computers

MDPI

Article

Depth-Aware Neural Style Transfer for Videos

Eleftherios Ioannou * and Steve Maddock *

Department of Computer Science, The University of Sheffield, Sheffield S1 4DP, UK

* Correspondence: eioannou1@sheffield.ac.uk (E.I.); s.maddock@sheffield.ac.uk (S.M.)

Abstract: Temporal consistency and content preservation are the prominent challenges in artistic video style transfer. To address these challenges, we present a technique that utilizes depth data and we demonstrate this on real-world videos from the web, as well as on a standard video dataset of three-dimensional computer-generated content. Our algorithm employs an image-transformation network combined with a depth encoder network for stylizing video sequences. For improved global structure preservation and temporal stability, the depth encoder network encodes ground-truth depth information which is fused into the stylization network. To further enforce temporal coherence, we employ ConvLSTM layers in the encoder, and a loss function based on calculated depth information for the output frames is also used. We show that our approach is capable of producing stylized videos with improved temporal consistency compared to state-of-the-art methods whilst also successfully transferring the artistic style of a target painting.

Keywords: neural style transfer; deep learning; depth estimation; temporal consistency

check for updates

Citation: Ioannou, E.; Maddock, S. Depth-Aware Neural Style Transfer for Videos. *Computers* **2023**, *12*, 69. https://doi.org/10.3390/computers12040069

Academic Editors: Martin J. Turner, Peter Vangorp and Edmond Prakash

Received: 14 February 2023
Revised: 16 March 2023
Accepted: 24 March 2023
Published: 27 March 2023

1. Introduction

Artistic neural style transfer [1] involves the style of an input image being transferred onto a second input 'content' image, for example, the style of a Van Gogh painting being transferred onto an everyday photograph. For video, the style transfer inputs are a style image and a content video. Current approaches to video stylization [2–5] mainly focus on improving upon the temporal coherence across sequential frames which is the prominent challenge that arises when stylizing videos rather than still images. To our knowledge, previous methods do not utilize ground truth depth during training, whilst only a few methods use three-dimensional information in the stylization process [6,7].

Stylizing videos can be useful for multiple creative processes, for visual art generation, and for the creative industries sector in general, such as film, TV, and computer games. This can lead to new artistic effects achieved with just one reference image.

For instance, the color grading process often materialized during post-production of films and games, is a significant tool that contributes to the fabrication of a particular, distinctive visual look and style. This is tightly connected to the mood and emotions that are intended to be evoked to viewers [8]. In recent years, such film production techniques (e.g., embedding cinema shots or abstract and artistic looks for the delivery of more immersive experiences) have become commonplace in computer games. Both photorealistic style transfer [9] and artistic style transfer can be utilized. This paper builds on the earlier work on depth-aware style transfer [10] by employing a system that considers 3D data for the artistic stylization of videos.

In contrast to current approaches, we aim to leverage ground truth three-dimensional information that is available from synthetic 3D videos to use in training. As the nature of the training data can significantly influence the efficiency and quality of neural style transfer algorithms, it is important to recognize and distinguish the different available data sources. In the three-dimensional domain, which this work is focused on, the data can either be synthetic (computer-generated) or recorded in the real-world. Real-world 3D videos can be videos that are accompanied by ground-truth depth data (e.g., recorded using *LiDAR* or *Microsoft Kinect* [11]) and videos for which depth information is estimated using a depth prediction network. Although it can be easier to access depth information of computer-graphics content, it is also possible to utilize synthetic videos with estimated depth. Therefore, synthetic 3D videos can also be differentiated depending on whether the depth data are ground-truth or calculated.

We present an approach that leverages ground truth depth and optic flow data from publicly available datasets in order to generate robust real-time artistic stylizations for videos. We adapt the standard feed-forward neural network architecture [12] to intercept encoded depth information to efficiently preserve the content and global structures of the input video frames (Figure 1). As the concepts of motion, optical flow and depth can be intertwined [13], we additionally propose a loss function that combines optical flow and depth to enhance temporal stability. The primary contributions of our work are:

- We develop a depth-aware neural network for artistic stylization that considers depth information infused by a depth encoder network.
- We implement a loss function that combines depth and optical flow to encourage temporal stability.
- We present qualitative and quantitative results that demonstrate that our algorithm's performance is superior when compared to state-of-the-art approaches.

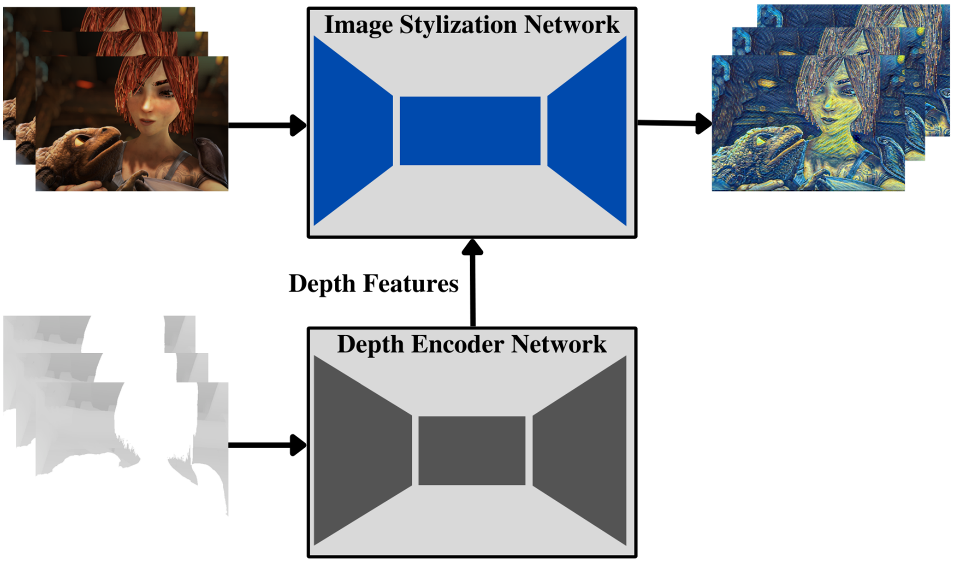

Figure 1. System Overview: A stylization network intercepts each frame and generates a stylized version of it considering depth encoded features infused by the depth encoder network that takes as input the depth information for each frame.

