Modeling Multi-style Portrait Relief from a Single Photograph

Yu-Wei Zhang*, Hongguang Yang, Ping Luo, Zhi Li
Faculty of Mechanical Engineering, Qilu University of Technology (Shandong Academy of Sciences)
Jinan,China

Hui Liu
School of Computer Science and Technology, Shandong University of Finance and Economics
Jinan,China

Zhongping Ji
School of Computer Science, Hangzhou Dianzi University
Hangzhou,China

Caiming Zhang
School of Computer Science and Technology, Shandong University
Jinan,China

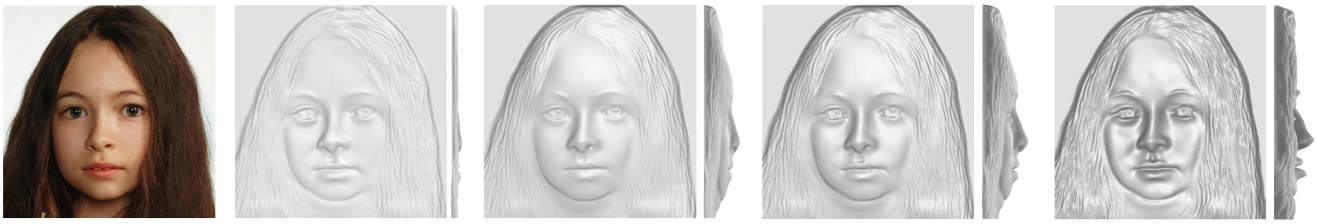


Figure 1. Given a photograph (a), our method is capable of modeling multi-style portrait reliefs with adjustable depth range and reasonable depth ordering (b-d). In contrast, the state-of-the-art method of [35] can only generate a bas-relief with small depth range. Increasing the depth range by linearly rescaling the bas-relief in depth direction produces unnatural geometrical details and unreasonable depth ordering (e).

Abstract

This paper aims at extending the method of Zhang et al.[35] to produce not only portrait bas-reliefs from single photographs, but also multi-style reliefs with adjustable depth range and reasonable depth ordering. We cast this task as a problem of style-aware photo-to-depth translation, where the input is a photograph conditioned by a style vector and the output is a portrait relief with desired depth style. To construct ground-truth data for network training, we first propose an optimization-based method to synthesize high-depth reliefs from 3D portraits. Then, we train a normal-to-depth network to learn the mapping from normal maps to relief depths. After that, we use the trained network to construct high-depth relief samples from the given normal maps of [35]. As each normal map has pixel-wise photograph, we are able to make correspondences between photographs and high-depth reliefs. Taking the bas-reliefs of [35], the new high-depth reliefs and their mixtures as target ground-truths, we finally train a encoder-to-decoder network to achieve style-aware relief modeling. Specially, the network is based on a U-shaped architecture, consisting of Swin Transformer blocks to process hierarchical deep features. Extensive experiments have proven the effectiveness of the proposed method. Comparisons with previous works have confirmed its flexibility and state-of-the-art performance.

Keywords: Portrait relief ; Multi-style modeling ; Swin

Transformer ; Photo-to-depth translation.

1. Introduction

Portrait relief is a 2.5D sculpture in which portrait elements are raised from the background with constrained depth range [39]. Traditional methods [29, 1] modeled portrait reliefs from single photographs by using SFS (Shape from Shading). Although geometrical details could be faithfully reconstructed, the resulting reliefs usually suffered from unnatural depth ordering. Template-based methods [31, 40] produced more realistic results by incorporating 3D priors, but a number of user interventions were required to guide depth reconstruction, which brought difficulties to common users. In the recent work of [35], Zhang et al. presented a neural-based solution to model portrait bas-reliefs from single photographs, without the need of any user interventions. The main contributions were the first photograph/bas-relief dataset and a CNN-based network for photo-to-depth translation, which was able to handle photos with various head poses, hairstyles and facial expressions. However, this method can only produce bas-reliefs within small depth range. To produce a high-depth relief, one straightforward way is to linearly rescale the bas-relief in depth direction, but this will exaggerate geometrical details and generate a high-depth relief with poor depth ordering, as shown the example in Fig.1e.

This paper aims at extending the work of Zhang et al.[35] to produce not only portrait bas-reliefs, but also multi-style reliefs with adjustable depth range and reasonable depth ordering, as shown the examples in Fig.1b-Fig.1d. We cast this task as a problem of style-aware photo-to-depth translation, where the input is a photograph conditioned by a style vector and the output is a portrait relief with desired depth style. The main challenge for this task is the lack of ground-truth data for network training. We notice that the dataset of [35] has provided pixel-wise normal map and bas-relief for each photograph. If we succeed in inferring high-depth reliefs from the normal maps, then the ground-truth data can be obtained by weighted depth interpolations. To this end, we first propose an optimization-based method to synthesize high-depth reliefs from 3D portraits. Then, we train a normal-to-depth network to learn the mapping from normal maps to relief depths. After that, we use the network to infer high-depth reliefs from the given normal maps of [35]. Taking the bas-reliefs, the inferred high-depth reliefs and their mixtures as ground-truth data, we finally train a neural network to achieve style-aware relief modeling. Specially, we design the network with a U-shaped encoder-decoder architecture, consisting of Swin Transformer blocks [18] to process hierarchical deep features. Due to the self-attention mechanism in modeling long-range dependency, the proposed network shows better performance than previous CNN-based architectures. We make extensive experiments to prove the effectiveness of the Transformer-based method. Comparisons with previous works have confirmed its flexibility and state-of-the-art performance. Our contributions are as follows:

(1) An optimization-based method for high-depth relief synthesis from 3D portrait.

(2) A multi-style solution for portrait relief modeling from single photograph.

(3) The first Transformer-based architecture for photo-to-relief translation.

2. Related works

Relief modeling. Relief modeling from 3D object or 2D image has been a hot topic since the pioneer work of [6]. As the depth values can be sampled from a 3D object with arbitrary view direction, object-based relief modeling comes down to a problem of nonlinear depth reconstruction. Instead of directly processing depth values, most of previous works [28, 22, 41, 11, 36] generated bas-reliefs by first attenuating large gradients at discontinuity edges, and then recovering new depth fields from the scaled gradients. In the current work of [10], Ji et al. proposed a learning-based method to produce bas-reliefs from 3D objects. The relief dataset was constructed by using the methods of [41] and [11]. Due to the fast convolution operations, the network was able to generate bas-reliefs in real time.

Modeling relief from 2D image is much more challenging than that from 3D object, because geometric priors, which is available in a 3D object, have been lost during the projection into 2D space. However, a 2D image is much easier and less expensive to capture, indicating wider application prospects. In the works of [29] and [1], SFS (Shape from Shading) were utilized to produce portrait bas-reliefs whose shadings approximate given photos. Although geometrical details could be faithfully constructed, the results suffered from unnatural depth ordering. To solve this problem, some works [31, 40] have used 3D templates to guide depth reconstruction and produced more realistic results. However, these methods required a number of user interventions to resolve shape ambiguities, which brought difficulties to common users.

