}^{\alpha} = \sum_{s=1}^{5}\frac{2^{s-1}}{5}\left\|L_{pyr}^{s}\left(\alpha_t\right) - L_{pyr}^{s}\left(\alpha_t^*\right)\right\|_1 \tag{5}$$

$$\mathcal{L}_{tc}^{\alpha} = \left\|\frac{\mathrm{d}\alpha_t}{\mathrm{d}t} - \frac{\mathrm{d}\alpha_t^*}{\mathrm{d}t}\right\|_2 \tag{6}$$

$$\mathcal{L}_{l1}^{F} = \left\|\left(a_t^* > 0\right) * \left(F_t - F_t^*\right)\right\|_1 \tag{7}$$

$$\mathcal{L}_{tc}^{F} = \left\|\left(a_t^* > 0\right) * \left(\frac{\mathrm{d}F_t}{\mathrm{d}t} - \frac{\mathrm{d}F_t^*}{\mathrm{d}t}\right)\right\|_2 \tag{8}$$

$$\mathcal{L}^{M} = \mathcal{L}_{l1}^{\alpha} + \mathcal{L}_{lap}^{\alpha} + 5\mathcal{L}_{tc}^{\alpha} + \mathcal{L}_{l1}^{F} + 5\mathcal{L}_{tc}^{F} \tag{9}$$

We add the continual learning and feature-aware constraints to the matting loss function to get the final HOI-matting loss function as follows:

$$\mathcal{L}_{\mathcal{T}_i}\left(f_\theta, h_\theta, h_{\theta_f}, \theta_f\right) = \gamma_1\mathcal{L}^{M} + \gamma_2\mathcal{L}^{feat} + \gamma_3\mathcal{L}^{reg} \tag{10}$$

where γ_1, γ_2 and γ_3 are the weighting parameters. We empirically set $\gamma_1 = 0.7$, $\gamma_2 = 0.1$ and $\gamma_3 = 0.2$.

Fig. 3 illustrates the computing graph of the HOI-matting loss function. A factory model is used to compute the two items of the continual learning constraint and the feature-aware constraint.

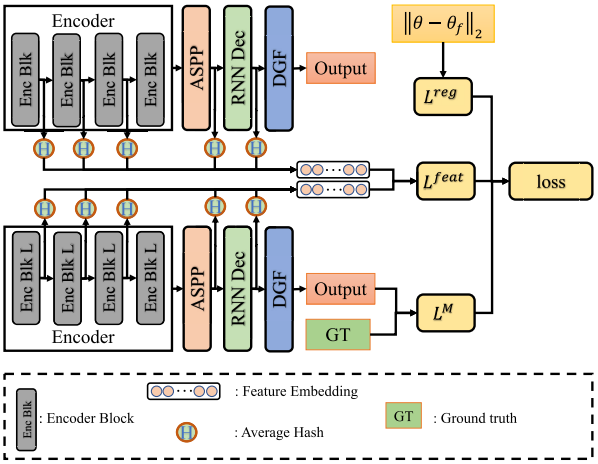


Fig. 3. Overview of the roles for the factory model, the target model and the losses.

D. MAML-Based Rule-Aware Learning (MAML-RAL) Algorithm

In real-time video HOI-Matting problem, the target model should perform well even meet unknown objects that are interacting with humans. Formally, suppose we have two HOI videos, V_1 and V_2, where the objects are completely different, we first use V_1 to train the model. The parameters that are trained with V_1 should also perform well on V_2, which satisfies

the MAML's training settings. So we propose a MAML-based rule-aware learning (MAML-RAL) algorithm to learn domain-invariant HOI rules.

MAML is a task-based algorithm. Carefully designed tasks enable the target model to learn more domain-invariant rules of HOI. In our design, a task sample consists of two parts, the training set and the testing set. We use a Meta Dataset to build the task samples. A task sample is constructed by randomly selecting two videos and using one of the videos as the training set and the other as the testing set. We construct the task dataset using 10 HOI videos, each with a green wall as backgrounds. Using the green wall as a background helps the model focus on learning HOI features. Also, the objects of the objects in the video need to avoid overlapping with the body. In order to facilitate the evaluation of algorithm performance, the objects we exhibit in the video have obvious domain gaps with those in the real-world HOI dataset. Finally, after permutation and combination, the task dataset contains a total of 90 training samples.

In MAML-RAL, we first use the human-only video matting data to train a factory model of video matting. Then the MAML-RAL algorithm forces the target model to learn the domain-invariant rules by following a MAML training mechanism with the constraints of continual learning and feature-aware constraints. Different from MAML, MAML-RAL does not randomly initialize the model parameters in the first step. Instead, MAML-RAL uses the parameters of the factory model θ_f to initialize the target model.

