2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

# Human-Robot Interactive Creation of Artistic Portrait Drawings

Fei Gao, Lingna Dai, Jingjie Zhu, Mei Du, Yiyuan Zhang, Maoying Qiao, Chenghao Xia, Nannan Wang, and Peng Li †

***Abstract*— In this paper, we present a novel system for *Human-Robot Interactive Creation of Artworks* (HRICA). Different from previous robot painters, HRICA allows a human user and a robot to alternately draw strokes on a canvas, to collaboratively create a portrait drawing through frequent interactions. The key is to enable the robot to understand human intentions, during the interactive creation process. We here formulate this as a mask-free image inpainting problem, and propose a novel method to estimate the complete version of a portrait drawing, after the human user has drawn some initial strokes. In this way, the robot can select some complementary strokes and draw them on the canvas. To train and evaluate our inpainting method, we construct a novel large-scale portrait drawing dataset, CelebLine, which composes of high-quality portrait line-drawings, with dense labels of both 2D semantic parsing masks and 3D depth maps. Finally, we develop a human-robot interactive drawing system with low-cost hardware, user-friendly interface, and interesting creation experience. Experiments show that our robot can stably cooperate with human users to create diverse styles of portrait drawings. In addition, our portrait drawing inpainting method significantly outperforms previous advanced methods. The code and dataset have been released at: https://github.com/fei-aiart/HRICA.**

## I. Introduction

Making robots create artworks like human is an interesting but challenging task. In recent years, inspired by the extraordinary success of deep learning based generative Artificial Intelligence (AI) [1], [2], there have been great progress in artistic robots. For now, robots are capable of creating watercolours [3], oil paintings [4]–[6], and line drawings [7], [8], conditioned on a photo or language descriptions provided by a human user, or even automatically [9]. However, during these collaborations, the human and the robot interact only once essentially. Specifically, a human first provides some inputs or requirements. The robot then conditionally generate an image and draw it on a canvas.

*This work was supported in part by the National Natural Science Foundation of China under Grants U22A2096 and 61971172, in part by the Technology Innovation Leading Program of Shaanxi under Grant 2022QFY01-15, in part by the Fundamental Research Funds for the Central Universities under GrantQTZX23042; in part by the Nanjing Science and Technology Plan under Grants Y23002ZX01.

Fei Gao is with the Hangzhou Institute of Technology, Xidian University, Hangzhou 311231, China. Lingna Dai, Jingjie Zhu, and Yiyuan Zhang are with AiSektcher Technology, Hangzhou 311200, China. Mei Du is with Hangzhou Dianzi University, Hangzhou 310018, China. Maoying Qiao is with The University of Technology, Sydney, Australia. Chenghao Xia is with The University of Sydney, Australia. Nannan Wang is with the ISN State Key Laboratry, Xidian University, Xian 710126, China. Peng Li is with the Institute of Software, Chinese Academy of Sciences, Beijing 100190, China, and also with University of Chinese Academy of Sciences, Nanjing/Beijing/Hangzhou and Nanjing Institute of Software Technology. († Corresponding author: lipeng@iscas.ac.cn)

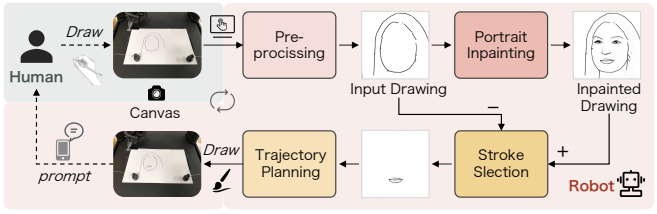


Fig. 1. The proposed Human-Robot Interactive Creation of Artworks (HRICA) framework.

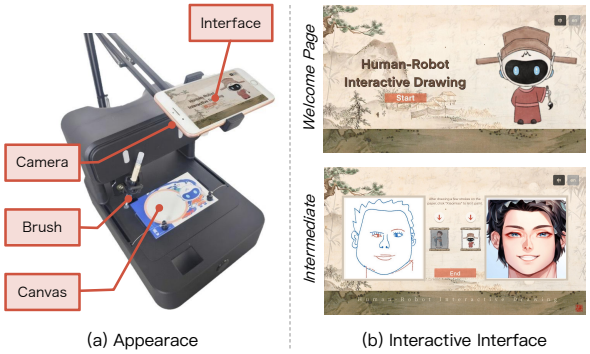


Fig. 2. Appearance and interactive interface of our robot.

In this work, we explore how to enable a robot to collaboratively create an artwork with a human, through multiple alternate interactions. To this end, we propose a novel *Human-Robot Interactive Creation of Artworks* (HRICA) framework (as shown in Fig. 1). At the beginning, a human first draws some strokes on the canvas and commands the robot to draw next. The robot then automatically draw complementary strokes on the canvas. After several rounds of such interactions, a portrait drawing is collaboratively created by the human and the robot. Recall that the creation procedure of colorful paintings, e.g. watercolours and oil-paintings, is time-consuming and demands fastidious attention for novices. We instead focus on portrait line-drawings. It's possible for the human and the robot co-create a vivid portrait drawing through only a small amount of interactions.

The key challenge of such a human-robot interactive creation system, is to enable the robot to understand human creating intentions. We re-format this challenge as inpainting an input portrait drawing to a complete version, with accurate facial geometry. To solve the problem, we propose a novel *Geometry-Aware Portrait Drawing Inpainting* (GAPDI) method. Unlike previous image inpainting methods [10], [11], we first predict the potential facial semantic geometry of an input portrait drawing, and then generate a complete drawing in a coarse-to-fine manner. In addition,

we constrain the generated drawing to be consistent with the target one, in terms of both 2D semantic structure and 3D geometry. As a result, the inpainted portrait drawing is expected to accurately present the geometry of the face that the human intends to create. Furthermore, the GAPDI method enables the robot collaborate with the human, with dynamic planning, during each interaction.