Several neural-based methods [24, 38, 9, 37] have been proposed to model reliefs from single images without the need of user interventions. However, none of these methods aimed at modeling portrait reliefs. Recently, Zhang et al.[35] presented a neural-based solution to model portrait bas-reliefs from single photographs. The main contribution was the first dataset that contains 23k pixel-wise photo/bas-relief samples. To construct such a dataset, the authors first synthesized normal maps from reference photos, and then generated bas-relief samples from the synthetic normal maps. Finally, they used a CNN-based network to achieve photo-to-depth translation. Due to the ver-

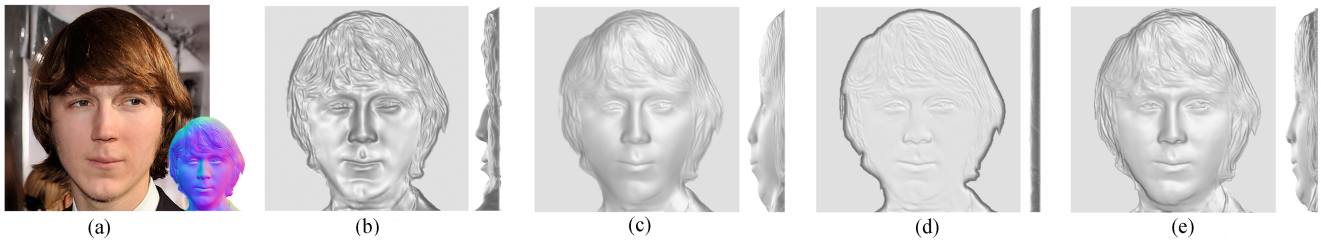


Figure 2. Limitation of [35] in generating high-depth relief (a) photograph and normal map in the dataset of [35]. (b) linear rescaling of the bas-relief [35] in depth direction leads to unreasonable depth ordering and exaggerated geometrical details. (c, d) neither removing the second term nor increasing the constant h in Eq.1 can generate comprising result due to the lack of geometrical constraints for depth reconstruction. (e) high-depth relief predicted by our normal-to-depth network has reasonable depth ordering and natural geometrical details.

satile training samples, the network is capable of handling photographs with various head poses, hairstyles and facial expressions. Besides the state-of-the-art performance, one limitation of this method is that the network can only produce portrait bas-reliefs with fixed depth style and small depth range. Thus, it is hard to apply it in specific scenarios such as art decoration and memorial sculpturing, where high-depth reliefs might be required to provide strong 3D perceptions. In this paper, we aim at extending the method of [35] to a multi-style version that allows users to freely adjust the depth style to meet different application requirement.

Image-to-Image Translation. Image-to-image translation (I2I) aims at learning the mapping from a source domain to a target domain while preserving the content/structure of the input. I2I with CNN-based architecture has been broadly applied in image synthesis [20, 4], segmentation [30, 32], style transfer [25, 5], inpainting [17, 23] and super-resolution [19, 34]. As an alternative to CNN, Vision Transformer [26] designed a self-attention mechanism to capture long-term dependencies between contexts and has shown comparable performance with CNN. Recently, Swin Transformer [18] have achieved state-of-the-art performance in many vision tasks as it integrated the advantages of both CNN and Transformer. Following the hierarchical design, many works have used Swin Transformer as backbone for dense I2I tasks [2, 33, 16, 15]. In this paper, we cast relief modeling from single photograph as a style-conditioning I2I problem, where a style vector is used to control the depth style of the output. Different from the method of [35] that used ResNet blocks [8] for feature representation, we design a U-shaped encoder-decoder architecture that takes Swin-Transformer block [18] as basic unit. We show in the experimental results that the Transformer-based network is able to achieve better performance than CNN-based architectures for photo-to-depth translation.

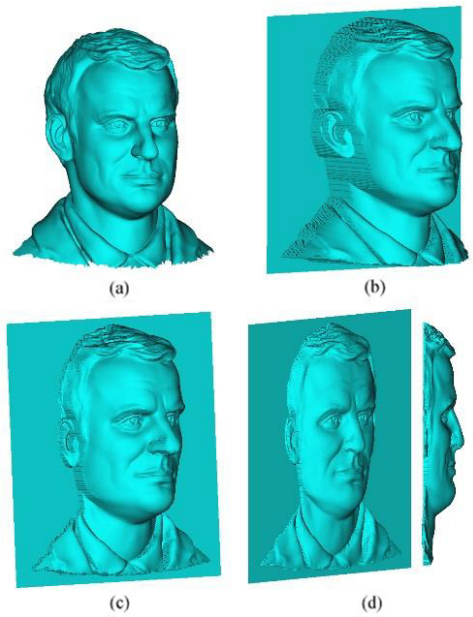


Figure 3. Pipeline of high-depth relief synthesis from 3D portrait. (a) input 3D portrait (b) original depth field with large depth discontinuities at portrait border (c) result of border descending (d) result of depth reconstruction.

3. Method

3.1. Motivation

As mentioned above, Zhang et al.[35] introduced the first relief dataset containing 22k pairs of photograph/bas-relief samples. To construct such a dataset, the authors first estimated normal maps from reference photos, and then generated bas-reliefs from the normal maps by minimizing:

$$\iint_{\Omega}((\nabla H - G)^2) + \mu \cdot (H - h)^2)dxdy \qquad (1)$$

where the first term was used to make the relief gradients ∇H as close as possible to the gradients G, which implicitly recovered geometrical details from the normals, and the second term was used to constrain the relief within depth

range h, which explicitly limited depth variations.

Taking the bas-relief samples as ground-truth data, the authors finally trained a neural network to achieve end-to-end photo-to-depth translation. Limited by the training data, the network can only produce bas-reliefs with small depth range. To produce a high-depth relief, one straightforward way is to apply a linear rescaling of the bas-relief in depth direction. Obviously, this does not work due to the unreasonable depth ordering and exaggerated geometrical details, as shown the example in Fig.2b. We also generate a high-depth relief either by removing the second term in Eq.1 or by increasing the value of h. As shown in Fig.2c and Fig.2d, none of the results are acceptable due to the lack of geometry constraints for depth reconstruction.

To ensure reasonable depth ordering, we propose an optimization-based method for high-depth relief synthesis, and train a neural network to learn the mapping from normal map to relief depth. After that, we use the network to construct high-depth relief samples from the given normal maps of [35] (Section 3.2). In this way, we are able to upgrade the dataset of [35] to contain not only bas-relief but also pairwise high-depth relief for each reference photograph. Taking the bas-reliefs, the new high-depth reliefs and their mixtures as target ground-truths, we finally train a photo-to-depth network to achieve multi-style relief modeling in real-time (Section 3.3).

3.2. High-depth Relief Data Construction

In this section, the main task is to upgrade the database of [35], making it contains not only bas-relief but also pairwise high-depth relief for each reference photograph. For this task, we first train a network to learn the mapping from normal map to relief depth, and then use the network to construct high-depth relief samples from the normal maps of [35].