2. Related Work

Our approach elicits ideas from three related areas: Video Style Transfer, Depth-aware Style Transfer and Style Transfer in Games. We develop a system in the intersection of these areas attempting to strengthen the connection between them.

2.1. Video Style Transfer

One of the main considerations of image style transfer approaches that have aimed to improve upon the initial work by Gatys et al. [1] has been speed, with methods emerging that claim real-time processing [12,14]. However, the critical issue these approaches face is the temporal incoherence that becomes visible across subsequent frames.

A few methods attempt to alleviate this issue by focusing on improving the stability of the transferred content and style features. Without taking consecutive frames into consideration during training, the method proposed by Deng et al. [15] works directly on the feature space. Their proposed network (*MCCNet*) learns to rearrange content and style features based on their correlation, resulting in stylizations that preserve the content structures and temporal continuity. In a similar manner, *CSBNet*, the system developed by Lu and Wang [16], achieves temporally stable arbitrary style transfer by introducing a Crystallization module (Cr), a Separation module (Sp), and a Blending module (Bd). These allow for generating hierarchical stability-aware content and style representations (Cr), whose features are then separated (Sp) and cross-blended again (Bd) for the generation of stable stylized videos. Their system also relies on a pair of component enhancement losses, ignoring any subsequent frames or optical flow information. Similarly, Liu et al. [17] disregard any optical flow information, and extend their image stylization network (*AdaAttN*) to work on videos by replacing softmax with cosine similarity for their attention map computation and by designing a regularization between two content frames during training.

Other approaches introduce a temporal consistency loss to enforce coherent stylization [2,3]. End-to-end systems have also been proposed [18–20], taking as input a sequence of content images $\{I_t | t = 1, ...n\}$ and producing a stylized output video sequence $\{O_t | t = 1, ...n\}$. In contrast, post-processing methods [21,22] rely on both the content frames and the generated stylized results to achieve temporal consistency. The proposed algorithms routinely utilize pre-computed or estimated optical flows to define the temporal consistency loss that ensures that output frames remain coherent across time. Gao et al. [4] further suggests the inclusion of a feature-map-level temporal loss to preserve high-level feature maps of subsequent frames and a luminance difference to encourage the luminance changes of the output frames to be analogous to the corresponding luminance changes of the original frames. More recently, Gao et al. [5] proposes a multi-instance normalization block to accommodate multiple styles and a system architecture that embeds *ConvLSTM* modules in the recurrent network so that at each stage the information from previous frames is taken into account.

While introducing modules in the generator network and loss functions to train them can help in achieving temporal coherence, the next section considers how taking into account depth information can also aid the structure preservation and further improve the quality of the results.

2.2. Depth-Aware Style Transfer

Liu and Lai [6] realized that the pre-trained networks used by style transfer methods to define content and style losses, are originally designed for object recognition and while focusing on the primary target in an image, they neglect much of the finer details of the background. Their algorithm introduces a depth loss function in addition to content and style loss in order to better preserve the depth and global structure of the stylized results. Integrating depth preservation requires the use of an image-depth prediction network [23] in order to measure the differences between the transformed image and the content image. Another approach aiming to better preserve the spatial distribution and structure of the dominant object in the image employs a global structure extraction network (represented by the depth map) and a local structure refinement network (represented by the image edges) [24]. Utilizing an edge detection system [25] along with a depth perception network [23] to define the edge loss and depth loss, respectively, their method generates results capable of capturing both global and local structures of the input content image. Similarly, the work of Liu and Zhu [7] involves their model being trained on a global content loss and a local region structure loss and they further introduce a temporal loss along with a cycle-temporal loss in order to achieve temporal consistency between adjacent frames and avoid motion blur in videos. Their system also adopts the *AdaIN* layer from [26] for improved flexibility. Depth information is mainly computed by these approaches, albeit it can be available from multiple sources such as 3D computer games.

2.3. Style Transfer in Games

Whilst our approach focuses on stylising videos, it can also be suitable for usage in game environments where 3D information can be accessed during the rendering process.

Unity [27] and Stadia [28] have respectively demonstrated that style transfer techniques can be incorporated in their gaming pipelines as a post-effect. In both cases, the multi-style neural network of Ghiasi et al. [29] is used. Unity's integration proposes slight modifications, such as the reduction of convolutions for upsampling and downsampling. In addition, they show how applying the algorithm as a post-effect can be exploited to improve temporal coherence and reduce artifacts. This is implemented by re-using information from previously rendered frames (applying image-space bidirectional scene reprojection [30]) to apply a temporalization scheme that restricts disocclusion errors and reduces artifacts. Similarly, the system of Stadia reduces the convolutional layers of the network and significantly increases the residual layers. Their attempt at achieving temporal coherence is based on the argument that incorporating an additional temporal loss function to the regular perceptual losses during training can enforce stability through runtime even if at runtime only the current frame is taken into account. Our method is also grounded on this idea, as our trained network can be utilized in a game engine at the end of the rendering pipeline.

3. Method

3.1. Overview

Our method is inspired by previous work that utilizes an Image Transformation Network with Instance Normalisation layers [10] which has shown improved results for depth and global structure preservation. Figure 2 provides a detailed system architecture of our approach. The system consists of a feed-forward image transformation network and a depth encoder. The image transformation network takes as input an RGB frame and outputs a stylized version of it. To accommodate for enhanced depth preservation and temporal coherence, the network is provided with additional depth encoded features from the depth encoder network. Ground truth data are processed by a depth encoder that is composed of convolutional and max-pooling layers.