To learn the GAPDI model, we construct a novel large-scale portrait drawing dataset, by transferring the facial photos in CelebAMask-HQ [12] to free-hand line-drawings [13] via AiSketcher [8] and Simplify [14]. In addition, we represent the 2D and 3D geometry of a portrait drawing, by using the semantic parsing masks (available in CelebAMask-HQ) and the depth map (estimated by [15]) of the corresponding photo, respectively. Our final dataset, named `CelebLine`, composes of 30,000 high-quality line-drawings, with dense labels of semantic parsing masks and depth maps.

Finally, we develop a human-robot interactive drawing system, with low-cost hardware and user-friendly interactive interface (Fig. 2). To enrich the fun of the interactive creation process, we transfer the co-created drawing to a colourful avatar via a pretrained ControlNet [16]. The avatar is simultaneously shown beside the portrait drawing on the interface (Fig. 2b). Our system is easy to use, and a portrait drawing can be co-created in minutes.

Our main contributions are summarized as follows:

- **Framework.** We propose a novel HRICA framework for human-robot interactive creation of artworks, with alternate and frequent interactions.
- **Dataset.** We construct a novel CelebLine dataset, which composes of 30,000 high-quality portrait line-drawings, with labels of semantic parsing masks and depth maps. We hope CelebLine will serve as a benchmark for downsteam visual analysis tasks.
- **Method.** We propose a novel portrait drawing inpainting method, GAPDI, to enable the robot to understand human creating intentions. Experiments show that GAPDI can precisely complete a portrait drawing, and significantly outperforms existing advanced methods.
- **System.** We develop a human-robot interactive drawing system, with low-cost hardware, user-friendly interface, fluent interactive creation process, and rich fun.

## II. Related Works

**Sketching Assistants.** Inspired by the great success in generative AI, there have been diverse creation or sketching assistants. These methods mainly follow the pipeline of multimodal image synthesis and editing [1]. Namely, a human can control the generated image through images [17]–[20], semantic masks [21], languages [5], [6], [22], etc. Some specially designed sketching assistants, try to modify the strokes drawn by human [11], [23], or provide drawing guidance to the human [24]–[26]. Differently, we aim to exactly preserve what a human subjectively creates, and enable the robot to draw complementary strokes.

**Artistic Portrait Drawing Generation.** *Artistic Portrait Drawing Generation* (APDG) has attracted numerous interests. Existing APDG methods focuses on translating a facial photo to an artistic style, such as pen-drawings [27], pencil-sketches [28]–[31], line-drawings [8], [32], and oil-paintings [20], [33], [34]. Existing methods to achieve this follow either the architecture of Generative Adversarial Networks (GANs) [27], [31], [32], [35], Neural Style Transfer (NST) [8], [36], or Diffusion models [20], [37]. However, a high-quality facial line-drawing dataset is lack. Existing line-drawing datasets focus on natural objects [38], [39]. To approach these challenges, in this work, we construct a novel portrait line-drawing dataset.

**Image Inpainting.** Image inpainting (or completion) has been a longstanding task in Computer Vision (CV). Existing work mainly focus on photos and rarely on portrait drawing yet [40]–[43]. Techniques to improve the quality of inpainted images include partial convolution [44], semantic modulation [41], context information [10], etc. Although these methods have achieved satisfactory results, they require the mask of missing area in the inference stage. However, this mask is unavailable in portrait drawing inpainting. There have been several sketch completion methods [11], [45], [46]. All these methods adopt stacked networks to inpaint an input sketch from coarse to fine, and focus on objects or animals, instead of human faces. In addition, they tend to fail [11], [46], or generate images [45], when too many strokes are missing. In contrast, we propose a novel geometry-aware sketch inpainting method, which can successfully predict the target portrait drawing conditioned on few strokes.

**Facial Geometry Estimation.** Face geometry estimation from facial photos, including semantic parsing [47]–[49] and 3D pose estimation, have been widely investigated in the past decades. The most successful methods are deep learning based, and explore aspects including multiscale features [50], context correlations [51]–[53], boundary constraints [47], [49], synthetic data [48], etc. However, these methods do not work for other styles of facial images, such as line-drawings or oil-paintings. In addition, the facial 3D pose estimation methods usually predict the 3D information of the facial area, without the regions of hair, neck, etc. Currently, several methods have been proposed for the semantic segmentation [54]–[57] and depth estimation [58] of sketches about natural scenes or objects. There has been no method for semantic parsing or 3D geometry of facial sketches.

## III. Dataset

Given a facial photo, previous works use the edges extracted by Canny operator, semantic boundaries [33], [34], or HED [33], [59], as portrait drawings. Such drawings are unappealing, and differ dramatically from those drawn by a human. Recently, Gao et al. [8] propose a novel line-drawing generation method, AiSketcher, based on neural style transfer. The resulting drawings present an appealing appearance, with detailed facial structures. We thus construct the portrait drawing dataset by using AiSketcher. In addition, to mimic the line strokes drawn by human, we further transfer the images generated by AiSketcher to binary line-drawings with smooth strokes, by using the Simplify method

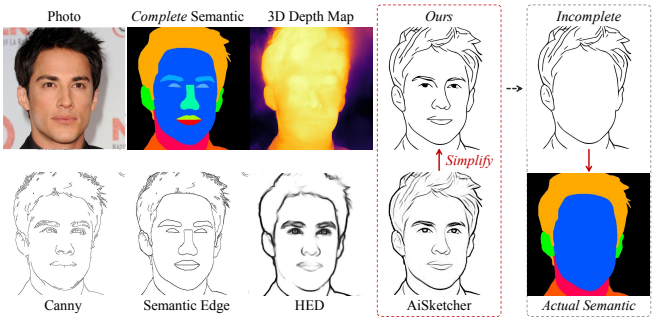


Fig. 3. Illustration of different types of artistic portrait drawings, and the incomplete portrait drawing with actual semantic masks (last column).

