# Stereoscopic Neural Style Transfer

Dongdong Chen[1*]    Lu Yuan[2]    Jing Liao[2]    Nenghai Yu[1]    Gang Hua[2]

[1]University of Science and Technology of China    [2]Microsoft Research

cd722522@mail.ustc.edu.cn,   {jliao, luyuan, ganghua}@microsoft.com,   ynh@ustc.edu.cn

## Abstract

*This paper presents the first attempt at stereoscopic neural style transfer, which responds to the emerging demand for 3D movies or AR/VR. We start with a careful examination of applying existing monocular style transfer methods to left and right views of stereoscopic images separately. This reveals that the original disparity consistency cannot be well preserved in the final stylization results, which causes 3D fatigue to the viewers. To address this issue, we incorporate a new disparity loss into the widely adopted style loss function by enforcing the bidirectional disparity constraint in non-occluded regions. For a practical real-time solution, we propose the first feed-forward network by jointly training a stylization sub-network and a disparity sub-network, and integrate them in a feature level middle domain. Our disparity sub-network is also the first end-to-end network for simultaneous bidirectional disparity and occlusion mask estimation. Finally, our network is effectively extended to stereoscopic videos, by considering both temporal coherence and disparity consistency. We will show that the proposed method clearly outperforms the baseline algorithms both quantitatively and qualitatively.*

## 1. Introduction

Stereoscopic 3D was on the cusp of becoming a mass consumer media such as 3D movies, TV and games. Nowadays, with the development of head-mounted 3D display (*e.g.*, AR/VR glasses) and dual-lens smart phones, stereoscopic 3D is attracting increasing attention and spurring a lot of interesting research works, such as stereoscopic inpainting [36, 27], video stabilization [15], and panorama [39]. Among these studies, creating stereoscopic 3D contents is always intriguing.

Recently, style transfer techniques used to reproduce famous painting styles on natural images become a trending topic in content creation. For example, the recent film "*Loving Vincent*" is the first animated film made entirely of oil

---

paintings by well-trained artists. Inspired by the power of Convolutional Neural Network (CNN), the pioneering work of Gatys *et al.* [13] presented a general solution to transfer the style of a given artwork to any images automatically. Many follow-up works [21, 19, 34, 12, 11] have been proposed to either improve or extend it. These techniques are also applied to many successful industrial applications (*e.g.*, Prisma [1], Ostagram [2], and Microsoft Pix [3]).

However, to the best of our knowledge, there are no techniques that apply style transfer to stereoscopic images or videos. In this paper, we address the need for this emerging 3D content by proposing the first stereoscopic neural style transfer algorithm. We start with a careful examination of naive application of existing style transfer methods to left and right views independently.

We found that it often fails to produce geometric consistent stylized texture across the two views. As a result, it induces problematic depth perception and leads to 3D fatigue to the viewers as shown in Figure 1. Therefore, we need to enable the method to produce stylized textures that are consistent across the two views. Moreover, a fast solution is required, especially for practical real-time 3D display (*e.g.*, AR/VR glasses). Last but not least, style transfer in stereoscopic video as a further extension should satisfy temporal coherence simultaneously.

In this paper, we propose the first feed-forward network for fast stereoscopic style transfer. Besides the widely adopted style loss function [13, 19], we introduce an additional disparity consistency loss, which penalizes the deviations of stylization results in non-occluded regions. Specifically, given the bidirectional disparity and occlusion mask, we establish correspondences between the left and right view, and penalize the stylization inconsistencies of the overlapped regions which are visible in both views.

We first validate this new loss term in the optimization-based solution [13]. As shown in Figure 1, by jointly considering stylization and disparity consistency in the optimization procedure, our method can produce much more consistent stylization results for the two views. We further incorporate this new disparity loss into a feedforward deep network that we designed for stereoscopic stylization.
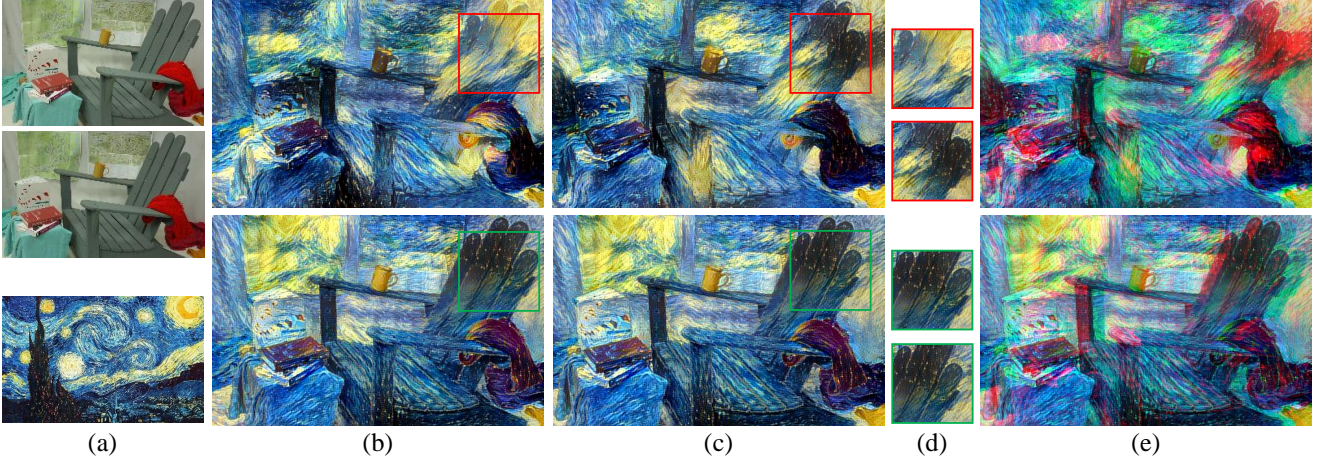
---

Figure 1. (a) Given a stereoscopic image pair and a style image, when the left and right view are stylized separately (first row), the left stylization result (b) will be inconsistent with that of the right view (c) in spatially corresponding areas (d).This will lead to undesirable vertical disparities and incorrect horizontal disparities, subsequently causing 3D visual fatigue in anaglyph images (e) . In contrast, by introducing a new disparity consistency constraint, our method (second row) can produce consistent stylization results for the two views.