We need to have enough ground-truth data for normal-to-depth translation. Inspired by the works of [41, 38], we utilize 3D portraits to synthesize high-depth reliefs. We have collected a set of 3D portraits models with various identities, hairstyles and expressions. For each model, we sample 2.5D depth fields and render normal maps from multiple viewing directions (pitch $\theta \in [-5°, 20°]$ and yaw $\varphi \in [-45°, 45°]$) with a resolution of 360×360. As shown in Fig.3b, the original depth field has large depth discontinuities at portrait border. Following the basic rule of relief creation,we hope to bring the portrait border closer to the background, while preserve the original depth structure as much as possible. Here, we propose a two-stage optimization-based method to achieve this target. In the first stage, we linearly descend the border vertices towards the background. The depth offsets inside the portrait caused by border descending are computed by solving a harmonic equation:

$$L \cdot \Delta d = 0 \tag{2}$$

where L is the Laplace–Beltrami matrix and Δd is the offset vector needs to be solved. For the border vertices, we set the descending values as 0.9 times of their original depths, and take them as boundary conditions to solve Eq.1. After that, we update the depth field by subtracting the offset vector from the original one, $d^* = d - \Delta d$, as shown the result in Fig.3c.

In the second stage, we generate a high-depth relief from the updated depth field by minimizing Eq.1, which equals to solve the following linear system:

$$\Delta H + \mu \cdot H = div\ G + \mu \cdot h \tag{3}$$

where ΔH is the depth Laplacian and $div\ G$ is the divergency of gradients that can be estimated from the known normal map. In contrast to Eq.1 that defines h as a constant, we redefine h as a non-constant vector: $h = \alpha \cdot d^*$, where α is the ratio for depth range scaling. Since h has provided pixel-wise depth constraint, the depth ordering can be well preserved in the final relief. We fix the background and portrait border, and take them as boundary conditions when solving Eq.3. By default, we set $\alpha = 0.4$ and $\mu = 0.005$. Fig.3d shows the final relief after depth reconstruction.

By using the above method, we obtain about 15k pairs of high-depth relief/normal map data, among which 13k pairs are used for network training and 2k pairs are used for network testing, as shown the examples in Fig.4. The normal-to-depth network is based on a UNet architecture [21], where the input is a three-channel normal map and the output is one-channel depth map. The encoder has four down-sampling modules with (32, 64, 128, 256) output channels, and the decoder has four up-sampling modules with (256, 128, 64, 1) output channels. Each down-sampling/up-sampling module is composed of two 5×5 convolution layers followed by ReLU activations. The loss function for network training is defined by:

$$L_{total} = L_d + \omega \cdot L_N \tag{4}$$

where L_d is the mean depth error between predicted depths and ground-truths, and L_N is the mean angular error between relief normals and ground-truths. We use $\omega = 0.01$ to balance the two loss terms. The network is trained by using Adam optimizer [14] with a batch size of 16.

We evaluate the performance of the normal-to-depth network by using mean depth error and mean normal angular error. The results are 1.03×10^{-4} and $3.42°$ respectively on the testing data. Fig.4 shows the comparisons of network predictions with ground-truths. The imperceptible differences indicate that the network is capable of predicting high-depth reliefs with good depth ordering even though it receives only single normal maps without known depth constraints.

Once the normal-to-depth network has been trained, we use it to construct high-depth relief samples from the normal maps of [35]. As the network runs very fast (60fps),

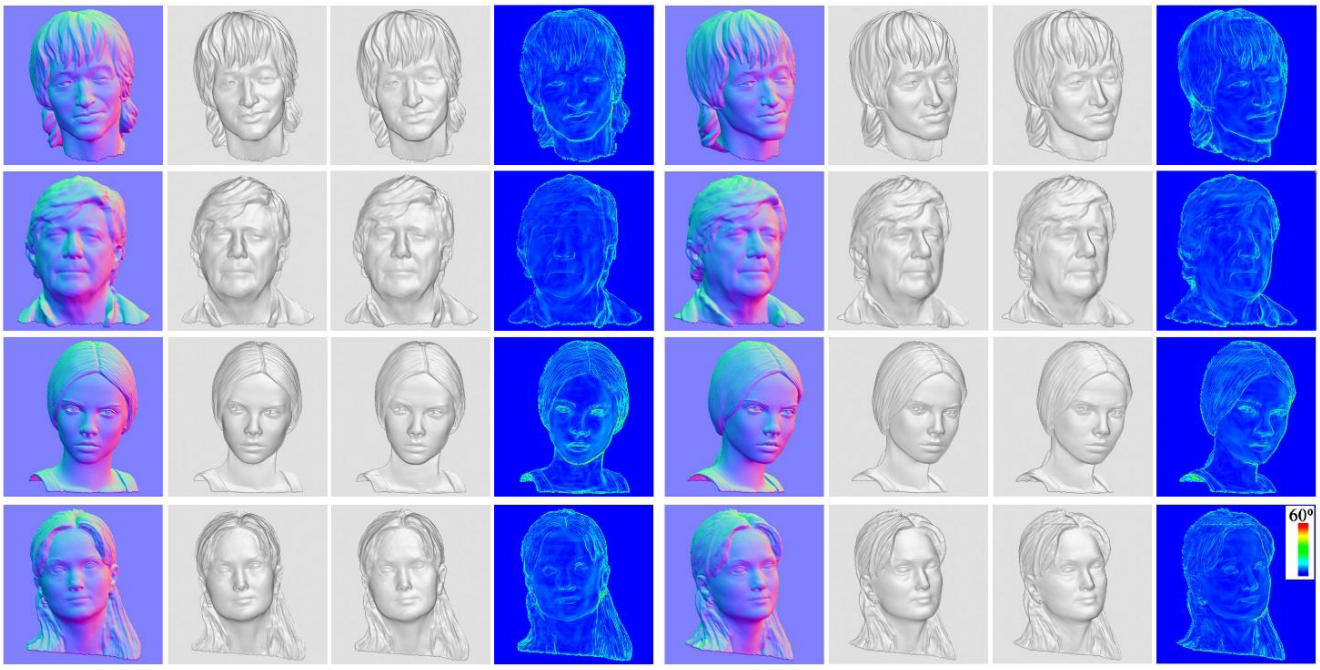


Figure 4. Normal-to-depth network evaluation. From left to right for each example: normal map rendered from 3D portrait, ground-truth relief generated by solving Eq.3, high-depth relief predicted by the normal-to-depth network, and normal angular error map.