Algorithm 1 MAML-Based Rule-Aware Learning (MAML-RAL)

Require: $p(\mathcal{T})$: distribution over tasks
Require: α, β: step size hyperparameters
1: Train a factory model with human-only video matting dataset and get the parameter θ_f
2: Initialize θ with factory model parameter θ_f
3: **while** not done **do**
4: Sample batch of tasks $\mathcal{T}_i \sim p(\mathcal{T})$
5: **for all** $\mathcal{T}_i$ **do**
6: Get hidden features h_{θ_f} of factory model
7: Evaluate the gradient $\nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta, h_\theta, h_{\theta_f}, \theta_f)$ by Eq. 10 with respect to K examples
8: Compute adapted parameters with gradient descent: $\theta_i' = \theta - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$
9: **end for**
10: Update $\theta \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_i \sim p(\mathcal{T})} \mathcal{L}^M_{\mathcal{T}_i}(f_{\theta_i'})$
11: **end while**

In each round of MAML training, we first perform a fast adaptation on the training data in each $\mathcal{T}_i$, and obtain the first updated parameter θ_i', respectively. Further, the testing set is used to update the initial θ, where a Hessian needs to be computed. The detailed algorithm is summarized in Alg. 1. In the fast adaptation stage which is shown in line 6 of Alg. 1, we first run the factory model. Afterwards, in line 7 we evaluate the HOI-Matting loss which is different from the video matting loss $\mathcal{L}^M$ that is used in line 10. In this way, we add HOI rules into the model while retaining the model's feature extraction capabilities of the human body.

E. Local Adaptation

We designed an additional local adaptation pipeline in case of challenging environments encountered during the real-world deployment and application. The pipeline consists of three components including local data capturing module, local data augmentation, sampling and group-based training (SGT) strategy.

1) Local Data Capturing Module: For local adaptation, we need to collect a small number of video frames and annotate the alpha mattes. Considering that fully manual annotation is impractical, we use the MatteAnything [51] model to perform offline annotations. MatteAnything is an interactive matting model based on SegmentAnything [53], which generates alpha mattes for the target objects through user clicks. Specifically, the user first collects a small set of images depicting human-object interactions using a camera, and then uses MatteAnything to generate the alpha mattes for these images. These data are ultimately used for the local adaptation of the real-time video matting model.

2) Local Data Augmentation: We first perform the background augmentation to the local adaptation dataset. The foreground information in the real-time applications keeps the same as the data during local adaptation. However, the background is easily changed by the movement of the camera. As a result, background augmentation plays an important role in local adaption, which is used to ensure the robustness of the model when the background changes. We first filter out 100 images from the MSCOCO dataset as image backgrounds. Then we collect 36 real-world video backgrounds. While training, we have a 50% chance to replace the video background with a new image or a video background.

In addition to background augmentation, we also perform regular data augmentation operations, including random flip,random erasing, random hue change, random contrast change, brightness change, random grayscale change, random sharpening, random smoothing, random pause and so on.

3) Sampling and Group-Based Training (SGT) Strategy: We perform local adaptation using the augmented model obtained by MAML-RAL. In this process, the loss function for local adaptation must match the loss function used for fast adaptation in the MAML-RAL algorithm, as shown in Eq. 10. The default learning rate (lr) for local adaptation is set to 0.1, though it can be adjusted to optimize practical application outcomes.

Local adaptation differs significantly from cloud-based model training. Given the limited computing resources of local devices, conventional training strategies can be time-intensive and impair user experience. Thus, optimizing the training strategy for local adaptation is crucial to reducing training time and ensuring stable convergence, both key factors for enhancing user experience. In this section, we describe our optimized approach to local adaptation, encompassing both a sampling-based training strategy and the SGT strategy.

For a sequential model based on an RNN structure, one input consists of multiple frames, even with a batch size

of 1. For instance, in the RVM model employed in this paper, the input during training includes k frames. On cloud servers, we use a sequential sampling strategy, selecting frames sequentially from a starting point and moving backward k frames per batch, iteratively processing until the entire video is covered.

However, local adaptation must be completed within a short timeframe. Traversing the entire video, as done on the server, is too time-consuming for local adaptation. To expedite this process, we developed a random sampling training strategy. Here, the batch size is set to 1 per epoch, but instead of sequentially traversing the entire video, we randomly sample k frames from the video as input each time. In RVM, k is set to 10. This sampling-based strategy better suits local adaptation scenarios, allowing the model to quickly access the video's global information and enabling users to halt training based on model performance without requiring full video traversal.

To further mitigate the impact of noisy data in local adaptation, we introduce a group-based adaptation strategy. During each training round, we sample 10 groups of data, optimizing the model separately for each group to yield 10 sets of optimized parameters. These parameters are then averaged to produce the final result for the epoch, which serves as the basis for the next training round. SGT thus implements a pseudo-batch gradient descent approach, reducing fluctuations from individual noisy data points. This approach stabilizes model convergence, even in the presence of label noise. Since group-based training relies on sampling, we refer to this combined approach as the Sampling and Group-based Training (SGT) strategy.

V. Experiments

Experiment environments. Our MAML-RAL model is trained on a cloud server equipped with Nvidia Tesla V100 32G GPU. For evaluation purposes, we utilize a desktop system comprising a commercially available configuration that includes 11th Generation Intel Core i7-11700 @ 2.50GHz x 16 CPU, NVIDIA 3060 GPU, and 16G RAM.

Baselines. We use existing open-source real-time video matting algorithms as baselines, including MODNet [5], RVM [6], VMFormer [27], and the trimap-based FTP-VM [22] method. Additionally, considering that our algorithm uses RVM as the target model, we also include the fine-tuned RVM model from the Meta Dataset as a baseline, denoted as RVM-F, to demonstrate the effectiveness of our method. RVM-F was fine-tuned for 1000 epochs on the Meta Dataset, building upon the original RVM model. We used the Adam optimizer, with the learning rate and other hyperparameters kept consistent with those in [6]. For FTP-VM, we randomly selected a frame from the video as the initial memory frame, and generated the corresponding trimap as the initial memory trimap. To simulate real-time applications, we kept the memory frame and trimap unchanged, as there is no opportunity to update them in real-time scenarios. For the MODNet, VMFormer, and RVM methods, we directly used the parameters provided in the official releases for evaluation.