Our network consists of two sub-networks. One is the stylization sub-network *StyleNet*, which employs the same architecture in [19]. The other is the disparity sub-network *DispOccNet*, which can estimate bidirectional disparity maps and occlusion masks directly for an input stereo image pair. These two sub-networks are integrated in a feature level middle domain. They are first trained on each task separately, and then jointly trained as a whole.

Our new disparity sub-network has two advantages: 1) it enables real-time processing, when compared against some state-of-the-art stereo matching algorithms [33, 22] that use slow global optimization techniques; 2) it is the first end-to-end network which estimates the bidirectional disparities and occlusion masks simultaneously, while other methods [26, 38] only estimate a single directional disparity map in each forward and need post-processing steps to obtain the occlusion mask. In Sec. 5.2, we will show that this bidirectional design is better than the single directional design.

Our network can also be easily extended to stereoscopic 3D videos by integrating the sub-networks used in [10]. In this way, the final stylization results can keep not only the horizontal spatial consistency at each time step, but also the temporal coherence between adjacent time steps. This work may inspire film creators to think about automatically turning 3D movies or TVs into famous artistic styles.

In our experiments, we show that our method outperforms the baseline both quantitatively and qualitatively. In summary, this paper consists of four main contributions:

- We propose the first stereoscopic style transfer algorithm by incorporating a new disparity consistency constraint into the original style loss function.
- We propose the first feed-forward network for fast stereoscopic style transfer, which combines styliza-

tion, bidirectional disparities and occlusion masks estimation into an end-to-end system.
- Our disparity sub-network is the first end-to-end network which simultaneously estimates bidirectional disparity maps and occlusion masks.
- We further extend our method to stereoscopic videos by integrating additional sub-networks to consider both disparity consistency and temporal coherency.

In the remainder of this paper, we will first summarize some related works. In our method, we validate our new disparity constraint using a baseline optimization-based method, and then introduce our feed-forward network for fast stereoscopic style transfer, and extend it to stereoscopic videos. Experiments will show the evaluation and comparison with other ablation analysis of our method. Finally, we conclude with further discussion.

## 2. Related Work

With the increasing popularity and great business potential of 3D movies or AR/VR techniques, stereoscopic image/video processing techniques have drawn much attention. Some interesting stereoscopic topics include image inpainting [36], object copy and paste [25], image retargeting [9, 5], image warping [28] and video stabilization [15]. In this work, we introduce a new topic which turns stereoscopic images or videos into synthetic artworks.

In the past, re-drawing an image in a particular style required a well-trained artist to do lots of time-consuming manual work. This motivated the development of a plenty of Non-photorealistic Rendering (NPR) algorithms to make the process automatic, such that everyone can be an artist. However, they are usually confined to the specific artistic

styles (*e.g.*, oil paintings and sketches). Gatys *et al.* [14] were the first to study how to use CNNs to reproduce famous styles on natural images. They leverage CNNs to characterize the visual content and style, and then recombine both for the transferring styles, which is general to various artistic styles.

To further improve the quality, many domain priors or schemes are used, including face constraints [31], MRF priors [21], or user guidances [8]. To accelerate the rendering process, a feed-forward generative network [19, 34, 12, 11] can be directly learnt instead, which were successfully deployed popular apps (*e.g.*, Prisma[1], Microsoft Pix [3]). However, the algorithms described above are designed only for monocular images. When they are independently applied to stereoscopic views, they inevitably introduce spatial inconsistency, causing visual discomfort (3D fatigue).

Generally, stereoscopic 3D techniques consider the disparity consistency in the objective function. Stereoscopic style transfer is not an exception. The first step is to estimate a high quality disparity map from the two input views, which is still an active research topic. Traditional methods often use sophisticated global optimization techniques and MRF formulations, such as [17, 37, 35]. Recently, some CNN-based methods have been explored. Zbontar *et al.* [38] adopted a Siamese network for computing matching distances between image patches. Mayer *et al.* [26] synthesized a large dataset and trained an end-to-end network for disparity estimation. However, it can only obtain a single directional disparity map in each forward and need post-processing steps to obtain the occlusion mask. Our disparity sub-network adopts a similar network architecture but with a different loss function. It can simultaneously estimate the bidirectional disparity maps and occlusion masks, which is essential to high-quality stereo image/video editing. To the best of our knowledge, it is the first end-to-end network for bidirectional disparities and occlusion masks estimation.

The most related work to ours is video style transfer. Previous methods [4, 29, 10, 16, 18, 30] all incorporated a new temporal consistency constraint in the loss function to avoid flickering artifacts. Analogous to the temporal consistency, stereoscopic style transfer requires spatial consistency between the left and right view. Different from sequential processing in video style transfer, two views should be processed symmetrically in stereoscopic style transfer. It is crucial to avoid the structure discontinuity near the occlusion boundary as shown in Figure 2. Our style transfer network is naturally extended to stereoscopic videos by considering both spatial consistency and temporal consistency at the same time.