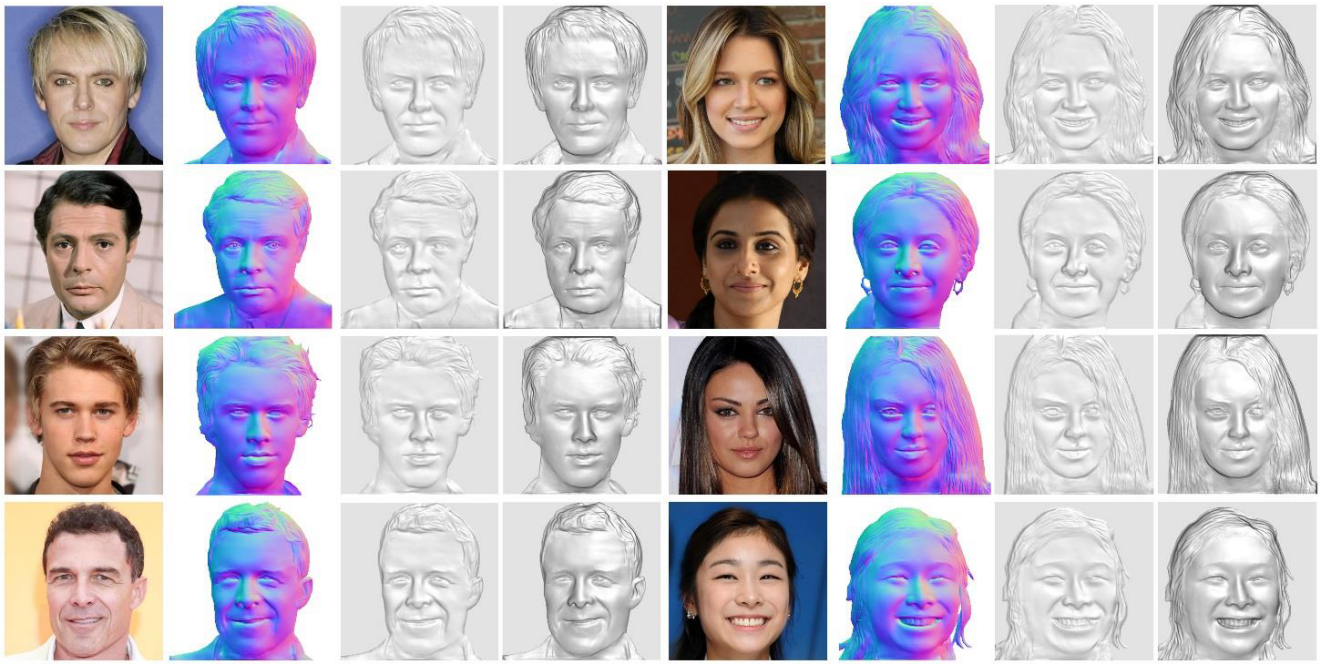


Figure 5. Examples from the upgraded dataset. From left to right for each example: reference photo, normal map, bas-relief in [35] and new high-depth relief predicted by our normal-to-depth network.

the whole process can be completed in less than one hour including the time for data storage. Thus, we succeed in upgrading the dataset of [35], which now contains not only portrait bas-relief but also pairwise high-depth relief for each reference photo, as shown the examples in Fig.5.

3.3. Multi-style Relief Modeling

In this section, the main task is to train a photo-to-depth network to enable portrait relief modeling with adjustable depth multi-style relief modeling. Inspired by the recent

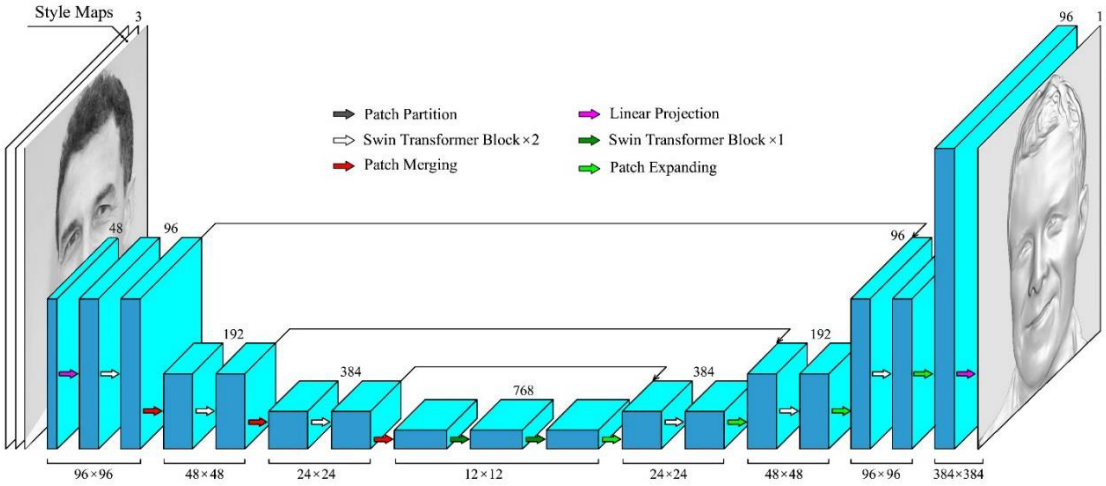


Figure 6. Transformer-based architecture for style-aware photo-to-depth translation

work of [2] that use Swin Transformer [18] as backbone for image segmentation, we design a similar Transformer-based network for style-aware photo-to-depth translation. The whole architecture is presented in Fig.6. The input is a three-channel image including one-channel grayscale photo and two-channel style maps, and the output is one-channel depth map with a resolution of 384×384. Specially, the style maps are generated by a two-element style vector and used to control the depth style of the output relief.

Encoder. Following the specification of Swin Transformer [18], we split the input image into non-overlapping patches with size of 4×4 at the beginning of the encoder. Thus, the raw-valued feature dimension of each patch becomes 4×4×3=48, and the patch token resolution turns to 96×96. After that, a linear embedding layer is applied to project the feature dimension from 48 to 96. The patch tokens then pass through several Swin Transformer blocks and patch merging layers for hierarchical feature transformation. Similar to the down-sampling operation in CNN, the patch merging layer is used to reduce the number of patch token (2× down-sampling). At the same time, it doubles the dimension of the input features.

Decoder. The decoder has symmetric structure with the encoder. It consists multi-scale Swin Transformer blocks and patch expanding layers. Different from the patch merging operation in the encoder, a patch expanding layer [33] is used to up-sample the resolution of patch tokens (2× up-sampling) and decrease feature dimension. Similar to the UNet architecture [21], skip connections are added to concatenate multi-scale features in the encoder and up-sampled features in the decoder. In the end, we apply a dual patch expanding layer [7] with Bilinear and Pixel-Shuffle up-sampling to transform the feature maps to the original resolution 384×384, and a linear projection layer to output one-channel depth field.

Swin Transformer block. We follow the structure of LayerNorm, multi-head self-attention module and residual connection in Swin Transformer block [18]. The window-based multi-head self-attention (W-MSA) module and the shifted window-based multi-head self-attention (SW-MSA) module are applied in two successive transformer blocks. Inspired by [27], we modify the two-layer MLP by adding a 3 ×3 depth-wise convolution between the first fully-connected layer and GELU. We find that this slight change is very effective in eliminating block artifacts.