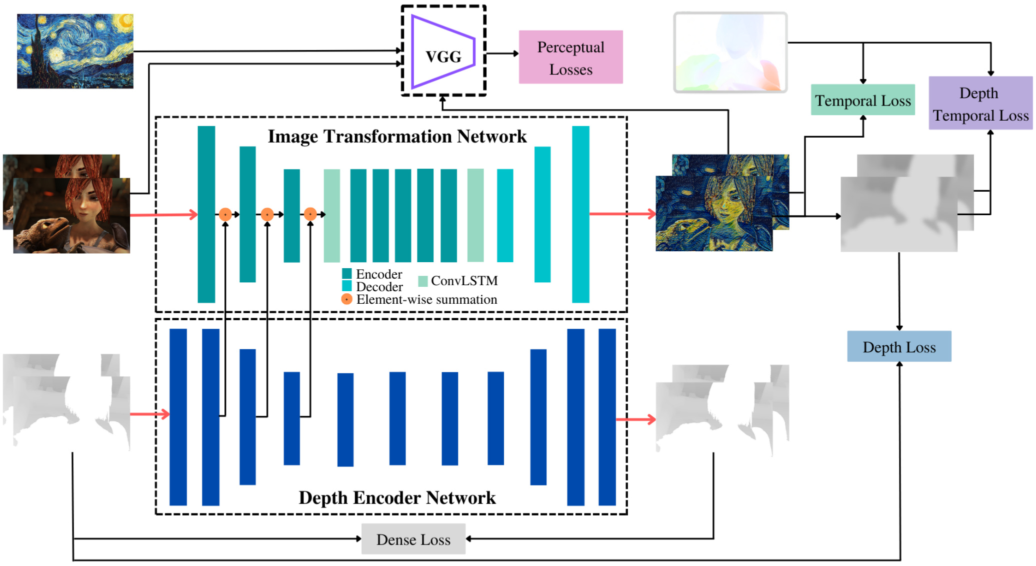

Figure 2. System Architecture: The depth encoder network is trained to minimize a depth reconstruction loss while it encodes depth features that are fused into the image transformation network. Two *ConvLSTM* layers are placed before and after the 5 residual layers in the generator network. The image transformation network is trained to minimize the perceptual losses (content and style), a temporal loss (calculated using optical flow), a depth temporal loss, and a depth loss.

3.2. Network Architecture

Our feed-forward image transformation network utilizes an encoder-decoder architecture, which is a one-style-per-model algorithm, adapted from [10] and similar to the works of Johnson et al. [12] and Gao et al. [4]. The encoder consists of three convolution layers followed by five residual layers which are placed between two *ConvLSTM* layers. The use of *ConvLSTM* modules allows better capture of the spatiotemporal information of the video sequences, as they are capable of generating a hidden state that is a compressed representation of the previous input sequence (previous input content frame) and can be utilized to forecast the current state (current input content frame). For the decoder, three convolution layers are used, with the first two being up-sampling convolutions. Instance normalization is used after each convolution process as it has been shown to improve significantly upon the results of style transfer algorithms [31]. We also use the *ReLU* activation function except for the final layer where the *tanh* activation function is used. Fusion layers are inserted after each convolutional block of the encoder to combine the encoded depth features provided by the depth encoder network with the encoded image frames.

Initially, we train our network on single-frame images from real-world data so that it generalizes to video sequences that are not necessarily computer-generated. We also take advantage of publicly available datasets that provide ground truth depth data instead of just RGB frames of computer-generated scenes. To leverage this data, we propose a simple depth encoder network comprising convolutional and max-pooling layers that process depth information before it is fused in the image transformation network. Our design is inspired by the *FuseNet* architecture that utilizes RGB-D data in order to improve upon the state-of-the-art semantic segmentation methods [32].

The depth encoder also follows a similar simple encoder–decoder architecture. It consists of four convolutional layers followed by three residual layers and again by four convolutional layers with the initial two being up-sample convolutions to enforce the reconstruction of the original depth map. Features from the 2nd, 3rd, and 4th convolutional layers are fused into the image transformation network. A mean square error loss is utilized for the training of the depth encoder that learns to reconstruct the input depth maps. This dense loss is defined as the difference between the original input depth map and the depth encoder's output. We find that by training the depth encoder's weights this results in better capturing of the features of the depth information.

3.3. Loss Functions

We define multiple loss functions to take into account temporal coherence, in addition to the perceptual information, for the training of the image transformation network. A two-frame synergic training mechanism is employed [4]: At each iteration during the training stage, the image transformation network generates stylized outputs for two subsequent frames, and feature maps of both the original content frames and the artistically stylized frames are inferred. Ground truth and calculated depth information is utilized for the encouragement of structural sustainment and temporal consistency.

3.3.1. Perceptual losses

We utilize the perceptual loss functions as introduced by the method of Johnson et al. [12]. A pre-trained image recognition network (*VGG-16* [33]) is used for the generation of feature representations of the original and transformed frames. Content loss is based on feature maps at the $j = relu3\_3$ layer. At each time frame t, we define the transformed stylized frame as $\hat{y}$ and original image frame as x:

$$\mathcal{L}^{\phi_0}_{content}(\hat{y}, x) = \frac{1}{C_j H_j W_j} \|\phi^j_0(\hat{y}) - \phi^j_0(x)\|^2_2 \tag{1}$$

where ϕ_0 is the image classification network and ϕ^j_0 represents the activations of the jth layer of ϕ_0 when processing an image with shape $H \times W \times C$, where H denotes the height, W the width and C the number of channels.