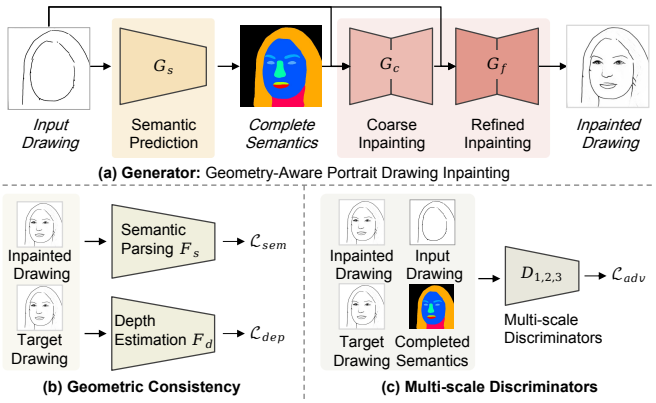


Fig. 4. Pipeline of the proposed *Geometry-Aware Portrait Drawing Inpainting* (GAPDI) method.

[14]. As shown in Fig. 3, the final line-drawings vividly present the corresponding photographic face, with a sparse set of smooth strokes. Moreover, our final line drawing shows distinct superiority over previous methods.

To construct a large scale portrait drawing dataset, we select 30,000 high-resolution facial photos from the CelebAMask-HQ dataset, and transfer them to line-drawings via AiSketcher [8] and Simplify [14]. It is notable that, the semantic parsing masks of each facial photo is available in the CelebAMask-HQ dataset. In addition, we use the LeRas-based model [15], [60] to estimate the depth map of each photo, to preset the 3D structure. Note that both AiSketcher and Simplify cause little or no geometric deformation. Thus, such semantic masks and depth maps can be used as dense labels of portrait drawings. As a result, each sample becomes a quartet, i.e., {portrait drawing $x$, photo $y$, semantic parsing masks $S$, depth map $D$}.

Following the original partition of CelebAMask-HQ, we divide the whole dataset into training, validation, and testing subsets, with the ratio of about 8:1:1. We refer to our portrait drawing dataset as `CelebLine`, and will release upon publish. As far as we are aware, this is the first portrait drawing dataset with such dense labels. Our dataset can be used for not only the generation of portrait drawings, but also semantic analysis or depth estimation.

## IV. Method

### A. Overview

In our setting, a human and a robot take turns to draw strokes on a canvas, and finally create a portrait drawing after several interactive iterations. During one iteration, formally, we denote the version of drawing with $x_h$ after human finish their creation. Then, the robot estimates the *complete* version of this drawing, $x$, and draw some complementary strokes on the canvas, obtaining the complemented drawing, $x_r$. Thus, it is crucial to complete the initial drawing, while preserving existing strokes and presenting reasonable facial structures.

To this end, we develop a novel *Geometry-Aware Portrait Drawing Inpainting* (GAPDI) method. Our method follows the architecture of GANs [61], [62]. We here propose a geometry-aware generator to inpaint the input drawing from coarse to fine. In addition, we constrain the generated drawings to be consistent with the target ones, in terms of both 2D semantics and 3D depths. Finally, we use multiscale discriminators to constrain the generated sketches to be natural and high quality. Details are presented below.

### B. Geometry-Aware Generator

As shown in Fig. 4(a), our generator $G$ composes of a semantic predictor, $G_s$, and two stacked inpainting networks, $G_c$ and $G_f$. Given a portrait drawing $x$, $G_s$ estimates the potentially *complete semantic masks* $G_s(x)$. Afterward, $G_c$ and $G_f$ generates the complete drawing in a coarse-to-fine manner, conditioned on both $x$ and $G_s(x)$.

**Complete Semantic Prediction.** To reasonably inpaint an incomplete drawing, we first estimate the potential geometry of the final portrait. In other words, our semantic predictor aims to estimate a complete semantic masks from an incomplete drawing. In the implementation, $G_s$ follows the architecture of PoolFormer [63], due to its inspiring performance in various tasks. The predicted *complete* semantic masks are:

$$\hat{S} = G_s(x_h). \tag{1}$$

During the training stage, $G_s$ is optimized by minimizing the L2 distance between the predicted semantic masks $G_s(x_h)$ and the corresponding target $s$, i.e.

$$\mathcal{L}_{G_s} = ||\hat{S} - S||_2^2 = ||G_s(x_h) - S||_2^2. \tag{2}$$

**Stacked Portrait Drawing Inpainting.** The architectures of our coarse inpainting network $G_c$ and refined inpainting network $G_f$ are similar to Pix2PixHD [62]. Specially, the architecture of $G_c$ is the same as Pix2PixHD, and composes of two encoding layers, two decoding layers, and four residual blocks. Differently, $G_f$ has four encoding layers, four decoding layers, and nine residual blocks. Here, an encoding/decoding layer consists of a downsampling/upsampling convolutional layer, followed by batch normalization (BN) and a ReLU activation layer. The stacked portrait drawing inpainting procedure is formulated as:

$$\hat{x}_f = G_f(\hat{x}_c, x_h, \hat{S}), \quad \text{with} \quad \hat{x}_c = G_c(x_h, \hat{S}). \tag{3}$$

### C. Geometric Consistency

A generated portrait drawing should accurately present the target facial structure. We therefore constrain the generated drawings to be consistent with the target ones, in terms of

both 2D semantics and 3D depths. As shown in Fig.4b, we use a semantic parsing network $F_s$ and a depth estimation network $F_d$ for computing geometric consistency losses.

**2D Semantic Consistency.** The face semantic parsing model $F_s$ is to estimate the *actual* semantic masks of a given portrait drawing. This is different from the previous semantic predictor $G_s$. Given an incomplete drawing $x_h$, $F_s$ is expected to predict the corresponding semantic masks, $S_h$, which precisely represent what have drawn in the input drawing. Instead, $G_s$ aims to estimate the complete semantic masks $s$ from what have drawn, to further estimate the complete portrait drawing. Fig. 3 illustrates differences between complete and actual semantic masks. The architecture of $F_s$ also follows PoolFormer [63]. $F_s$ is optimized by minimizing the semantic reconstruction loss, i.e.

$$\mathcal{L}_{F_s} = ||F_s(x_h) - S_h||_2^2. \tag{4}$$

Afterward, the learned $F_s$ is fixed and used for calculating the 2D semantic consistency loss, i.e.

$$\mathcal{L}_{sem} = ||F_s(\hat{x}_c) - S||_2^2 + ||F_s(\hat{x}_f) - S||_2^2, \tag{5}$$

to optimize the generator. In this way, the inpainted portrait drawings, $\hat{x}_c$ and $\hat{x}_f$, would present the target 2D structure.