## 3. Disparity Loss for Spatial Consistency

Since previous neural style transfer methods only tackle monocular views, the naive extension fails to preserve spa-
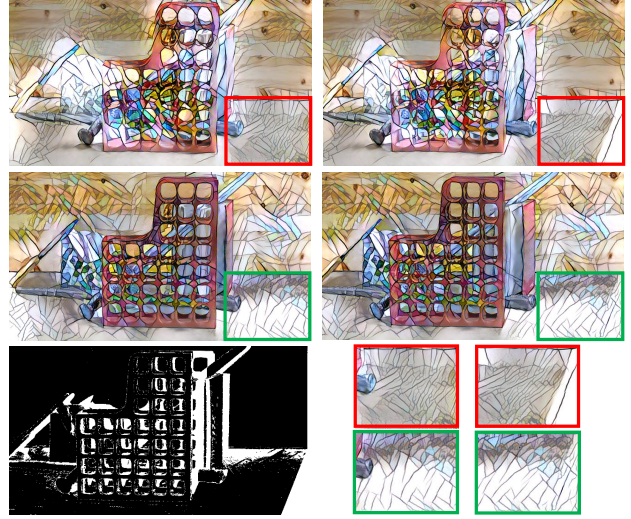


Figure 2. Comparison results of stylizing left and right view sequentially (top row) or jointly (middle row) with the disparity consistency constraint. The former method will often generate texture discontinuity near occlusion mask boundary. Bottom row is the right view occlusion mask and enlarged stylization patches.

tial consistency between the left and right view images, as shown in in Figure 1. How to enforce spatial consistent constraint across two views is worth exploring. Video style transfer is also confronted by similar consistency preservation problem, but in the temporal axis.

To reduce inconsistency, some methods [10, 29, 30] use the stylization results of previous frames to constrain that of the current frame. By analogy, we can firstly stylize the left view, and then use it to constrain the stylization of the right view with disparity consistency, or vice versa. In this way, we can obtain spatially consistent results in visible overlapping regions, but the continuity of stylization patterns near the occlusion boundary is often damaged, as shown in Figure 2. The underlying reason is that the left stylization result is fixed during the optimization procedure of the right view. To avoid it, we should regard the left and right view in a symmetric way and jointly process them.

Therefore, we add a new term by enforcing symmetric bidirectional disparity constraint, to the stylization loss function, and jointly optimize left and right views. We first validate the new loss in optimization-based style transfer framework [13], which is formulated as an energy minimization problem. In the next section, we will further incorporate the proposed loss to our feed-forward network.

Given a stereoscopic image pair $I_l, I_r$ and a style image $S$, the left and right view stylization results $O_l, O_r$ are iteratively optimized via gradient descent. The objective loss function $\mathcal{L}_{total}$ consists of three components: content loss $\mathcal{L}_{cont}^v$, style loss $\mathcal{L}_{sty}^v$ and disparity loss $\mathcal{L}_{disp}^v$, where $v \in \{l, r\}$ represents the left or right view, *i.e.*,

$$\mathcal{L}_{total} = \sum_{v \in \{l,r\}} \left( \alpha \mathcal{L}_{cont}^v(O_v, I_v) + \beta \mathcal{L}_{sty}^v(O_v, S) \right.$$
$$\left. + \gamma \mathcal{L}_{disp}^v(O_v, D_v, M_v) \right). \quad (1)$$

Both the content loss and style loss are similar to [19]:

$$\mathcal{L}_{cont}^v(O_v, I_v) = \sum_{i \in \{l_c\}} \|F^i(O_v) - F^i(I_v)\|^2,$$
$$\mathcal{L}_{sty}^v(O_v, S) = \sum_{i \in \{l_s\}} \|G(F^i(O_v)) - G(F^i(S))\|^2, \quad (2)$$

where $F^i$ and $G$ are feature maps and Gram matrix computed from the layer $i$ of a pre-trained VGG-19 network [32]. $\{l_c\}, \{l_s\}$ are VGG-19 layers used for content representation and style representation, respectively.

The new term of disparity loss enforces the stylization result at one view to be as close as possible to the warped result from the other view in the visible and overlapping regions (*i.e.*, non-occludded regions). It is defined as:

$$\mathcal{L}_{disp}^v(O_v, D_v, M_v) = (1 - M_v) \odot \|O_v - \overleftarrow{\mathcal{W}}(O_{v^*}, D_v)\|^2 \quad (3)$$

where $v^*$ is the opposite view of $v$ (if $v$ is the left, then $v^*$ is the right). $\overleftarrow{\mathcal{W}}(O_{v^*}, D_v)$ is the backward warping function that warps $O_{v^*}$ using the disparity map $D_v$ via bilinear interpolation, namely $\overleftarrow{\mathcal{W}}(O_{v^*}, D_v)(p) = O_{v^*}(p + D_v(p))$. $M^v$ is the occlusion mask, where $M_v(p) = 0$ for pixel $p$ visible in both views and $M_v(p) = 1$ for pixel $p$ occluded in the opposite view. Given the left and right disparity map, the occlusion mask $M$ can be obtained by a forward-backward consistency check, which is also used in [29].

Note that the loss $\mathcal{L}_{disp}$ is symmetric for both views, and relies on the bidirectional disparities and occlusion masks. As shown in Figure 1, compared to the baseline (*i.e.*, processing each view independently), our method achieves more consistent results and can avoid 3D visual fatigue in final anaglyph images. Jointly optimizing the left and right view in a symmetric way can further avoid the discontinuity near the occlusion boundary as shown in Figure 2, .

## 4. Stereroscopic Style Transfer Network

In this section, we propose a feed-forward network for fast stereoscopic style transfer. The whole network consists of two sub-networks: the *StyleNet* which is similar to existing style transfer networks [10, 11, 12, 16], and the *DispOccNet* which simultaneously estimates bidirectional disparity maps and occlusion masks. we integrate these two sub-networks in a feature level middle domain, making the left view and right view completely symmetric.

*StyleNet.* We use the default style network structure firstly proposed by [19] and used extensively in the other

works [10, 11, 12, 16]. The architecture basically follows an image auto-encoder, which consists of several strided convolution layers (encoding the image into feature space), five residual blocks, and fractionally strided convolution layers (decoding feature to the image). In our implementation, we follow the same designing as [10], where the layers before the third residual block (inclusive) are regarded as the encoder, and the remaining layers are regarded as the decoder.