Metrics. We evaluate matte outputs using metrics from MAD (mean of absolute difference), MSE (mean squared error), Grad (spatial-gradient metric) [54], Conn (connectivity) [54] for quality. We additionally use a mIoU (mean intersection over union) [55] score to evaluate the performance of HOI recognition both on IBHOI and VBHOI datasets. To compute the an mIoU metrics, we first set a threshold T to classify the pixels in the alpha matte α to generate a classification result C. We have $C_i = 0$ if $\alpha_i < T$ and $C_i = 1$ if $\alpha_i \leq T$, where i means the ith pixel in the alpha matte. Then we use C to calculate the mIoU metric.

A. Quantitative Comparison Between Different Methods

To evaluate the performance of our proposed learning algorithm, we use RVM as our target model which is the SOTA real-time video matting model. Then we apply the proposed MAML-RAL algorithm to the target model with the VideoMatte240K and the Meta Datasets. Finally, we evaluate the performance of MAML-RAL on the proposed VBHOI dataset and compare our method with existing approaches including MODNet, RVM, VMFomer, FTP-VM, and HOI-fine-tuned model by transfer learning.

The performance of different methods evaluated on VBHOI dataset is shown in Table II, where RVM-F is the RVM model fine-tuned by transfer learning and MAML-RAL is our proposed algorithm. MAD and MSE are scaled by $1e^3$ and Conn is scaled by $1e^{-3}$ for better readability. The results demonstrate that the MAML-RAL algorithm can significantly improve the HOI-matting performance in the real-world scences. We can observe that our proposed MAML-RAL method outperforms existing real-time video matting techniques across all metrics. Notably, the mIoU metric demonstrates the superior performance of our method in recognizing objects involved in human-object interactions. Additionally, it is evident that the impact of RVM-F on improving the performance of HOI-Matting is limited. This is primarily because the objects interacting with humans in the VBHOI dataset were not present in the Meta Dataset used for fine-tuning. This indicates that merely fine-tuning existing models with HOI data is insufficient to handle diverse scenarios effectively. In contrast, our model leverages the MAML-RAL algorithm to actively guide the model in learning HOI knowledge, thereby enhancing its understanding of HOI-Matting tasks and enabling it to better handle various human-object interaction scenarios.

To further validate the performance improvement of MAML-RAL on the target model in more unknown scenarios, we additionally collected 200 image samples to test the target model's performance on HOI-Matting. Although the selected target model RVM is not specifically designed for image matting, the comparison of the target model's factory model, fine-tuned model(RVM-F), and the MAML-RAL-enhanced model still validates the effectiveness of MAML-RAL. The results of these experiments are presented in Table III. Our method consistently outperformed both RVM and RVM-F across all metrics. Additionally, we observed that RVM-F showed a slight improvement in the mIoU metric compared to the native RVM; however, its performance declined on other metrics. This decline can be attributed to the Meta Dataset used for HOI-finetuning, which is a few-shot HOI-Matting

TABLE II
PERFORMANCE OF DIFFERENT METHODS ON VBHOI

Dataset	Method	MAD↓	MSE↓	Grad ↓	Conn ↓	mIoU(%) ↑
VBHOI	FTP-VM	61.76	49.60	38.62	54.15	83.91
	VMFormer	82.90	72.04	22.74	55.29	71.32
	MODNet	103.72	94.01	27.84	39.17	78.42
	RVM	34.13	24.43	19.22	16.57	89.05
	RVM-F	32.68 ± 11.59	22.96 ± 11.14	17.74 ± 1.91	17.38 ± 0.07	90.43 ± 1.37
	MAML-RAL(Ours)	**25.58** ± 3.88	**15.87** ± 3.78	**15.66** ± 1.13	**17.44** ± 0.29	**92.30** ± 0.86

TABLE III
PERFORMANCE OF OUR METHODS ON IBHOI

Dataset	Method	MAD	MSE	Grad	Conn	mIoU(%)
IBHOI	RVM	61.03	49.88	25.35	55.62	84.57
	RVM-F	64.08	54.79	36.12	71.33	85.98
	MAML-RAL(Ours)	**31.58**	**20.45**	**22.45**	**31.66**	**92.79**

TABLE IV
PERFORMANCE OF DIFFERENT METHODS ON NON-HOI VIDEOS

Dataset	Method	MAD	Grad	Conn
VideoMatte240K	RVM	3.02	7.46	1.13
	RVM-F	708.23	31.12	125.36
	MAML-RAL(Ours)	3.50	8.40	1.289

TABLE V
PERFORMANCE OF LOCAL ADAPTATION(LA)

Dataset	Method	MAD	Grad	Conn	mIoU(%)
VBHOI	Ours w/ LA	13.52	10.44	14.31	95.84

dataset. While conventional training with such a dataset can enhance HOI-Matting performance, it may also reduce the model's generalization ability. Our method addresses this by introducing continual and feature-aware constraints combined with the task design of MAML, which allows the model to quickly learn HOI-Matting knowledge while minimizing the loss of its pre-existing matting knowledge. This ensures maintained performance on conventional human matting tasks.

B. Evaluate the Performance on Non-HOI Videos

We evaluated our method's performance on conventional human matting tasks using the VideoMatte240K dataset [8], which is a video matting dataset containing only humans. Table IV compares our method with the HOI-fine-tune method. As mentioned above, we can see the RVM-F model obtained by finetuning the model using HOI data has an obvious decrease in the performance on human-only data. However, the model trained using the MAML-RAL learning algorithm still maintains high performance on human-only data.