**3D Structural Consistency.** In addition, a portrait drawing should correctly convey the corresponding 3D facial structure. We thus constrain to reconstruct the input depth map from the inpainted drawings. To this end, we first pretrain a network, $F_d$, for estimating the depth map of a portrait drawing. In the implementation, the architecture of $F_d$ is the same as Pix2PixHD [62]. $F_d$ is learned from pairs of complete drawings and depth maps, $\{x, D\}$, and is optimized by minimizing the depth reconstruction loss, i.e.

$$\mathcal{L}_{F_d} = ||F_d(x) - D||_1. \tag{6}$$

Afterward, the learned $F_d$ is fixed and used for calculating the depth consistency loss:

$$\mathcal{L}_{dep} = ||F_d(\hat{x}_c) - D||_1 + ||F_d(\hat{x}_f) - D||_1. \tag{7}$$

### D. Multiscale Discriminators

Following Pix2PixHD [62], we use multiscale discriminators $D_{1,2,3}$ to boost the quality of generated sketches. These discriminators have the same network structure but operate at different image scales. They downsample the images and label masks by a factor of 2 and 4 (denoted by $\downarrow_2$ and $\downarrow_4$ in Fig. 4c) to create pyramids of 3 scales. These discriminators enforce the generated portrait drawings to be consistent with the semantic masks, as well as the 3D structures (i.e. depth maps), at different scales.

### E. Training Details

To train our models, we build *pseudo* paired training data based on our CelebLine dataset. Specially, given a sample $\{x, y, S, D\}$, we generate an incomplete portrait drawing as $x_h$ by randomly removing some strokes from the original complete drawing $x$. The corresponding *incomplete* parsing mask $S_h$ is obtained by changing the corresponding labels in $S$ to the category indices of skin or background. In addition, we use a binary mask $M$ to denote the incomplete regions. In this way, we obtain a training sample $\{x_h, S_h, M, x, S, D\}$. Fig. 3 illustrates some complete and incomplete portrait drawings, as well as the corresponding semantic masks.

**Loss Functions.** In addition to previous geometric consistency losses, $\mathcal{L}_{sem}$ and $\mathcal{L}_{dep}$, we also use a weighted stroke reconstruction loss for optimizing the generator. Specially, we use a weighted L1 distance between the generated drawing and the target drawing, as the reconstruction loss:

$$\begin{aligned}\mathcal{L}_{rec} = \gamma_1||(x - \hat{x}_c) \odot M||_1 + \gamma_2||(x - \hat{x}_c) \odot \bar{M}||_1 \\ + \gamma_1||(x - \hat{x}_f) \odot M||_1 + \gamma_2||(x - \hat{x}_f) \odot \bar{M}||_1,\end{aligned} \tag{8}$$

where $\odot$ denotes the Hadamard product; $\gamma_1$ and $\gamma_2$ are weighting factors. In the binary mask $M$, 1 denotes missing regions; $\bar{M}$ denotes the logical not of $M$. In the implementation, we set $\gamma_1 = 10$ and $\gamma_2 = 40$, thus our generator will emphasize on generating the missing strokes.

We also use the integrated adversarial loss $\mathcal{L}_{adv}$ from the multiscale discriminators, the feature matching loss $\mathcal{L}_{feat}$, and the perceptual loss $\mathcal{L}_{vgg}$, following Pix2PixHD [62], during training. The total loss is computed by:

$$\begin{aligned}\mathcal{L}_{total} =& \mathcal{L}_{rec} + \lambda_1\mathcal{L}_{adv} + \lambda_2\mathcal{L}_{dep} + \lambda_3\mathcal{L}_{sem} \\ &+ \lambda_4\mathcal{L}_{feat} + \lambda_5\mathcal{L}_{vgg},\end{aligned} \tag{9}$$

where $\lambda_{1,2,...,5}$ are weighting factors and set to 1, 3, 50, 10, 10, respectively, in the implementation. During training, we alternately optimize the generator and discriminators.

**Implementation.** In the training stage, we use the Adam optimizer. The momentum parameters are $\beta_1 = 0.5, \beta_2 = 0.999$. We use a learning rate of 0.0002, a batch size of 1, and an epoch of 40. We conduct experiments on a single RTX 3090. In the testing stage, it costs about 0.14 seconds to estimate the complete version of a given portrait drawing, via the learned generator.

## V. System

The appearance of our interactive drawing system is as shown in Fig. 2. It mainly composes of a drawing robot, a canvas, a camera, and an interface. As shown in Fig. 1, in the beginning, a human user draws some strokes on the canvas and commands the robot to draw.

**Image Preprocessing.** The robot will then take a picture of the canvas, and processes this picture to obtain the initial portrait drawing. To this end, we first automatically detect the region of canvas [64], and then transform it to a standard rectangle by affine transformation. Since the drawing area is fixed on the canvas, we crop the patch of initial drawing empirically. To alleviate the impact of illumination or shadows, we also enhance the cropped image, and transfer it to binary.

**Drawing by Robot.** Afterward, we input the enhanced image $x_h$ into our learned GAPDI model, and estimate the complete portrait drawing $\hat{x}_f$. Based on the differences between $\hat{x}_f$ and $x_h$, the robot randomly selects some complementary strokes and draws them on the canvas. Then there will be a prompt to remind the user to continue or to end the interactive creation process.