*DispOccNet.* Recently, Mayer *et al.* [26] introduced an end-to-end convolution network called *DispNet* for disparity estimation. However, it can only predict single directional disparity map $D_l(l \rightarrow r)$ in each forward. Here, we use the similar network structure, but add three more branches in the expanding part for each resolution (1/64,...,1/2). These three branches are used to regress disparity $D_r$ and bidirectional occlusion masks $M_l, M_r$. The loss function for each resolution is:

$$\mathcal{L} = \sum_{v \in \{r,l\}} \mathcal{L}_d(M_v^g, D_v, D_v^g) + \lambda \mathcal{L}_o(W_v, M_v, M_v^g),$$
$$\mathcal{L}_d(M_v^g, D_v, D_v^g) = (1 - M_v^g) \odot \|D_v - D_v^g\|,$$
$$\mathcal{L}_o(W_v, M_v, M_v^g) = -\frac{1}{n} \sum_i W_v(i)[M_v^g(i)log(M_v(i))$$
$$+ (1 - M_v^g(i))log(1 - M_v(i))], \quad (4)$$

where the superscript $g$ denotes the ground truth. Different from the original loss in [26], we remove the disparity deviation penalty in occluded regions in $\mathcal{L}_d$, where the disparity values are undefined in real scenarios. $W_v$ is a pixel-wise class balance weight map, where values in occluded regions are the ratio of non-occluded and occlusion pixel number $\frac{\#non\_occ}{\#occ}$, while values in non-occluded regions are 1. Note that the ground truth of $D_v, M_v$ at each resolution are bilinearly interpolated from that of the original image resolution. The losses for each resolution are summed. Please refer to the supplementary material for details of the sub-network.

**Middle Domain Integration.** Since the left and right view are completely symmetric, we also consider the *StyleNet* and *DispOccNet* in a symmetric way. In fact, for a stereoscopic image pair, the overlapping regions visible in both views can be defined in a intermediate symmetric middle domain [6, 23]. If we know the occlusion mask for each view, we can stylize the overlapping regions and occluded regions respectively, then compose the image based on the occlusion mask. In this way, the final stylization results would naturally satisfy disparity consistency.

Image level composition often suffers from flow or disparity errors, and produces blurring and ghosting artifacts. As demonstrated in [10], feature level composition, followed by a decoder back to the image space, is more tol-
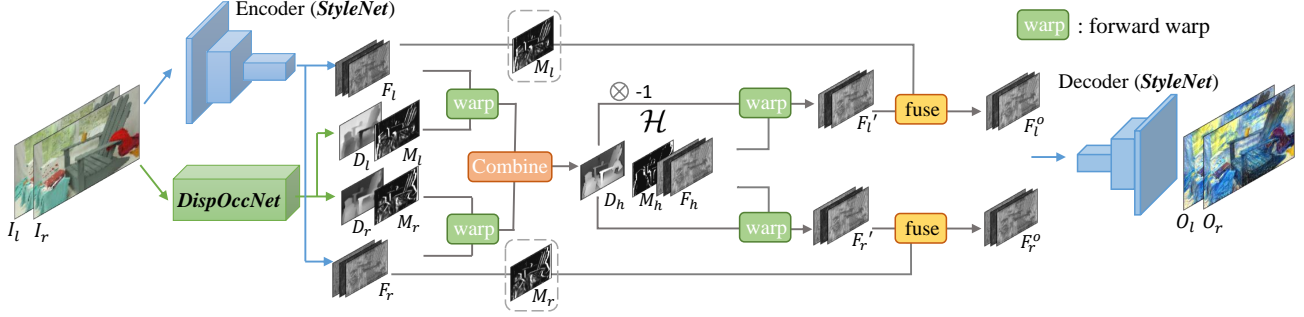
Figure 3. The overall network structure for fast stereoscopic image style transfer. It consists of two sub-networks: *StyleNet* and *DispOccNet*, which are integrated in the feature level middle domain $\mathcal{H}$.

erant to errors. Therefore, we integrate the *StyleNet* and *DispOccNet* in a new feature level middle domain $\mathcal{H}$. The overall Network architecture is shown in Figure 3.

**Network Overview.** Specifically, We first encode $I_l, I_r$ into feature maps $F_l, F_r$ with the encoder of the *StyleNet* and predict the bidirectional disparity maps and occlusion masks $D_l, D_r, M_l, M_r$ with the *DispOccNet*, which are bilinearly resized to match the resolution of $F_l, F_r$. Then for each view $v$, we warp $F_v$ and the halved $D_v$ to the middle domain using a forward warping function, which allows warping from each view to the middle domain without knowing the middle view disparity. The warped two views are combined, generating the middle disparity $D_h$, feature map $F_h$, and hole mask $M_h$. Similar to [23], the value $D_h(p)$ of a point $p$ is defined as the symmetric shift distance to the correspondence point in the left and right view (left: $p - D_h(p)$, right: $p + D_h(p)$). The hole masks generated in the left and right view forward warping are combined as $M_h$, and the corresponding pixel values in $F_h$ are excluded in the following warping, *i.e.*,

$$
\begin{aligned}
D_l &:= \frac{D_l}{2}, \quad D_r := \frac{D_r}{2}, \\
D_h &= \frac{-\overrightarrow{\mathcal{W}}(D_l, D_l, M_l) + \overrightarrow{\mathcal{W}}(D_r, D_r, M_r)}{2} \\
F_h &= \frac{\overrightarrow{\mathcal{W}}(F_l, D_l, M_l) + \overrightarrow{\mathcal{W}}(F_r, D_r, M_r)}{2},
\end{aligned}
\tag{5}
$$

where $\overrightarrow{\mathcal{W}}(x, y, m)$ is the forward warping function that warps $x$ using the disparity map $y$ guided by the occlusion mask $m$. Namely, if $z = \overrightarrow{\mathcal{W}}(x, y)$, then

$$
z(p) = \frac{\sum_q w_q \times x(q + y(q))}{\sum_q w_q}, \forall q : q + y(q) \in \mathcal{N}^8(p)
\tag{6}
$$

where $\mathcal{N}^8(p)$ denotes the eight-neighborhood of $p$, $w_q$ is the bilinear interpolation weight, making $z$ both differentiable to $x$ and $y$. All the occluded pixels $q$ in $m$ are excluded in

the forward warping procedure, which avoids the "many-to-one" mapping problem.