To compute style loss, features are extracted from the J = *{relu1_2, relu2_2, relu3_3, relu4_3}* layers. The definition of style loss is based on Gram-based style representations of the transformed image $\hat{y}$ and the style image y:

$$\mathcal{L}_{style}^{\phi_0,j}(\hat{y},y) = \|G_j^{\phi_0}(\hat{y}) - G_j^{\phi_0}(y)\|_F^2 \tag{2}$$

where G is used for Gram matrix. This is summed up for all the layers j in J.

In addition, for each time frame t we adopt the total variation regularizer ($\mathcal{L}_{tv}$) as described in [12].

3.3.2. Depth Loss

To retain depth information, we employ the *MiDaS* [34] depth estimation network (ϕ_1) to define an additional depth reconstruction loss function as proposed in [10].

$$\mathcal{L}_{depth}^{\phi_1}(\hat{y},x) = \frac{1}{C_j H_j W_j}\|\phi_1(\hat{y}) - \phi_1(x)\|_2^2 \tag{3}$$

3.3.3. Temporal Losses

We employ an output-level temporal loss that makes use of the ground truth optical flow and occlusion mask. This is defined as the warping error between the two subsequent frames:

$$\mathcal{L}_{temp,o}(t-1,t) = \sum_{t=2}^{T} M_t\|Warp(O_{t-1}, Flow_t) - O_t\|_2^2 \tag{4}$$

where O_t is the t-th frame, $Flow_t$ is the optical flow at time t and M_t is the ground-truth forward occlusion mask.

In addition to the temporal loss defined at the RGB-level, we attempt to enforce coherence on a depth level. Using the predicted depth of the stylized frames (D), computed using *MiDaS* [34], we define a depth-level temporal loss:

$$\mathcal{L}_{temp,d}(t-1,t) = \|Warp(D_{t-1}, Flow_t) - D_t\|_2^2 \tag{5}$$

This smooths out the changes in estimated depth of the subsequent frames, producing consecutive frames that preserve much of the depth information (Equation (3)) while being temporally stable.

The overall temporal loss is thus summarized as:

$$\mathcal{L}_{temp}(t-1,t) = \mathcal{L}_{temp,o}(t-1,t) + \mathcal{L}_{temp,d}(t-1,t) \tag{6}$$

3.4. Dataset

We first train without the temporal losses on the MS COCO dataset [35], composed of 80,000 images. We train the image transformation network along with the depth encoder that encodes computed depth maps of the input images. At this stage, we do not optimize the *ConvLSTM* layers and we only train on the perceptual and depth losses. We then use ground truth optical flow, motion boundaries and depth provided by the *Scene Flow* dataset [36]. Specifically, we train on the *Monkaa* and *FlyingThings3D* scenes, as they are both computer-generated 3D animations of complicated scenes. At this stage, the *ConvLSTM* layers are optimized and the networks are trained to minimize the perceptual losses, the depth loss and the temporal losses.

3.5. Training Details

We use the Adam optimizer [37] with a learning rate of 1×10^{-3} and train with a batch size of 2 for 2 epochs. We found the optimal weights for the content, style and depth loss to be 1×10^4, 1×10^5 and 1×10^3, respectively. The weight for the output-level temporal loss is 2×10^5 and the weight for the depth temporal loss is 1×10^3. Training takes approximately 5 hours (2.5 hours per epoch) on an NVIDIA Tesla V100 SXM2.

4. Results and Discussion

Our system takes as input sequential video frames and their corresponding depth maps and produces stylized frames in real-time. The depth maps, if not available, are computed using the depth prediction network by Ranftl et al. [34]. For each input content–

depth frame pair, the depth encoder is used to produce encoded features of the input depth map which are fused into the image transformation network that generates consistent depth-aware stylizations. We demonstrate results of improved quality and provide qualitative and quantitative comparisons against state-of-the-art approaches in video stylization. In addition to stylizations of real-world videos, we present results on synthetic 3D videos where the ground truth depth maps are available.

4.1. Qualitative Results

Figure 3 demonstrates results on real-world videos retrieved from Videvo [38]. We use the depth prediction network *MiDaS* [34] to compute the depth maps of the video frames (this can be performed offline or in real-time) and the stylizations produced are of high quality, with the global structure preserved and temporal coherence maintained. Results for synthetic videos with available depth maps are included in Figure 4. Similarly, the texture and color information of the style is captured adequately and the temporal inconsistencies across consecutive frames are avoided.

Visual qualitative side-to-side comparison with state-of-the-art methods is shown in Figures 5 and 6. In comparison with the FVMST [5] and ReCoNet [4] approaches (Figure 5), our method manages to better preserve the depth effect of the original content frames and sustain the global structure of the input video. While achieving similar aesthetic effect, our system avoids temporal inconsistencies. The hand of the subject in Figure 5 demonstrates this–the other approaches introduce artifacts in the form of white spots, while the frames generated using our method are more coherent. Similarly, in comparison with MCCNet [15] and CSBNet [16], the two approaches that do not consider optical flow and temporal information during training, our method produces visually better results and arguably captures the style patterns of the reference style image better. The close-ups in Figure 6 reveal an undesired halo effect generated by MCCNet, while CSBNet fails to preserve the details on the face and hair. Our approach manages to retain the fine details along with the depth information.