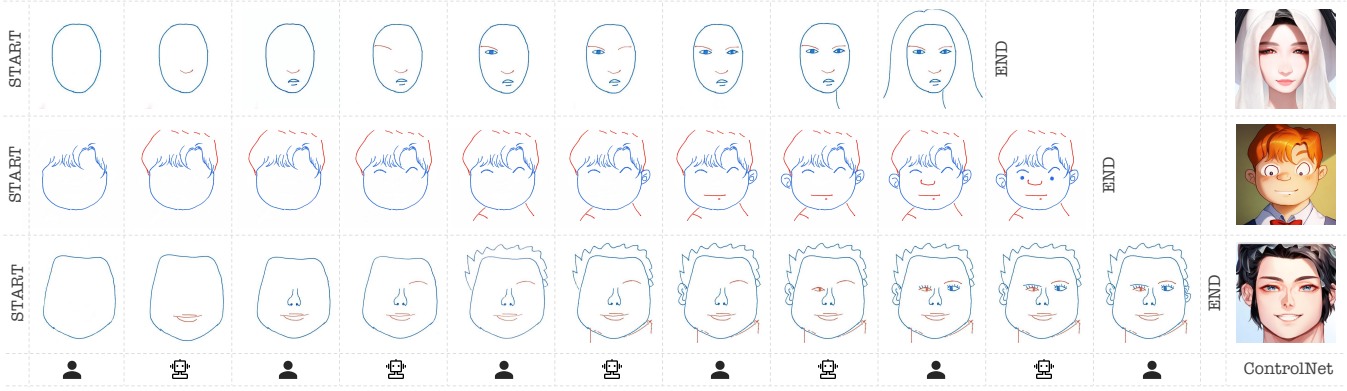


Fig. 5. Illustrations of human-robot interactive creation of portrait drawings. The blue and red strokes are drawn by the human and the robot, respectively.

**Hardware.** In the implementation, we use a hardware chip of the STM43F405RGT6 Microcontroller Unit, and develop based on the STM32 HAL library framework and RT-Thread operating system. For motion control, we used a three-axis linkage architecture to control the brush. The brush is fixed by a pen clip, which can move stably in the $z$-axis. The motor controls the brush to move freely in three-dimensional space, by turning gears and belts. Moreover, we can control the direction and speed of the brush, making the robot draw steadily. Our robot provides an $3.9 \times 5.8$ inch canvas board for the human user and robot to draw on it. For the facility, we use a mobile phone to work as the camera and to present the interactive interface. Note that most of our algorithms run on a server, we thus don't require much on the configuration of the mobile phone or the robot. These design strategies significantly decrease the cost of our robot.

**Interactive Interface.** We provide an interactive interface to facilitate interactions between human and robots (as shown in Fig. 1). A human user can start or end the interactive drawing process by pushing the corresponding buttons on the interface. In addition, we use the pretrained ControlNet [16] to generate the corresponding avatars conditioned on the created drawing, to improve the fun of the interactive drawing process. Both the portrait drawing created by human-robot and the generated avatar are shown on the interface.

## VI. Experiments

### A. Settings

We optimize our portrait drawing inpainting model, GAPDI, on the training set of our CelebLine dataset. Afterward, we embed the learned model to our whole human-robot interactive drawing system. To evaluate the performance of our system, we require several human participants to interact with our robot, and create diverse portrait drawings. In addition, we analyze GAPDI and compare with state-of-the-art (SOTA) methods on the testing set of CelebLine.

### B. Human-Robot Interactive Creation of Portrait Drawings

Fig. 5 and Fig. 6 illustrate diverse portrait drawings interactively created by different human users and our robot. In addition, we show the intermediate results alternately drawn by the human and the robot in Fig. 5. Obviously, the

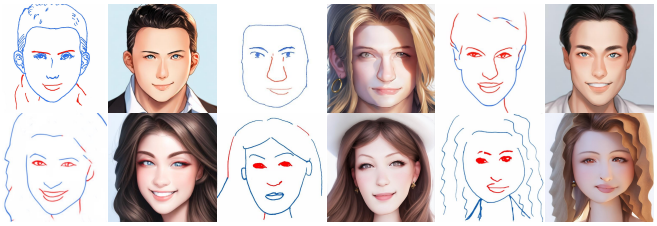

Fig. 6. Illustrations of completed portrait drawings, and the corresponding avatars generated by ControlNet. The blue and red strokes are drawn by the human and the robot, respectively.

strokes created by our robot are complement to those drawn by human users, with consistent geometry and style. Namely, our robot creates compelling strokes at the human users' skill level. The harmony of the final portrait drawings demonstrate that our robot learns to understand human creating intentions. In addition, the avatars generated by ControlNet are appealing and present consistent facial geometry.

In our experiments, a portrait drawing usually consists of 20-40 strokes, and is typically finished after 3-6 rounds of human-robot interactions. During each interaction, a human user or the robot draws 1-5 strokes in most cases. It costs about 5 minutes for a human user and our robot to interactively create a portrait drawing. Such a fast interactive creation process allows our robots to serve more users.

**Limitations:** Our system occasionally fails to complement the hair region. In our CelebLine dataset, there are usually a sparse set of strokes over the hair regions. It's challenging for the inpainting model to precisely predict the hair geometry and to generate accurate strokes.

### C. Analysis of Portrait Drawing Inpainting

In this part, we apply the learned portrait drawing inpainting model to the testing set of CelebLine, and compare with several SOTA image inpainting methods, i.e. MIST [65], MST [66], and ILVR-ADM [67], and classic image-to-image (I2I) translation methods, i.e. Pix2Pix [61], Pix2PixHD [62], and CycleGAN [35]. All these methods are trained and test under the same experimental settings as ours.