Next, we further forward warp $F_h$ back to the original left and right view, and fuse them with $F_l, F_h$ based on $M_l, M_r$ respectively, *i.e.*,

$$
\begin{aligned}
F_l' &= \overrightarrow{\mathcal{W}}(F_h, -D_h, M_h), \quad F_r' = \overrightarrow{\mathcal{W}}(F_h, D_h, M_h) \\
F_v^o &= M_v \odot F_v + (1 - M_v) \odot F_v', v \in \{l, r\}
\end{aligned}
\tag{7}
$$

Finally, the fused feature maps $F_l^o, F_r^o$ are fed into the decoder of the *StyleNet* to obtain the final stylization results $O_l, O_r$.

### 4.1. Extension to Stereoscopic Videos

To extend our network for stereoscopic videos, similar to [10, 30], we incorporate one more temporal coherence term for each view $v$ into our objective function, *i.e.*,

$$
L_{cohe} = \sum_{v \in \{l,r\}} (1 - M_v^t) \odot \|O_v^t - \mathcal{W}_{t-1}^t(O_v^{t-1})\|^2
\tag{8}
$$

where $\mathcal{W}_{t-1}^t(.)$ is the function to warp $O_v^{t-1}$ to time step $t$ using the ground truth backward flow as defined in [10].

Inspired by [10], we further add an additional flow sub-network and mask sub-network (together referred to as the "*Temporal* network") into the original network. We show the basic working flow in the left part of Figure 4. For a view $v$, two adjacent frames $I_v^{t-1}, I_v^t$ are fed into the flow sub-network to compute the feature flow $w_v^t$, which warps the input feature map $F_v^{o,t-1}$ to $F_v^{t\prime}$. Next the difference $\Delta F_v^t$ between the new feature map $F_v^t$ computed from $I_v^t$ and $F_v^{t\prime}$ is fed into the mask sub-network, generating the composition mask $M$. The new feature map $F_v^{u,t}$ is the linear combination of $F_v^t$ and $F_v^{t\prime}$ weighted by $M$.

We integrate the above stereoscopic image style transfer network (referred to as "*Stereo* network") with *Temporal* network in a recurrent formulation. Specifically, for each view $v$, we recursively feed the previous disparity consistent and temporal coherent feature maps $F_v^{o,t-1}$, together with adjacent frames $I_v^{t-1}, I_v^t$, into the *Temporal* network, generating the temporal coherent feature map $F_v^{u,t}$. Then $F_l^{u,t}$
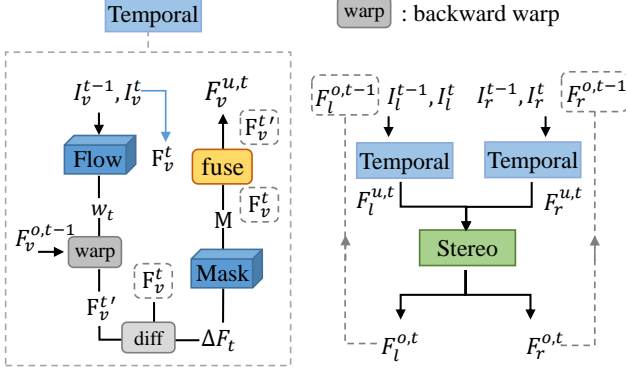
Figure 4. The overall structure for stereoscopic video style transfer. The left part is the simplified working flow for *Temporal* network. The right part is the recurrent formulation for combining the above *Stereo* network and the left additional *Temporal* network.

and $F_r^{u,t}$ are further fed into the *Stereo* network to guarantee disparity consistency. At this point, the output feature maps $F_l^{o,t}, F_r^{o,t}$ are both temporal coherent and disparity consistent and fed to the decoder of *StyleNet* to get the stylization results. Note that for $t = 1$, $F_l^{u,1}, F_r^{u,1}$ are just the output feature maps of encoder of the *StyleNet*.

## 5. Experiments

### 5.1. Implementation Details

For fast stereoscopic image style transfer, the overall network contains two sub-networks: the *StyleNet* and the *DispOccNet*. In our implementation, these two sub-networks are first pretrained separately, then jointly trained. For the *StyleNet*, we adopt the pretrained models released by [19] directly, which is trained on Microsoft COCO dataset[24]. To pretrain *DispOccNet*, we adopt the same training strategy in [26] with the synthetic dataset *FlyingThings3D*, which contains 21818 training and 4248 test stereo image pairs. During training, we use the bidirectional consistency check to obtain the ground truth occlusion mask as in [29].

These two sub-networks are jointly trained with a batch size of 1 (image pair) for 80k iterations. The Adam optimization method [20] is adopted with an initial learning rate 0.0001, which is decayed by 0.1 at 60k iterations. Because the style loss is around $10^7$ times larger than the disparity loss, the corresponding gradient of *DispOccNet* coming from *StyleNet* is scaled with $10^{-7}$ to balance. By default, $\gamma = 500$ is used for all the styles, and $\alpha, \beta$ remain unchanged as the pretrained style models.

For stereoscopic video style transfer, we add an additional *Temporal* network (one flow sub-network and one mask sub-network) to the above *Stereo* network. In our default implementation, the additional *Temporal* network is trained using the same method as [10], and directly integrated with a well-trained *Stereo* network in a recurrent formulation as show in Figure 4.

| Method | MPI Sintel clean | FlyingThings3D | Time |
|---|---|---|---|
| *DispNet*[26] | 4.48 | 1.76 | 0.064s |
| *DispOccNet** | **4.16** | **1.68** | 0.07s |
| *DispOccNet-SD* | 4.66 | 1.69 | 0.067s |
| *DispOccNet-OL* | 5.43 | 1.99 | 0.07s |

Table 1. Non-occluded endpoint errors for *DispOccNet* and its variants. The test time is for a 960x540 image on GTX TitanX.

| Method | Disparity loss | | | | Time |
|---|---|---|---|---|---|
| | *Candy* | *La_muse* | *Mosaic* | *Udnie* | |
| baseline [19] | 0.0624 | 0.0403 | 0.0668 | 0.0379 | 0.047s |
| finetuned [19] | 0.0510 | 0.0325 | 0.0597 | 0.0317 | 0.047s |
| our method | 0.0474 | 0.0301 | 0.0559 | 0.0285 | 0.07s |
| our method †† | 0.0481 | 0.0284 | 0.0570 | 0.0284 | 0.067s |
| | Perceptual loss | | | | |
| baseline [19] | 531745.9 | 249705.4 | 351760.7 | 136927.1 | |
| our method | 515511.6 | 250943.0 | 379825.3 | 124670.6 | |
| our method †† | 529979.8 | 260230.5 | 399216.3 | 135765.8 | |

Table 2. Comparison results of different methods of disparity loss and perceptual loss on *FlyingThings3D* test dataset. The test time is for 640x480 image pair on GTX TitanX.

