The Visual Computer
https://doi.org/10.1007/s00371-025-03970-1

RESEARCH

Instance-guided anime editing with a curated large-scale dataset

Jian Lin[1] · Chengze Li[1] · Xueting Liu[1] · Zhongping Ge[1]

Accepted: 24 April 2025

Abstract
Anime content, such as Japanese-style illustrations, manga, and animation, is popular worldwide among diverse audiences. However, editing and repurposing this content for an enhanced viewing experience is complex and relies heavily on manual processes, due to the challenge of automatically identifying individual character instances. Therefore, automated and precise segmentation of these elements is essential to enable various anime editing applications such as visual style editing, motion decomposition and transfer, and depth estimation. Most state-of-the-art segmentation methods are designed for natural photographs and do not capture the intricate aesthetics of anime-style characters, which reduces segmentation quality. The primary challenges are the lack of high-quality anime-dedicated datasets and the absence of competent models for high-resolution instance extraction on anime. To address these issues, we introduce a high-quality dataset of over 100k paired high-resolution anime-style images and their instance labeling masks. We also present an instance-aware image segmentation model that generates accurate, high-resolution segmentation masks for characters in a wide variety of anime-style images. Furthermore, we show that our approach supports segmentation-dependent editing applications such as 3D Ken Burns effects, text-guided style editing, and puppet animation from illustrations and manga.

Keywords Anime segmentation · Anime editing · Visual style manipulation

1 Introduction

Anime content, encompassing Japanese-style illustrations, manga, and animations, has gained immense popularity worldwide. Along with this popularity, anime editing has garnered significant interest among both professional illustrators and hobbyists, enabling unprecedented freedom in derivative creation and expansion of original anime works. In modern workflows, enhanced visual effects, such as the parallax pop-up effect, movement effects, and visual style translation, are commonly applied to enrich visual experiences from single anime-style images. However, to successfully apply these and other post-production operations, illustrators must first precisely separate the subjects from the image through accurate segmentation. This process is crucial across a range of tasks, from filling missing background pixels for parallax effects, to ensuring proper depth estimation and subject-background distinction during style editing. In multiple post-editing tasks for anime contents, imprecise boundaries can result in broken structures or distorted appearances, especially when magnified or viewed up close. Manual segmentation, although an option, is often tedious, time-consuming, and error-prone, making it impractical in demanding workflows.

Automatic segmentation methods are therefore essential for streamlining these processes. However, current segmentation models, despite numerous advancements [9, 17, 23], are typically focused on natural images and prioritize generalization and content recognition over fine-grained understanding on anime contents with aesthetic considerations. Unlike natural images, anime illustrations and manga feature explicit structural lines and delicate shading styles and therefore require much higher-quality segmentation to accurately capture their unique characteristics. Inaccuracies in segmentation often result in severe artifacts during post-processing, such as incorrect depth estimations, aliasing, and broken silhouettes. Consequently, existing segmentation methods fall short of the high-resolution demands of anime editing, highlighting the need for purpose-built, high-quality automatic segmentation solutions tailored to the unique characteristics of anime.

✉ Chengze Li
czli@sfu.edu.hk

[1] Saint Francis University, New Territories, Hong Kong SAR, China

Published online: 23 May 2025

Fig. 1 Given an input anime-style image, our method extracts high-quality instance masks for the subjects in the image. The instance extraction enables a variety of segmentation-dependent applications, such as the 3D Ken Burns effect and visual style editing

To address this specific requirement, we propose a large-scale anime segmentation dataset consisting of 98.6k paired high-resolution anime-style images with their corresponding ground truth subject instance masks. This dataset is created using a novel method that extracts character and object instances from chroma-keying anime videos (Fig. 2) and still illustrations, simulating the actual composition of anime illustrations and animations. Notably larger than existing datasets, it covers a wide range of complex and challenging cases, including multiple characters, occlusions, extreme lighting conditions, and intricate compositions. In addition to the dataset, we also introduce a novel two-stage instance segmentation approach. We first employ a low-resolution segmentation stage to locate a rough, occlusion-free mask for each subject. After that, we introduce a high-resolution refinement stage that uses the rough mask as guidance, to output precise high-resolution segmentation instance masks. This novel two-stage design emphasizes efficiency and performance in high-resolution anime-style images, which is crucial to capture delicate details such as hair, clothing, and intricate decorations that existing solutions cannot handle correctly.