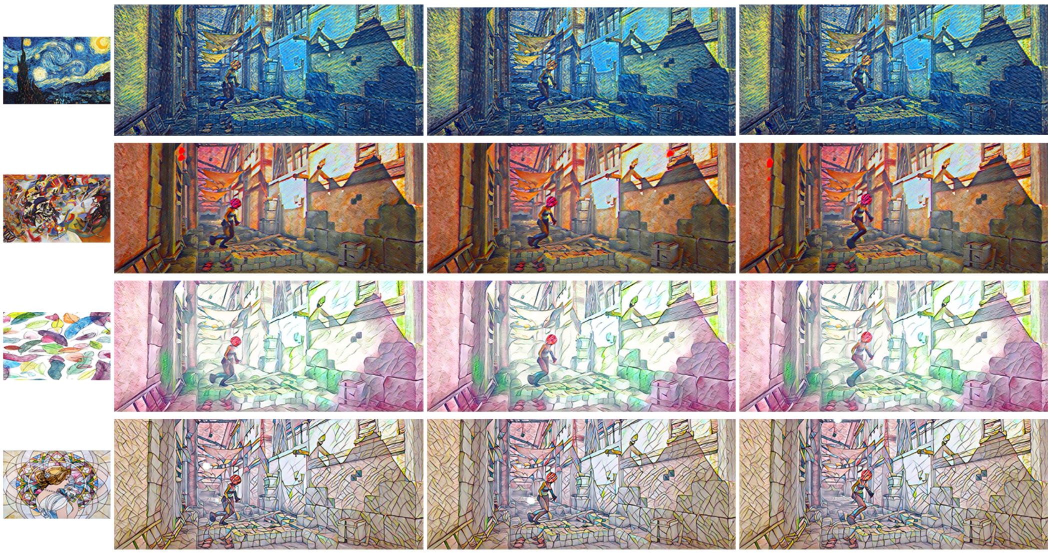

Figure 3. Results for different style images and for subsequent frames. The first column (small images) presents the reference style image, followed by the corresponding stylized consecutive frames in the next columns.

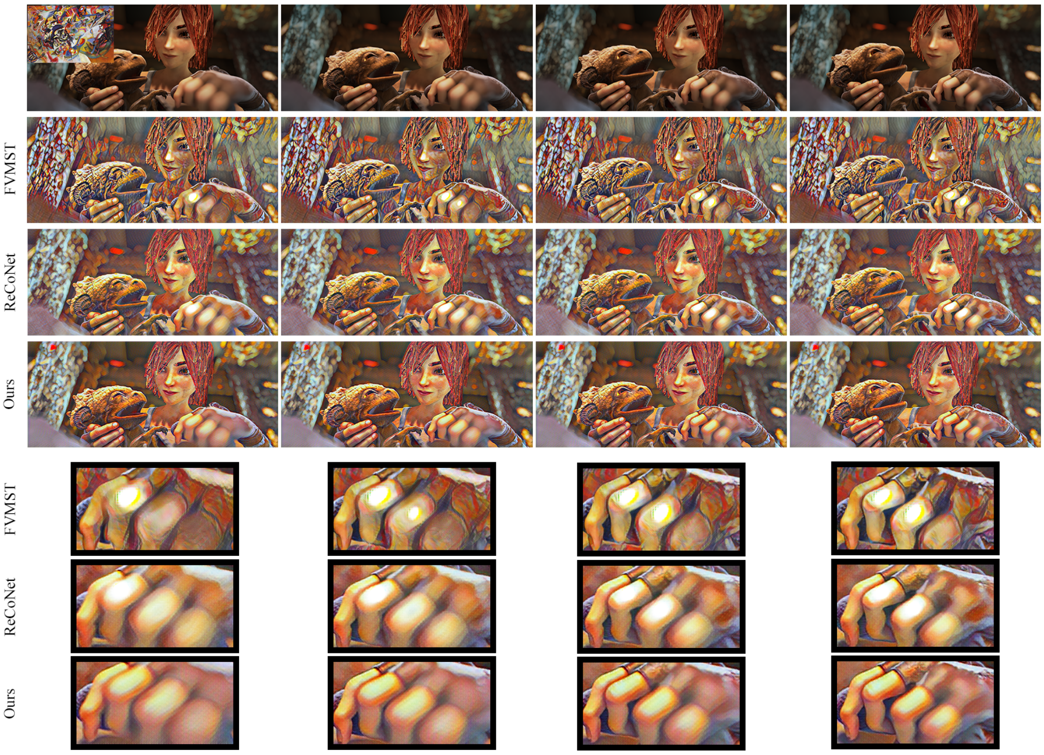

Figure 4. Comparison of our results against the methods: Fast Video Multi-Style Transfer (FVMST) [5] and Real-time Coherent Video Style Transfer Network (ReCoNet) [4]. The first row includes the original content frames and the last three rows are corresponding close-ups of the frames in rows 2–4. Our method does not introduce the white artifacts that are visible in the results of the first two methods.

