StyleBin: Stylizing Video by Example in Stereo (Supplementary Material)

Michal Kučera
CTU in Prague, FEE
Prague, Czech Republic
kucerm22@fel.cvut.cz

David Mould
Carleton University, SCS
Ottawa, Canada
mould@scs.carleton.ca

Daniel Sýkora
CTU in Prague, FEE
Prague, Czech Republic
sykorad@fel.cvut.cz

1 INTRODUCTION

In this supplementary material we present settings of all tunable parameters (see Table 1) that were used to generate results presented in Figures 5 and 6 in the main paper.

Table 1: Settings of all tunable parameters mentioned in the main paper.

parameter	description	value
$\lvert s' \rvert$	source patch size	5x5
w_{dis}	disparity coherence weight	1.0
w_{tex}	textural coherence weight	1.0
w_{color}	color guide weight	8.0
w_{edge}	edge guide weight	4.0
w_{pos}	positional guide weight	10.0
w_{stereo}	stereo coherence weight	1.0
w_{temp}	temporal coherence weight	8.0
w_{uni}	patch uniformity weight	1.0

In addition we also provide quantitative and qualitative evaluation of our method by comparing it with two alternative baseline approaches that can be used to stylize video by example in stereo.

2 EVALUATION

We first evaluate the *stylize-and-warp* method where the example-based video stylization framework of Jamriška et al. [2019] is initially applied to the original monocular video sequence and then the final stereoscopic output is produced by warping. We then consider the *warp-and-stylize* approach, in which we first warp the monocular sequence to produce left and right views and then apply the technique of Jamriška et al. to each view separately. In the following sections we quantitatively and qualitatively compare our approach with these two baseline techniques.

2.1 Stylize-and-Warp

To quantitatively compare our method with the *stylize-and-warp* approach, we measure how well these two techniques reproduce the user-specified stylized keyframe(s). To do that, for each stylized frame O_i and its corresponding keyframe S_k we compute the following style consistency metric:

$$\begin{aligned} \mathrm{M}_{\text{style}}(S_k, O_i^L, O_i^R) = & \sum_{\hat{t}^L \in O_i^L} \min_{\hat{s}^L \in S_k} \mathrm{E}^L_{\text{style}}(\hat{s}^L, \hat{t}^L) \\ & + \sum_{\hat{t}^R \in O_i^R} \min_{\hat{s}^R \in S_k} \mathrm{E}^R_{\text{style}}(\hat{s}^R, \hat{t}^R), \end{aligned} \tag{1}$$

where S_k is the style exemplar (corresponding to the keyframe T_k), O_i^L and O_i^R are left/right views of the stylized output frame O_i, and $\hat{s}^L$, $\hat{s}^R$, $\hat{t}^L$, $\hat{t}^R$ are left/right source/target patches. Finally, the style consistency measure $\mathrm{E}^V_{\text{tex}}$ is defined as follows:

$$\mathrm{E}^V_{\text{style}}(\hat{s}, \hat{t}) = \sum_{s \in \hat{s},\, t \in \hat{t}} |S_k(s) - O_i^V(t)|^2. \tag{2}$$

Here s and t denote source/target pixels within patches $\hat{s}$ and $\hat{t}$ and V stands for a viewpoint, either L (left) or R (right). The style consistency metric corresponds to a texture coherence term used by texture synthesis algorithms [Kwatra et al. 2005; Wexler et al. 2007] to ensure that the synthesized texture resembles the exemplar.

Quantitative comparison of our approach with the *stylize-and-warp* is shown in Fig. 1. Lower values of M_{tex} correspond to smaller error. All graphs indicate that namely at disocclusions the output produced by our method preserve the original style exemplar more faithfully than the *stylize-and-warp* approach. Qualitatively this difference is mainly manifested by smearing artifacts visible in Fig. 2, however, there can also be subtle changes in shape. In contrast, our approach fills these critical areas with a consistent texture.

