

MODNet-V: Improving Portrait Video Matting via Background Restoration

Jiayu Sun^{1,2†} Zhangan Ke^{1†} Lihe Zhang² Huchuan Lu² Rynson W.H. Lau¹

¹Department of Computer Science, City University of Hong Kong

²School of Information and Communication Engineering, Dalian University of Technology

Abstract

To address the challenging portrait video matting problem more precisely, existing works typically apply some matting priors that require additional user efforts to obtain, such as annotated trimaps or background images. *In this work, we observe that instead of asking the user to explicitly provide a background image, we may recover it from the input video itself.* To this end, we first propose a novel background restoration module (BRM) to recover the background image dynamically from the input video. BRM is extremely *lightweight* and can be easily integrated into existing matting models. By combining BRM with a recent image matting model, MODNet, we then present MODNet-V for portrait video matting. Benefited from the strong background prior provided by BRM, MODNet-V has only 1/3 of the parameters of MODNet but achieves comparable or even better performances. Our design allows MODNet-V to be trained in an end-to-end manner on a single NVIDIA 3090 GPU. Finally, we introduce a new patch refinement module (PRM) to adapt MODNet-V for high-resolution videos while keeping MODNet-V lightweight and fast.

1. Introduction

Portrait video matting aims to extract the foreground portraits from a given video sequence. Accurately addressing this task in real time is essential for many practical applications, such as special effect generation in interactive video editing and background replacement in online meetings. Since video matting is a challenging problem due to the ill-posed nature of this task, early works [12, 4, 49, 43] typically expect the user to provide a pixel-wise trimap for each frame or for several consecutive key frames, which is a useful prior with great matting results. However, this trimap is unavailable in real-time applications as it requires human annotation. To alleviate this problem, Background Matting [31] proposes to use a pre-captured background

image as an additional input. Although Background Matting achieves real-time performance, it requires users to provide the additional background prior and will fail in scenarios with dynamic backgrounds (*i.e.*, videos with changing background contents). Instead of requiring auxiliary matting priors, some recent methods [48, 35, 32, 24] attempt to predict a matte from just the input image. However, these methods often produce unsatisfactory alpha mattes in difficult cases, *e.g.*, images with complex backgrounds, since missing matting priors increases model learning difficulty.

In this work, we observe that through dynamically accumulating as well as updating the background contents across consecutive video frames, we may recover a meaningful background image, which content changes dynamically as the background of the video changes, to serve as a prior for the matting task. Specifically, in a video sequence, the posture and location of the foreground portrait are usually changing continuously across frames, which means that the background contents in these frames are usually different and complementary. By accumulating and completing the missing background content frame-by-frame, we may produce a background image for the next frame. Compared with Background Matting [31], our approach does not require any user efforts. In addition, if we allow the background content to be updated continuously, we may handle videos with dynamic backgrounds.

Motivated by our observation, we propose a novel background restoration module (BRM) to help recover the image background dynamically. BRM maintains two latent variables: (1) *background features* to represent the restored background, and (2) a *binary mask* to indicate the recovered background pixels. At each frame time, BRM performs two steps. First, it uses the background features from the current frame as a prior for the matting model to produce a predicted alpha matte. Second, it extracts the new background pixels from the current input frame and accumulates these background pixels into the background features. In this step, BRM will also update the existing background pixels, in order to handle the video sequences with dynamic backgrounds. The design of BRM is extremely lightweight, with negligible computational overhead. BRM can be easily

[†]Equal Contribution

*This manuscript is a preprint.

integrated into the existing matting models without modifying the original model architectures.

If we apply background restoration as a matting prior, we may regard it as a sub-objective of the portrait video matting objective in MODNet [24], and we combine BRM with MODNet to form a new video matting model, named MODNet-V. Benefiting from the background prior provided by BRM, the proposed MODNet-V has only 1/3 of the parameters but achieves comparable or even better performance compared with MODNet. Moreover, we also introduce a new patch refinement module (PRM) to adapt MODNet-V to high-resolution videos. In summary, we make three-fold contributions. First, we observe that the background of a video frame can be restored from its previous frames, and we verify the feasibility of using the restored background as a video matting prior. Second, we propose a novel background restoration module (BRM) to restore the background prior for video matting. Third, We present MODNet-V, a new light-weight model that contains the aforementioned BRM and a newly proposed patch refinement module (PRM). Our new model achieves notable improvements in model size, inference speed, and matting performance/stability.

2. Related Work

2.1. Image Matting

Matting is an inherently under-constrained problem as its formula contains three unknown variables (F , B , and α) but only one known variable (I), as:

$$I = \alpha F + (1 - \alpha)B. \quad (1)$$

To solve Eq.(1), most of existing matting methods take an auxiliary trimap as a prior. The earlier traditional methods can be classified into sampling-based methods [13, 15, 16, 19, 22, 23, 37] and affinity-based methods [1, 2, 3, 9, 17, 28, 29]. However, these methods only utilise the low-level properties of image pixels. Instead, recent deep learning based methods leverage the high-level semantic information from neural networks and improve the matting results significantly. For example, Cho et al. [10] propose to combine neural network with close-form matting [29] and KNN matting [9]. Xu et al. [45] introduce the first end-to-end neural network for image matting. Besides, some methods promote the accuracy of alpha predictions by combining extra information (e.g., optimal trimaps [7], foreground information [20], etc.) learned from an additional network.

