Heterogeneous Avatar Synthesis Based on Disentanglement of Topology and Rendering

Nan Gao[0000−0001−9547−1304], Zhi Zeng[0000−0003−2150−2088], GuiXuan Zhang†[0000−0002−1072−8279], and ShuWu Zhang[0000−0002−6013−6351]

Institute of Automation Chinese Academy of Sciences, Beijing, China
{nan.gao,zhi.zeng,guixuan.zhang,shuwu.zhang}@ia.ac.cn

Abstract. There are obviously structural and color discrepancies among different heterogeneous domains. In this paper, we explore the challenging heterogeneous avatar synthesis (HAS) task considering topology and rendering transfer. HAS transfers the topology as well as rendering styles of the referenced face to the source face, to produce high-fidelity heterogeneous avatars. Specifically, first, we utilize a Rendering Transfer Network (RT-Net) to render the grayscale source face based on the color palette of the referenced face. The grayscale features and color style are injected into RT-Net based on adaptive feature modulation. Second, we apply a Topology Transfer Network (TT-Net) to conduct heterogeneous facial topology transfer, where the image content of RT-Net is transferred based on AdaIN controlled by heterogeneous identity embedding. Comprehensive experimental results show that the disentanglement of rendering and topology is beneficial to the HAS task, and our HASNet has comparable performance compared with other state-of-the-art methods.

Keywords: Image synthesis · Style transfer · Disentanglement representation learning.

1 Introduction

Avatar means another stylized identity in the heterogeneous domain. As for face images, there are various topology patterns for facial components, such as 3D cartoon, 2D anime, sketch, nir, real-world or other domains. Style transfer methods [11], [13], [19], [26], [20] change the textural style and preserve the content of the source image, guided by the referenced image. These methods are universal to different heterogeneous domains but ignore the facial topology transfer. To adapt to the target face distribution, some GAN-based methods [28], [17], [24] are proposed to address two-domain face translations. However, these methods have obvious topology and color distortions.

The high-fidelity HAS task is also a style transfer task, which is supposed to possess high-fidelity facial topology and global color consistencies with the

† Corresponding author

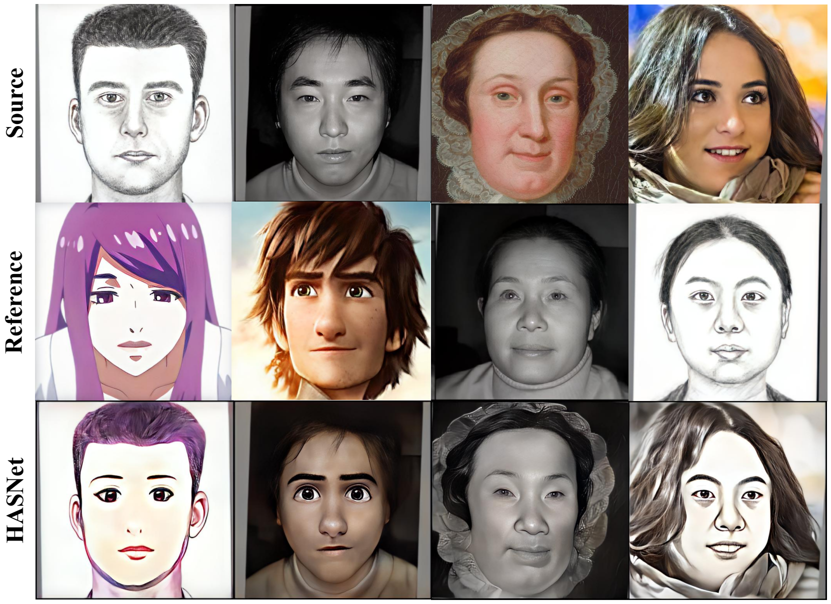

Fig. 1. HASNet realizes high-fidelity heterogeneous avatar synthesis based on topology and rendering transfer, where HASNet handles various heterogeneous domains.

reference-domain face, while preserving as much as other content and attributes of the source face, e.g., facial pose, expression, background and hair. Heterogeneous faces have diverse facial component topology patterns. Moreover, since the different domains might have a large color distribution discrepancy, it poses us a non-trivial challenge to implement HAS task without considering rendering consistency. Therefore, to produce more vivid heterogeneous avatars, both rendering and topology styles deserve to be manipulated. As for color rendering, ColorThief is used to extract prominent color palette values. As for topology adaptation, the pretrained identity extractor is capable of capturing discriminative facial component shapes, e.g., Arcface extracts the prominent structural feature of a face. In our paper, we conduct rendering transfer based on the color palette from ColorThief [1], and implement topology transfer based on the manifold of Arcface [6]. This disentanglement framework realizes high-fidelity heterogeneous avatar synthesis. As shown in Figure 1, the results of HASNet have a good performance on style transfer of global color and facial topology, while preserving other contents and attributes of the source faces.

Our contributions are three folds as follows.

- We propose a two-stage framework to explore the challenging and meaningful heterogeneous avatar synthesis task. We propose an effective rendering transfer framework. By considering adaptive modulation of color and spa-

[1] http://lokeshdhakar.com/projects/color-thief/

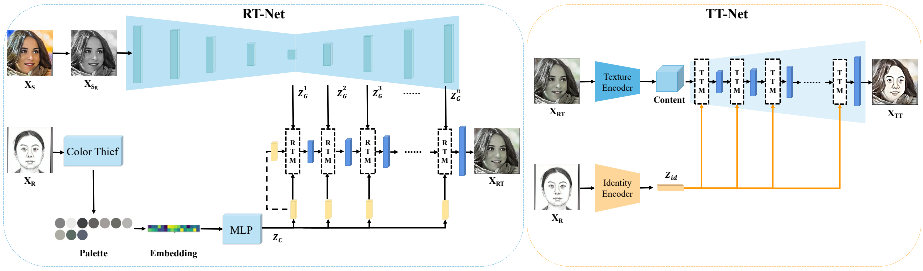

Fig. 2. Two-stage framework of HASNet. RT-Net helps HASNnet implement higher-fidelity heterogeneous rendering. TT-Net refines the facial topology style of the heterogeneous avatars.

tial content, RT-Net achieves topology-aware and high-quality colorization in various heterogeneous domains.