2.2 Warp-and-Stylize

The *warp-and-stylize* approach has texture dissimilarity scores M_{tex} comparable or even better than those of our technique. This is mainly due to the unconstrained optimization which does not take into account the view-independent disparity. A key issue here, however, is that when those independently stylized images are viewed in stereo, the view inconsistency may cause the observer to experience unpleasant eye strain, the extent of which can be quantitatively measured using the following stereo consistency metric:

$$\mathrm{M}_{\text{stereo}}(S_k, O_i^L, O_i^R) = \sum_{\hat{l} \in O_i^L} \mathrm{E}^L_{\text{stereo}}(\hat{l}) + \sum_{\hat{r} \in O_i^R} \mathrm{E}^R_{\text{stereo}}(\hat{r}). \tag{3}$$

Here $\hat{l}$ and $\hat{r}$ are all patches taken from left O_i^L and right O_i^R output image respectively. Stereo consistency over each patch is measured for left $\mathrm{E}^L_{\text{stereo}}$ and right $\mathrm{E}^R_{\text{stereo}}$ view separately:

$$\mathrm{E}^L_{\text{stereo}}(\hat{l}) = \sum_{l \in \hat{l}} |O_i^L(l) - O_i^R(l + D_i^L(l))|^2 \tag{4}$$

and

$$\mathrm{E}^R_{\text{stereo}}(\hat{r}) = \sum_{r \in \hat{r}} |O_i^L(r - D_i^R(r)) - O_i^R(r)|^2, \tag{5}$$

where l and r denote individual pixels of patches $\hat{l}$ and $\hat{r}$. D_i^L is left and D_i^R right disparity map that stores relative shift vectors of the corresponding pixels in the opposite view.