### 5.2. Evaluation and Analysis of the *DispOccNet*

To evaluate the performance of our *DispOccNet*, we test our model on the *MPI Sintel* stereo dataset [7] and *FlyingThings3D* test dataset, respectively. To fully understand the effects of each modification, we further train two variant networks. The *DispOccNet-SD* only regresses single directional disparity and occlusion mask rather than bidirectional. The *DispOccNet-OL* is trained with the original loss function in [26] without removing the penalty for occluded regions. Since we only care about the disparity precision in non-occluded regions, we use the endpoint error (EPE) in non-occluded regions as the error measure.

As shown in Table 1, our *DispOccNet* is only $9.3\%$ slower than the original *DispNet* while predicts the bidirectional disparity maps and occlusion masks simultaneously. With the modified loss function and network structure, it achieves even better disparity both on the *MPI Sintel* dataset and the *FlyingThings3D* dataset. Compared to the variant network *DispOccNet-SD*, which only trains the single directional disparity and occlusion mask, *DispOccNet* is also better. We believe that feeding bidirectional disparities and occlusion masks helps the network to learn the symmetric property of the left and right disparities, thus learn more meaningful intermediate feature maps.

Besides disparity, occlusion mask is also very important for image or feature composition. In fact, there are two different ways to obtain the occlusion mask. The first is to make the network learn the occlusion mask directly (such as our method). The other is to run post bidirectional consistency check after getting accurate bidirectional disparity maps. However for the latter method, one needs $2\times$ forward time or retrains one network for bidirectional disparities similar to ours. Moreover, when the disparity map is not sufficiently good, the occlusion mask generated by the latter

Figure 5. Comparison of occlusion masks: ground truth (top left), our method (top right), and post bidirectional consistency check(bottom left). The occlusion mask generated by post bidirectional consistency check contains more false positives and noise.

method will contain more false alarms and noises, as shown in Figure 5. We also compare the F-score of the predicted occlusion masks by these two methods on the *FlyingThings3D* test dataset, our method is much better (F-score: 0.887) than the latter method (F-score: 0.805).

## 5.3. Evaluation for Stereoscopic Style Transfer

**Quantitative Evaluation.** To validate the effectiveness of our method, we use two different quantitative evaluation metrics: 1) the perceptual loss $\alpha\mathcal{L}_{cont} + \beta\mathcal{L}_{sty}$ to represent the faithfulness to the original styles, and 2) the disparity loss $L_{disp}$ to evaluate the disparity consistency. We compare three different methods on the *FlyingThings3D* test dataset for four different styles: the baseline monocular method [19], and *StyleNet* [19] finetuned with disparity loss but test without *DispOccNet*, our method (finetuned *StyleNet + DispOccNet*).

As shown in Table 2, compared to the baseline method [19], our results are more disparity consistent while keeping the original style faithfulness (*i.e.*, similar perceptual loss). When testing finetuned *StyleNet* [19] without *DispOccNet*, the disparity loss also decreases a lot. This shows that the stability of the original style network is actually improved a lot after joint training with the new disparity loss.

We further conduct an user study to compare our method with the baseline method [19]. Specifically, we randomly select 5 stereoscopic image pairs and 2 videos from the MPI-Sintel and kitty dataset for 4 different styles then ask 20 participants to answer "which is more stereoscopic consistent?". Our method wins $94.8\%$ of the time while [19] only wins $5.2\%$ of the time.

**Qualitative Evaluation.** In Figure 7, we show the comparison results with our baseline (stylizing each view independently) for a real street view stereoscopic image pair. The top row with red marked boxes is the baseline results, where the stylized textures in the corresponding regions between the left and right views are often inconsistent. When watching these results with 3d devices, these inconsistencies will make it more difficult for our eyes to focus, causing
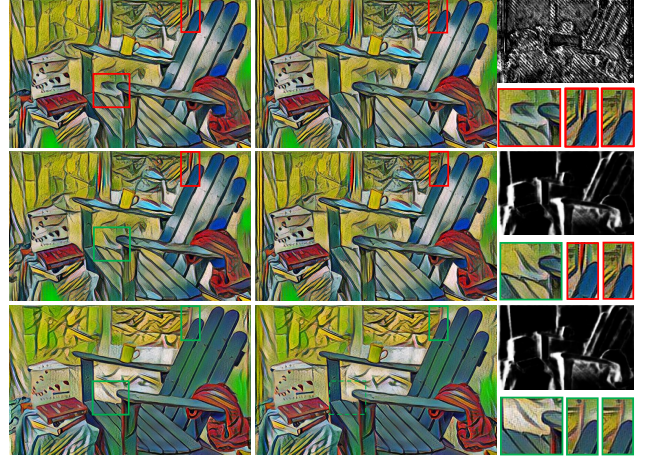


Figure 6. Comparison results using a similar variant of [10](top row), which suffers from both ghost artifacts and stylization inconsistency. The middle row is the results with composition mask replaced by our method, the ghost artifacts disappear but inconsistencies still exist. By contrast, our results (bottom row) do not have the above problems.

3d fatigue. In contrast, our results are more consistent.