However, annotating a trimap is costly and time-consuming in practice. To address this problem, some recent methods either directly predict the alpha matte from the input image [39, 28, 50, 8, 32, 35, 48] or adopt the auxiliary prior [38]. Among these methods, there are two

different inputs for these methods: the natural images with diverse object categories and the portrait images. In terms of the natural images with diverse object categories, Qiao et al. [35] apply attention mechanism and Zhang et al. [48] integrate the possibility maps of foreground and background for alpha prediction. With regard to the portrait image matting, Shen et al. [39] apply a fully-convolutional network to generate a pseudo trimap before using an image matting layer to learn the Laplacian matrix. Chen et al. [8] first predict a low-resolution segmentation map, and then use it as matting guidance. Liu et al. [32] make use of coarsely annotated data in a three-stage coupled pipeline. Sengupta et al. [38] supply an extra background image as an alternative auxiliary cue to predict the alpha matte and foreground.

2.2. Video Matting

Although we can simply apply aforementioned algorithms on per frame for video matting, it results in temporal incoherence. By taking temporal relationship of frames into consideration, previous trimap-based video matting methods [12, 4, 43, 49] created coherent trimaps or mattes of video sequence by performing spatio-temporal optimisations. For example, Chuang et al. [12] propagate trimaps from a limited number of user-defined key-frames to the entire sequence by optical flow and predict alpha mattes via using Bayesian matting. Bai et al. [4] improve temporal coherence by a temporal matte filter while preserving matte structures on individual frames. Zou et al. [49] connect spatial and temporal relationship through non-local structure constructed by a sparse and low-rank representation. In addition to the methods employing low-level colour features, some learning-based methods employ deep structural and semantic features. Some methods [11, 30, 52] utilise non-local matting Laplacian of multiple frames to encode temporal coherency. Moreover, Sun et al. [43] introduce a trimap propagation network by spatio-temporal feature aggregation module.

Recently, trimap-free video matting methods draw much attention. Ke et al. [24] divide the matting objective into three sub-objectives and learn the consistency of each sub-objective for real-world portrait matting. However, this method fails in the coherence of prediction among frames as the temporal relationship has not been comprehensively considered. Following the design of Sengupta et al. [38], Lin et al. [31] also supply extra background image for real-time portrait video matting. This extra background is a strong prior for improving video matting performance, which has been demonstrated by the research of Lin et al [31]. However, using a captured background image as the prior has two main drawbacks which greatly limit the application of Background Matting in practice. First, background have to be static if using a captured background image, which means even small disturbances in the back-