C. Evaluate the Performance on Real-World Videos

We present a portion of the qualitative comparisons conducted on real-world video datasets. As depicted in Fig. 4, our method effectively recognizes objects interacting with the human while also filtering out background noise interference. Additionally, as shown in Fig. 5, our method maintains the same level of performance as the SOTA model in capturing the details of the human body.

In comparison, the RVM-F model, fine-tuned with HOI data, demonstrates the ability to recognize certain objects; however, its performance degrades significantly when the characteristics of objects interacting with humans deviate markedly from those in the training set. While RVM, MODNet, and VMFormer are capable of generating high-quality alpha mattes for humans, they exhibit suboptimal performance in HOI-Matting tasks. The FTP-VM method offers some proficiency in addressing HOI-Matting challenges, but due to its reliance on Trimap-propagation, propagation errors accumulate over time in real-time applications, resulting in artifacts in the matting output. Compared to these approaches, our method delivers more robust and reliable HOI-Matting results in real-time scenarios.

D. Evaluate Performance of Local Adaptation

Table V shows the results of local adaptation based on the MAML-RAL-augmented model, from which we can see that all the metrics are significantly larger than the best results in Table II.

Since a new local model is generated for each video after local adaptation, we are more concerned about the performance distribution, which is more indicative of the effectiveness of the method. Fig. 6 shows the CDF of mIoUs that different methods yield in evaluating videos. For MAML-RAL augmented model, there are more than 80% of the videos have an accuracy higher than 90%. The performance is much better than that of the factory RVM model and HOI-fine-tuned RVM model, MODNet, and the HOI-fine-tuned RVM model. More importantly, after local adaptation, there are more than 95% of the videos have an accuracy higher than 90%. The proportion of mIoU greater than 90% is significantly improved by 15% after local adaptation.

We count the number of epochs required to achieve the mIoU of more than 90% for each video using the random

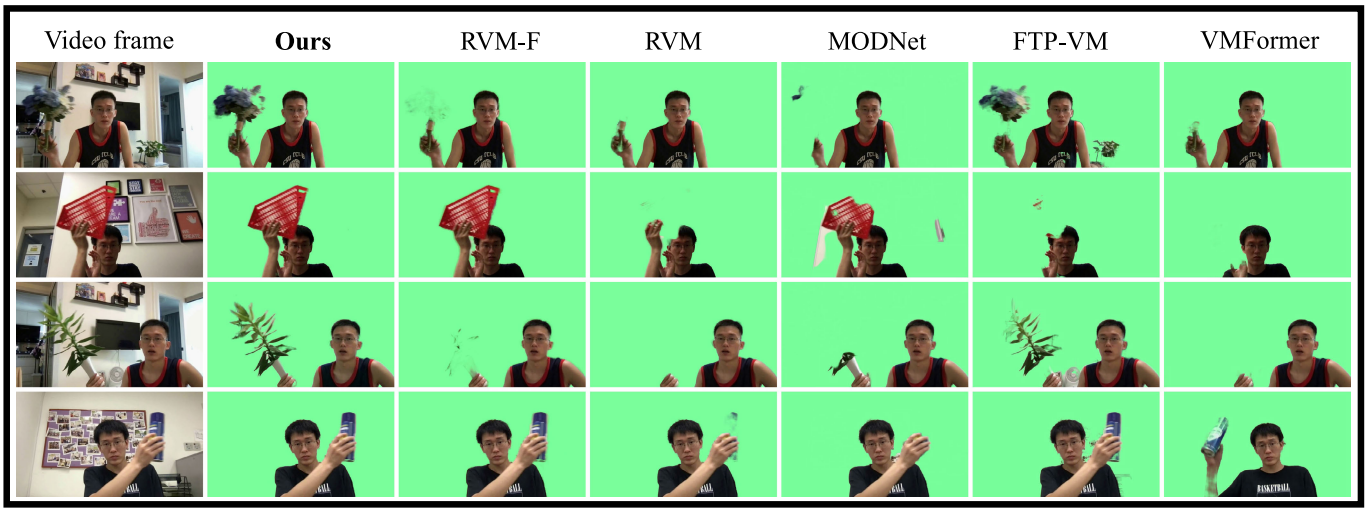


Fig. 4. Performance on real-world continuous video streaming. Our method produces a more accurate alpha matte compared with all state-of-the-art methods and HOI-fine-tuned RVM (RVM-F). A dynamic video demo can be found at https://jerryxin1994.github.io/MAML-RAL/.

Fig. 5. The model enhanced by MAML-KAL still maintains hair-strand-level matting performance.

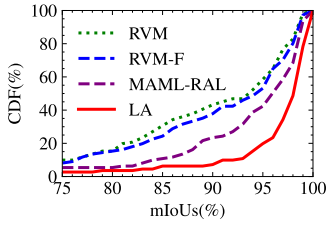


Fig. 6. CDF for the performance of each local video.

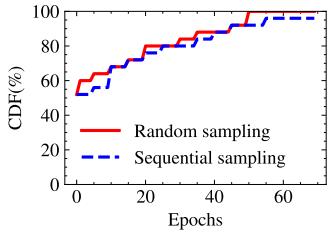


Fig. 7. Epochs for local adaptation to get the best result for each video.