In Figure 8, we show the comparison results for two adjacent stereoscopic image pairs of a stereoscopic video. By incorporating the additional *Temporal* network, our method can obtain both disparity consistent and temporal coherent stylization results (More visual results can be found on youtube [1]).

## 5.4. More Comparison

**Single Directional** *vs*. **Bidirectional** We have also designed an asymmetric single directional stereoscopic image style transfer network for feature propagation and composition. Different from the symmetric bidirectional network structure shown in Figure 3, we directly warp the left view feature $F_l$ to the right view using the right disparity map $D_r$, then conduct composition with $F_r$ based on the right occlusion mask $M_r$. As shown in Table 2, for the four test styles, the single directional method (marked with ††) can obtain similar stable errors, but all higher perceptual loss than our default bidirectional design. Furthermore, by experiment, we find it more difficult to jointly train *StyleNet* and *DispOccNet* for this asymmetric design, because the gradient for the left and right view is very unbalanced, making *DispOccNet* diverge easily.

**Comparison with Variant of [10]** In the monocular video style transfer method [10], a flow sub-network is utilized to guarantee temporal coherence, and the composition mask is implicitly trained with a mask sub-network. We have also designed a similar network structure with flow sub-network replaced by *DispNet* [26]. As shown Figure 6,

---
[1] https://www.youtube.com/watch?v=7py0Nq8TxYs

Figure 7. Comparison with our baseline for a real street view stereoscopic image pair. The top row with red marked boxes is the baseline results, and the bottom row with corresponding green marked boxes is our results. Obviously, our results are more disparity consistent.
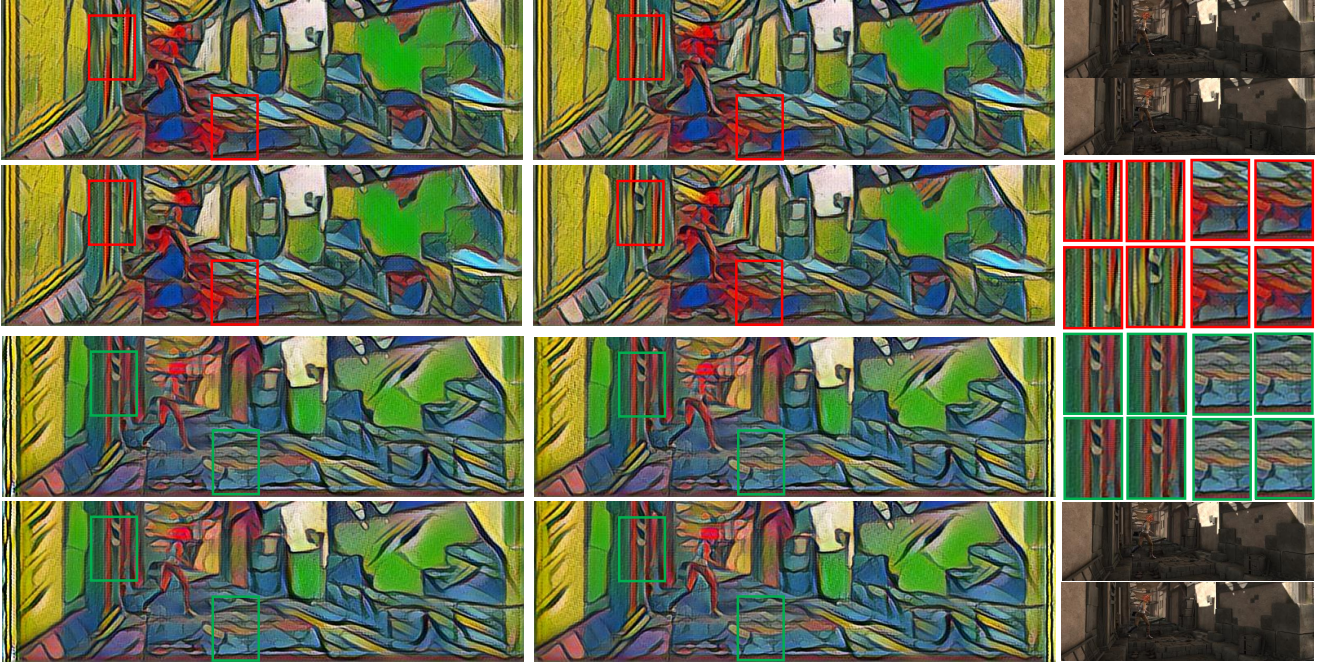


Figure 8. Comparison with our baseline for two adjacent stereoscopic image pairs. The top two rows are the baseline results, and the bottom two rows are our results. Compared to our baseline, our results can guarantee both disparity consistency and temporal coherency. The top-rightest is the input stereoscopic image pair of time step $t-1$, and the bottom-rightest is that of time step $t$.

it suffers from both ghosting artifacts and stylization inconsistency. The ghosting artifacts are resulted from the undefined disparity in occluded regions. When the occlusion mask is unknown, or the final composition mask is not good enough, the incorrectly warped feature will be used in the final composite feature, causing ghost artifacts. We also visualize the implicitly trained composition mask $M$, which is clearly worse than the composition mask $M_l$, $M_r$(occlusion mask) from our proposed *DispOccNet*.

For further validation, we replace the final composition mask with the occlusion mask $M_l$ from our *DispOccNet*, the ghost artifacts disappear. But inconsistencies still exist, because the original style sub-network is fixed in [10], which is sensitive to small perturbations. By contrast, our style sub-network is more stable after joint training with the disparity consistency loss.

## 6. Conclusion

In this paper, we present the first stereoscopic style transfer algorithm by introducing a new disparity consistency loss. For a practical solution, we also propose a feed-forward network by jointly training a stylization sub-network and a disparity sub-network. To the best of our knowledge, our disparity sub-network is the first end-to-end network that enables simultaneous estimation of the bidirectional disparity maps and the occlusion masks, which can potentially be utilized by other stereoscopic techniques. To further extend our method for stereoscopic videos, we incorporate an additional *Temporal* network [10] into our

*Stereo* network.