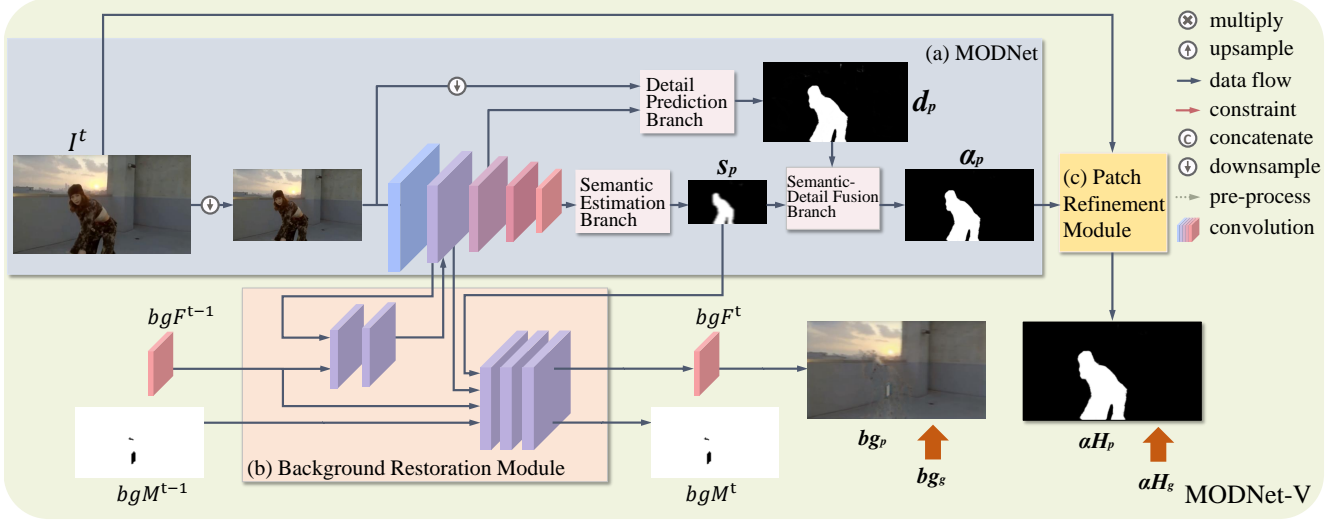


Figure 1. Architecture of MODNet-V. Our framework consists of a MODNet component adopted from [24], a background restoration module (BRM), and a patch refinement module (PRM). In each timestamp t , BRM provide a restored background feature bgF^{t-1} as a matting prior and update bgF under the guidance of a binary mask bgM . After the MODNet component predicts the alpha matte α_p from a downsample version of I^t , PRM detects the wrong patches in α_p and refines these patches in the original input resolution.

ground (e.g., light changes and slight jitter) will affect the matting results. Second, the background image requires additional user efforts to capture, and this process must be done carefully to ensure that the obtained background image is aligned with the video sequence. To eliminate the above issues, in this paper, we propose a novel BRM module, which restores background from the historical frames and then use it as a matting prior for coming frames.

2.3. Scene background modeling

Scene background modeling [6, 33] is the task of estimating a clear background image without foreground objects from a video sequence, which helps to identify the moving foreground objects in video. The scene background modeling methods always follow the same scheme. They predict an unreliable background model using the first few frames as their first step, and then gradually update background model by the analysis of the extracted foreground objects from the video sequences in an online process. The most traditional background modeling approaches are based on Gaussian distribution [34, 5]. Then, Stauffer and Grimson design a modeling of pixel-level color intensity variations based probability density functions which is a mixture of Gaussians [41, 42]. Following the Gaussian mixture model, Zivkovic [51] introduces an adaptive Gaussian mixture model, which is an iterative approach between median blending and spatial segmentation. Some methods [27, 40] further improve the Gaussian mixture model for background modeling. In addition, Kim et al. [25] establish the modeling of the background

by a codebook method. Laugraud et al. [26] generate background based on motion detection. Javed et al. [21] formulate background-foreground modeling as a low-rank matrix decomposition problem through Robust Principal Component Analysis (RPCA). Recently, some deep-learning-based methods also emerge. These methods apply weightless neural network [14], convolutional neural network [18, 36], and auto-encoder [46, 47] to the task of scene background modeling. The performance of the scene background modeling is profoundly affected by the presence of dynamic backgrounds, illumination changes, long-time static foreground, as well as global motions caused by camera jitter. In this paper, we try lessening the effects of dynamic backgrounds, illumination changes and other factors by the design of BRM.

3. Method

3.1. Review of MODNet

MODNet [24] proposed to decompose the challenging matting objective into three sub-objectives, including semantic estimation, detail prediction, and semantic-detail fusion. As shown in Fig. 1 (a), in each timestamp t , MODNet (denoted by f_{mod} here) predicts an alpha matte α_p^t from the input image frame I^t , as:

$$\alpha_p^t = f_{mod}(I^t) = \mathcal{F}(s_p^t, d_p^t) = \mathcal{F}(\mathcal{S}(I^t), \mathcal{D}(\mathcal{S}(I^t), I^t)), \quad (2)$$

where \mathcal{S} , \mathcal{D} , and \mathcal{F} are the three branches in MODNet. s_p^t and d_p^t are the predicted semantic and detail respectively. However, Eq. (2) does not take any matting prior as the in-

Algorithm 1 The Proposed Background Restoration Method.

Require: timestamp t ; AND operation $\&$; pixel-by-pixel multiplication \otimes

Require: background feature bgF ; background binary mask bgM ; background information bgI ; predicted semantic s_p

```

1:  $bgM_{new}^t = (bgM^{t-1} == 0) \& ((1 - s_p^t) > 0.5)$  // mask of background newly appeared in timestamp  $bgI$ 
2:
3:  $bgM_{avg}^t = (bgM^{t-1} == 1) \& ((1 - s_p^t) > 0.5)$  // mask of background appeared in both  $bgF$  and  $bgI$ 
4:
5:  $bgF_{tmp}^t = bgF^{t-1} + bgM_{new}^t \otimes bgI^t$  // add newly appeared background into  $bgF$ 
6:
7:  $bgF^t = (1 - bgM_{avg}^t) \otimes bgF_{tmp}^t + bgM_{avg}^t \otimes ((bgF_{tmp}^t + bgI^t)/2)$  // average background appeared in both  $bgF$  and  $bgI$ 
8:
9:  $bgM^t = bgM^{t-1} + bgM_{new}^t$  // update  $bgM$ 

```

put and overlooks the temporal information, leading to unsatisfactory alpha mattes predicted under complex scenarios and inconsistent alpha mattes predicted among video frames.

3.2. Overview of MODNet-V

To overcome the aforementioned problems of MODNet, we present MODNet-V, a light-weight video matting model contains newly proposed BRM and PRM.