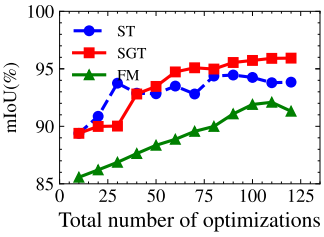


Fig. 8. Performance comparison between normal training and SGT.

sampling strategy and the sequential sampling strategy, respectively. From Fig. 7, we find that the sampling strategy does not decrease the training accuracy. When training epochs are less than 10, the sampling approach is able to achieve a performance greater than 90% for more than 60% of the videos, which is significantly higher than sequential sampling. However, by applying the sampling strategy, the computation overhead is reduced since we use fewer frames to train the model in each epoch. Besides, more than 90% videos can achieve mIoU of 96% within 50 training epochs, which means the time cost for local adaptation is low.

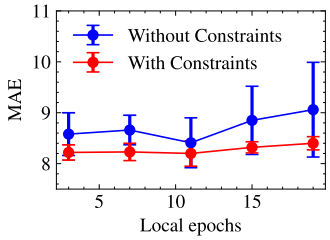


Fig. 9. Impact of the proposed constraints.

Aiming to show the effect of SGT, we select 6 videos from the real-world HOI-Matting dataset that requires local adaptation, and train on them in a regular way and SGT respectively. Fig. 8 shows that the SGT strategy converges slower at the first stage than the normal training which is trained with sampling strategy, but it can converge stably to a higher accuracy. With SGT, all videos can achieve the best adaptation performance within **120** seconds on a GPU of Nvidia 3060ti. In contrast, the regular training method suffers from performance fluctuations.

In Fig. 8, we further compare the performance of local adaptation using the augmented model and directly using the factory model. Obviously, the augmented model gets a higher accuracy. Moreover, there is a faster convergence speed when performing local adaptation which is due to the MAML's fast learning capability.

E. Performance in Dynamic Backgrounds

Fig. 10 illustrates the performance of our algorithm in a dynamic scene, demonstrating that our method can

Fig. 10. Performance in Dynamic Backgrounds. The camera follows the movement of the subject, resulting in a dynamically changing background.

TABLE VI
REAL-TIME PERFORMANCE OF DIFFERENT METHODS

	Ours	RVM	MODNet	FTP-VM	VMFormer	D-CL-RN
FPS	97	98	34	28	24	≤ 2

effectively recognize both humans and interacting objects within such environments. This example further substantiates that HOI-Matting challenges can be addressed by recognizing human-object interaction rules without relying on auxiliary input.

F. Real-Time Performance of the Proposed Method

In this paper, we did not design a new model for the HOI-matting problem. Instead, we propose a novel HOI-rule-aware learning framework. Theoretically, this framework can supervise any real-time video matting model to learn HOI rules, as long as the model can provide a high-quality alpha matte for humans. Since no additional computations were introduced into the model, it maintains real-time performance, as shown in Table VI. We evaluated the real-time performance of the relevant methods on a consumer-grade Nvidia 3060 GPU. Additionally, we compared our method with D-CL-RN, a depth-estimation-based HOI-Matting approach. As observed, the FPS of D-CL-RN is less than 2, which is far from achieving 30 FPS.

VI. Ablation Study

A. Role of Continual and Feature-Aware Constraints

To test the influence of the continual constraint(CC) and feature-Aware constraint(FC), we first remove the CC from the loss function and retrain the model with the MAML-RAL learning algorithm. Then we remove the CC from the loss and repeat the experiment again. We compare the results with the full MAML-RAL algorithm which add both CC and FC to the loss function.

The values in Table VII demonstrate that both CC and FC are important to MAML-RAL. After removing CC and FC, the performance of the trained model decreases significantly.

B. Forgetting Test for Local Adaptation

The roles of CC and FC in the local adaptation stage are to protect the model's ability to recognize humans at a fine-grained level (such as the recognition of hair strands) from being forgotten when the model is learning new knowledge. To evaluate whether these two constraints work during the local adaptation, we perform a forgetting test. Specifically, we first select one of the local HOI videos to perform 20 rounds of local adaptation training with CC and FC constraints. We then remove these two constraints from the loss as another control experiment. Moreover, we use a human-only video to evaluate the MAE of the two models with the number of training rounds. We repeat the experiments five times, and the averaged results are shown in Fig. 9.

From Fig. 9, it can be seen that the model yields a lower MAE after introducing CC and FC. It indicates that the two constraints successfully prevent the model from forgetting the original human matting knowledge because adding these two constraints yields better results. Besides, it can be seen from the error bar that the constraint-free method suffers from instability. It would be exacerbated by the increasing training epoch.

TABLE VII
PERFORMANCE OF LOCAL ADAPTATION(LA)

Dataset	Method	MAD ↓	Grad ↓	Conn ↓	mIoU(%) ↑
VBHOI	MAML-RAL	25.58	15.66	17.44	92.30
	w/o CC	29.25	18.71	19.90	91.73
	w/o FC	28.51	18.51	19.59	91.87

VII. Discussion

A. Comparison between MAML-RAL and HOI-Fine-Tuned Method

Directly using the HOI-fine-tuned method proves inadequate due to the domain gap between the objects interacting with humans in HOI videos and those in real-world scenarios. Conventional deep learning algorithms rely on classifying pixels based on local features of humans and objects in video frames, which leads to limitations in detecting global relationships between humans and objects. When the training data does not encompass the objects present in real-world videos, the model struggles to recognize them. In contrast, the proposed MAML-RAL algorithm focuses on domain-invariant HOI rules, such as the connections between humans and objects, by leveraging meta-task design, meta-training settings, and two constraints: continual and feature-aware. These enhancements enable the model to better identify global relationships and improve its performance in real-world scenarios. The HOI rules learned through MAML-RAL are primarily determined by the design of the meta tasks. In this paper, the data constituting the meta tasks consist of two entirely different objects, with the only commonality being that both are held in a person's hand. Through extensive training, the model internalizes this rule, enabling it to perform HOI-Matting.