**Evaluation metrics.** We use six metrics as the criteria, including the SSIM [68], FSIM [69], and LPIPS [70] between an inpainted portrait drawing and the corresponding

TABLE I

PERFORMANCE INDICES IN PORTRAIT DRAWING INPAINTING.

| | SSIM↑ | FSIM↑ | LPIPS↓ | Precision↑ | Recall↑ | F-Measure↑ |
|---|---|---|---|---|---|---|
| Pix2Pix [61] | 70.76 | 70.00 | 0.353 | 98.04 | 96.72 | 97.37 |
| Pix2PixHD [62] | 67.11 | 67.76 | 0.284 | 97.27 | 95.58 | 96.41 |
| CycleGAN [35] | 61.15 | 62.64 | 0.389 | 95.59 | 96.67 | 96.12 |
| MISF [65] | 69.12 | 69.12 | 0.377 | 97.96 | 96.50 | 97.22 |
| MST [66] | 77.24 | 74.97 | 0.355 | **99.30** | 96.51 | **98.00** |
| ILVR-ADM [67] | 56.50 | 56.41 | 0.427 | 97.18 | 94.69 | 95.92 |
| GAPDI (Ours) | **77.57** | **75.61** | **0.260** | 98.82 | **97.00** | 97.90 |

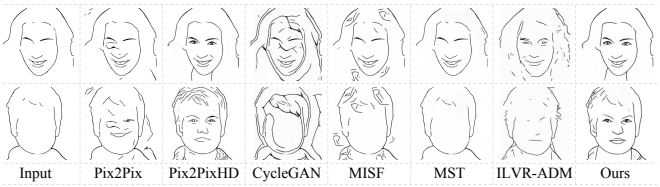


Fig. 7. Comparison with SOTA methods in portrait drawing inpainting.

target one. In addition, we calculate the Precision, Recall, and F-measure in a pixel-by-pixel comparison manner. Lower values of LPIPS (denoted by ↓), but higher values of the other criteria (denoted by ↑), indicate better performance. We report the average values across all the testing images.

**Comparison with SOTAs.** As shown in Table I, our method achieves the best performance in terms of four indices, and the second-best performance in the other two. Notably, our method achieves a significantly lower LPIPS, indicating that our inpainted portrait drawings consist well with the target ones in terms of both semantic and style. In addition, Fig. 7 shows that our method can complete portrait drawings with consistent structure and high-quality strokes. In contrast, existing image inpainting methods, MISF, MST, and ILVR-ADM, fail to complete portrait drawings. Both Pix2Pix and CycleGAN produce chaotic strokes. Pix2PixHD achieves the second-best performance qualitatively, but inferior performance indices. This implies that the portrait drawings inpainted by Pix2PixHD diverse from the target ones in geometric details. These results demonstrate that our method achieves the best performance in portrait drawing inpainting, both quantitatively and qualitatively.

**Ablation Study.** We further conduct a series of ablation study to analyse the impacts of our proposed techniques, including the depth consistency loss $\mathcal{L}_{dep}$, the semantic consistency loss $\mathcal{L}_{sem}$, the complete semantic predictor $G_s$, and the stacked inpainting strategy (*stack*). To this end, we build several model variants by gradually adding these components to our base model, i.e. Pix2PixHD [62]. As shown in Table II, the corresponding quantitative performance progressively improves as we use more proposed components. Especially, using $G_s$ significantly improve both SSIM and FSIM. Such improvement verify the significance of our geometry-aware strategy in the generator. In addition, our full model, i.e. GAPDI, achieves the best overall performance. These observations demonstrate that all the proposed components contribute to the performance improvement.

**Geometry Estimation of Portrait Drawings.** We finally

TABLE II

PERFORMANCE OF MODEL VARIANTS IN THE ABLATION STUDY.

| | SSIM↑ | FSIM↑ | LPIPS↓ | Precision↑ | Recall↑ | F-Measure↑ |
|---|---|---|---|---|---|---|
| Base | 67.11 | 67.76 | 0.284 | 97.27 | 95.58 | 96.41 |
| + $\mathcal{L}_{dep}$ | 75.88 | 73.72 | **0.249** | 98.29 | 96.99 | 97.63 |
| . + $L_{sem}$ | 76.27 | 74.16 | 0.263 | 98.69 | 96.96 | 97.81 |
| .. + $G_s$ | 77.10 | 74.63 | 0.254 | 98.80 | 96.98 | 97.88 |
| ... + stack (*full*) | **77.57** | **75.61** | 0.260 | **98.82** | **97.00** | **97.90** |

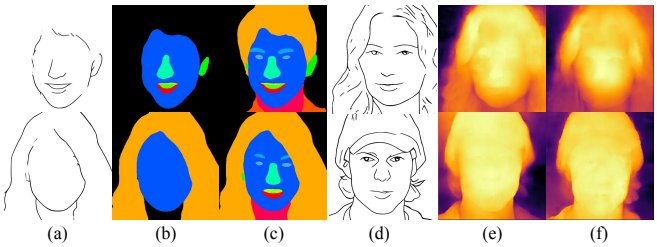


Fig. 8. Illustration of facial sketch semantic analysis. (a) Input sketch, (b) predicted semantic masks by $F_s$, (c) estimated *complete* semantic masks by $G_s$; (d) input sketch, (e) estimated depth map by $F_d$, and (e) the pseudo target depth map of the corresponding facial photo.

analyze the performance of geometry estimation for portrait drawings, including the semantic parsing, complete semantic prediction, and depth estimation. As illustrated in Fig. 8, our semantic parsing model $F_d$ can accurately estimate the actual semantic masks of an incomplete portrait drawing. While the semantic predictor $G_s$ can predict reasonable complete masks, which present natural facial structures. In addition, the estimated depth maps are highly consistent with the target 3D structures of corresponding facial photos. Such high precision of geometry estimation contributes significantly to our portrait drawing inpainting performance. In addition, these results imply the potential values of our dataset in dense analysis tasks of portrait drawings.