BRM restores the background information from the historical frames. In MODNet-V, we use the background information restored by BRM as a matting prior, which is input to the semantic branch \mathcal{S} along with the coming frame. Such a prior potentially contains the temporal information that is essential for video matting. We connect BRM to \mathcal{S} by using:

$$s_p^t = \mathcal{S}(I^t, \mathcal{B}(I^t)), \quad (3)$$

where \mathcal{B} denotes BRM. Another module, PRM, refines the alpha matte predicted by Eq. (2) in a patch-based manner. As a result, MODNet-V can process high-resolution (e.g., 1080p or 2k) images with much less computational overhead than MODNet. We append PRM as:

$$\alpha_p^t = \mathcal{P}(\mathcal{F}(s_p^t, d_p^t)), \quad (4)$$

where \mathcal{P} denotes PRM. By combining Eq. (2)(3)(4), we can obtain the inference formula of MODNet-V:

$$\alpha_p^t = \mathcal{P}\left(\mathcal{F}\left(\mathcal{S}(I^t, \mathcal{B}(I^t)), \mathcal{D}(\mathcal{S}(I^t, \mathcal{B}(I^t)), I^t)\right)\right). \quad (5)$$

Benefit from BRM and PRM, we can also simplify the MODNet backbone used in MODNet-V to further reduce computational overhead. Specifically, **we make three modifications: (1) reducing the input resolution of \mathcal{S} , (2) deleting the last encoder block (the 32x downsampling block), and (3) removing several convolutional layers in \mathcal{D} .**

In the following two subsections, we will delve into BRM and PRM respectively.

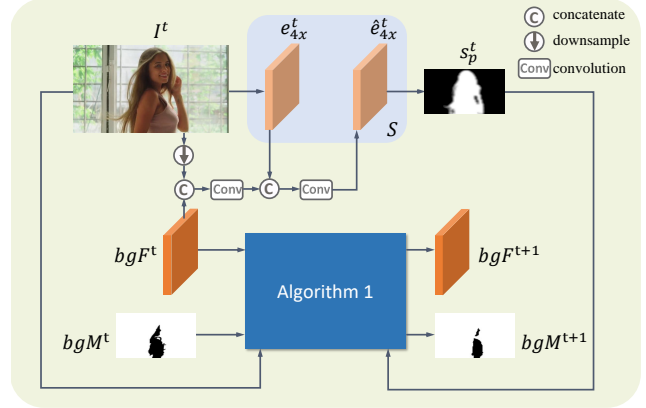


Figure 2. Illustration of BRM.

3.3. Background Restoration Module (BRM)

Fig. 2 illustrates the architecture of BRM. At each timestamp, BRM has two responsibilities. First, BRM should provide a background prior for predicting the alpha matte of the current frame. Second, BRM should extract the background information from the current frame and accumulate it to improve the restored background. In our design, BRM achieves these by maintaining (1) a background feature bgF to store the background information and (2) a binary mask bgM to indicate the pixels with restored background.

For the first responsibility, the background feature bgF^{t-1} from the previous timestamp will be regarded as the prior for the current frame I^t . In practice, we concatenate bgF^{t-1} with I^t and the 4x downsampled low-level feature e_{4x}^t from \mathcal{S} to predict alpha matte α^t . For the second responsibility, BRM extracts the background information bgI^t of the current frame by using:

$$bgI^t = (1 - s_p^t) \otimes e_{4x}^t, \quad (6)$$

where \otimes is pixel-by-pixel multiplication. bgI^t is then accumulated to bgF under the guided of bgM and s_p , as summarized in Alg. 1. In this way, bgF , which is continuously delivered and improved in temporal, can provide a strong prior for alpha matte prediction without any user effort or

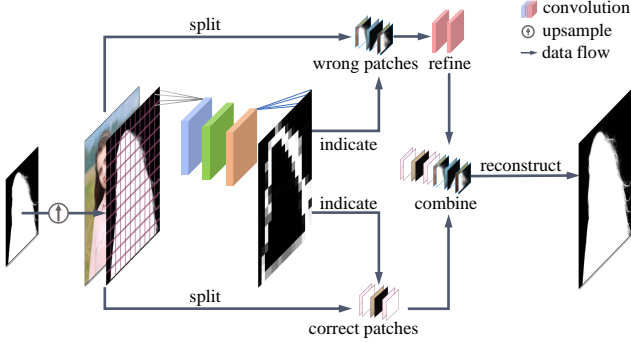


Figure 3. Illustration of PRM.

auxiliary inputs. Note that both bgF and bgM are initialized by 0 for the first video frame.