B. Selection of the Target Model

We experimented with several algorithms, including MODNet, VMFormer, and RVM, ultimately selecting RVM as our target model. Although our method incorporates a

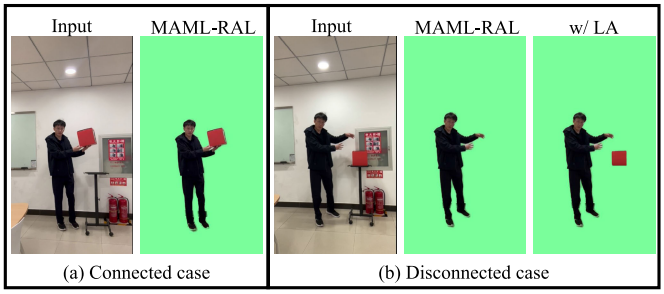


Fig. 11. Results for Connected and Disconnected Human-Object Interactions.

Model-Agnostic training strategy, we assume that the target model must be capable of generating high-quality alpha mattes of humans. If the target model produces alpha mattes with flaws, such as large regions of misclassification, it would hinder our algorithm's ability to learn HOI knowledge and could even degrade the model's performance. After extensive experimentation, we chose RVM as the target model because it consistently provides high-quality human alpha mattes in video sequences, making it the most compatible with our MAML-RAL algorithm.

C. Role of Random Sampling and SGT

The random sampling strategy is particularly advantageous for sequence models as it reduces the time overhead associated with training such models. This sampling approach significantly enhances the efficiency of local adaptation in scenarios where video content changes gradually. However, for models such as MODNet, where the input consists of a single frame, it is possible to use all frames in the video as an image dataset and apply the batch gradient descent algorithm. In this case, random sampling is equivalent to a shuffling operation.

The use of SGT mitigates the risk of performance degradation arising from anomalous data. SGT ensures stable convergence of the model by averaging the parameters after simultaneously training multiple data groups. This stability of convergence is crucial in practical applications, as fluctuating model performance makes it challenging to ascertain if the model has been appropriately trained.

D. Impact of Local Adaptation on Real-Time Performance

Although local adaptation may introduce some time overhead due to data reparation and training, we assert that this overhead does not compromise the algorithm's real-time performance. Local adaptation is a preparatory step conducted before the model begins its formal operation, and therefore, it does not affect the model's speed during real-time execution. For example, performing adaptation before a video meeting begins does not impact the matting speed of the algorithm on real-time video frames once the video meeting starts.

E. Further Potential Applications of Real-Time HOI Matting

For applications in Augmented Reality (AR) and Virtual Reality (VR), MAML-RAL's real-time matting significantly enhances user immersion by accurately blending real-world users with virtual environments, even during complex human-object interactions. For example, in video-based human-centric 3D reconstruction tasks [56], [57], incorporating our method enables precise separation of individuals and their interacting objects within the video, allowing for accurate 3D reconstruction. This approach enhances the flexibility and accuracy of 3D reconstructions. In the field of video representation research, video content is typically structured by representing the foreground and background as two distinct canonical images, as demonstrated in methods like CoDeF [58]. This approach facilitates video editing by employing image-based techniques. Our method further advances this concept by enabling more effective separation of foreground and background in human-object interaction scenarios, yielding more precise canonical images.

F. Limitations

In theory, the model enhanced with MAML-RAL can recognize only objects that are connected to the human body. Therefore, when objects are not connected to the human body, relying solely on the MAML-RAL-enhanced model cannot achieve effective HOI-Matting. However, our proposed local adaptation strategy can compensate for this limitation, as shown in Fig. 11. The trade-off is that some preparatory work is required before the model's formal operation, including gathering a small amount of data for local adaptation and conducting a brief local training session on the model. Developing more efficient, real-time, end-to-end algorithms for non-connected HOI-Matting could be a potential topic for future research.

VIII. Conclusion

In this paper, we first introduce two HOI-Matting datasets, VBHOI and IBHOI, which serve as benchmarks for evaluating the performance of HOI-matting. Subsequently, we propose a MAML-based HOI-rule-aware learning approach, named MAML-RAL, that efficiently trains the real-time video matting model to learn HOI rules via a few designed HOI meta-tasks. Finally, we present a fast local adaptation pipeline for the HOI-Matting task, leveraging an SGT strategy to customize model enhancements in challenging situations. Notably, the SGT strategy guarantees fast and stable convergence during local adaptation. Our proposed method demonstrates significant performance improvements for real-time video HOI matting, thus contributing to the advancement of this field.

References

[1] Q. Chen, D. Li, and C.-K. Tang, "KNN matting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2175–2188, Sep. 2013.

[2] Y. Aksoy, T. O. Aydin, and M. Pollefeys, "Designing effective inter-pixel information flow for natural image matting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 228–236.

[3] B. Xu, H. Huang, C. Lu, Z. Li, and Y. Guo, "Virtual multi-modality self-supervised foreground matting for human-object interaction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 428–437.