- Based on the first stage, we propose TT-Net to realize high-fidelity and controllable facial topology transfer across various heterogeneous domains.
- We observe HASNet variants and conduct comparative experiments with state-of-the-art universal and GAN-based style transfer methods. Experimental results demonstrate a good performance of HASNet both in visual and quantitative manners.

2 Related Work

2.1 Universal Style Transfer

There are some universal style transfer methods [11], [13], [19], [26]. [11] explores the domain-aware characteristics from the texture and topology features of referenced images to handle both artistic and photo-realistic style transfer. AdaIN [13] makes the variance and mean values of the source features aligned with those of the reference features, which achieves real-time universal style transfer. WCT [19] enables universal style transfer based on disentangled whitening and coloring operations. SANet [26] integrates the style patterns according to the semantic spatial attentional map of the source content.

2.2 GAN-based Methods

StyleGANs [15], [16], [14] synthesize high-resolution images based on noise disentanglement and style modulation. UGATIT [17] modulates the generator applying the discriminative domain-aware style embedding and proposes AdaILN. NICE-GAN [5] reuses the discriminator to implement unsupervised image-to-image translation. [7] proposes an unsupervised face rotation method considering disentangling the facial shape and identity. Toonify [28] trains the generator and discriminator for 3D cartoon and real-world domain, respectively. Then the base and transferred models are used to create the interpolated translation model. FS-Ada [24] proposes a transfer learning schedule to impose the adaption to diverse target domains from the source model, based on limited data.

2.3 Cross-domain Image Synthesis

As for VIS-NIR translation, [8], [9] demonstrate that dual heterogeneous face mapping is beneficial to recognizing heterogeneous faces. [33] utilizes the inter-supervision and intra-supervision style transfer for Sketch-VIS task. [32] designs a memorized Grayscale-VIS colorization model with AdaIN modulation. As for heterogeneous datasets for research, CASIA NIR-VIS 2.0 [18], Oulu-CASIA NIR-VIS [4] and BUAA-VisNir Face [12] databases are widely used in the NIR-VIS task. CUHK Face Sketch FERET (CUFSF) [34], IIIT-D Viewed Sketch database [2] and XM2VTS database [22] are established in Sketch-VIS research. [17] proposes the selfie2anime dataset in the anime domain. IIIT-CFW [23] provides some faces with exaggerated drawings. Tufts dataset [25] presents subjects captured in different imaging domains, e.g. RGB, NIR and Thermal.

3 Approach

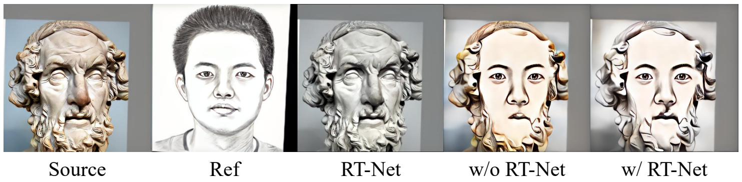

Fig. 3. RT-Net helps HASNnet implement higher-fidelity heterogeneous rendering.

We design a two-stage framework to synthesize heterogeneous avatars, as shown in Figure 2. In stage I, given a source face X_S, we obtain the grayscale image X_{Sg} by means of the transformation from RGB to Lab space. We apply ColorThief to extract the dominant color palette of the reference face X_R. Based on the spatial features of X_{Sg} and color style z_C, we modulate RT-Net to obtain the rendering transfer face X_{RT}. In stage II, we conduct topology transfer based on X_{RT} to synthesize X_{TT} whose facial topology distribution is consistent with X_R. In this way, HASNet generates high-fidelity heterogeneous avatars.

3.1 RT-Net

As shown in Figure 3, the color consistency between the facial area and the background is higher for w/ RT-Net than w/o RT-Net, which indicates the importance of RT-Net for HAS task. More examples are shown in Figure 5.

In RT-Net, X_{Sg} is mapped to the multi-scale spatial feature maps $z_G = \{z_G^1, z_G^2, ..., z_G^n\}$ via an Unet [29]. A multi-FC mapping network MLP is leveraged to convert the extracted color palette embedding $C_{X_R} \in \mathbb{R}^{20 \times 3}$ to color style $z_C \in \mathbb{R}^{256}$, which is used to modulate multi-scale decoder of RT-Net. We integrate these two embeddings adopting the Rendering Transfer Module (RTM) in an adaptive alignment manner.

Specifically, the input feature $h^i \in \mathbb{R}^{C_h^i \times H^i \times W^i}$ of RTM_i is aligned with guidance of two groups of learning parameters from z_G^i and z_C, respectively. It is formulated as:

$$h_I^i = \frac{h^i - \mu_I^i}{\sigma_I^i}, \tag{1}$$

$$\{G, C\}_I^i = \gamma^i_{\{G,C\}_I} \odot h_I^i + \beta^i_{\{G,C\}_I}, \tag{2}$$

where μ_I^i and σ_I^i, i.e., the mean and standard deviation of h^i, are used as instance-aware normalization parameters. Let $z_G^i \in \mathbb{R}^{C_G^i \times H_G^i \times W_G^i}$ and $z_C \in \mathbb{R}^{C_C \times 1}$ be the grayscale and color embedding, respectively. Furthermore, affine transform parameters $\gamma^i_{G_I}$ and $\beta^i_{G_I} \in \mathbb{R}^{C_h^i \times H^i \times W^i}$ are obtained from z_G^i using a convolutional layer, similar with SPADE [27]. Meanwhile, $\gamma^i_{C_I}$ and $\beta^i_{C_I} \in \mathbb{R}^{C_h^i \times H^i \times W^i}$ are mapped from z_C using a full connection transform, inspired by AdaIN [13]. To adaptively integrate the spatial and color modulation, the attentional maps $M^i_{G_I}$ and $M^i_{C_I}$ from h_I^i are generated after passing a convolution layer followed by a sigmoid operation. The instance denormalization result H_I^i is denoted as

$$H_I^i = G_I^i \odot M^i_{G_I} + C_I^i \odot M^i_{C_I}, \tag{3}$$

where $\odot$ represents the element-wise product, and $M^i_{G_I} + M^i_{C_I} = 1$.