3.4. Patch Refinement Module (PRM)

Although MODNet is carefully designed for reducing inference time, it is still slow when processing images of 1080p or a higher resolution. In order to speed up the matting process of high-resolution images, Background Matting V2 [31] proposed to first generate a coarse alpha matte from a downsampled input image and then upsample it to the original size for refinement. In the refinement stage, the image with original resolution is cropped into 8×8 patches, and the refinement model only processes the patches with a high error probability. However, we notice that such a strategy has two possible issues when images has 2k or a higher resolution. **First, the total number of patches is proportional to image resolution, which means patch selection and cropping will be more time-consuming under a high resolution. Second, as the image resolution increases, the information contained in each 8×8 region will greatly decrease. As a result, artifacts are likely to appear in the edges of refined patches.**

In this work, we proposed a new patch refinement module (PRM). As shown in Fig. 3, different from the size-fixed patch used in Background Matting V2, we divide the input images into a fixed number of patches. Among these patches, the wrong patches will be refined and combined with the other correct patches to form a new alpha matte.

Formally, for an input image of size (h, w) , PRM first downsamples it and predict an initial coarse alpha matte. PRM then upsamples the initial coarse alpha matte into the original image size and divides it into $k \times k$ patches of size $(\frac{h}{k}, \frac{w}{k})$. Meanwhile, PRM applies an adaptive pooling layer and two convolutional layers to the initial coarse alpha matte to predict a flaw map (proposed by GCT) of size (k, k) , where each pixel value corresponds to the flaw probability of each patch. Finally, PRM only refines the patch whose flaw probability is higher than a predefined threshold ξ . In practice, for images with a resolution of

4k and below, we set $k = 16$ (the maximum patch size is only 256×256). For images with a resolution higher than 4k, we set $k = 32$. We set $\xi = 0.01$, which constrains only 15% of all patches has a flaw probability higher than ξ . This means that PRM can reduce the refinement overhead by ~ 6.5 times, making it possible for MODNet-V to process high-resolution images.

Our PRM has a consistent time-consuming for patch selection under different input resolutions. In addition, benefit from the larger patch resolution, our PRM can effectively reduce artifacts on fine boundaries (such as the hair regions).

3.5. Training Losses

We apply loss \mathcal{L}_α that is the same as MODNet to learn alpha matte. In addition, we propose \mathcal{L}_{bg} as an explicit constraint to measure the difference between the ground truth background image bg_g^t and the background image bg_p^t predicted from bgF^t , as:

$$\mathcal{L}_{bg} = \sum_{t=1}^N \gamma \sqrt{(bg_p^t - bg_g^t)^2 + \epsilon^2}, \quad (7)$$

where ϵ is a small constant value, and γ is a binary mask that equals to 4 in the portrait boundaries otherwise 1. The purpose of γ is to let \mathcal{L}_{bg} focus more on the difficult boundary regions. We also calculate the constraint $\mathcal{L}_{\alpha H}$ between the ground truth high-resolution alpha matte αH_g and the predicted high-resolution alpha matte αH_p for PRM, as:

$$\mathcal{L}_{\alpha H} = \gamma \sqrt{(\alpha H_p - \alpha H_g)^2 + \epsilon^2}, \quad (8)$$

where ϵ and γ has the same meaning as Eq. (7).

4. Experiments

4.1. Datasets

We follow MODNet to train our network on both labeled and unlabeled datasets. We use the VideoMatte240K dataset [31] proposed by Background Matting V2 as the labeled foregrounds and ~ 1000 video clips from the internet as the unlabeled data. Note that we also annotate one frame in each video as the labeled data.

To prepare a labeled training sample, we randomly take 10 consecutive frames from a video sequence and composite them with a dynamic background sequence that is generated from a single background image by applying translation and affine operations. For each video sequence, we composite it with 20 background images, and we finally construct the dataset with 9680 videos.

4.2. Training Strategy

Our MODNet-V applies a two-stage training strategy.