Along this direction, there is much future work worth investigating. For example, motivated by our *DispOccNet*, the flow sub-network used for temporal coherence can potentially also be extended to simultaneously predict the bidirectional flow and occlusion masks if suitable dataset exists. Furthermore, how to unify the flow and disparity into one network remains an open question, and worth exploring.

# References

[1] Prisma labs. prisma: Turn memories into art using artificial intelligence, 2016. 1, 9. 1, 3

[2] https://ostagram.ru/. 2016. 1

[3] Ai with creative eyes amplifies the artistic sense of everyone. 2017. 1, 3

[4] A. G. Anderson, C. P. Berg, D. P. Mossing, and B. A. Olshausen. Deepmovie: Using optical flow and deep neural networks to stylize movies. *arXiv preprint arXiv:1605.08153*, 2016. 3

[5] T. Basha, Y. Moses, and S. Avidan. Geometrically consistent stereo seam carving. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1816–1823. IEEE, 2011. 2

[6] P. N. Belhumeur and D. Mumford. A bayesian treatment of the stereo correspondence problem using half-occluded regions. In *Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on*, pages 506–512. IEEE, 1992. 4

[7] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *Proc. ECCV*, pages 611–625. Springer, 2012. 6

[8] A. J. Champandard. Semantic style transfer and turning two-bit doodles into fine artworks. *arXiv preprint arXiv:1603.01768*, 2016. 3

[9] C.-H. Chang, C.-K. Liang, and Y.-Y. Chuang. Content-aware display adaptation and interactive editing for stereoscopic images. *IEEE Transactions on Multimedia*, 13(4):589–601, 2011. 2

[10] D. Chen, J. Liao, L. Yuan, N. Yu, and G. Hua. Coherent online video style transfer. *arXiv preprint arXiv:1703.09211*, 2017. 2, 3, 4, 5, 6, 7, 8

[11] D. Chen, L. Yuan, J. Liao, N. Yu, and G. Hua. Stylebank: An explicit representation for neural image style transfer. In *Proc. CVPR*, 2017. 1, 3, 4

[12] V. Dumoulin, J. Shlens, and M. Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. 1, 3, 4

[13] L. A. Gatys, A. S. Ecker, and M. Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015. 1, 3

[14] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman. Controlling perceptual factors in neural style transfer. *arXiv preprint arXiv:1611.07865*, 2016. 3

[15] H. Guo, S. Liu, S. Zhu, and B. Zeng. Joint bundled camera paths for stereoscopic video stabilization. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 1071–1075. IEEE, 2016. 1, 2

[16] A. Gupta, J. Johnson, A. Alahi, and L. Fei-Fei. Characterizing and improving stability in neural style transfer. *arXiv preprint arXiv:1705.02092*, 2017. 3, 4

[17] H. Hirschmuller. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2):328–341, 2008. 3

[18] H. Huang, H. Wang, W. Luo, L. Ma, W. Jiang, X. Zhu, Z. Li, and W. Liu. Real-time neural style transfer for videos. CVPR, 2017. 3

[19] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. *arXiv preprint arXiv:1603.08155*, 2016. 1, 2, 3, 4, 6, 7

[20] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[21] C. Li and M. Wand. Combining markov random fields and convolutional neural networks for image synthesis. *arXiv preprint arXiv:1601.04589*, 2016. 1, 3

[22] L. Li, S. Zhang, X. Yu, and L. Zhang. Pmsc: Patchmatch-based superpixel cut for accurate stereo matching. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016. 2

[23] J. Liao, R. S. Lima, D. Nehab, H. Hoppe, P. V. Sander, and J. Yu. Automating image morphing using structural similarity on a halfway domain. *ACM Transactions on Graphics (TOG)*, 33(5):168, 2014. 4, 5

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Proc. ECCV*, pages 740–755. Springer, 2014. 6

[25] W.-Y. Lo, J. Van Baar, C. Knaus, M. Zwicker, and M. Gross. Stereoscopic 3d copy & paste. *ACM Transactions on Graphics (TOG)*, 29(6):147, 2010. 2

[26] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4040–4048, 2016. 2, 3, 4, 6, 7

[27] T.-J. Mu, J.-H. Wang, S.-P. Du, and S.-M. Hu. Stereoscopic image completion and depth recovery. *The Visual Computer*, 30(6-8):833–843, 2014. 1

[28] Y. Niu, W.-C. Feng, and F. Liu. Enabling warping on stereoscopic images. *ACM Transactions on Graphics (TOG)*, 31(6):183, 2012. 2

[29] M. Ruder, A. Dosovitskiy, and T. Brox. Artistic style transfer for videos. In *Proc. GCPR*, pages 26–36. Springer, 2016. 3, 4, 6

[30] M. Ruder, A. Dosovitskiy, and T. Brox. Artistic style transfer for videos and spherical images. *arXiv preprint arXiv:1708.04538*, 2017. 3, 5

[31] A. Selim, M. Elgharib, and L. Doyle. Painting style transfer for head portraits using convolutional neural networks. *ACM Trans. Graph. (Proc. of SIGGRAPH)*, 35(4):129, 2016. 3

[32] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[33] T. Taniai, Y. Matsushita, Y. Sato, and T. Naemura. Continuous 3D Label Stereo Matching using Local Expansion Moves. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017. (accepted). 2

[34] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. *arXiv preprint arXiv:1603.03417*, 2016. 1, 3

[35] L. Wang, H. Jin, and R. Yang. Search space reduction for mrf stereo. *Computer Vision–ECCV 2008*, pages 576–588, 2008. 3

[36] L. Wang, H. Jin, R. Yang, and M. Gong. Stereoscopic inpainting: Joint color and depth completion from stereo images. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 1, 2

[37] K. Yamaguchi, T. Hazan, D. McAllester, and R. Urtasun. Continuous markov random fields for robust stereo estimation. *Computer Vision–ECCV 2012*, pages 45–58, 2012. 3

[38] J. Zbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, 17(1-32):2, 2016. 2, 3

[39] F. Zhang and F. Liu. Casual stereoscopic panorama stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2002–2010, 2015. 1