As for the objective functions, in the training stage of RT-Net, we use Huber loss where $\triangle_{X_{RT}} = |X_{RT} - X_S|$, to constrain the colored image reconstruction via

$$\mathcal{L}_{rec} = \begin{cases} \frac{1}{2}\triangle^2_{X_{RT}} & \triangle_{X_{RT}} \leq \delta \\ \delta \cdot \left[\triangle_{X_{RT}} - \frac{1}{2}\delta\right] & otherwise \end{cases}. \tag{4}$$

Moreover, we optimize a discriminator for the generated X_{RT} with conditions of X_{Sg} and C_{X_S}. The total loss of RT-Net is as follows

$$\mathcal{L}_{RT-Net} = \mathcal{L}_{adv}(X_{Sg}, C_{X_S}, X_{RT}) + \lambda_{rec}\mathcal{L}_{rec}. \tag{5}$$

In the test time, the color style is extracted from the heterogeneous reference face X_R, to lessen the color gap between the source and reference domains.

3.2 TT-Net

After stage I, the global rendering style has been transferred to the reference heterogeneous domain. Stage II designs TT-Net to further transfer the facial topology style to produce vivid HAS results.

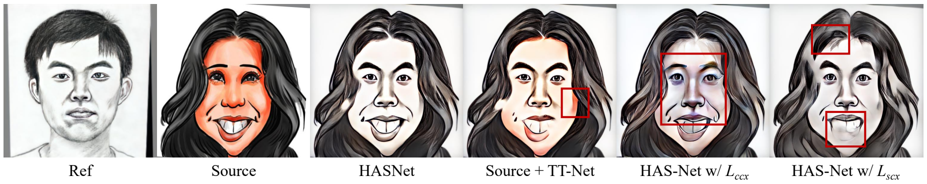

Fig. 4. $\mathcal{L}_{SCX}$ helps HASNnet implement higher-fidelity topology transfer.

In the multi-level topology transfer modulation (TTM) module, we inject z_{id} to each feature level of TT-Net by implementing layer-aware AdaIN operations, where we use StyledConv block of StyleGAN [16]. It is denoted as:

$$AdaIN(Con^i, \gamma_{id}^i, \beta_{id}^i) = \gamma_{id}^i \odot \frac{Con^i - \mu^i}{\sigma^i} + \beta_{id}^i, \tag{6}$$

where $Con^i \in \mathbb{R}^{C_{Con}^i \times H^i \times W^i}$ is the spatial information of TTM_i. σ^i and μ^i are the standard deviation and mean values of Con^i. γ_{id}^i and $\beta_{id}^i \in \mathbb{R}^{C_{Con}^i \times H^i \times W^i}$ based on z_{id} are sent to address the instance denormalization. The initial content feature Con^0 is extracted by VGG [30] for X_{RT}.

As for the objective functions, we use the reconstruction loss as the pixel-level supersision between X_{TT}^{ij} and X_S^i, when the source face X_S^i and reference face X_R^j are same. It is denoted as

$$\mathcal{L}_{rec} = \begin{cases} \frac{1}{2}\left\|X_{TT}^{ij} - X_S^i\right\|_2^2 & if\ i = j \\ 0 & otherwise \end{cases}. \tag{7}$$

We use identity consistency loss on X_{TT}^{ij} for HAS face synthesis as follows. Specifically, we measure the distance between the Arcface manifolds of X_{TT}^{ij} and X_R^j via

$$\mathcal{L}_{id} = 1 - \langle z_{id}(X_{TT}^{ij}), z_{id}(X_R^j)\rangle, \tag{8}$$

where $\langle\cdot,\cdot\rangle$ means cosine similarity, z_{id} is the pretrained Arcface [6].

Furthermore, TT-Net utilizes the contextual similarity loss [21] for X_{TT}^{ij} and X_R^j to improve the contextual distribution perception. It is denoted as

$$\mathcal{L}_{SCX} = -\log(CX(F_{vgg}^l(X_{TT}^{ij}), F_{vgg}^l(X_R^j))), \tag{9}$$

where l means $relu_{3\_2}$ and $relu_{4\_2}$ layers of the pretrained VGG19 model [30]. At the same time, we introduce a content contextual similarity loss between X_{TT}^{ij} and X_S^i for content preservation of the source face.

$$\mathcal{L}_{CCX} = -\log(CX(F_{vgg}^l(X_{TT}^{ij}), F_{vgg}^l(X_S^i))). \tag{10}$$

As shown in Figure 4, only with $\mathcal{L}_{CCX}$ fails to synthesize the topology of the sketch domain, while only with $\mathcal{L}_{SCX}$ transfers excessive reference styles ignoring