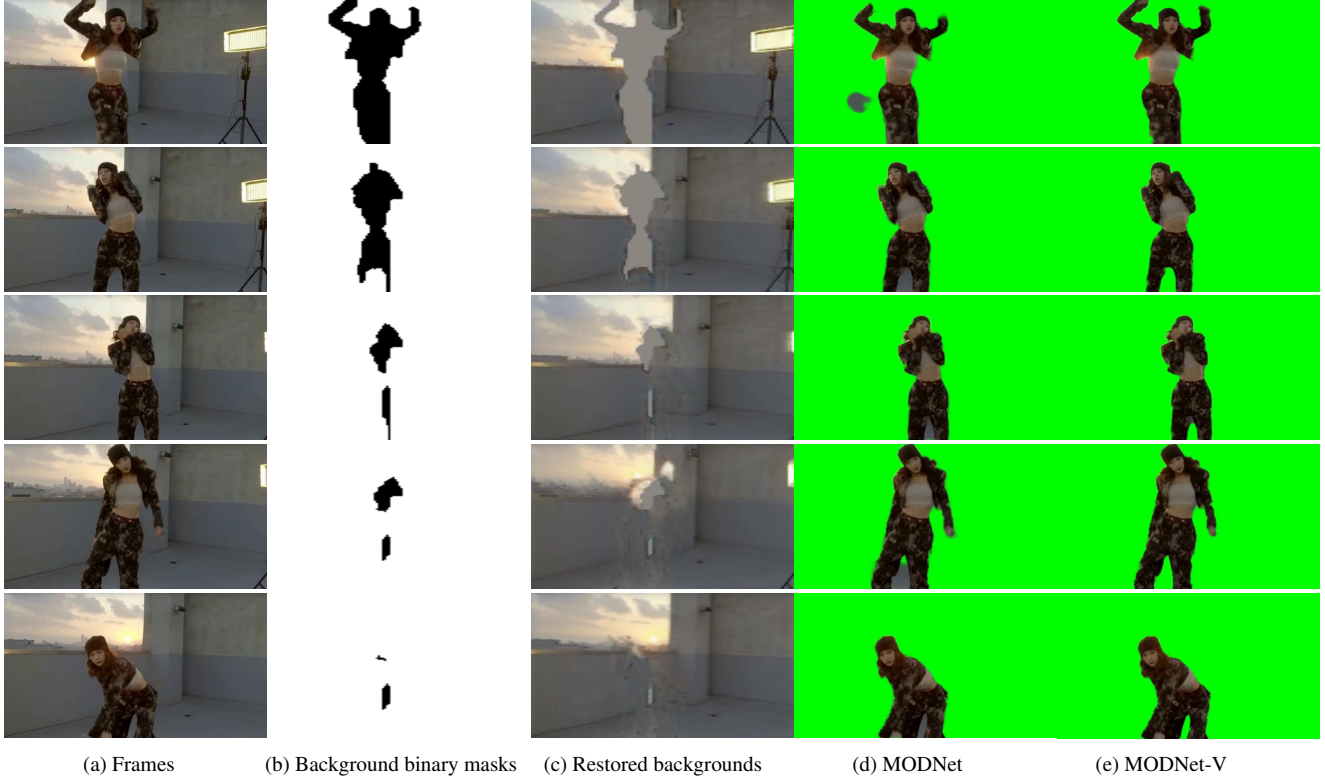


Figure 4. Visual Comparison of MODNet and MODNet-V. The proposed BRM restores a background image (c), which helps MODNet-V achieve better and more stable portrait matting results.

In the first stage, we train the MODNet component and BRM in an end-to-end manner. To save the GPU memory, in each iteration, we select a random number $T \in [1, 10]$ to indicate the forward frames, and we only calculate the loss of the 1_{th} frame and the T_{th} frame for optimization, *i.e.*, no gradients are cached and backward among the $[2, T - 1]$ frames. As a result, our first training stage can be performed in a single RTX 3090 GPU (24G memory) with a batch size of 16. We train the first stage for 30 epochs by using the SGD optimizer. The initial learning rate is 0.01 and will be decreased by multiplying 0.1 in the 15_{th} and 25_{th} epochs.

In the second stage, we train PRM by using the high-resolution labeled samples only. In this stage, we fix the weights of both the MODNet component and BRM. We use the SGD optimizer with the learning rate of 0.001 to train this stage for 10 epochs, and we set the batch size to 1.

4.3. Performance Comparison

Table 1 shows the evaluations on a real-world portrait video matting dataset [44] which contains 19 videos with 711 labeled frames. We can see that the MODNet-V outperforms MODNet in terms of MAD and MSE. In Fig. 4, we visualize the matting results of MODNet and MODNet-V for a video sequence.

Table 1. Quantitative results of MODNet and MODNet-V.

Methods	MAD(10^{-4})	MSE(10^{-4})
MODNet	83.30	34.05
MODNet-V	65.63	32.39

5. Conclusion

In this paper, we presented MODNet-V for portrait video matting. We designed its architecture based on the observation that the background of a video frame can be restored by accumulating the background information from historical frames. In MODNet-V, we proposed a light-weight BRM that performs iteratively with MODNet to construct background as the prior for obtaining better matting results. Besides, we introduced a new PRM module that can adapt our method to high-resolution videos. The results on various real-world videos show that the proposed MODNet-V can significantly improve the quality and stability of , compared with previous SOTA like MODNet.

References

- [1] Yağiz Aksoy, Tae-Hyun Oh, Sylvain Paris, Marc Pollefeys, and Wojciech Matusik. Semantic soft segmentation. *ACM*