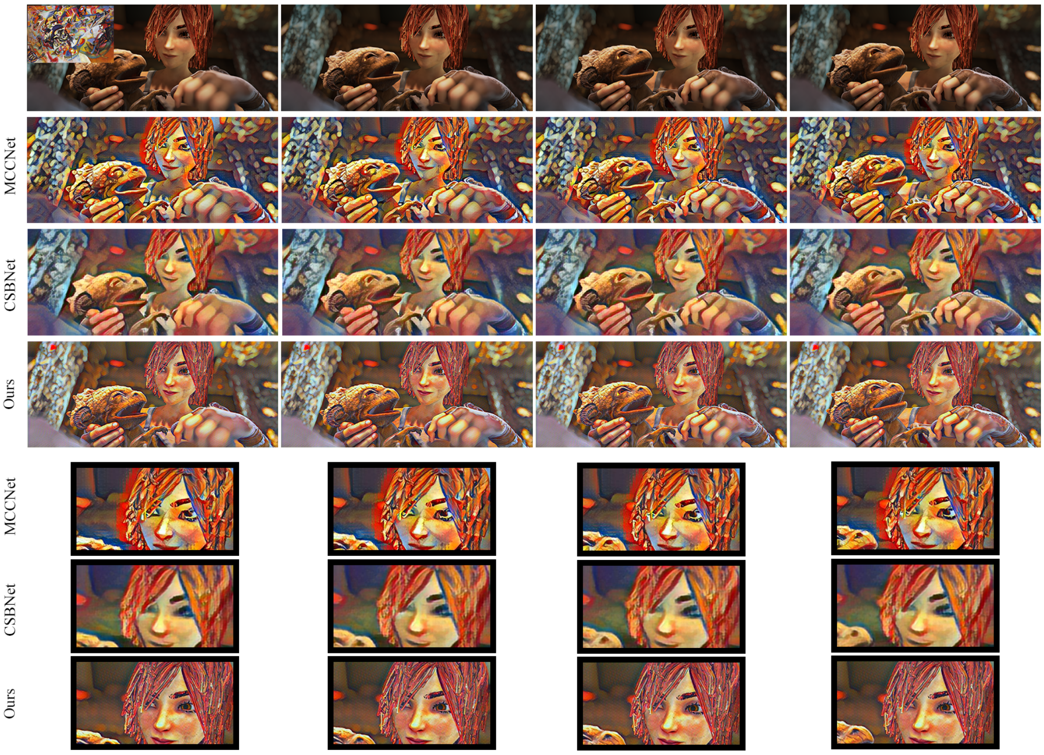

Figure 5. Comparison of our results against MCCNet [15] and CSBNet [16]. The first row includes the original content frames and the last three rows are corresponding close-ups of the frames in rows 2–4. Our method avoids the halo effect around the subject's face and retains the details and content information.

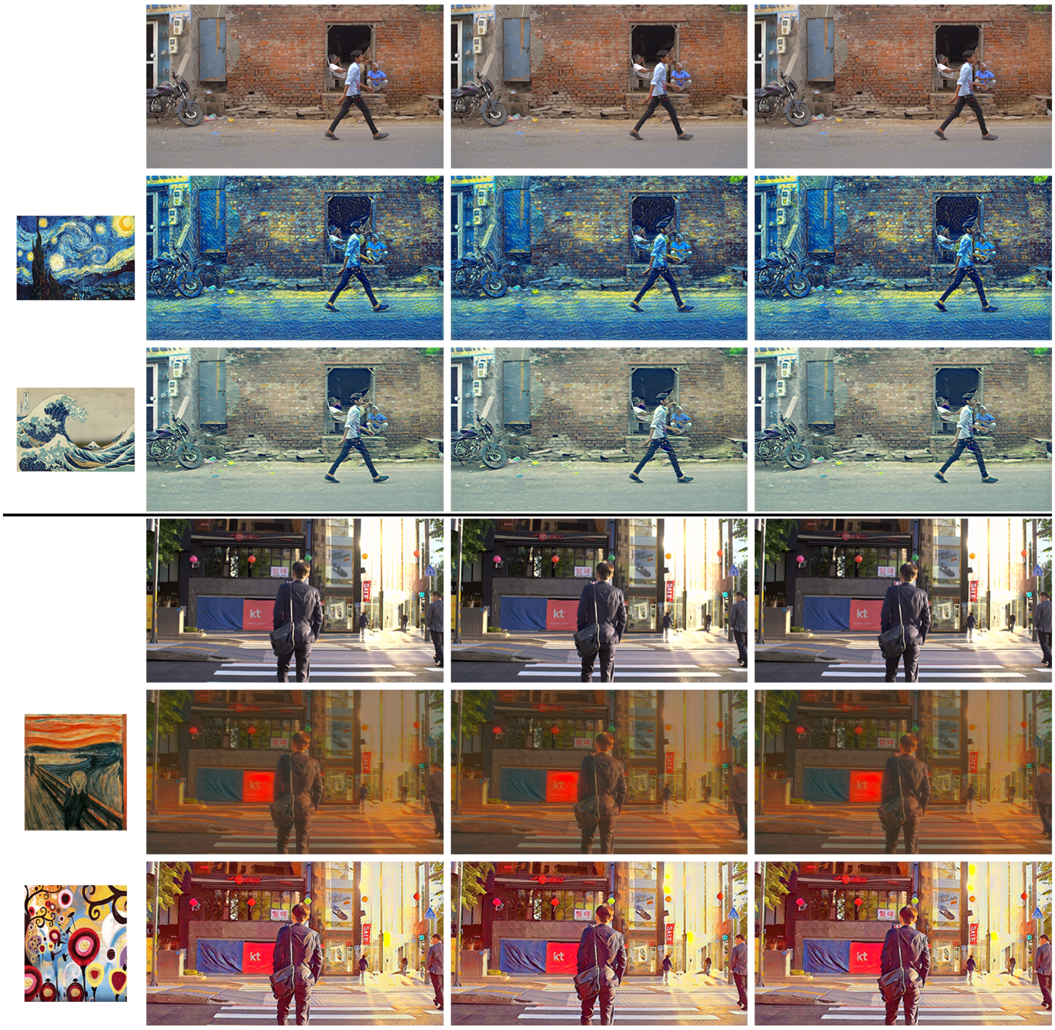

Figure 6. Results on real-world videos retrieved from Videvo [38] for different style images and for sequences of frames. The first column presents the reference style image, followed by the corresponding stylized consecutive frames in the next columns. Rows 1 and 4 show the respective original content frames.

4.2. Quantitative Results

We quantify the effectiveness of our results using the metrics of temporal coherence, depth preservation, and perceptual similarity. These criteria are used in a comparison with state-of-the-art approaches. All the metrics are computed for multiple style images using videos from MPI Sintel [39] as the evaluation dataset. The results of the quantitative evaluation are shown in Table 1.

The warping error calculates the difference between a warped next frame and the ground truth next frame. It is the most reliable way to quantify the robustness in achieving temporal consistency, as we use the ground truth flow data provided in the dataset. For the depth comparison, we also use the ground truth depth maps and compare against the computed depth images of the stylized frames using the MiDaS [34] depth prediction network. We calculate the depth difference for all the subsequent frames and average the results. In addition, we compute the perceptual similarity metric [40] since it is shown to be the closest to human visual perception. We compute the Learned Perceptual Image Patch Similarity (LPIPS), comparing the original content frames and the corresponding stylized frames in order to evaluate the effectiveness of the stylization in terms of retaining the structure and perceptual details—a higher LPIPS score means more difference, whilst a lower score means more similarity.