Training. We divide the database (Section 3.2) into two groups, 20k for network training and 3k for network evaluation. Each photograph has pixel-wise bas-relief and high-depth relief. We define a set of style vectors to guide style-aware relief modeling. The two basic depth styles, i.e. bas-relief and high-depth relief, are specified with style vectors [1.0, 0.0] and [0.0, 1.0] respectively. Assuming the vector elements be s_1 and s_2, the depth values of bas-relief and high-depth relief be D_b and D_h, the ground-truth relief corresponding to $[s_1, s_2]$ can be computed by weighed depth interpolation:$D_{new} = s_1 \cdot D_b + s_2 \cdot D_h$. During training, we combine eight photos into a group, and assign each photo with three kinds style vectors: [1.0, 0.0], [0.0, 1.0] and $[s_1, s_2]$, which correspond to D_b, D_h and D_{new} respectively. The values of s_1 and s_2 are randomly selected from $\{0.5, 1, 1.5, 2\}$ and $\{0.2, 0.4, 0.6, 0.8, 1.0\}$ respectively. Thus, the batch size for network training becomes: 8×3=24. We train the Transformer-based network using the same loss function in Eq.4. We employ Adam optimizer [7] for 80 epochs using a cosine decay learning rate scheduler. An initial learning rate of 0.001 and a weight decay of 0.05 are used for back propagation.

Table 1. Network evaluations with multiple style vectors

		Style vector								
		[1.0,0.0]	[1.0,0.5]	[1.0,1.0]	[1.5,0.0]	[1.5,0.5]	[1.5,1.0]	[2.0,0.0]	[2.0,0.5]	[2.0,1.0]
MDE($\times 10^{-4}$)	ResUnet	0.266	5.305	19.353	2.987	6.006	22.854	3.939	7.252	27.002
	Our network	**0.188**	**2.294**	**7.970**	**0.593**	**3.016**	**8.902**	**0.950**	**4.148**	**10.531**
MNAE(°)	ResUnet	7.370	11.523	14.902	11.16	13.416	16.288	12.777	15.108	17.602
	Our network	**6.789**	**10.372**	**13.542**	**9.435**	**12.268**	**14.950**	**11.628**	**14.012**	**16.285**

Figure 7. Perceptual evaluations. From left to right in each example: input photo, result of ResUnet, result of our Transformer-based network and ground-truth. The style vector is set by [1.0, 1.0].

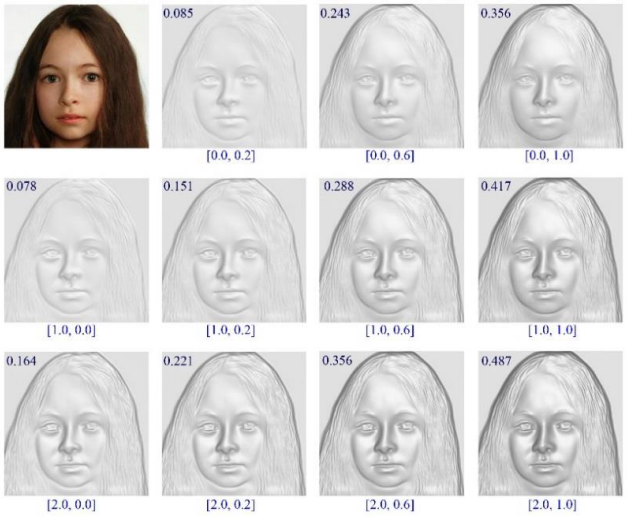


Figure 8. Effect of style vector with different element combination. Depth range value is shown on the top-left of each figure. By default, the x- and y- coordinate are limited in the range of [-1.0, 1.0]

4. Experimental Results

Network evaluation. In the work of [35], Zhang et al. designed several CNN-based architectures to model portrait bas-reliefs from single photographs. Through qualitative and quantitative comparisons, they chose ResUnet as the final network which gave the best results. To verify the advantage of the Transformer-based architecture, we compare it with the CNN-based ResUnet in terms of mean depth error (MDE) and mean normal angular error (MNAE). For fair comparison, we add two-channel style maps to the input of ResUnet, and re-train the network using the same data, loss function in Section 3.3.

As shown in Table 1, the Transformer-based network outperforms both on MDEs and MNAEs with less parameters (20.30M vs 95.92M). Particularly, the mean depth errors (MDEs) have been reduced by a large margin. We argue that the performance improvement is mainly caused by the long-range self-attention mechanism, which can better learn feature representation for photo-to-depth translation. We also make perceptual comparisons by taking style vector [1.0, 1.0] as input. As shown in Fig.7, both results have promising depth ordering, but the geometrical details

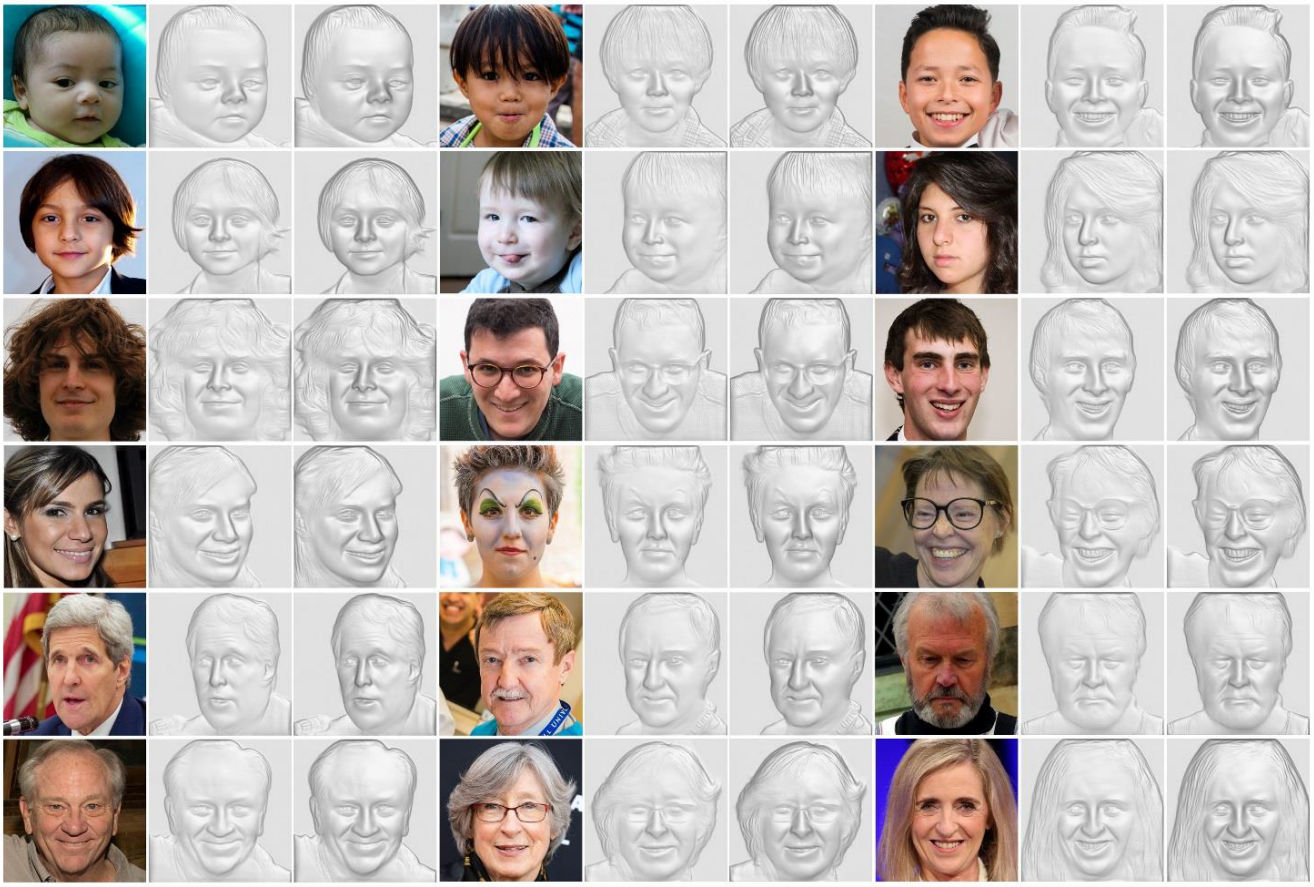


Figure 9. Modeling on FFHQ dataset [12]. Input style vectors are [1.0, 0.5] and [1.0, 1.0] respectively.