We evaluated our proposed method on a diverse set of Japanese-style illustrations, manga, and anime frames, demonstrating a clear improvement over existing approaches in both qualitative and quantitative assessments. Performance gains are a direct result of the synergy between our large-scale data set and the novel segmentation approach; neither element alone could achieve the observed advances. We also demonstrate a range of downstream applications enabled by our precise high-resolution segmentation, including 3D Ken Burns pop-up effects, text-guided anime-style editing, and puppet animation for still illustrations and manga.

Our contributions can be summarized as follows:

- To our knowledge, this is the first approach specifically designed for high-resolution subject instance segmentation in anime-style contents.
- We introduce a large-scale anime segmentation dataset, generated through a novel reverse engineering process by remixing instances from chroma-keying videos and illustrations.
- Our high-quality instance segmentation results significantly streamline a variety of anime production workflows and enable enhanced applications that rely on precise segmentation.

2 Related work

2.1 Image and instance segmentation

Instance segmentation identifies and labels individual objects in an image, distinguishing different instances of the same object class. This is crucial for anime editing and various downstream tasks. Among generic instance segmentation methods, Faster R-CNN [22] and Mask R-CNN [9] output object masks alongside bounding boxes, but they produce low-resolution masks and are computationally heavy. Advances such as RTMDet-Ins [17] use efficient architectures like CSPDarkNet [3] and instance-aware mask heads [31], improving performance without RoI proposals. However, these models generally predict masks at low resolution, resulting in coarse, polygon-like instance masks, which are not suitable for anime editing tasks that require precise instance boundaries. Recently, the Segment Anything Model (SAM) [11], trained on a large dataset, demonstrates improved generalization across media types, including some anime-style content. However, SAM requires additional detector input, which prevents fully automatic segmentation.

Grounded SAM [16] addresses this limitation by using text input to infer initial bounding boxes. However, the quality of its segmentation depends on the initial extraction stage, which can lead to missed instances or false positives. In addition, their dataset focuses on natural photos, which may cause inaccuracies in predicting anime-style content. The second version of SAM [21] extends to video frame segmentation, but its main focus is on consecutive frames, and there is almost no performance gain of SAMv2 over SAM on anime content.

Besides generic models, an anime-tailored segmentation model YAAS [37], derived from CondInst [31] and SOLOv2 [32], was introduced with a specific dataset of 945 anime images. However, due to the scarcity of high-quality domain-specific data and the limitations of its model architecture, it struggles with accurate instance mask prediction, understanding complex multi-character anime scenes and handling very high-resolution segmentations. There is also related work on segmenting western comics [2], which uses domain adaptation to transfer comics into the photographic domain for segmentation and depth estimation. However, anime content is more diverse and visually distinct from photographs than western comics, making domain adaptation more challenging. We believe a larger-scale dataset with pixel-level labeling should provide more direct and effective supervision for anime segmentation compared to implicit adaptation techniques. Matting techniques can also extract high-quality foreground objects. Sun et al. [30] uses Mask R-CNN for trimap generation and instance matting, but its reliance on natural and synthetic data, along with low-resolution R-CNN output, limits its use for anime-style content.

In contrast, our approach addresses these limitations by providing a robust solution for high-quality anime instance segmentation. Our improved framework generates accurate, high-resolution segmentation for anime content, offering superior results.

2.2 Datasets for cartoon segmentation

Datasets are vital for deep learning in anime editing due to the significant domain gap between natural photographs and anime-style content. For example, models trained on natural photos like MSCOCO [14] often perform poorly on anime illustrations and manga, highlighting the need for a tailored anime-specific dataset. Currently, there is a lack of such datasets with detailed high-resolution instance labeling. Among existing datasets, AniSeg [12] presents a small dataset of 945 ground truth character-mask pairs and a synthetic dataset generated from these pairs. However, the limited diversity of the data restricts the generalization ability of the models trained on it, especially when the input images feature multiple character instances. Chen et al. [4] present a synthetic dataset, combining 18.5k foreground images from the Danbooru dataset [1] with 8.1k background images from Danbooru and Pixiv. However, these foreground images contain still illustrations only, with different visuals from actual animations and manga. Also, this publicly unavailable dataset may not support high-resolution segmentation due to its limited data scale. There are also fine-grained large-scale anime datasets such as annotated manga109 [7] and Unlocking Comics [8], but manga109 lacks instance-level segmentation annotations, and Unlocking Comics focuses on western comics, which may not generalize to Japanese-style anime. In this work, we propose a dataset with 98.6k high-resolution anime-style image–mask pairs, to ensure quality in segmentation.