The metrics show that our method performs significantly well in terms of temporal consistency, depth, and structure preservation. Our method is the best in retaining depth information of the input content frames. Only CSBNet [16] is better in warping error, while MCCNet [15] is the best in LPIPS score. Although a further user evaluation study is required, it is arguable from the qualitative results that our method is better at capturing the artistic influences of the style image and transferring the colors and texture more efficiently.

Table 1. Quantitative Results: Warping error, depth loss, and perceptual loss are measured for multiple style images that were used to generate stylizations for 250 frames (for 5 different video sequences of the MPI Sintel dataset each containing 50 frames).

Method	Warping Error	Depth Loss	LPIPS
FVMST [5]	0.2589	69.6017	0.5289
ReCoNet [4]	0.2686	52.3109	0.5156
CSBNet [16]	**0.2335**	46.422	0.5004
MCCNet [15]	0.2516	52.6037	**0.4160**
Ours	0.2472	**38.7580**	0.5038

4.3. Discussion

The two main factors that we have introduced are fusing depth features into the generator network (Figure 2) and training with a depth temporal loss that combines depth and optical flow (Equation (5)). To test the quality of our method and the extent to which our contributions improve the results, we investigate these two primary factors for a more comprehensive understanding of their effectiveness. More precisely, we examine the trade-off between the extent that depth information is preserved and temporal coherence.

In order to inspect the efficacy of the depth fusion and the depth temporal loss in further detail, we train our system under different configurations. For each configuration, the warping error and depth loss are computed. Specifically, we train without optimizing the depth encoder and without encoding depth features to fuse into the image transformation network (w/o depth fusion), and also without the depth temporal loss (Equation (5)) (w/o depth temp loss). The results of these metrics are illustrated in Table 2.

Table 2. Quantitative evaluation for different configurations. Warping error and Depth loss is computed for the different configurations and training settings of our approach. The metrics are calculated for our method as described in Section 3 (Ours), for our method without depth fusion (W/O Depth fusion) and for our method without the depth temporal loss as defined in Equation (5) (W/O Depth Temp loss).

Configuration	Warping Error	Depth Loss
Ours	0.2472	38.7580
W/O Depth fusion	0.2335	43.4186
W/O Depth temp loss	0.2537	30.4074

Although not fusing the depth encoded features into the generator network results in slightly better performance in terms of temporal stability, the global structure of the content frames is not so well preserved and depth information is lost. In contrast, when the depth temporal loss is not used to train the network, depth information is better preserved but temporal coherence is lost.

This shows that for our proposed system there is a trade-off between depth preservation and temporal stability. Fusing the depth encoded features in the generator network leads to better global structure and style contrast across the image but has a negative effect on temporal coherence. The opposite is true when introducing a training loss function that connects depth with optical flow–temporal stability is improved but depth information is lost. Therefore, omitting one of the two factors (depth fusion or depth temporal loss) in training is not optimal. Our initially proposed setup (Ours) that combines depth fusion with the depth temporal loss is therefore the most suitable, and, as shown in Table 2, generates more robust results.

A final point is that our approach relies on additional information for both training and inference in comparison with the rest of the approaches. Specifically, during inference, it requires depth information to generate the stylized results. The rest of the approaches do not require any additional information during inference, except for the content video frames. In addition, our approach utilizes depth and optical flow information during training, whereas FVMST [5] and ReCoNet [4] only make use of optical flow and MCCNet [15] and CBSNet [16] avoid relying on additional information.

5. Conclusions

We have presented a method for artistic stylization of videos. Our system is capable of preserving the depth and global structure of an input video's frames by considering depth information that is encoded and infused into the transformation network. The results demonstrate improved temporal coherence in comparison to state-of-the-art approaches, as a result of our proposed loss function that combines depth and optical flow. The performance of the technique is evaluated both qualitatively and quantitatively.

Our system can stylize 3D video frames in real-time for both computer-generated videos with accompanying depth maps or for real-world videos with depth maps computed using state-of-the-art depth prediction methods. However, our system is limited to reproducing one style per trained network. Utilizing adaptive instance normalization layers in future work could increase flexibility. In addition, improving the time the system requires to output a stylized frame for each input frame would make it more suitable for use in real-time post-process effects of computer games. Lastly, evaluation must be considered further. Currently, qualitative and quantitative evaluation approaches rely on subjective visual comparisons and a variety of dissimilar metrics. Finding a robust and effective evaluation approach remains a challenge.

Author Contributions: Methodology, E.I. and S.M.; Software, E.I.; Writing—original draft, E.I.; Writing—review & editing, E.I. and S.M.; Supervision, S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the EPSRC.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data were presented in the main text.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gatys, L.A.; Ecker, A.S.; Bethge, M. A Neural Algorithm of Artistic Style. *arXiv* **2015**, arXiv:1508.06576.
2. Huang, H.; Wang, H.; Luo, W.; Ma, L.; Jiang, W.; Zhu, X.; Li, Z.; Liu, W. Real-time neural style transfer for videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 783–791.
3. Ruder, M.; Dosovitskiy, A.; Brox, T. Artistic style transfer for videos and spherical images. *Int. J. Comput. Vis.* **2018**, *126*, 1199–1219. [CrossRef]
4. Gao, C.; Gu, D.; Zhang, F.; Yu, Y. Reconet: Real-time coherent video style transfer network. In Proceedings of the Computer Vision—ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; Revised Selected Papers, Part VI 14; Springer: Berlin/Heidelberg, Germany, 2019; pp. 637–653.
5. Gao, W.; Li, Y.; Yin, Y.; Yang, M.H. Fast video multi-style transfer. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 3222–3230.
6. Liu, X.C.; Cheng, M.M.; Lai, Y.K.; Rosin, P.L. Depth-aware neural style transfer. In Proceedings of the Symposium on Non-Photorealistic Animation and Rendering, Los Angeles, CA, USA, 29–30 July 2017; pp. 1–10.
7. Liu, S.; Zhu, T. Structure-Guided Arbitrary Style Transfer for Artistic Image and Video. *IEEE Trans. Multimed.* **2022**, *24*, 1299–1312. [CrossRef]
8. Zabaleta, I.; Bertalmío, M. Photorealistic style transfer for cinema shoots. In Proceedings of the 2018 Colour and Visual Computing Symposium (CVCS), Gjovik, Norway, 19–20 September 2018; pp. 1–6.