## VII. CONCLUSIONS

In this paper, we propose a novel framework, HRICA, for human-robot interactive creation of artistic portrait drawings. To this end, we construct a novel portrait drawing dataset CelebLine, propose a novel portrait drawing inpainting method GAPDI, and develop an interactive drawing robot system with a user-friendly interactive interface. Experimental results demonstrate the effectiveness of our framework, method, and system, as well as the potential values of our dataset. In the future, we'll explore richer interactions and more styles of drawings. In addition, it's meaningful to explore dense analysis tasks, e.g. semantic segmentation and depth estimation, of artistic drawings. Finally, we are planning to explore multi-modality driven human-robot interactive creation of artworks, and explore its practical applications in education and digital entertainment.

## REFERENCES

[1] F. Zhan, Y. Yu, R. Wu, J. Zhang, S. Lu, L. Liu, A. Kortylewski, C. Theobalt, and E. Xing, "Multimodal image synthesis and editing: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[2] J. Qin, X. Sun, and W. Xu, "A state-of-art review on intelligent systems for drawing assisting," in *International Conference on Human-Computer Interaction*. Springer, 2023, pp. 583–605.

[3] L. Scalera, S. Seriani, A. Gasparetto, and P. Gallina, "Watercolour robotic painting: a novel automatic system for artistic rendering," *Journal of Intelligent & Robotic Systems*, vol. 95, pp. 871–886, 2019.

[4] A. Karimov, E. Kopets, S. Leonov, L. Scalera, and D. Butusov, "A robot for artistic painting in authentic colors," *Journal of Intelligent & Robotic Systems*, vol. 107, no. 3, p. 34, 2023.

[5] P. Schaldenbrand, J. McCann, and J. Oh, "Frida: A collaborative robot painter with a differentiable, real2sim2real planning environment," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 712–11 718.

[6] C. Guo, Y. Dou, T. Bai, X. Dai, C. Wang, and Y. Wen, "Artverse: A paradigm for parallel human–machine collaborative painting creation in metaverses," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 4, pp. 2200–2208, 2023.

[7] P. Tresset and F. F. Leymarie, "Portrait drawing by paul the robot," *Computers & Graphics*, vol. 37, no. 5, pp. 348–363, 2013.

[8] F. Gao, J. Zhu, Z. Yu, P. Li, and T. Wang, "Making robots draw a vivid portrait in two minutes," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 9585–9591.

[9] E. C. Ferrer, I. Berman, A. Kapitonov, V. Manaenko, M. Chernyaev, and P. Tarasov, "Gaka-chu: a self-employed autonomous robot artist," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 583–11 589.

[10] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, and J. Jia, "Mat: Mask-aware transformer for large hole image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 758–10 768.

[11] F. Liu, X. Deng, J. Song, Y.-K. Lai, Y.-J. Liu, H. Wang, C. Ma, S. Qin, and H. Wang, "Sketchmaker: Sketch extraction and reuse for interactive scene sketch composition," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 12, no. 3, pp. 1–26, 2022.

[12] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5549–5558.

[13] P. Xu, T. M. Hospedales, Q. Yin, Y.-Z. Song, T. Xiang, and L. Wang, "Deep learning for free-hand sketch: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 285–312, 2022.

[14] E. Simo-Serra, S. Iizuka, K. Sasaki, and H. Ishikawa, "Learning to simplify: fully convolutional networks for rough sketch cleanup," *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–11, 2016.

[15] S. M. H. Miangoleh, S. Dille, L. Mai, S. Paris, and Y. Aksoy, "Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9685–9694.

[16] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models."

[17] C. Chan, F. Durand, and P. Isola, "Learning to generate line drawings that convey geometry and semantics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 7915–7925.

[18] Y. Vinker, E. Pajouheshgar, J. Y. Bo, R. C. Bachmann, A. H. Bermano, D. Cohen-Or, A. Zamir, and A. Shamir, "Clipasso: Semantically-aware object sketching," *ACM Transactions on Graphics (TOG)*, vol. 41, no. 4, pp. 1–11, 2022.

[19] X. Xing, C. Wang, H. Zhou, J. Zhang, Q. Yu, and D. Xu, "Diffsketcher: Text guided vector sketch synthesis through latent diffusion models," *arXiv preprint arXiv:2306.14685*, 2023.

[20] S. Huang, J. An, D. Wei, J. Luo, and H. Pfister, "Quantart: Quantizing image style transfer towards high visual fidelity," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5947–5956.

[21] L. Huang, D. Chen, Y. Liu, Y. Shen, D. Zhao, and J. Zhou, "Composer: Creative and controllable image synthesis with composable conditions," *arXiv preprint arXiv:2302.09778*, 2023.

[22] Z. Qu, T. Xiang, and Y.-Z. Song, "Sketchdreamer: Interactive text-augmented creative sketch ideation," *arXiv preprint arXiv:2308.14191*, 2023.

[23] J. Xie, A. Hertzmann, W. Li, and H. Winnemöller, "Portraitsketch: Face sketching assistance for novices," in *Proceedings of the 27th annual ACM symposium on User interface software and technology*, 2014, pp. 407–417.

[24] A. Shesh and B. Chen, "Smartpaper: An interactive and user friendly sketching system," *Computer Graphics Forum*, vol. 23, no. 3, pp. 301–310, 2004.

[25] Z. Huang, Y. Peng, T. Hibino, C. Zhao, H. Xie, T. Fukusato, and K. Miyata, "dualface: Two-stage drawing guidance for freehand portrait sketching," *Computational Visual Media*, vol. 8, pp. 63–77, 2022.

[26] J. Choi, H. Cho, J. Song, and S. M. Yoon, "Sketchhelper: Real-time stroke guidance for freehand sketch retrieval," *IEEE Transactions on Multimedia*, pp. 1–1, 2019.

[27] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "Apdrawinggan: Generating artistic portrait drawings from face photos with hierarchical gans," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 743–10 752.

[28] N. Wang, S. Zhang, C. Peng, J. Li, and X. Gao, *Face Sketch Recognition via Data-Driven Synthesis*, 2017.

[29] S. Zhang, R. Ji, J. Hu, Y. Gao, and L. C.-W., "Robust face sketch synthesis via generative adversarial fusion of priors and parametric sigmoid," in *IJCAI*, 2018, pp. 1163–1169.