[4] J. Sun, Z. Ke, L. Zhang, H. Lu, and R. W. H. Lau, "MODNet-V: Improving portrait video matting via background restoration," 2021, *arXiv:2109.11818*.

[5] Z. Ke, J. Sun, K. Li, Q. Yan, and R. W. Lau, "MODNet: Real-time trimap-free portrait matting via objective decomposition," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 1, pp. 1140–1147.

[6] S. Lin, L. Yang, I. Saleemi, and S. Sengupta, "Robust high-resolution video matting with temporal guidance," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 3132–3141.

[7] S. Sengupta, V. Jayaram, B. Curless, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Background matting: The world is your green screen," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2288–2297.

[8] S. Lin, A. Ryabtsev, S. Sengupta, B. L. Curless, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Real-time high-resolution background matting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 8762–8771.

[9] C. Geng, S.-J. Huang, and S. Chen, "Recent advances in open set recognition: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3614–3631, Oct. 2020.

[10] Y. Xu, B. Liu, Y. Quan, and H. Ji, "Unsupervised deep background matting using deep matte prior," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4324–4337, Jul. 2022.

[11] F. Zhou, Y. Tian, and Z. Qi, "Attention transfer network for nature image matting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 6, pp. 2192–2205, Jun. 2021.

[12] B. Peng, M. Zhang, J. Lei, H. Fu, H. Shen, and Q. Huang, "RGB-D human matting: A real-world benchmark dataset and a baseline method," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 4041–4053, Aug. 2023.

[13] Y. Wang, L. Tang, Y. Zhong, and B. Li, "From composited to real-world: Transformer-based natural image matting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2097–2111, Apr. 2024.

[14] H. Lu, Y. Dai, C. Shen, and S. Xu, "Indices matter: Learning to index for deep image matting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3265–3274.

[15] Y. Sun, C.-K. Tang, and Y.-W. Tai, "Semantic image matting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11115–11124.

[16] Y. Liu et al., "Tripartite information mining and integration for image matting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7555–7564.

[17] C. Liu, H. Ding, and X. Jiang, "Towards enhancing fine-grained details for image matting," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 385–393.

[18] C.-L. Hsieh and M.-S. Lee, "Automatic trimap generation for digital image matting," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Oct. 2013, pp. 1–5.

[19] V. Gupta and S. Raman, "Automatic trimap generation for image matting," in *Proc. Int. Conf. Signal Inf. Process. (IConSIP)*, Oct. 2016, pp. 1–5.

[20] Y. Sun, G. Wang, Q. Gu, C.-K. Tang, and Y.-W. Tai, "Deep video matting via spatio-temporal alignment and aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6971–6980.

[21] H. Seong, S. W. Oh, B. Price, E. Kim, and J.-Y. Lee, "One-trimap video matting," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 430–448.

[22] W.-L. Huang and M.-S. Lee, "End-to-end video matting with trimap propagation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14337–14347.

[23] Y. Zhou, L. Zhou, T. L. Lam, and Y. Xu, "Sampling propagation attention with trimap generation network for natural image matting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 10, pp. 5828–5843, Apr. 2023.

[24] L. Hu, Y. Kong, J. Li, and X. Li, "Effective local-global transformer for natural image matting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 8, pp. 3888–3898, Aug. 2023.

[25] Y. Qiao et al., "Attention-guided hierarchical structure aggregation for image matting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13673–13682.

[26] C.-C. Lin et al., "Adaptive human matting for dynamic videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10229–10238.

[27] J. Li, V. Goel, M. Ohanyan, S. Navasardyan, Y. Wei, and H. Shi, "VMFormer: End-to-end video matting with transformer," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 6664–6673.

[28] Y. Zhong and I. Zharkov, "Lightweight portrait matting via regional attention and refinement," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 4146–4155.

[29] O. Ulutan, A. S. M. Iftekhar, and B. S. Manjunath, "VSGNet: Spatial attention network for detecting human object interactions using graph convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13617–13626.

[30] X. Zhong, C. Ding, X. Qu, and D. Tao, "Polysemy deciphering network for robust human–object interaction detection," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1910–1929, Jun. 2021.

[31] Y. Liao, S. Liu, F. Wang, Y. Chen, C. Qian, and J. Feng, "PPDM: Parallel point detection and matching for real-time human-object interaction detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 482–490.

[32] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning human-object interaction detection using interaction points," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4116–4125.

[33] J. Li, J. Zhang, and D. Tao, "Referring image matting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2023, pp. 22448–22457.

[34] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, Aug. 2017, pp. 1126–1135.

[35] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 403–412.

[36] R. Geirhos et al., "Shortcut learning in deep neural networks," *Nature Mach. Intell.*, vol. 2, no. 11, pp. 665–673, Nov. 2020.

[37] M. Caccia et al., "Online fast adaptation and knowledge accumulation (OSAKA): A new approach to continual learning," in *Proc. NIPS*, 2020, pp. 16532–16545.

[38] X. Song et al., "Rapidly adaptable legged robots via evolutionary meta-learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 3769–3776.

[39] R. Kaushik, T. Anne, and J.-B. Mouret, "Fast online adaptation in robotics through meta-learning embeddings of simulated priors," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 5269–5276.

[40] B. Li et al., "Invariant information bottleneck for domain generalization," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, Jun. 2022, pp. 7399–7407.

