Enhancing the Realism of Sketch and Painted Portraits with Adaptable Patches

Supplementary Document

1. Alignment of face, neck and hair regions

In the preprocessing step, we apply the Stacked Trimmed Active Shape Model (STASM) [MN08] to finding 61 facial feature points. The positions of these points can be manually adjusted to avoid failure results. Based on these points, we construct the continuous contours along feature points with Catmull-Rom splines, and we sample additional feature points along splines for warping. Besides facial feature points, 6 neck feature points and 24 hair feature points are marked by users. These assignment processes can be replaced by other advanced contour extraction methods. Sup.Fig.1(b) and Sup.Fig.2(a) show the feature points of face, neck and hair regions.

To align a face from photo sets with an input face, the perspective projection is first used to warp the photo data. The step globally approximates the projective transformation between two feature point sets. We further apply 2D local warping to amend the local differences between pairs of feature points.

1.1. Perspective projection

A source image (photo) and a target image (input painting) are denoted by I and I', respectively. A feature point i in I is denoted by $p_i=(u_i,v_i)^T$ and its destination position in I' is denoted by $q_i=(x_i,y_i)^T$, $i=1,\ldots,n$. The goal of perspective warping is to find a transformation $f_{per}(p_i)= q_i'$, such that q_i' is close to q_i.

The transformation can be represented in terms of a linear system by a projection matrix A:

$$\begin{bmatrix} wx \\ wy \\ w \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}, \tag{1}$$

Each point correspondence forms two equations:

$$\begin{aligned} x(A_{31}u + A_{32}v + A_{33}) &= A_{11}u + A_{12}v + A_{13} \\ y(A_{31}u + A_{32}v + A_{33}) &= A_{21}u + A_{22}v + A_{23} \end{aligned} \tag{2}$$

A projective mapping has 8 degrees of freedom and it needs at least 4 correspondences to find a deterministic solution. After rearrangement, the linear system with n point correspondences can be written as follows:

$$\begin{bmatrix} u_1 & v_1 & 1 & 0 & 0 & 0 & -x_1u_1 & -x_1v_1 & -x_1 \\ 0 & 0 & 0 & u_1 & v_1 & 1 & -y_1u_1 & -y_1v_1 & -y_1 \\ u_2 & v_2 & 1 & 0 & 0 & 0 & -x_2u_2 & -x_2v_2 & -x_2 \\ 0 & 0 & 0 & u_2 & v_2 & 1 & -y_2u_2 & -y_2v_2 & -y_2 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ u_n & v_n & 1 & 0 & 0 & 0 & -x_nu_n & -x_nv_n & -x_n \\ 0 & 0 & 0 & u_n & v_n & 1 & -y_nu_n & -y_nv_n & -y_n \end{bmatrix} \begin{bmatrix} A_{11} \\ A_{12} \\ A_{13} \\ A_{21} \\ A_{22} \\ A_{23} \\ A_{31} \\ A_{32} \\ A_{33} \end{bmatrix} = 0, \tag{3}$$

It is an overdetermined system and can be regarded as a least-square problem. We find the optimal projection matrix A through Singular Value Decomposition (SVD). Hence, for a pixel p, we can find its warped position $q'=f_{per}(p)$.

1.2. Local field warping

For a feature point p_i in the source image, its perspective projected position is q_i', but the destination of p_i is q_i. There is an offset ($q_i - q_i'$). The concept of the local warping is similar to scattering sparse data. For a warped point q', if its original position p is close to a feature point p_i, its amending offset should be close to ($q_i - q_i'$). There are various

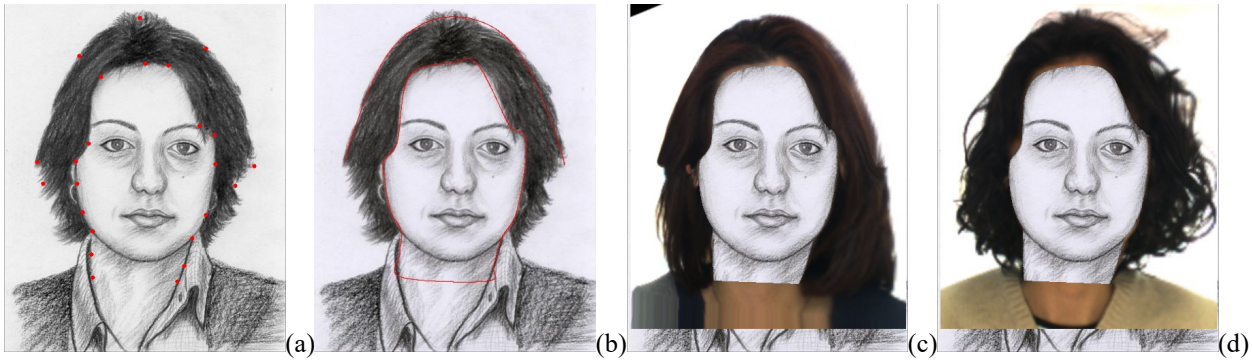
methods to generate a smooth offset field. In our current system, we choose weighted combination of amending offsets of feature points. The equation is as follows:

$$f_{loc}(q') = q' + \sum_{i=1}^{n} \eta_i (q_i - q_i') \tag{4}$$

$$\hat{\eta}_j = \frac{1}{|p - p_i|^2 + c} \text{ and } \eta_i = \frac{\hat{\eta}_i}{\sum_{h=1}^{n} \hat{\eta}_h} , \tag{5}$$

where c is a small constant to prevent zero division, and η_i is the weight about the influence of feature point i. The warping method can be approximated and speeded up by using mesh-based interpolation. It can be replaced by other approaches, such as the work by Beier and Neely, Feature-based image metamorphosis in *Proc. SIGGRAPH'92* or the

Sup. Figure 1: *Feature points and alignment of faces and necks. (a) A training photo in AR face database. (b) 61 face and 6 neck feature points for the face in (a). (c) An input sketch. (d) The face and neck feature points of (c). (e) Aligning the face in (a) with feature points in (d).*



Sup. Figure 2: *Feature points and alignment of hair. (a) 24 hair feature points. (b) The hair contour estimated by Catmull-Rom spline. (c) An example of aligned hair. (d) Another example of aligned hair.*

work by Ruprecht and Müller, Image Warping with Scattered Data Interpolation in *IEEE CG&A 1995*.

2. Voting Scores in User Evaluations

As mentioned in 6.4, 24 volunteers participated in 18 sets of test data. For each face in a set, a volunteer evaluated the similarity score, realism score, and overall preference vote. In this subsection, we listed the average scores of each method in these 18 test sets.

Average similarity scores

	The proposed	*Regular-patch*	*Best-match*
1	7.8	5.5	4.5
2	7.1	5.4	7.7
3	7.5	6.3	4.6
4	7.1	6.2	6.5
5	7.6	6.4	6
6	5.5	4.7	6.4
7	7.6	5.5	4.9
8	6.8	6.3	6
9	7.8	5.3	4.1
10	7.9	5.5	4.1
11	7.3	5.7	7
12	6.8	7.4	6.2
13	7.4	5.8	5.5
14	6.8	6.6	5.5
15	6.4	6.5	6.8
16	7.1	7.3	7.2
17	8.3	7	6
18	6.8	5.9	5

Average realism scores

The proposed	*Regular-patch*	*Best-match*
7.7	6	6.8
7.1	5.2	7.6
7.6	6.2	5.1
7.1	6.5	8.3
7.1	6.8	7.1
5.8	5.6	6.8
7.7	5.7	6.1
7.5	6.1	7.3
7	6.5	6
7.6	5.1	5
6.5	4.8	7
7	7.4	7.1
7.4	6	5.9
7.2	6.4	7.3
5.8	6	7.6
7.1	7.1	7.8
7.6	7.6	7.7
7.9	5.9	5.9