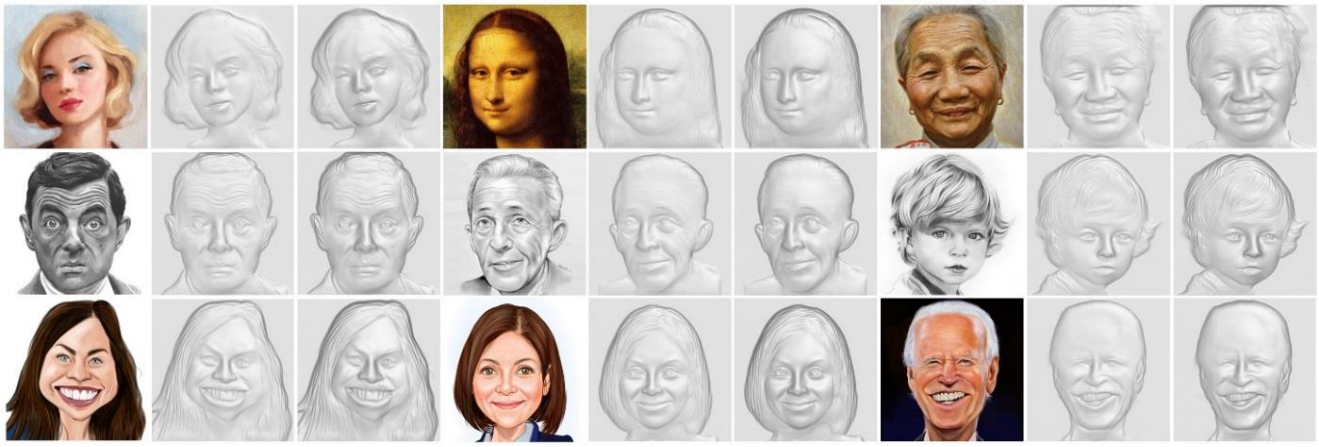


Figure 10. Modeling on stylized images. Input style vectors are [1.0, 0.5] and [1.0, 1.0] respectively.

are weaker than those of the ground-truths. Compared to the ground-truths, our predictions contain less geometrical noises at the face regions.

Effect of Style Elements. The style vector $[s_1, s_2]$ is used to control the depth style of the output. Fig.8 shows the reliefs produced by different element combinations. It can be seen that the depth range and depth ordering vary smoothly with the adjustment of s_1 and s_2, while the same depth range might be produced by different style vectors. For instance, the reliefs produced by the style vectors [0.0, 1.0] and [2.0, 0.6] have same depth range (0.356), but the depth ordering looks quite different, particularly at the face re-

Figure 11. Reliefs fabricated by NC machining with different material. Top: plaster. Bottom-left: acrylic. Bottom-right: copper plated with gold.

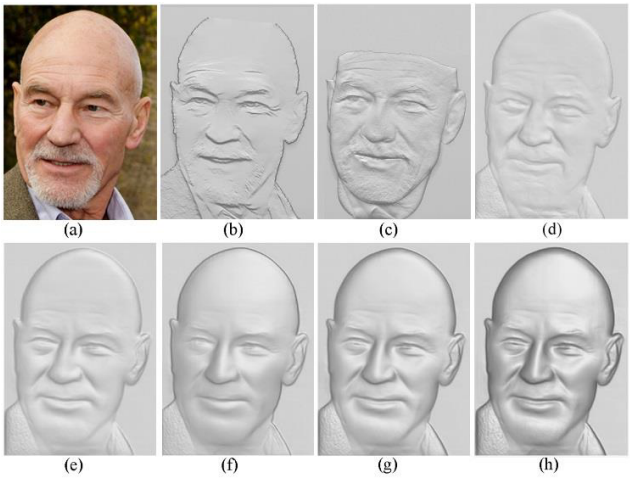


Figure 12. Comparisons with previous methods. (a) input photo. (b) result of Wu et al.[29]. (c) result of Zhang et al.[40]. (d) result of Zhang et al.[35]. (e-h) our results with input style vectors [1.5, 0.0], [0.0, 1.0], [1.5, 0.5] and [2.0, 1.0] respectively.

gion. In each row, the depth range gradually rises with the increasement of s_2. In each column, although the depth range also rises with the increasement of s_1, the change of depth ordering is more impressive.

In real applications, the values of $s_1 \in [0.0, 2.0]$ and $s_2 \in [0.1, 1.0]$ can be determined in accordance to the material used for relief fabrication. For example, a middle s_1 and a small s_2 are suitable to produce a metallic bas-relief. For the applications of art decoration and memorial sculpturing, a large s_2 would be appropriate to produce a high-depth relief. In case that the relief is made of low-reflective material such as stone and plaster, a large s_1 can be applied to enhance the contrast of facial features. In case that the relief is made by high-reflective metallic material, it is better to reduce the value of s_1 to avoid excessive surface distortion.

Modeling on FFHQ dataset. Fig.9 shows the experimental results on FFHQ database [12]. Different from the method of [35] that requires a mask map for network prediction, our method does not need any portrait mask. Supervised by the training data, the network is able to identify the foreground, and output a portrait relief with zero background. Due to the versatile training data, the network can handle photos with various skin colors, hairstyles, expressions, poses and illumination conditions.

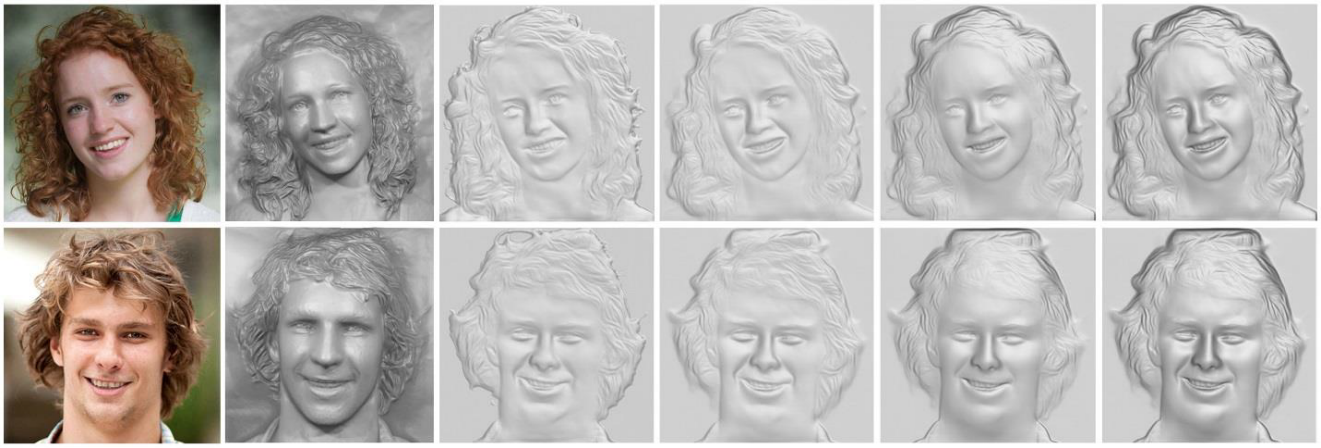
Modeling on Stylized Images. The proposed network can also model portrait reliefs from stylized images such as paintings, pencil drawings and caricatures. As shown in Fig.10, although the stylized images differ a lot from real-world photos, and some portraits even have exaggerated facial expressions, the network is still able to produce reliefs with promising appearances.