[41] C. Chen, J. Li, X. Han, X. Liu, and Y. Yu, "Compound domain generalization via meta-knowledge encoding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7119–7129.

[42] Y. Shu, Z. Cao, C. Wang, J. Wang, and M. Long, "Open domain generalization with domain-augmented meta-learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9624–9633.

[43] M. Bui, T. Tran, A. Tran, and D. Q. Phung, "Exploiting domain-specific features to enhance domain generalization," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 21189–21201.

[44] Y. Zhao et al., "Learning to generalize unseen domains via memory-based multi-source meta-learning for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6277–6286.

[45] L. Chen, Q. Li, C. Shen, J. Zhu, D. Wang, and M. Xia, "Adversarial domain-invariant generalization: A generic domain-regressive framework for bearing fault diagnosis under unseen conditions," *IEEE Trans. Ind. Informat.*, vol. 18, no. 3, pp. 1790–1800, Mar. 2022.

[46] Y. Wang, B. Xu, Z. Li, H. Huang, C. Lu, and Y. Guo, "Video object matting via hierarchical space-time semantic guidance," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 5109–5118.

[47] Y. Sun, C.-K. Tang, and Y.-W. Tai, "Ultrahigh resolution image/video matting with spatio-temporal sparsity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14112–14121.

[48] Z. Gu, W. Xian, N. Snavely, and A. Davis, "FactorMatte: Redefining video matting for re-composition tasks," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–14, Aug. 2023.

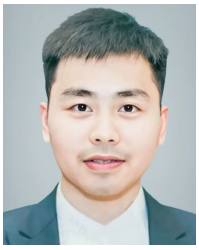
[49] J. Li, R. Henschel, V. Goel, M. Ohanyan, S. Navasardyan, and H. Shi, "Video instance matting," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 6654–6663.

[50] J. Li, M. Ohanyan, V. Goel, S. Navasardyan, Y. Wei, and H. Shi, "VideoMatt: A simple baseline for accessible real-time video matting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 2177–2186.

[51] J. Yao, X. Wang, L. Ye, and W. Liu, "Matte anything: Interactive natural image matting with segment anything model," *Image Vis. Comput.*, vol. 147, Jul. 2024, Art. no. 105067.
[52] K. James et al., "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, Mar. 2017.
[53] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.
[54] C. Rhemann, C. Rother, J. Wang, M. Gelautz, P. Kohli, and P. Rott, "A perceptually motivated online benchmark for image matting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1826–1833.
[55] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.
[56] P. Wu, X. Lu, J. Shen, and Y. Yin, "Clip fusion with bi-level optimization for human mesh reconstruction from monocular videos," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 105–115.
[57] C. Guo, T. Jiang, X. Chen, J. Song, and O. Hilliges, "Vid2Avatar: 3D avatar reconstruction from videos in the wild via self-supervised scene decomposition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12858–12868.
[58] H. Ouyang et al., "CoDeF: Content deformation fields for temporally consistent video processing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 8089–8099.

Jiang Xin (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science from Central South University, China, in 2017 and 2020, respectively, where he is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering. His research interests include computer vision, computer graphics, and generative AI.

Sheng Yue received the B.Sc. degree in mathematics and the Ph.D. degree in computer science from Central South University, China, in 2017 and 2022, respectively. He is currently a Post-Doctoral Researcher with the Department of Computer Science and Technology, Tsinghua University, China. His research interests include network optimization, distributed learning, and reinforcement learning.



Jinrui Zhang (Student Member, IEEE) received the B.Sc. and Ph.D. degrees in computer science from Central South University, Changsha, China, in 2016 and 2023, respectively. He is currently a Post-Doctoral Researcher with the Department of Computer Science and Technology, Tsinghua University, China. His research interests include mobile computing, edge AI, and operating system.

Ju Ren (Senior Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in computer science from Central South University, Changsha, China, in 2009, 2012, and 2016, respectively. He is currently an Associate Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. Prior to joining Tsinghua University, he was a Professor with the School of Computer Science and Engineering, Central South University. His research interests include the Internet-of-Things, edge computing, distributed and embedded AI, and operating systems. He is a member of ACM. He was recognized as a highly cited researcher by Clarivate in 2020 and 2022.

Feng Qian (Member, IEEE) received the B.S. degree from Shanghai Jiao Tong University, China, and the Ph.D. degree from the University of Michigan, Ann Arbor, MI, USA. He is currently an Associate Professor with the Ming Hsieh Department of Electrical and Computer Engineering, Viterbi School of Engineering, University of Southern California (USC), with a joint appointment at the Thomas Lord Department of Computer Science. Prior to joining USC, he held positions at AT&T Labs, Indiana University, and the University of Minnesota, Twin Cities. His research interests include mobile systems, augmented/virtual reality (AR/VR), mobile networking, wearable computing, real-world system measurements, and system security. He is a member of ACM.

Yaoxue Zhang (Senior Member, IEEE) received the B.Sc. degree from the Northwest Institute of Telecommunication Engineering, Xi'an, China, in 1982, and the Ph.D. degree in computer networking from Tohoku University, Sendai, Japan, in 1989. He is currently a Professor with the Department of Computer Science and Technology, Tsinghua University, Beijing, China. He has authored over 200 papers in peer-reviewed IEEE/ACM journals and conferences. His research interests include computer networking, operating systems, and transparent computing. He serves as the Editor-in-Chief for *Chinese Journal of Electronics* and is a fellow of Chinese Academy of Engineering.