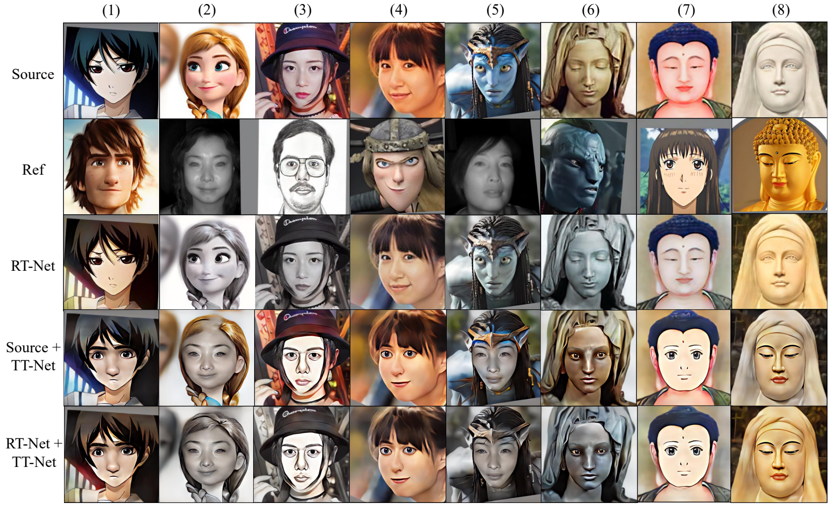

Fig. 5. Stage I and stage II of HASNet. RT-Net helps HASNet implement higher-fidelity heterogeneous rendering.

the spatial content preservation of the source face. More ablation analyses are shown in Figure 9. Therefore, TT-Net combines both $\mathcal{L}_{CCX}$ and $\mathcal{L}_{SCX}$.

The total loss of TT-Net is as follows

$$\begin{aligned}\mathcal{L}_{TT-Net} = \mathcal{L}_{adv} + \lambda_{id}\mathcal{L}_{id} + \lambda_{rec}\mathcal{L}_{rec} + \\ \lambda_{CCX}\mathcal{L}_{CCX} + \lambda_{SCX}\mathcal{L}_{SCX}.\end{aligned} \quad (11)$$

4 Experiment

4.1 Dataset

We collect three kinds of heterogeneous domains to conduct HAS task as follows:

- Lighting condition variants. We select daytime portraits with variable attributes and identities from FFHQ [15]. We collect some night portraits taken in the evening or at night. And we randomly select some samples in CASIA NIR-VIS 2.0 dataset [18].
- Art drawing variants. 3D cartoons, anime images, sketches, exaggerated drawings and oil paintings are collected from the Toonify dataset [28], selfie2 anime dataset [17], CUHK Face Sketch FERET (CUFSF) [34], IIIT-CFW [23] and MetFaces [1], respectively. Moreover, there are some sculptures of famous people (only for research). And we collect some role faces of Beijing Opera. Note that we obtain the super-resolution counterparts using [31] for the low-quality face samples.

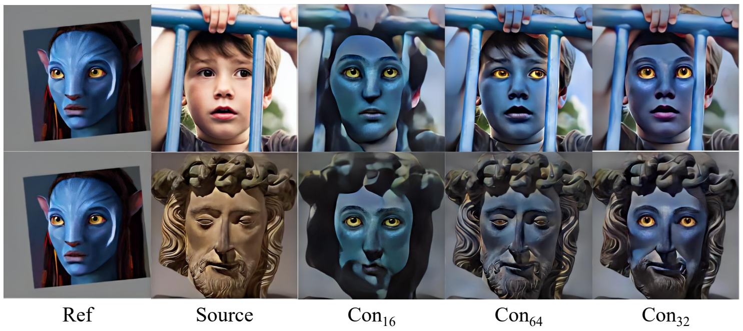

Fig. 6. Ablation study of HAS-Net concerning the dimension of initial content feature in TT-Net.

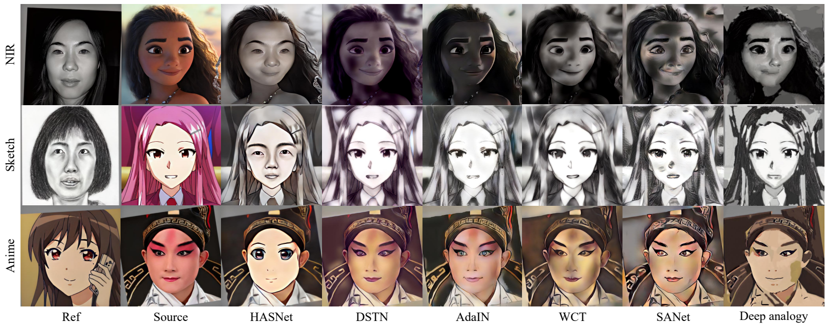

Fig. 7. Qualitative comparison with other state-of-the-art universal style transfer methods in NIR, sketch and anime style domains.

- Life dimension variants. We collect some Buddha statues images (only for research) from the Internet. We collect the faces from the movie *Avatar*. And we select some drawings of ancient people in Chinese culture.

Note that we take all kinds of heterogeneous faces, i.e. about 4000 images, as a whole training set in Stage I, where the true color palette of the source face is used to modulate RT-Net. While TT-Net takes random different heterogeneous-domain faces as the source and reference faces, respectively. There are 13 kinds of heterogeneous domains considered in the single generator of TT-Net, which is different from other GAN-based methods [28], [24], [17] that focus on certain two-domain translations.