9. Luan, F.; Paris, S.; Shechtman, E.; Bala, K. Deep photo style transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4990–4998.
10. Ioannou, E.; Maddock, S. Depth-aware neural style transfer using instance normalization. In Proceedings of the Computer Graphics & Visual Computing (CGVC), Cardiff, UK, 15–16 September 2022.
11. Microsoft. 2022. Available online: https://www.microsoft.com/en-au/windows-server (accessed on 14 March 2023).
12. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 694–711.
13. Baraldi, P.; De Micheli, E.; Uras, S. Motion and Depth from Optical Flow. In Proceedings of the Alvey Vision Conference, Reading, UK, 25–28 September 1989; pp. 1–4.
14. Dumoulin, V.; Shlens, J.; Kudlur, M. A Learned Representation For Artistic Style. *arXiv* **2016**, arXiv:1610.07629.
15. Deng, Y.; Tang, F.; Dong, W.; Huang, H.; Ma, C.; Xu, C. Arbitrary video style transfer via multi-channel correlation. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021; Volume 35, pp. 1210–1217.
16. Lu, H.; Wang, Z. Universal video style transfer via crystallization, separation, and blending. In Proceedings of the International Joint Conferences on Artificial Intelligence Organization (IJCAI), Vienna, Austria, 23–29 July 2022; Volume 36, pp. 4957–4965.
17. Liu, S.; Lin, T.; He, D.; Li, F.; Wang, M.; Li, X.; Sun, Z.; Li, Q.; Ding, E. Adaattn: Revisit attention mechanism in arbitrary neural style transfer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 6649–6658.
18. Jamriška, O.; Sochorová, Š.; Texler, O.; Lukáč, M.; Fišer, J.; Lu, J.; Shechtman, E.; Sỳkora, D. Stylizing video by example. *ACM Trans. Graph.* **2019**, *38*, 107. [CrossRef]
19. Gupta, A.; Johnson, J.; Alahi, A.; Fei-Fei, L. Characterizing and improving stability in neural style transfer. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4067–4076.
20. Chen, T.Q.; Schmidt, M. Fast patch-based style transfer of arbitrary style. *arXiv* **2016**, arXiv:1612.04337.
21. Lai, W.S.; Huang, J.B.; Wang, O.; Shechtman, E.; Yumer, E.; Yang, M.H. Learning blind video temporal consistency. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 170–185.
22. Dong, X.; Bonev, B.; Zhu, Y.; Yuille, A.L. Region-based temporally consistent video post-processing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 714–722.
23. Chen, W.; Fu, Z.; Yang, D.; Deng, J. Single-Image Depth Perception in the Wild. In *Proceedings of the Advances in Neural Information Processing Systems*; Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2016; Volume 29.
24. Cheng, M.M.; Liu, X.C.; Wang, J.; Lu, S.P.; Lai, Y.K.; Rosin, P.L. Structure-Preserving Neural Style Transfer. *IEEE Trans. Image Process.* **2020**, *29*, 909–920. [CrossRef] [PubMed]
25. Xie, S.; Tu, Z. Holistically-nested edge detection. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1395–1403.
26. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1501–1510.
27. Deliot, T.; Guinier, F.; Vanhoey, K. Real-Time Style Transfer in Unity Using Deep Neural Networks. 2020. Available online: https://blog.unity.com/engine-platform/real-time-style-transfer-in-unity-using-deep-neural-networks (accessed on 14 March 2023).
28. Poplin, R.; Prins, A. Behind the Scenes with Stadia's Style Transfer ML. 2019. Available online: https://stadia.google.com/gg/(accessed on 14 March 2023).
29. Ghiasi, G.; Lee, H.; Kudlur, M.; Dumoulin, V.; Shlens, J. Exploring the structure of a real-time, arbitrary neural artistic stylization network. *arXiv* **2017**, arXiv:1705.06830.
30. Yang, L.; Tse, Y.C.; Sander, P.V.; Lawrence, J.; Nehab, D.; Hoppe, H.; Wilkins, C.L. Image-based bidirectional scene reprojection. In Proceedings of the 2011 SIGGRAPH Asia Conference, Hong Kong, China, 12–15 December 2011; pp. 1–10.
31. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv* **2016**, arXiv:1607.08022.
32. Hazirbas, C.; Ma, L.; Domokos, C.; Cremers, D. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Proceedings of the Asian Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 213–228.
33. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2015. Available online: http://xxx.lanl.gov/abs/1409.1556 (accessed on 14 March 2023).
34. Ranftl, R.; Lasinger, K.; Hafner, D.; Schindler, K.; Koltun, V. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1623–1637. [CrossRef] [PubMed]
35. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
36. Mayer, N.; Ilg, E.; Häusser, P.; Fischer, P.; Cremers, D.; Dosovitskiy, A.; Brox, T. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
37. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
38. Videvo. 2019. Available online: https://www.videvo.net/ (accessed on 14 March 2023).

39. Butler, D.J.; Wulff, J.; Stanley, G.B.; Black, M.J. A naturalistic open source movie for optical flow evaluation. In Proceedings of the European Conference on Computer Vision (ECCV), Florence, Italy, 7–13 October 2012; pp. 611–625; Part IV, LNCS 7577.
40. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.