3. Additional experiments with a sketch-to-photo method

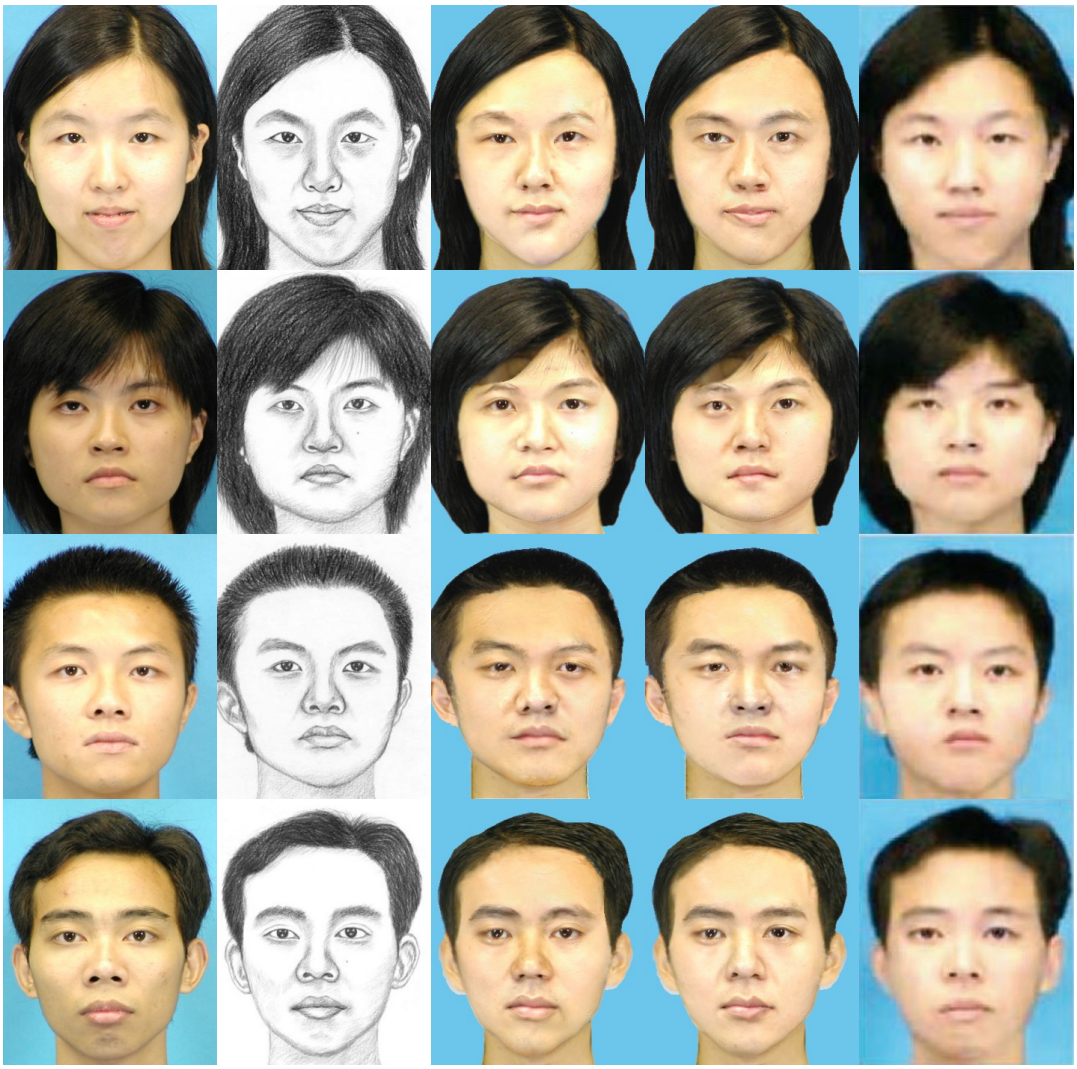
To evaluate our method for different kinds of targets and to compare our method with the state-of-the-art sketch-to-photo method, we conducted additional experiments with two public and free datasets used in [WT09].

The first dataset is CUHK Face Sketch (CUFS) database. There are 88 training faces (54 males and 34 females) selected by Wang and Tang [WT09]. For each training face, there are a frontal sketch drawn by an artist and the corresponding photo. The sketch-to-photo method matched the patches of the input sketch with those of training sketches. After the best sketch patches from the training set were found, the corresponding photo patches were applied to photo synthesis. By contrast, our method used the photo data only. We compared the four target faces demonstrated in [WT09], as shown in Sup.Fig.3(b). During the experiment, we found that these four faces were also included in the training list.

Hence, we temporarily excluded the input face from the training data when realizing that face.