Fabrication. Once a desired portrait relief has been generated, it can be converted into the form of 3D mesh, and fabricated by NC machining or 3D printing. Fig.11 shows several examples fabricated by NC machining. From top to bottom, the reliefs are made of non-metallic material (plaster), transparent material (acrylic) and metallic material (copper plated with gold) respectively. It can be seen that all results have natural depth orderings and rich geometrical details.

Animation. The Transformer-based network runs very fast (60fps) and can be used to produce real-time relief animations from portrait videos. Please refer to the animations in the supplemental material. In each frame of the animation, we render the relief using Blinn-Phone shading algorithm.

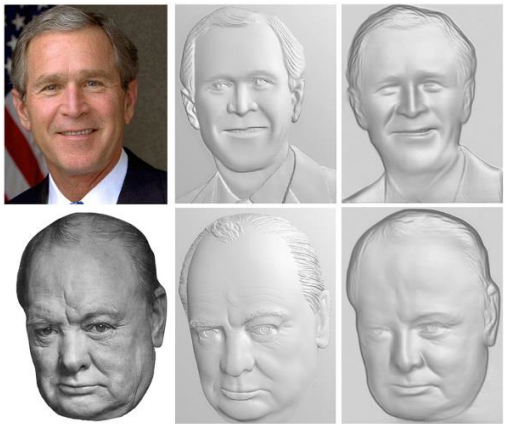
Discussion. It should be noted that some other strategies can also be applied to reach the target of multi-style relief modeling. For example, one can first train a neural network using our relief data to infer a high-depth relief from single photograph, and then use the method of [35] to produce a pixel-wise bas-relief. After that, a desired style of relief can be generated by weighted depth interpolation. For another example, one can also train a neural network to infer a high-depth relief, followed by taking advantage of previous model-based methods such as [40] and [36] to generate a portrait relief with desired depth range and depth ordering. However, these alternatives are much more time-consuming than our single-pass solution in this paper, which runs network only once without any additional computation.

Comparisons with previous methods. We now compare our method with previous works of [35, 29, 40], which also aim at modeling portrait reliefs from singe photographs. As shown in Fig.12, the method of Wu et al. [29] succeeds in recovering facial details, but the resulting relief has unnatural depth ordering. The result of Zhang et al.[40] has good depth ordering and fine facial details due to the template-based depth optimization. However, only the face region



(a) (b) (c) (d) (e) (f)

Figure 13. Hair comparisons. (a) result of Chai et al.[13]. (b) result of Zhang et al.[35]. (c-f) our reliefs produced with style vectors [1.0, 0.0], [0.0, 1.0] and [1.0, 1.0] respectively.



(a) (b) (c)

Figure 14. Comparisons with hand-made reliefs. (a) input photos. (b) reliefs created by artists. (c) our results.

can be constructed in the final bas-relief. The method of Zhang et al. [35] models full head features through a CNN-based network, but the result is limited within a small depth range. Instead, we extend the method of [35] to a generalized version, which is able to produce not only bas-reliefs, but also portrait reliefs with adjustable depth style, as shown the results in Fig.12.

Next, we compare our method with the work of [3], which focuses on hair modeling from a single photo. By combining depth clues and hair priors in an optimization framework, the method of Chai et al.[3] is able to construct high-fidelity hair geometry, as shown in Fig.13b. However, it requires user interventions to segment the hair region in a pre-processing stage, thus bringing difficulties to users. In contrast, the method of [35] constructs hair geometry through a neural network, without the need of hair segmentation. However, it requires a mask map [13] to identify the foreground. In contrast, our method does not need any mask information. It implicitly extracts hair features from the input photo, and outputs hair geometry naturally fused with the face region and the background. As shown in Fig.13d, our results generated by style vector [1.0, 0.0] have comparable hair quality with the ones in Fig.13c. In case a new style vector is fed to the network, the strength of geometrical details can be adaptively adjusted, as shown in Fig.13e and Fig.13f.

Finally, we compare our results with reliefs created by artists. Given a reference photo, artists usually utilize multiple sculpturing tools to model a portrait relief. On one hand, manual operation is flexible in dealing with different type of head features, even the ones with poor image quality. On the other hand, the modeling quality highly depends on the skills of the artists. Differently, our multi-style solution models portrait reliefs in an end-to-end way. Once a reference photo and a style vector are fed into the network, it takes advantage of shading clues in the input photo and automatically outputs a relief that mimics the appearance of the input. As shown in Fig.14, our method produces more realistic results than those created by artists.

Artistic evaluation. We have invited 5 skilled artists to evaluate the quality of our experimental results. All artists have over three years of experiences in modeling portrait reliefs. We provide each artist 100 photographs from FFHQ dataset [3] and corresponding relief models produced by four types of style vectors [1.0, 0.0], [1.0, 0.5], [1.0, 1.0] and [2.0, 1.0]. The artists are asked to score the reliefs from four attributes respectively: face quality, hair quality, depth ordering and feature sharpness. The score varies from 2 to 5, indicating poor (2), medium (3), good (4) and excellent (5) respectively.

Table 2. Artistic evaluation

	Face quality	Hair quality	Depth order	Feature sharpness
Artist #1	4.5	4.0	4.0	3.0
Artist #2	4.0	3.8	4.2	3.2
Artist #3	4.2	4.0	4.0	3.6
Artist #4	4.0	4.0	4.2	3.2
Artist #5	4.0	3.6	3.8	3.0
Mean score	4.14	3.88	4.04	3.2

Evaluation results are reported in Table 2, where the mean scores are 4.14, 3.88, 4.04 and 3.2 respectively. All artists have given positive feedbacks on the face quality, hair quality and depth ordering. They report that the high-depth reliefs look more impressive than the bas-reliefs with small depth range. They also point out that the sharpness of some portrait features such as eyes and cloth collars should be enhanced. All artists support the multi-style strategy, which increases the flexibility of relief creation and provides users more choices to meet different application requirement.