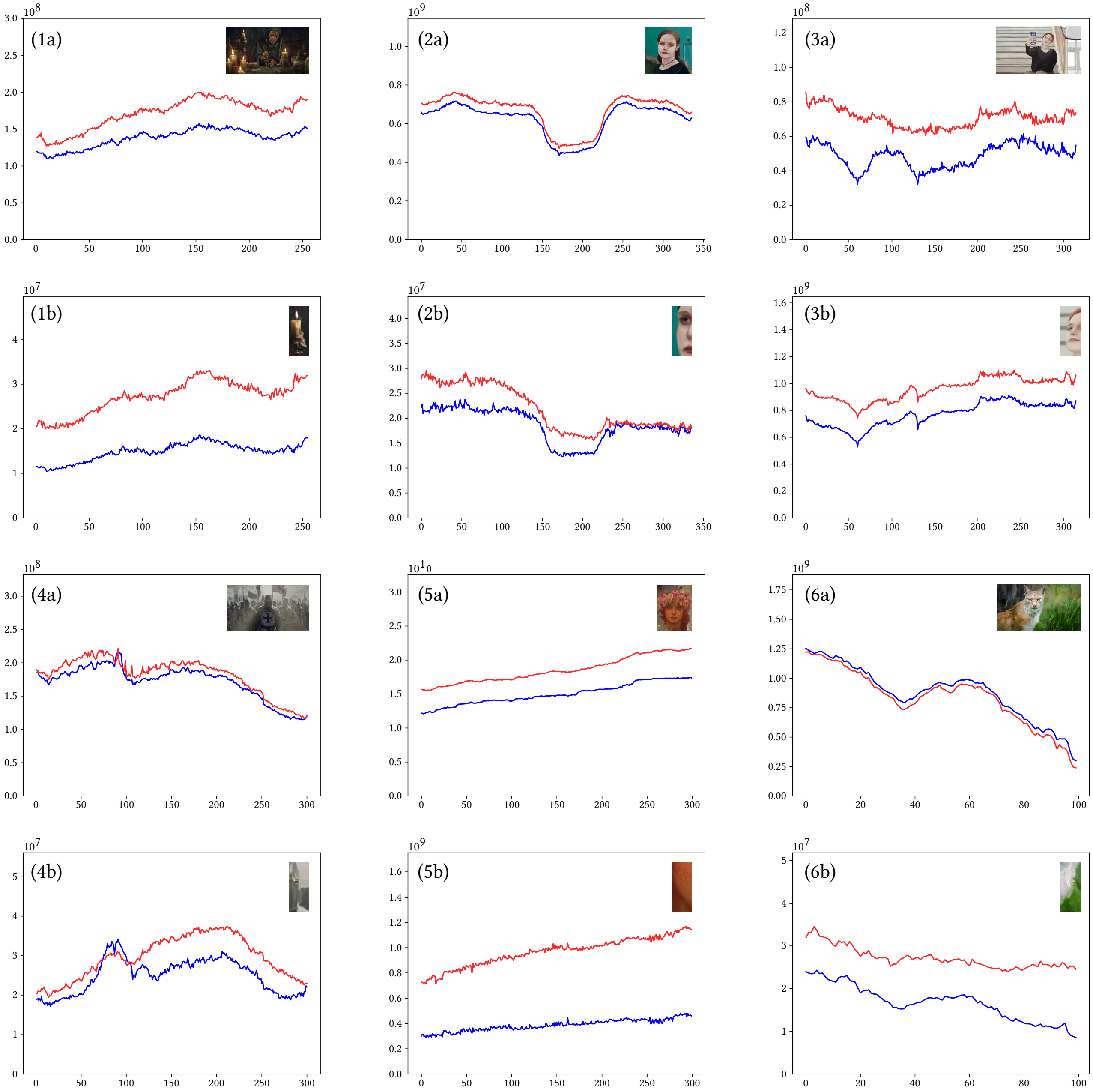

Figure 1: Quantitative evaluation of the ability to reproduce the original style using the *stylize-and-warp* method (red curve) and our approach (blue curve)—values of M_{tex} (y-axis) were measured over the entire image (a) and over the disoccluded parts only (b) of all frames (x-axis) in all sequences presented in the main paper: *Alchemist* (1), *Jana* (2), *Selfie* (3), *Knights* (4), *Lili* (5), and *Lynx* (6). Higher values indicate higher M_{tex} error. Note how at disocclusions the difference is notably more prominent. Those are the most sensitive locations where our approach outperforms the *stylize-and-warp* method. See also Fig. 2 for qualitative evaluation.

Results of quantitative evaluation of the *warp-and-stylize* with respect to our approach are shown in Fig. 3. In all presented graphs, the lower values of M_{stereo} indicate that our approach has notably better stereo consistency than the *warp-and-stylize* method. Qualitatively we evaluated this fact during the informal user study where