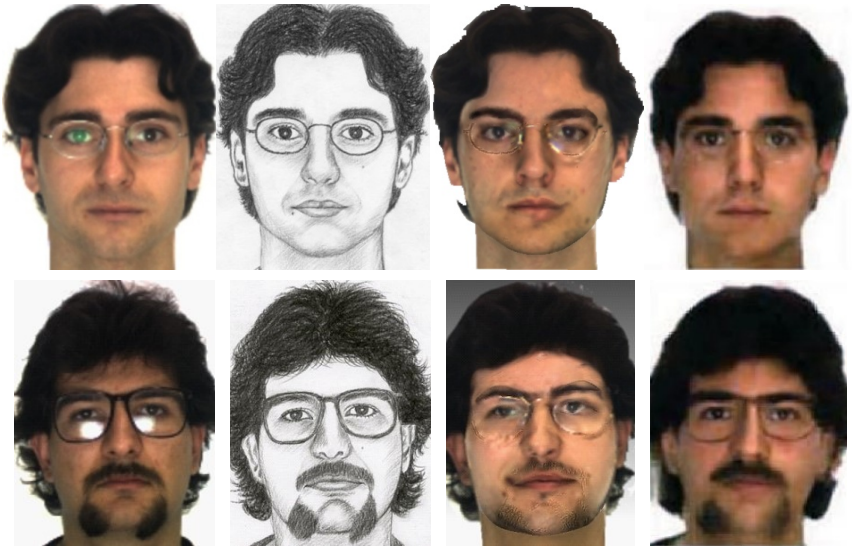
In our first trial, we noticed that when we synthesized a female sketch, the result might look masculine due to the imbalance numbers of male and female training faces, as shown in Sup.Fig.3(d). That is because male faces accounted for up to 60 percent of the training data, so there were more chances to select a block from male faces during histogram-based selection. Therefore, we separated the training data into male and female sets and enhanced a facial sketch with only the training data in the corresponding gender. As shown in Sup.Fig.3(c), the results with gender-dependent training data are more plausible than results with training data of both genders. Sup.Fig.3(a) shows the ground truth photos, and the results from [WT09] are shown in Sup.Fig.3(e).

The second dataset, AR Face dataset [MB98], comprises 126 pairs of facial photos and sketches. Different from our collected female photos or CUFS dataset, there are several faces with glasses, beards or grins. The intensities and edges of these additional features may be inconsistent between photos and sketches. For instance, as shown in the top row of Sup.Fig.4(a) and (b), artists used dark strokes to represent the light silver glasses and the reflection on glasses were



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

Sup. Figure 3: *Experiments with the CUFS dataset. (a) Ground truth photos. (b) Input sketches. (c)Results of the proposed method with training photos of the corresponding gender. (d)Results of the proposed method with training photos of both genders. (e)Results in* [WT09], where sketch-photo data were used for training.



(a) (b) (c) (d)

Sup. Figure 4: *Experiments for targets with glasses or beards in the AR Face dataset. (a) Ground truth photos. (b) Input sketches. (c)Results of the proposed method using training photos with labels, top: {male, with glasses}, bottom:{male, with glasses, with a beard}. (d)Results in* [WT09]*, where sketch-photo data were used for training.*

omitted. Unlike [WT09], we do not have the correspondences between sketches and photos. Our histogram-based block selection may not select blocks with glasses due to the large histogram differences. To avoid discarding adequate blocks in a too early step, we assigned three special labels: *with/without glasses*, *with/without a beard*, *with/without a grin*, to each training face. Given an input face, only the training faces with identical labels were included in the collected photo data. We compared results using AR face dataset shown in [WT09]. As shown in Sup.Fig.4, with the glasses label, the proposed patch formation can automatically synthesize eye glasses without any training sketch. However, the beard and mouth of the resulting face in the bottom of Sup.Fig.4 were relatively unsatisfactory. That was due to the cluttered Canny edges and unexpected block selection around the beard. The results of [WT09] are shown in Sup.Fig.4(d), in which slight block effects are visible. That can be due to the low image quality of their published document or it may result from the use of rectangular patches.

Sup.Fig.5 shows results of two female targets. Similar to the cases in CUFS dataset, female results with training photos of both genders show certain masculine characteristics, such as stubble. Noteworthily, as shown in Sup.Fig.5 (d), the stubble on the chin looks similar to the chin shadow in the sketch. We think this issue is beyond the appearance-based synthesis and requires a higher-level label for distinction.

In this supplementary section, we faithfully demonstrated results synthesized by the proposed method for CUFS and AR datasets. Our synthesis method relies on measures in appearance. It may not distinguish the semantics or implication of features on a face. Hence, we can find our system using stubble to imitate the chin shadow. To overcome this problem, we assigned each face a characteristic label set and synthesized a face with photos of the corresponding category. We found that even without sketch training data, parts of our results could be comparable to results of the state-of-the-art sketch-to-photo method. Some other cases may be less satisfactory. We think the success or unexpected outcomes illustrate the strengths and limitations of the proposed appearance-based method. They can point out future work of face realization research.

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

Sup. Figure 5: *Experiments with the AR Face dataset. (a) Ground truth photos. (b) Input sketches. (c)Results of the proposed method using training photos with label {female, without glasses} . (d) Results of the proposed method using training photos with labels {both genders, without glasses, without a beard}. (e)Results in* [WT09]*, where sketch-photo data were used for training.*