[1] https://github.com/postite/metfaces-dataset

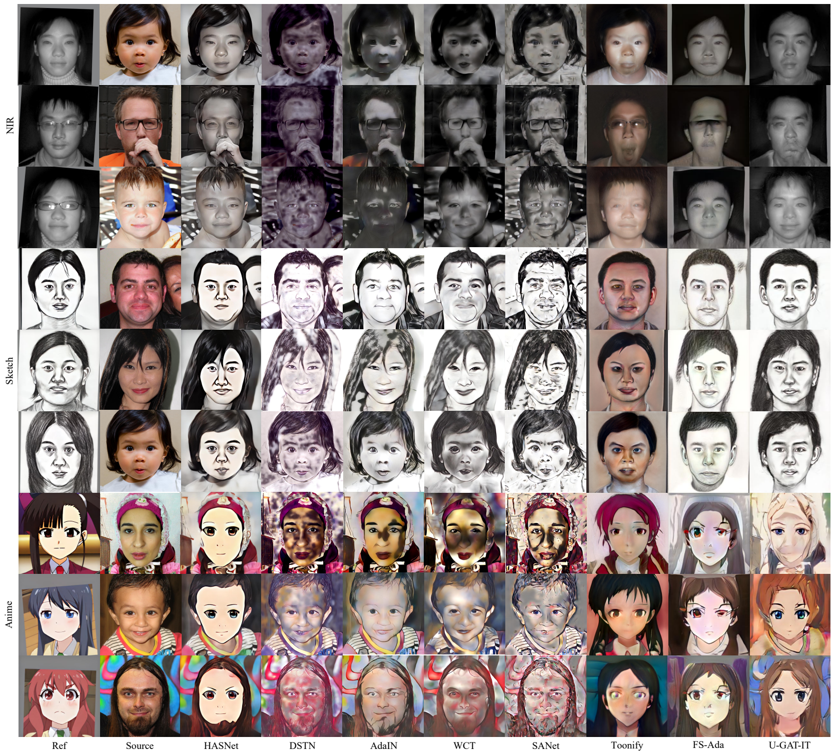

Fig. 8. Additional comparison results of HASNet and other methods. HASNet realizes high-fidelity HAS based on topology and rendering transfer, where the identity and attributes are more controllable with better background preservation.

4.2 Implementation Details

The source and target faces of our dataset are aligned and cropped based on 5 facial landmarks [31]. We reshape all samples to 512×512 resolution. We set $Con^0 \in \mathbb{R}^{512\times32\times32}$, $z_{id} \in \mathbb{R}^{512\times1}$, $z_{Palette} \in \mathbb{R}^{20\times3}$, and $z_C \in \mathbb{R}^{256\times1}$. There are 7 RTM and 10 TTM modules. In Equation 5, $\lambda_{rec} = 10$. In Equation 11, $\lambda_{adv} = \lambda_{SCX} = 1$, $\lambda_{rec} = 100$, $\lambda_{id} = 10$, and $\lambda_{CCX} = 0.5$. We set batchsize to 4. RT-Net and TT-Net are trained separately. We adapt the architecture of our discriminator networks from [1] and [15] for RT-Net and TT-Net, respectively.

In the architecture of StyleGAN [15], the coarse-resolution (4×4-32×32) layers (1-7) are used to control shape modulation. Moreover, we conduct a toy experiment and find that the dimensions of the initial content feature of X_{RT} have an important impact on the TT-Net. As shown in Figure 6, Con_{16} has more background distortion, which is not enough to conduct the high-fidelity HAS task that considers background topology preservation. Furthermore, Con_{64} provides more constraints to the source content, so that it fails to transfer the shape style of eyes in the Avatar style domain. Our HASNet uses the initial Con^0 with 32×32 resolution and shows visually satisfying HAS behavior.

4.3 Qualitative Evaluation

As shown in Figure 7, DSTN [11] synthesizes distorted and messy faces with the wrong global color style in NIR and sketch domains. AdaIN [13] achieves a good textural transfer in the global view but is not competent to transferring shape styles of eyes, nose and mouth. WCT [19] also has obvious color distortion, especially in the NIR domain. SANet [26] has some artifacts on the face area and background. Deep analogy [20] is another universal style transfer method that conducts patch matching based on vgg features, which is time-consuming and easy to cause matching errors. Furthermore, although the local component is more vivid than [11], [13], [19], [26], the eye size is still similar to that of the source faces. HASNet realizes sufficient topology and rendering transfer. RT-Net bridges the color gap between the source and reference images. And TT-Net modifies the source facial topology to the reference component distribution.

As for the GAN-based method, we compare HASNet with Toonify [28], FS-Ada [24] and U-GAT-IT [17], as shown in Figure 8. Toonify finds an optimized latent code of the source, and then feeds this code to the reference domain generator. Its results have obvious background distortions (rows 7, 9), and bad heterogeneous rendering effects (rows 1-6). And the nose of avatars are not identifiable enough compared with HASNet. FS-Ada has easily detectable artifacts concerning topology and color, especially in anime domains. U-GAT-IT has an unstable cross-domain translation performance, e.g., the topological and color artifacts on the facial area (rows 2&7), and does not respect the facial occlusions (row 2). HASNet has higher topology and rendering consistencies with the reference faces, while better preserving the background content of the source images. From the above qualitative evaluations, we find a user study is deserved to be surveyed, considering that human eyes are highly sensitive to the topology and color appearance. More details are introduced in Section 4.4.

4.4 User Study

User study plays an important role in the quality evaluation of the HAS task. First, we briefly explain the HAS task, and invite 10 users to carefully observe the source, reference and HAS faces. Each type of face has 30 samples where NIR, sketch and anime domains have 10 samples respectively. These observers need to give scores recorded as 1-10, from four aspects: (a) transferred topology perception, (b) content preservation of the source face, (c) color transfer of the reference face, and (d) overall preference, where the higher HAS quality is reflected by higher scores. Finally, we collect 300 score tables, where each table contains 52 human decisions about different methods and indexes. The average scores are displayed in Table 1.

Specifically, transferred topology perception measures the identification degree and topological completeness of the HAS results referenced to the target

Table 1. User study results of HASNet variants, as well as the universal and GAN-based style transfer methods.