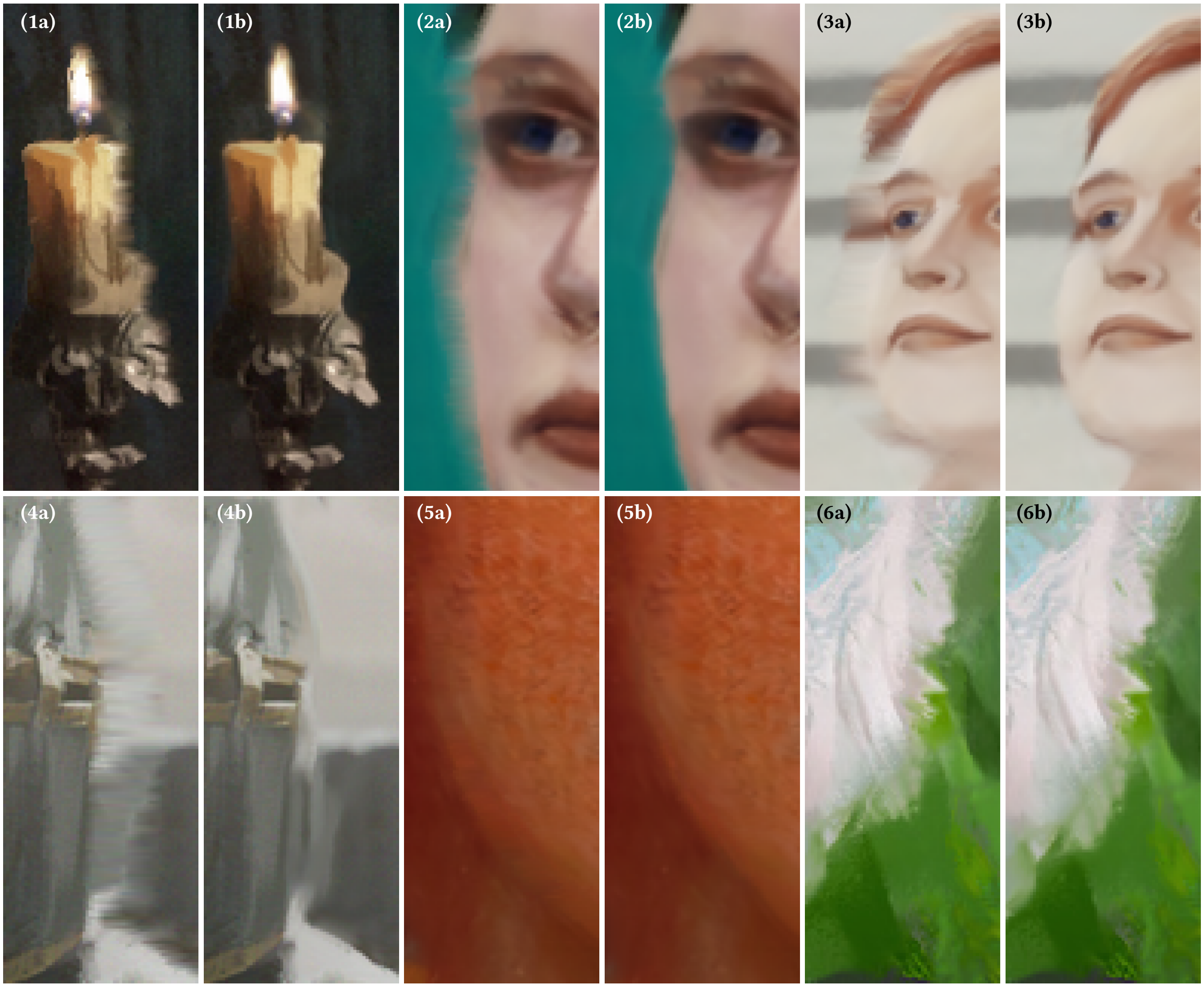

Figure 2: Qualitative evaluation of the ability to reproduce the original style using *stylize-and-warp* method (a) and our approach (b)—a selection of zoom-ins taken from the sequences presented in the main paper: *Alchemist* (1), *Jana* (2), *Selfie* (3), *Knights* (4), *Lili* (5), and *Lynx* (6). At strong discontinuities the *stylize-and-warp* approach produces visible smearing artifacts that can either distort the texture of the disoccluded area or deform the shape of the object’s boundary. Note, e.g., a subtle chin malformation in (5).

we asked a few first participants to evaluate also sequences produced by the *warp-and-stylize* method. Those were, however, so distracting (causing heavy eye discomfort) that we decided to not show them to other participants. All those observations indicate that stereo consistency is an essential quality that needs to be preserved in the final stylized sequence.

REFERENCES

Ondřej Jamriška, Šárka Sochorová, Ondřej Texler, Michal Lukáč, Jakub Fišer, Jingwan Lu, Eli Shechtman, and Daniel Sýkora. 2019. Stylizing Video by Example. *ACM Transactions on Graphics* 38, 4 (2019), 107.

Vivek Kwatra, Irfan A. Essa, Aaron F. Bobick, and Nipun Kwatra. 2005. Texture optimization for example-based synthesis. *ACM Transactions on Graphics* 24, 3 (2005), 795–802.

Yonatan Wexler, Eli Shechtman, and Michal Irani. 2007. Space-Time Completion of Video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 3 (2007), 463–476.