In this paper, we present a multi-style solution for portrait relief modeling from a single photograph. To provide ground-truth data for network training, we upgrade the database of [35], making it not only contain bas-relief sample, but also pixel-wise high-depth relief for each photograph. Taking the two types of reliefs and their mixtures as target ground-truths, we finally train a photo-to-depth network to achieve style-aware relief modeling. The Transformer-based network allows users to freely adjust the depth style to meet different application requirement. Experimental results and comparisons with previous methods has proved the state-of-the-art performance of the proposed method.

Our method still has some limitations. Similar to the work of [35], it fails to handle photographs with blurring portrait features and hard shadows. Currently, it is not effective in modeling hats, eyeglasses, beards and collars due to the lack of training data. We plan to solve these problems in the future.

Acknowledgements

We would like to thank the anonymous reviewers for their reviews and valuable suggestions. We also thank the artists who participate quality evaluations. This work was supported in part by the National Natural Science Foundation of China (Grant No. 61772293, No. 62072274).

References

[1] M. Alexa and W. Matusik. Reliefs as images. *ACM Transactions on Graphics*, 29(4), 2010. 2

[2] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *ECCV Medical Computer Vision Workshop*, 2021. 3, 6

[3] M. Chai, L. Luo, K. Sunkavalli, N. Carr, S. Hadap, and K. Zhou. High-quality hair modeling from a single portrait photo. *ACM Transactions on Graphics*, 34(6):1–10, 2015. 10

[4] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 3

[5] M. J. Chong and D. Forsyth. Jojogan: One shot face stylization. *arXiv preprint arXiv:2112.11641*, 2021. 3

[6] P. Cignoni, C. Montani, and R. Scopigno. Computer-assisted generation of bas-and high-reliefs. *Journal of graphics tools*, 2(3):15–28, 1997. 2

[7] C.-M. Fan, T.-J. Liu, and K.-H. Liu. Sunet: Swin transformer unet for image denoising. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2022. 6

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[9] M. Hudon, M. Grogan, A. Smolic, et al. Deep normal estimation for automatic shading of hand-drawn characters. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 246–262, 2018. 2

[10] Z. Ji, W. Feng, X. Sun, F. Qin, Y. Wang, Y.-W. Zhang, and W. Ma. Reliefnet: fast bas-relief generation from 3d scenes. *Computer-Aided Design*, 130:102928, 2021. 2

[11] Z. Ji, W. Ma, and X. Sun. Bas-relief modeling from normal images with intuitive styles. *IEEE transactions on visualization and computer graphics*, 20(5):675–685, 2013. 2

[12] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 8, 9

[13] Z. Ke, K. Li, Y. Zhou, Q. Wu, X. Mao, Q. Yan, and R. W. Lau. Is a green screen really necessary for real-time portrait matting? *arXiv preprint arXiv:2011.11961*, 2020. 10

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[15] F. Li, H. Zhang, S. Liu, L. Zhang, L. M. Ni, H.-Y. Shum, et al. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *arXiv preprint arXiv:2206.02777*, 2022. 3

[16] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 1833–1844, 2021. 3

[17] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100, 2018. 3

[18] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 2, 3, 6

[19] X. Luo, Y. Xie, Y. Zhang, Y. Qu, C. Li, and Y. Fu. Latticenet: Towards lightweight image super-resolution with lattice block. In *European Conference on Computer Vision*, pages 272–289. Springer, 2020. 3

[20] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 3

[21] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4, 6

[22] C. Schüller, D. Panozzo, and O. Sorkine-Hornung. Appearance-mimicking surfaces. *ACM Transactions on Graphics*, 33(6), 2014. 2

[23] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[24] W. Su, D. Du, X. Yang, S. Zhou, and H. Fu. Interactive sketch-based normal map generation with deep neural networks. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 1(1), 2018. 2

[25] M. Tomei, M. Cornia, L. Baraldi, and R. Cucchiara. Art2real: Unfolding the reality of artworks via semantically-aware image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5849–5859, 2019. 3

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[27] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao. Pvtv2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3), 2022. 6

[28] T. Weyrich, J. Deng, C. Barnes, S. Rusinkiewicz, and A. Finkelstein. Digital bas-relief from 3d scenes. *ACM transactions on graphics*, 26(3), 2007. 2

[29] J. Wu, R. R. Martin, P. L. Rosin, X.-F. Sun, Y.-K. Lai, Y.-H. Liu, and C. Wallraven. Use of non-photorealistic rendering and photometric stereo in making bas-reliefs from photographs. *Graphical Models*, 76(4):202–213, 2014. 2, 9

[30] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018. 3

[31] Z. Yang, B. Chen, Y. Zheng, X. Chen, and K. Zhou. Human bas-relief generation from a single photograph. *IEEE Transactions on Visualization & Computer Graphics*, 28(12):4558–4569, 2022. 2

[32] C. Yu, Y. Shao, C. Gao, and N. Sang. Condnet: Conditional classifier for scene segmentation. *IEEE Signal Processing Letters*, 28:758–762, 2021. 3

[33] B. Zhang, S. Gu, B. Zhang, J. Bao, D. Chen, F. Wen, Y. Wang, and B. Guo. Styleswin: Transformer-based gan for high-resolution image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11304–11314, 2022. 3, 6

[34] K. Zhang, J. Liang, L. Van Gool, and R. Timofte. Designing a practical degradation model for deep blind image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4791–4800, 2021. 3

[35] Y.-W. Zhang, P. Luo, H. Zhou, Z. Ji, H. Liu, Y. Chen, and C. Zhang. Neural modeling of portrait bas-relief from a single photograph. *IEEE Transactions on Visualization & Computer Graphics*, pages 1–16, 2022. 1, 2, 3, 4, 5, 7, 9, 10, 11

[36] Y.-W. Zhang, B.-b. Qin, Y. Chen, Z. Ji, and C. Zhang. Portrait relief generation from 3d object. *Graphical Models*, 102:10–18, 2019. 2, 9

[37] Y.-W. Zhang, J. Wang, W. Long, H. Liu, C. Zhang, and Y. Chen. A fast solution for chinese calligraphy relief modeling from 2d handwriting image. *The Visual Computer*, 36(10):2241–2250, 2020. 2

[38] Y.-W. Zhang, J. Wang, W. Wang, Y. Chen, H. Liu, Z. Ji, and C. Zhang. Neural modelling of flower bas-relief from 2d line drawing. *Computer Graphics Forum*, 40:288–303, 2021. 2, 4

[39] Y.-W. Zhang, J. Wu, Z. Ji, M. Wei, and C. Zhang. Computer-assisted relief modelling: A comprehensive survey. *Computer Graphics Forum*, 38:521–534, 2019. 2

[40] Y.-W. Zhang, C. Zhang, W. Wang, Y. Chen, Z. Ji, and H. Liu. Portrait relief modeling from a single image. *IEEE transactions on visualization and computer graphics*, 26(8):2659–2670, 2019. 2, 9

[41] Y.-W. Zhang, Y.-Q. Zhou, X.-L. Li, H. Liu, and L.-L. Zhang. Bas-relief generation and shape editing through gradient-based mesh deformation. *IEEE transactions on visualization and computer graphics*, 21(3):328–338, 2014. 2, 4