Methods	Topology↑	Content↑	Color↑	Preference↑
DSTN [11]	1.2	1.9	7.8	2.7
AdaIN [13]	0.9	4.4	8.3	3.2
WCT [19]	1.8	3.7	8.4	3.3
SANet [26]	1.9	1.7	7.9	2.5
Deep analogy [26]	2.2	4.6	8.1	4.5
Toonify [28]	8.0	4.5	3.5	7.1
FS-Ada [24]	6.5	4.9	7.6	7.9
U-GAT-IT [17]	7.2	5.5	5.8	8.4
HASNet	**8.9**	**9.1**	**8.5**	**9.0**
w/o RT-Net	8.8	8.9	5.1	8.4
w/ $\mathcal{L}_{CCX}$	7.3	8.9	4.7	4.7
w/ $\mathcal{L}_{SCX}$	8.5	7.1	7.0	3.3
w/o $\mathcal{L}_{CX}$	6.9	6.5	3.1	2.0

style images. Content preservation of the source face focuses on the face pose, expression, hair and other background content. These contents are supposed to be faithful to the source image to some extent. HAS task mainly transfers the global color and facial component topology. The color index represents the color obedience guided by the color style of the reference faces, as well as the diffusion degree of the colorization. The preference score means the overall preference degree for users. HASNet has the best scores on topology, content and color aspects. Moreover, HASNet has better perception scores than Toonify and U-GAT-IT. We show more cross-domain HAS results in Figure 10, which indicates that our approach synthesizes controllable heterogeneous avatars.

4.5 Ablation Study

As shown in Figure 9, we show some diverse HAS results of variants of HASNet. Specifically, RT-Net successfully transfers the color style to the source faces based on the reference images. As shown in row 5, HASNet has higher consistency between the facial area and the background (cols 4&5). While direct TT-Net based on the original source image will result in a large color discrepancy between the results and the reference faces, which hinders high-fidelity heterogeneous avatar synthesis. w/ $\mathcal{L}_{CCX}$ maintains more spatial grayscale information of the faces of RT-Net (col 4), which stays far from the topology style of the referenced sketch face. Moreover, this variant produces low-fidelity cartoon eyes (cols 6-8). w/ $\mathcal{L}_{SCX}$ transfers excessive reference style, e.g., hair (cols 1-3), and carries uncontrollable artifacts (cols 5-8). w/o $\mathcal{L}_{CX}$ means HASNet without $\mathcal{L}_{SCX}$ and $\mathcal{L}_{CCX}$, which results in more color artifacts.

4.6 Quantitative Evaluation

Quantitative results are only the evaluation reference for performance. It is more scientific to consider both the objective score and user study. If one kind of

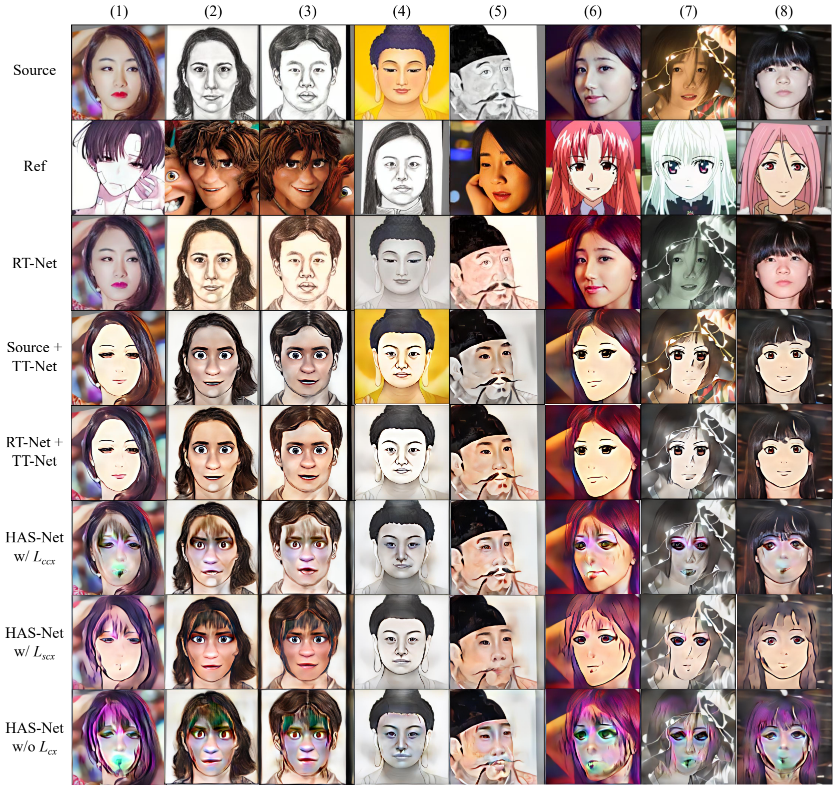

Fig. 9. Ablation study of HAS-Net. We show the results of RT-Net, Source+TT-Net, RT-Net+TT-Net (HASNet), only w/ L_{CCX}, only w/ L_{SCX} and w/o L_{CX}.

heterogeneous domain is the reference domain, we select 10 images from each of other 12 kinds of domains as the source dataset in the test stage of HAS task. We employ two evaluation metrics considering the image fidelity (FID [10], KID [3]). They are used in [17] to measure the cross-domain synthesis quality. As for the evaluation of topology transfer, we measure the identity distance of HAS results and the reference face based on Arcface. Moreover, we calculate the color distance of the first 20 prominent color values between the HAS result and the reference face based on ColorThief. As shown in Table 2, we compare HASNet with good-performance style transfer approaches including fine-tuned DSTN [11], AdaIN [13], WCT [19] and SANet [26] in three distinctive domains, i.e., NIR, sketch and Anime. Note that Deep analogy [20] needs to consume lots of inference time, so we only show the visual comparisons in Fig 7.