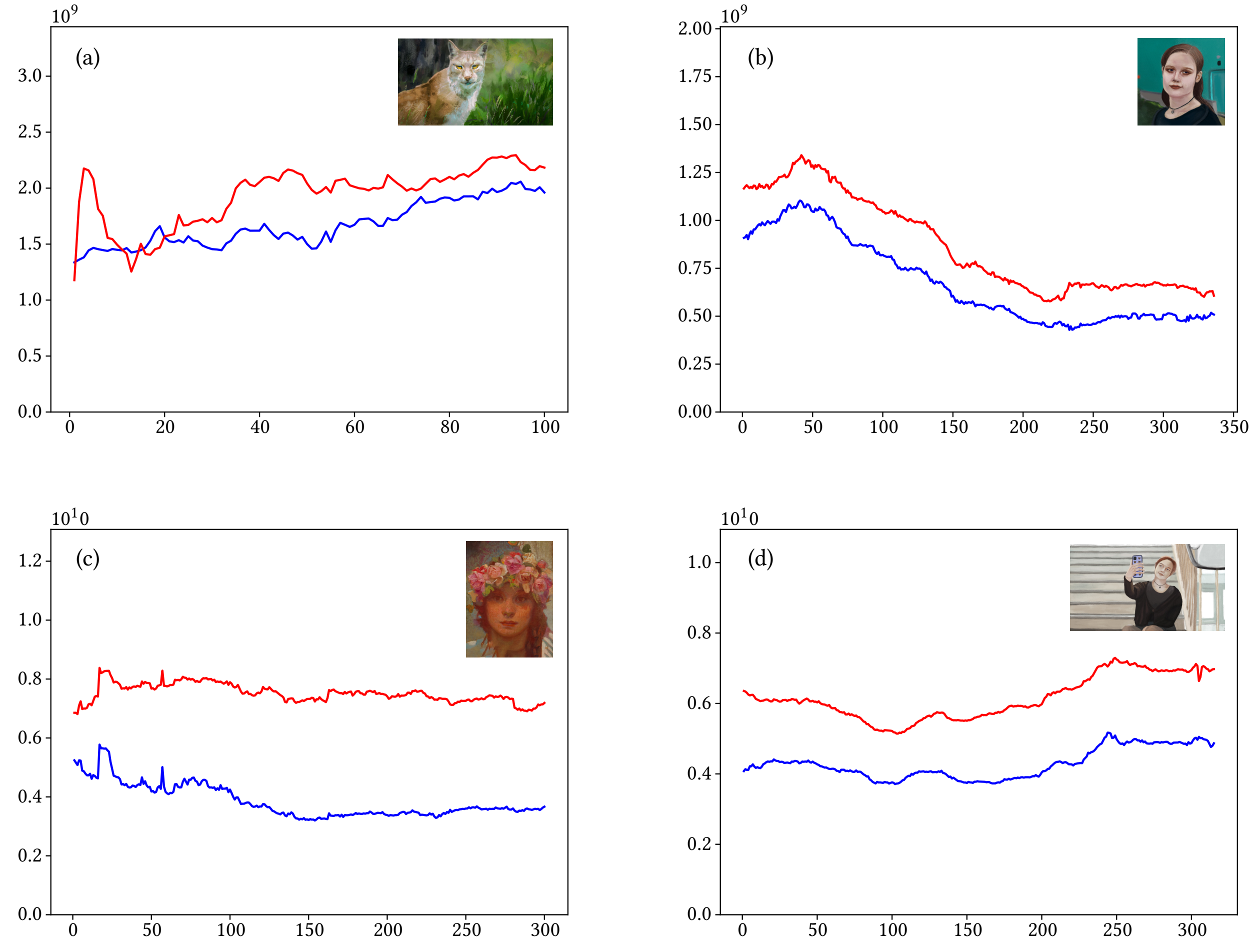

Figure 3: Quantitaive evaluation of the ability to preserve stereo consistency using the *warp-and-stylize* method (red curve) and our approach (blue curve)—values of M_{stereo} (y-axis) were measured over all pixels in all output frames (x-axis) of selected sequences presented in the main paper to which the *warp-and-stylize* method can be applied: *Lynx* (a), *Jana* (b), *Lili* (c), and *Selfie* (d). Higher values indicate higher M_{stereo} error.