[30] M. Zhu, C. Liang, N. Wang, X. Wang, Z. Li, and X. Gao, "A sketch-transformer network for face photo-sketch synthesis," in *International Joint Conference on Artificial Intelligence*, 2021.

[31] J. Yu, S. Shi, F. Gao, D. Tao, and Q. Huang, "Towards realistic face photo-sketch synthesis via composition-aided GANs," *IEEE TCYB*, vol. 51, no. 9, pp. 4350–4362, 2021.

[32] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "Quality metric guided portrait line drawing generation from unpaired training data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 905–918, 2022.

[33] P. Zhang, B. Zhang, D. Chen, L. Yuan, and F. Wen, "Cross-domain correspondence learning for exemplar-based image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5143–5153.

[34] C. Jiang, F. Gao, B. Ma, Y. Lin, N. Wang, and G. Xu, "Masked and adaptive transformer for exemplar based image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 418–22 427.

[35] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[36] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.

[37] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[38] C. Gao, Q. Liu, Q. Xu, L. Wang, J. Liu, and C. Zou, "Sketchycoco: Image generation from freehand scene sketches," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5174–5183.

[39] C. Zou, Q. Yu, R. Du, H. Mo, Y.-Z. Song, T. Xiang, C. Gao, B. Chen, and H. Zhang, "Sketchyscene: Richly-annotated scene sketches," in *Proceedings of the european conference on computer vision (ECCV)*, 2018, pp. 421–436.

[40] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4CD, pp. 107.1–107.14, 2017.

[41] H. Liu, Z. Wan, W. Huang, Y. Song, X. Han, and J. Liao, "Pd-gan: Probabilistic diverse gan for image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9371–9381.

[42] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 461–11 471.

[43] H. Xiang, Q. Zou, M. A. Nawaz, X. Huang, F. Zhang, and H. Yu, "Deep learning for image inpainting: A survey," *Pattern Recognition*, vol. 134, p. 109046, 2023.

[44] 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)*, 2018, pp. 85–100.

[45] A. K. Bhunia, S. Khan, H. Cholakkal, R. M. Anwer, F. S. Khan, J. Laaksonen, and M. Felsberg, "Doodleformer: Creative sketch drawing with transformers," in *European Conference on Computer Vision*. Springer, 2022, pp. 338–355.

[46] F. Liu, X. Deng, Y.-K. Lai, Y.-J. Liu, C. Ma, and H. Wang, "Sketchgan: Joint sketch completion and recognition with generative adversarial network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5830–5839.

[47] Y. Liu, H. Shi, H. Shen, Y. Si, X. Wang, and T. Mei, "A new dataset and boundary-attention semantic segmentation for face parsing," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 637–11 644.

[48] E. Wood, T. Baltrušaitis, C. Hewitt, S. Dziadzio, T. J. Cashman, and J. Shotton, "Fake it till you make it: face analysis in the wild using synthetic data alone," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3681–3691.

[49] Q. Zheng, J. Deng, Z. Zhu, Y. Li, and S. Zafeiriou, "Decoupled multi-task learning with cyclical self-regulation for face parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4156–4165.

[50] P. Luo, X. Wang, and X. Tang, "Hierarchical face parsing via deep learning," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2480–2487.

[51] S. Liu, J. Shi, J. Liang, and M.-H. Yang, "Face parsing via recurrent propagation," *arXiv preprint arXiv:1708.01936*, 2017.

[52] P. Huang, J. Han, D. Zhang, and M. Xu, "Clrnet: Component-level refinement network for deep face parsing," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[53] G. Te, W. Hu, Y. Liu, H. Shi, and T. Mei, "Agrnet: Adaptive graph representation learning and reasoning for face parsing," *IEEE Transactions on Image Processing*, vol. 30, pp. 8236–8250, 2021.

[54] L. Li, H. Fu, and C.-L. Tai, "Fast sketch segmentation and labeling with deep learning," *IEEE computer graphics and applications*, vol. 39, no. 2, pp. 38–51, 2018.

[55] F. Wang, S. Lin, H. Wu, H. Li, R. Wang, X. Luo, and X. He, "Spfusionnet: Sketch segmentation using multi-modal data fusion," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2019, pp. 1654–1659.

[56] L. Yang, J. Zhuang, H. Fu, K. Zhou, and Y. Zheng, "Sketchgcn: Semantic sketch segmentation with graph convolutional networks," *arXiv preprint arXiv:2003.00678*, vol. 5, no. 6, 2020.

[57] Y. Zheng, J. Xie, A. Sain, Y.-Z. Song, and Z. Ma, "Sketch-segformer: Transformer-based segmentation for figurative and creative sketches," *IEEE Transactions on Image Processing*, 2023.

[58] G. Berardi, S. Salti, and L. Di Stefano, "Sketchydepth: from scene sketches to rgb-d images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2414–2423.

[59] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1395–1403.

[60] W. Yin, J. Zhang, O. Wang, S. Niklaus, L. Mai, S. Chen, and C. Shen, "Learning to recover 3d scene shape from a single image," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (CVPR)*, 2021.

[61] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[62] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8798–8807.

[63] W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, J. Feng, and S. Yan, "Metaformer is actually what you need for vision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 819–10 829.

[64] J. Li, Y. Xu, T. Lv, L. Cui, C. Zhang, and F. Wei, "Dit: Self-supervised pre-training for document image transformer," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 3530–3539.

[65] X. Li, Q. Guo, D. Lin, P. Li, W. Feng, and S. Wang, "Misf: Multi-level interactive siamese filtering for high-fidelity image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1869–1878.

[66] C. Cao and Y. Fu, "Learning a sketch tensor space for image inpainting of man-made scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 509–14 518.

[67] J. Choi, S. Kim, Y. Jeong, Y. Gwon, and S. Yoon, "Ilvr: Conditioning method for denoising diffusion probabilistic models," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021, pp. 14 347–14 356.

[68] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[69] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[70] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.