Concretely, [11], [13], [19], [26] have a slightly better color consistency score than HASNet. However, as for DSTN [11] in the NIR domain, the results have purple color, which has a worse perception in human eyes. This demonstrates

Table 2. Quantitative evaluation on CASIA NIR-VIS 2.0 [18], CUHK [34] and selfie2anime [17] datasets. Our model has better ID scores and comparable color scores. As for FID, due to the specificity of the HAS problem, we need to preserve the background, which has an impact on FID because FID score is calculated based on the whole image. In our opinion, a controllable HAS task is supposed to consider both the vivid avatar synthesis and the topology preservation of the background. Otherwise, the essential heterogeneous scene information of the source will be seriously lost (Fig 8).

Domains	Methods	ID × 10↓	Color × 10↓	FID ↓	KID × 100 ↓	Overall↓
NIR	DSTN [11]	9.3	0.3	185.48	**14.6**	209.68
	AdaIN [13]	9.3	0.4	212.52	19.01	241.23
	WCT [19]	9.1	**0.3**	210.72	18.54	238.66
	SANet [26]	9.2	0.5	199.25	18.17	227.12
	HASNet	4.5	0.6	195.7	18.11	218.91
	w/o RT-Net	4.4	1.2	206.89	19.41	231.9
	w/ $\mathcal{L}_{CCX}$	3.3	1.4	206.95	19.13	230.78
	w/ $\mathcal{L}_{SCX}$	2.8	0.5	**177.26**	16.28	**196.84**
	w/o $\mathcal{L}_{CX}$	**2.6**	1.2	210.72	21.11	235.63
CUHK	DSTN [11]	8.7	0.4	**162.74**	**12.3**	**184.14**
	AdaIN [13]	9	0.5	187.42	16.93	213.85
	WCT [19]	8.8	0.5	185.07	16.22	210.59
	SANet [26]	8.6	**0.4**	166.71	15.2	190.91
	HASNet	5	0.7	166.73	15	187.43
	w/o RT-Net	4.9	1.3	177.37	16.14	199.71
	w/ $\mathcal{L}_{CCX}$	3.8	1.2	177.81	15.8	198.61
	w/ $\mathcal{L}_{SCX}$	3.1	0.6	173.09	17.07	193.86
	w/o $\mathcal{L}_{CX}$	**2.8**	1.3	186.43	18.1	208.63
Anime	DSTN [11]	7.9	0.4	138.5	7.45	154.25
	AdaIN [13]	7.6	0.6	115.29	**5.56**	129.05
	WCT [19]	7.4	**0.3**	124.65	7.72	140.07
	SANet [26]	7.6	0.7	113.47	6.09	127.86
	HASNet	4.6	0.7	112.3	7.36	124.96
	w/o RT-Net	4.7	1.2	111.92	7.24	125.06
	w/ $\mathcal{L}_{CCX}$	2.5	1.2	155.46	12.11	171.27
	w/ $\mathcal{L}_{SCX}$	2.1	0.5	**101.8**	6.19	**110.59**
	w/o $\mathcal{L}_{CX}$	**1.8**	1.1	128.75	9.90	141.55

the necessity of the user study. These state-of-the-art universal style transfer methods have worse identity consistency scores than HASNet. DSTN has better FID and KID scores than that of HASNet in NIR and sketch domains, but with messy visual imaging, as shown in Figure 7. w/o RT-Net has a color score drop, as well as obvious performance degradation of FID and KID in NIR and sketch domains. w/ $\mathcal{L}_{CCX}$ and w/o $\mathcal{L}_{CX}$ cause many color artifacts (Figure 9) because of no constraint of $\mathcal{L}_{SCX}$, which get worse FID and KID scores. However, only $\mathcal{L}_{SCX}$ has lower FID and KID scores, e.g., in the anime domain, but with severe topology degradations (Figure 9).

We further evaluate the performance compared with GAN-based methods, as shown in Table 3. There are 100 real-world faces from FFHQ as the source images. Toonify and U-GAT-IT have good FID and KID scores, and our HASNet has better identity and color consistency scores.

5 Conclusion

We explore the challenging heterogeneous avatar synthesis (HAS) task considering topology and rendering transfer. RT-Net and TT-Net alleviate the color and

Table 3. Quantitative evaluation with GAN-based methods on CASIA NIR-VIS 2.0 [18], CUHK [34] and selfie2anime [17] datasets. However, a better FID score does not prove better performance in the HAS task. U-GAT-IT has better FID scores, but with visual distortions of color, shape, and background (Fig 8).

Domains	Methods	ID × 10↓	Color × 10↓	FID ↓	KID × 100 ↓
NIR	Toonify [28]	8.3	0.9	**109.73**	**7.36**
	FS-Ada [24]	8.4	1.2	191.53	19.5
	U-GAT-IT [17]	7.7	0.8	118.94	9.9
	HASNet	**4.3**	**0.7**	195.46	17.94
CUHK	Toonify [28]	7.3	1.7	105.06	8.47
	FS-Ada [24]	7.3	0.3	117.46	7.82
	U-GAT-IT [17]	7.6	**0.3**	**89.84**	**6.77**
	HASNet	**4.8**	0.8	160.61	13.64
Anime	Toonify [28]	5.7	1.3	173.3	13.65
	FS-Ada [24]	5.2	1.6	192.43	17.61
	U-GAT-IT [17]	4.7	1.3	**98.48**	**3.79**
	HASNet	**4.5**	**0.9**	125.21	8.67

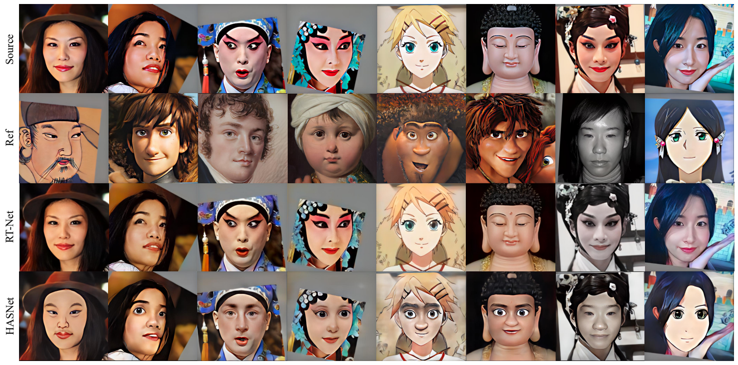

Fig. 10. More HAS results across diverse heterogeneous domains.

structural discrepancies between the HAS results and the reference faces, while preserving other source contents and attributes. Comprehensive experimental results in various heterogeneous domains show that the disentanglement of rendering and topology is beneficial to the HAS task. HASNet produces controllable and high-fidelity heterogeneous avatars.

