Figure 7. Exemplary frames (first column) of the VIM50 benchmark with corresponding ground truth alpha mattes and instance information visualized in terms of the color encoding (second column).

Appendix

This supplementary material elaborates on further aspects of our work regarding the benchmark VIM50 and the model MSG-VIM. In Appendix A, we show additional video frames and ground truth data of the VIM50 test set. Appendix B shows more qualitative results of MSG-VIM. Additional details on the matting model architecture of MSG-VIM are provided in Appendix C. Appendix D provides the parameter study we used to set the hyperparameters of MSG-VIM for the experiments presented in Section 5.

A. VIM50 Benchmark

To supplement the VIM50 samples presented in Section 3.2, we show clips from five sequences of the VIM50 benchmark in Figure 8. They depict two to four human instances as foreground objects with some frames containing heavy occlusions. Additional samples of the benchmark with corresponding individual ground truth alpha mattes are shown in Figure 7. Ground truth alpha mattes belonging to the same person are colored consistently across video frames.

B. Additional Qualitative Results

We visualize in Figure 9 additional qualitative results on selected video frames of VIM50. Similar to the qualitative comparison conducted for Figure 4, we use mask guidance from MaskTrackRCNN [47] and compare MSG-VIM with our transformation of MGMatting and InstMatt to video instance matting. The results show visually a significant advantage of MSG-VIM over the baseline methods.

Layers	Output Size	MSG-VIM
Convolution	256 × 256	[3×3 Conv] × 2
ResNet Block (1)	128 × 128	[3×3 Conv 3×3 Conv] × 3
ResNet Block (2)	64 × 64	[3×3 Conv 3×3 Conv] × 4
ResNet Block (3)	32 × 32	[3×3 Conv 3×3 Conv] × 4
ResNet Block (4)	16 × 16	[3×3 Conv 3×3 Conv] × 2
ASPP	16 × 16	dilations = [1, 2, 4, 8]
Upsample Block (1)	32 × 32	[3×3 Conv 3×3 Conv] × 2
Upsample Block (2)	64 × 64	[3×3 Conv 3×3 Conv] × 3
Upsample Block (3)	128 × 128	[3×3 Conv 3×3 Conv] × 3
Upsample Block (4)	256 × 256	[3×3 Conv 3×3 Conv] × 2
ConvRNN	256 × 256	TFG
Deconvolution	512 × 512	4×4 Deconv

Table 8. The detailed architecture of the matting network U used in MSG-VIM. TFG denotes temporal feature guidance presented in Section 4.3. Pooling layers and normalization layers are omitted for simplicity.

C. Matting Architecture

In Table 8 we present additional details on the architecture of the encoder-decoder-based matting network MSG-VIM (compare Section 4.1). The encoder is adopted from the modified ResNet-34 [49]. Each of its ResNet blocks contains consecutive 3×3 convolution layers with a final average pooling layer to downsample the feature maps. The decoder has multiple upsample blocks. Each one consists of consecutive 3×3 convolution layers with a final 2D nearest neighbor upsampling layer. The temporal feature guidance (TFP) module, which is implemented using a Convolution-based RNN (ConvRNN) network, is applied to the second largest feature map with a resolution of 256×256 in the decoder, as described in Section 4.3. At the first frame $t = 0$, we initialize the internal state via $h_0 = \tanh(\text{Conv}(F_0^S))$. After this stage, a 4×4 deconvolution layer with stride 2 is used to upsample the feature map to size 512×512 for the final predictions of the alpha mattes.

D. Parameter Study: Chunk Length

Given a video sequence of length T, we split the input (video and mask guidance) into chunks of t consecutive frames for inference. Then, we run inference on each chunk independently and concatenate results together. Identity in-

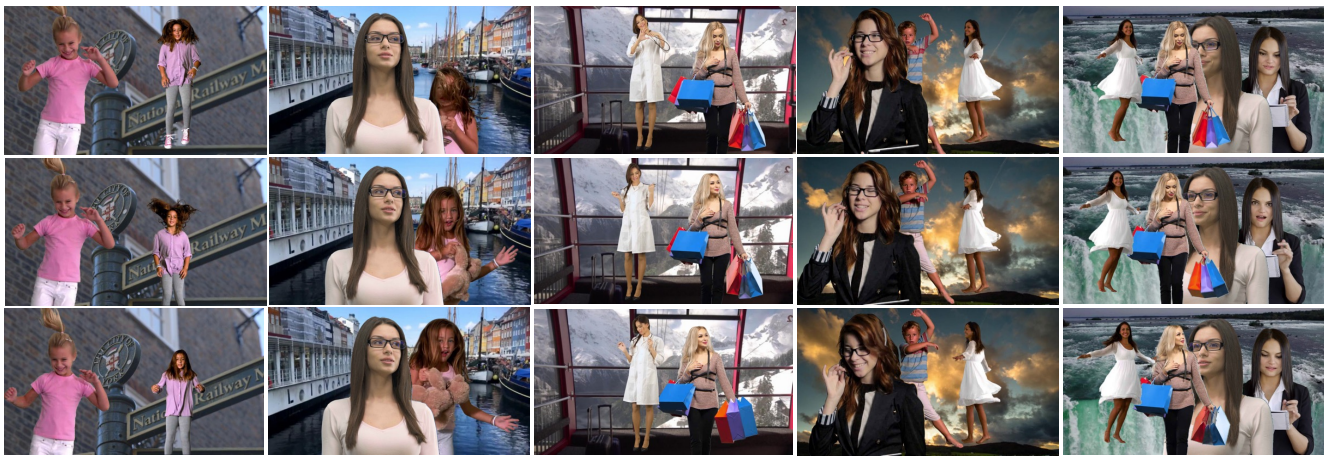

Figure 8. Each column shows exemplary frames of one sequence of the VIM50 benchmark. Some frames contain heavy occlusions between persons, making it challenging for current VIS/VM/VIM methods to obtain accurate alpha matte predictions.

Length of Chunk	RQ↑	TQ↑	MQ_{mse} ↑	VIMQ_{mse} ↑	MQ_{mad} ↑	VIMQ_{mad} ↑	MQ_{dtssd} ↑	$\text{VIMQ}_{\text{dtssd}}$ ↑
t = 1	72.12	92.03	54.74	36.33	39.10	25.95	27.02	17.93
t = 5	72.26	**93.27**	56.15	38.06	40.31	27.32	28.36	19.22
t = 10	**72.72**	93.17	**56.52**	**38.29**	**40.49**	**27.43**	**28.51**	**19.32**

Table 9. Analysis on the chunk length used during inference of MSG-VIM with mask sequence guidance from MaskTrackRCNN [47]. **Bold** numbers indicate best performance among all models.

formation across chunks are maintained from the underlying mask sequence generator, *e.g.* MaskTrackRCNN.

In Table 9, we analyze the impact of the video length that is processed during the inference of one chunk. The results show that the performance of the model improves with longer length t. The proposed temporal feature guidance module is thus an effective approach to exploit temporal information. Accordingly, we have set $t = 10$ in all experiments presented in Section 5. Note that we could not process chunks with a length larger than $t = 10$ due to memory limitations.

Figure 9. Video instance matting results of different models on VIM50. For each row, the first column shows the input frame, column 2-5 show the matting result of the respective frame and method. Difficult cases are highlighted with red boxes. Please zoom in for details.