- TOG*, 37(4):1–13, 2018. 2
- [2] Yagiz Aksoy, Tunc Ozan Aydin, and Marc Pollefeys. Designing effective inter-pixel information flow for natural image matting. In *CVPR*, pages 29–37, 2017. 2
 - [3] Xue Bai and Guillermo Sapiro. A geodesic framework for fast interactive image and video segmentation and matting. In *ICCV*, pages 1–8. IEEE, 2007. 2
 - [4] Xue Bai, Jue Wang, and David Simons. Towards temporally-coherent video matting. In *International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications*, pages 63–74. Springer, 2011. 1, 2
 - [5] Yannick Benezeth, Pierre-Marc Jodoin, Bruno Emile, Hélène Laurent, and Christophe Rosenberger. Review and evaluation of commonly-implemented background subtraction algorithms. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008. 3
 - [6] Thierry Bouwmans, Lucia Maddalena, and Alfredo Petrosino. Scene background initialization: A taxonomy. *Pattern Recognition Letters*, 96:3–11, 2017. 3
 - [7] Shaofan Cai, Xiaoshuai Zhang, Haoqiang Fan, Haibin Huang, Jiangyu Liu, Jiaming Liu, Jiaying Liu, Jue Wang, and Jian Sun. Disentangled image matting. In *ICCV*, pages 8819–8828, 2019. 2
 - [8] Quan Chen, Tiezheng Ge, Yanyu Xu, Zhiqiang Zhang, Xinxin Yang, and Kun Gai. Semantic human matting. In *ACM MM*, pages 618–626, 2018. 2
 - [9] Qifeng Chen, Dingzeyu Li, and Chi-Keung Tang. Knn matting. *IEEE TPAMI*, 35(9):2175–2188, 2013. 2
 - [10] Donghyeon Cho, Yu-Wing Tai, and Inso Kweon. Natural image matting using deep convolutional neural networks. In *ECCV*, pages 626–643. Springer, 2016. 2
 - [11] Inchang Choi, Minhaeng Lee, and Yu-Wing Tai. Video matting using multi-frame nonlocal matting laplacian. In *ECCV*, pages 540–553. Springer, 2012. 2
 - [12] Yung-Yu Chuang, Aseem Agarwala, Brian Curless, David H Salesin, and Richard Szeliski. Video matting of complex scenes. In *SIGGRAPH*, pages 243–248, 2002. 1, 2
 - [13] Yung-Yu Chuang, Brian Curless, David H Salesin, and Richard Szeliski. A bayesian approach to digital matting. In *CVPR*, volume 2, pages II–II. IEEE, 2001. 2
 - [14] Massimo De Gregorio and Maurizio Giordano. Background estimation by weightless neural networks. *Pattern Recognition Letters*, 96:55–65, 2017. 3
 - [15] Xiaoxue Feng, Xiaohui Liang, and Zili Zhang. A cluster sampling method for image matting via sparse coding. In *ECCV*, pages 204–219. Springer, 2016. 2
 - [16] Eduardo SL Gastal and Manuel M Oliveira. Shared sampling for real-time alpha matting. In *Computer Graphics Forum*, volume 29, pages 575–584. Wiley Online Library, 2010. 2
 - [17] Leo Grady, Thomas Schiwietz, Shmuel Aharon, and Rüdiger Westermann. Random walks for interactive alpha-matting. In *Proceedings of VIIP*, volume 2005, pages 423–429, 2005. 2
 - [18] Ibrahim Halfaoui, Fahd Bouzaraa, and Onay Urfalioglu. Cnn-based initial background estimation. In *ICPR*, pages 101–106. IEEE, 2016. 3
 - [19] Kaiming He, Christoph Rhemann, Carsten Rother, Xiaoou Tang, and Jian Sun. A global sampling method for alpha matting. In *CVPR*, pages 2049–2056. IEEE, 2011. 2
 - [20] Qiqi Hou and Feng Liu. Context-aware image matting for simultaneous foreground and alpha estimation. In *ICCV*, pages 4130–4139, 2019. 2
 - [21] Sajid Javed, Arif Mahmood, Thierry Bouwmans, and Soon Ki Jung. Background-foreground modeling based on spatiotemporal sparse subspace clustering. *IEEE Transactions on Image Processing*, 26(12):5840–5854, 2017. 3
 - [22] Jubin Johnson, Ehsan Shahrian Varnousfaderani, Hisham Cholakkal, and Deepu Rajan. Sparse coding for alpha matting. *IEEE TIP*, 25(7):3032–3043, 2016. 2
 - [23] Levent Karacan, Aykut Erdem, and Erkut Erdem. Image matting with kl-divergence based sparse sampling. In *ICCV*, pages 424–432, 2015. 2
 - [24] Zhanghan Ke, Kaican Li, Yurou Zhou, Qihua Wu, Xiangyu Mao, Qiong Yan, and Rynson WH Lau. Is a green screen really necessary for real-time portrait matting? *arXiv preprint arXiv:2011.11961*, 2020. 1, 2, 3
 - [25] Kyungnam Kim, Thanarat H Chalidabhongse, David Harwood, and Larry Davis. Background modeling and subtraction by codebook construction. In *2004 International Conference on Image Processing, 2004. ICIP'04.*, volume 5, pages 3061–3064. IEEE, 2004. 3
 - [26] Benjamin Laugraud, Sébastien Piérard, and Marc Van Droogenbroeck. Labgen: A method based on motion detection for generating the background of a scene. *Pattern Recognition Letters*, 96:12–21, 2017. 3
 - [27] Dar-Shyang Lee. Effective gaussian mixture learning for video background subtraction. *IEEE transactions on pattern analysis and machine intelligence*, 27(5):827–832, 2005. 3
 - [28] Anat Levin, Dani Lischinski, and Yair Weiss. A closed-form solution to natural image matting. *IEEE TPAMI*, 30(2):228–242, 2007. 2
 - [29] Anat Levin, Alex Rav-Acha, and Dani Lischinski. Spectral matting. *IEEE TPAMI*, 30(10):1699–1712, 2008. 2
 - [30] Dingzeyu Li, Qifeng Chen, and Chi-Keung Tang. Motion-aware knn laplacian for video matting. In *ICCV*, pages 3599–3606, 2013. 2
 - [31] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting. In *CVPR*, pages 8762–8771, 2021. 1, 2, 5
 - [32] Jinlin Liu, Yuan Yao, Wendi Hou, Miaomiao Cui, Xuansong Xie, Changshui Zhang, and Xian-sheng Hua. Boosting semantic human matting with coarse annotations. In *CVPR*, pages 8563–8572, 2020. 1, 2
 - [33] Alfredo Petrosino, Lucia Maddalena, and Thierry Bouwmans. Scene background modeling and initialization. *Pattern Recognition Letters*, 96:1–2, 2017. 3
 - [34] Massimo Piccardi. Background subtraction techniques: a review. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, volume 4, pages 3099–3104. IEEE, 2004. 3
 - [35] Yu Qiao, Yuhao Liu, Xin Yang, Dongsheng Zhou, Mingliang Xu, Qiang Zhang, and Xiaopeng Wei. Attention-guided hier-