Acknowledgments This work is supported by the National Key R&D Program of China (2019YFB1406202).

References

1. Bahng, H., Yoo, S., Cho, W., et al.: Coloring with words: Guiding image colorization through text-based palette generation. In: Proceedings of ECCV. pp. 431–447 (2018)

2. Bhatt, H.S., Bharadwaj, S., Singh, R., et al.: Memetically optimized mcwld for matching sketches with digital face images. IEEE TIFS **7**(5), 1522–1535 (2012)
3. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. In: ICLR (2018)
4. Chen, J., Yi, D., Yang, J., et al.: Learning mappings for face synthesis from near infrared to visual light images. In: 2009 IEEE Conference on CVPR. pp. 156–163. IEEE (2009)
5. Chen, R., Huang, W., Huang, B., et al.: Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In: Proceedings of CVPR. pp. 8168–8177 (2020)
6. Deng, J., Guo, J., Xue, N., et al.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of CVPR. pp. 4690–4699 (2019)
7. Duan, B., Fu, C., Li, Y., et al.: Cross-spectral face hallucination via disentangling independent factors. In: Proceedings of CVPR. pp. 7930–7938 (2020)
8. Fu, C., Wu, X., Hu, Y., et al.: Dual variational generation for low shot heterogeneous face recognition. In: NIPS (2019)
9. Fu, C., Wu, X., Hu, Y., et al.: Dvg-face: Dual variational generation for heterogeneous face recognition. IEEE Transactions on PAMI (2021)
10. Heusel, M., Ramsauer, H., Unterthiner, T., et al.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NIPS (2017)
11. Hong, K., Jeon, S., Yang, H., et al.: Domain-aware universal style transfer. In: Proceedings of ICCV. pp. 14609–14617 (2021)
12. Huang, D., Sun, J., Wang, Y.: The buaa-visnir face database instructions. School Comput. Sci. Eng., Beihang Univ., Beijing, China, Tech. Rep. IRIP-TR-12-FR-001 **3** (2012)
13. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of ICCV. pp. 1501–1510 (2017)
14. Karras, T., Aittala, M., Laine, S., et al.: Alias-free generative adversarial networks. In: NIPS (2021)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of CVPR. pp. 4401–4410 (2019)
16. Karras, T., Laine, S., Aittala, M., et al.: Analyzing and improving the image quality of stylegan. In: Proceedings of CVPR. pp. 8110–8119 (2020)
17. Kim, J., Kim, M., Kang, H., et al.: U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In: ICLR (2020)
18. Li, S., Yi, D., Lei, Z., et al.: The casia nir-vis 2.0 face database. In: Proceedings of the IEEE conference on CVPR workshops. pp. 348–353 (2013)
19. Li, Y., Fang, C., Yang, et al.: Universal style transfer via feature transforms. In: NIPS. pp. 386–396 (2017)
20. Liao, J., Yao, Y., Yuan, L., et al.: Visual attribute transfer through deep image analogy. SIGGRAPH (2017)
21. Mechrez, R., Talmi, I., Zelnik-Manor, L.: The contextual loss for image transformation with non-aligned data. In: Proceedings of ECCV. pp. 768–783 (2018)
22. Messer, K., Matas, J., Kittler, J., et al.: Xm2vtsdb: The extended m2vts database. In: Second international conference on audio and video-based biometric person authentication. vol. 964, pp. 965–966. Citeseer (1999)
23. Mishra, A., Rai, S.N., Mishra, A., et al.: Iiit-cfw: A benchmark database of cartoon faces in the wild. In: ECCV. pp. 35–47. Springer (2016)
24. Ojha, U., Li, Y., Lu, J., et al.: Few-shot image generation via cross-domain correspondence. In: Proceedings of CVPR. pp. 10743–10752 (2021)

25. Panetta, K., Wan, Q., Agaian, S., et al.: A comprehensive database for benchmarking imaging systems. IEEE transactions on PAMI **42**(3), 509–520 (2018)
26. Park, D.Y., Lee, K.H.: Arbitrary style transfer with style-attentional networks. In: Proceedings of CVPR. pp. 5880–5888 (2019)
27. Park, T., Liu, M.Y., Wang, T.C., et al.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of CVPR. pp. 2337–2346 (2019)
28. Pinkney, J.N., Adler, D.: Resolution dependent gan interpolation for controllable image synthesis between domains. arXiv preprint arXiv:2010.05334 (2020)
29. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
30. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: 3rd ICLR (2015)
31. Wang, X., Li, Y., Zhang, H., et al.: Towards real-world blind face restoration with generative facial prior. In: CVPR (2021)
32. Yoo, S., Bahng, H., Chung, S., et al.: Coloring with limited data: Few-shot colorization via memory augmented networks. In: Proceedings of CVPR. pp. 11283–11292 (2019)
33. Zhang, M., Wang, R., Gao, X., et al.: Dual-transfer face sketch–photo synthesis. IEEE Transactions on Image Processing **28**(2), 642–657 (2018)
34. Zhang, W., Wang, X., Tang, X.: Coupled information-theoretic encoding for face photo-sketch recognition. In: CVPR 2011. pp. 513–520. IEEE (2011)