- archical structure aggregation for image matting. In *CVPR*, pages 13676–13685, 2020. 1, 2
- [36] Zhenshen Qu, Shimeng Yu, and Mengyu Fu. Motion background modeling based on context-encoder. In *2016 Third International Conference on Artificial Intelligence and Pattern Recognition*, pages 1–5. IEEE, 2016. 3
- [37] Mark A Ruzon and Carlo Tomasi. Alpha estimation in natural images. In *CVPR*, volume 1, pages 18–25. IEEE, 2000. 2
- [38] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *CVPR*, pages 2291–2300, 2020. 2
- [39] Xiaoyong Shen, Xin Tao, Hongyun Gao, Chao Zhou, and Jiaya Jia. Deep automatic portrait matting. In *ECCV*, pages 92–107. Springer, 2016. 2
- [40] Atsushi Shimada, Daisaku Arita, and Rin-ichiro Taniguchi. Dynamic control of adaptive mixture-of-gaussians background model. In *2006 IEEE International Conference on Video and Signal Based Surveillance*, pages 5–5. IEEE, 2006. 3
- [41] Chris Stauffer and W Eric L Grimson. Adaptive background mixture models for real-time tracking. In *Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149)*, volume 2, pages 246–252. IEEE, 1999. 3
- [42] Chris Stauffer and W. Eric L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):747–757, 2000. 3
- [43] Yanan Sun, Guanzhi Wang, Qiao Gu, Chi-Keung Tang, and Yu-Wing Tai. Deep video matting via spatio-temporal alignment and aggregation. In *CVPR*, pages 6975–6984, 2021. 1, 2
- [44] Tiantian Wang, Sifei Liu, Yapeng Tian, Kai Li, and Ming-Hsuan Yang. Video matting via consistency-regularized graph neural networks. In *ICCV*. IEEE, 2021. 6
- [45] Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. Deep image matting. In *CVPR*, pages 2970–2979, 2017. 2
- [46] Pei Xu, Mao Ye, Xue Li, Qihe Liu, Yi Yang, and Jian Ding. Dynamic background learning through deep auto-encoder networks. In *ACM MM*, pages 107–116, 2014. 3
- [47] Pei Xu, Mao Ye, Qihe Liu, Xudong Li, Lishen Pei, and Jian Ding. Motion detection via a couple of auto-encoder networks. In *ICME*, pages 1–6. IEEE, 2014. 3
- [48] Yunke Zhang, Lixue Gong, Lubin Fan, Peiran Ren, Qixing Huang, Hujun Bao, and Weiwei Xu. A late fusion cnn for digital matting. In *CVPR*, pages 7469–7478, 2019. 1, 2
- [49] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 1, 2
- [50] Bingke Zhu, Yingying Chen, Jinqiao Wang, Si Liu, Bo Zhang, and Ming Tang. Fast deep matting for portrait animation on mobile phone. In *ACM MM*, pages 297–305, 2017. 2
- [51] Zoran Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *ICPR*, volume 2, pages 28–31. IEEE, 2004. 3
- [52] Dongqing Zou, Xiaowu Chen, Guangying Cao, and Xiaogang Wang. Unsupervised video matting via sparse and low-rank representation. *IEEE TPAMI*, 42(6):1501–1514, 2019. 2