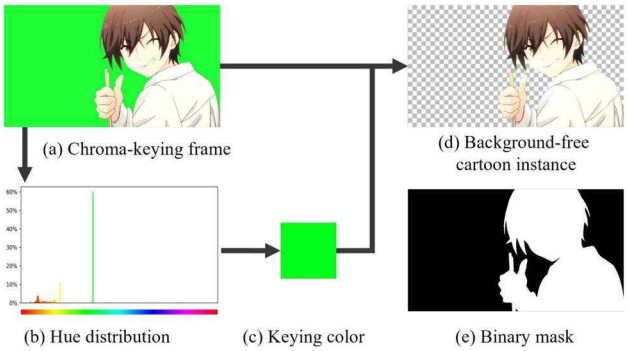

Fig. 2 Keying-based foreground preparation. We estimate the keying colors based on hue distribution, to produce the background-free instance image and its corresponding mask

3 Anime segmentation dataset

3.1 Preparation of foregrounds and backgrounds

As discussed previously, manually preparing a large-scale anime segmentation dataset is impractical due to the significant time and labor required for human labeling. To overcome this bottleneck, we propose a reverse engineering method that constructs the dataset efficiently by extracting instances from chroma-keying videos and drawings and then synthesizing them to replicate actual anime compositions, thereby eliminating the need for manual annotation.

Chroma-keying materials, as shown in Fig. 2a, provide a practical and scalable source for extracting subject instances without human labeling. These materials feature animated characters or objects against a uniform keying color background, commonly used for animation reuse or expansion. The keying color in each frame (Fig. 2a) can be automatically identified (Fig. 2b, c) and used as an alpha channel. This enables automatic extraction of background-free characters or objects (Fig. 2d) along with their corresponding pixel-based binary masks (Fig. 2e). Besides the native ability to represent segmentation masks, we also find them critical

to improving the generalization of the model in segmenting real-life anime video frames, compared to still illustrations only. We include the detailed algorithms in data preprocessing, keying color detection, and the extraction of instance masks in the supplemental material.

We collect 1064 redistributable chroma-keying videos from the online anime video community and supplement them with still subject illustrations from Danbooru2021 [1]. For these illustrations, we use images labeled with *solo* and *transparent_background* tags, ensuring a single, clearly masked subject. Our dataset contains 65,752 subjects: 26,844 from chroma-keying videos and 38,908 from Danbooru2021. We deliberately include decorations, accessories, and items attached to characters, treating them as part of the instance rather than segmenting them separately. This decision reflects anime editing workflows, where such elements (e.g., a staff, bag, or hair ornaments) are considered integral to the character's appearance and narrative contribution and are typically manipulated together during editing.

For background preparation, the quality and variety of the collected background images play a crucial role in the synthesis of the dataset. Owing to the focus on segmentation, motion backgrounds are excluded. Consequently, we opt for high-quality anime-style stills, known as *CG backgrounds*, as our image set. Our collection comprises 8163 scenic backgrounds from the Danbooru2021 dataset, filtered by *scenery* and *no_humans* tags, and 8,057 anime backgrounds from the bizarre pose estimator [4]. The vast majority (over 80%) of images in our dataset exceed a resolution of 1024×768, satisfying our high-quality prerequisite.

3.2 Dataset synthesis

After isolating the foreground and background elements, we synthesize anime-like images and their segmentation labels to create the final dataset. Careful subject positioning and effective augmentation are key to avoid sparse or crowded layouts and to generate sufficient training data. As shown in Fig. 3, the dataset synthesis uses a 720×720 canvas. For each image, we sample N_{subj} subjects from a Poisson distribution ($\lambda = 2.5$) (Fig. 3a, b) and position them by sampling the relative bounding box layout from randomly selected MSCOCO [14] photos with at least N_{subj} *person* labels to mimic natural multi-character positioning (Fig. 3c). We crop backgrounds randomly and place the subject instances in the bounding boxes (Fig. 3d) arbitrarily. We further optimize the composited image via histogram matching and color quantization (Fig. 3f) for more harmonized colors and shadings, which resemble the appearances in anime. This process produces a high-quality anime instance segmentation dataset of 98.6k image–mask pairs. Refer to the Supplementary Material for details on the synthesis of the data set.

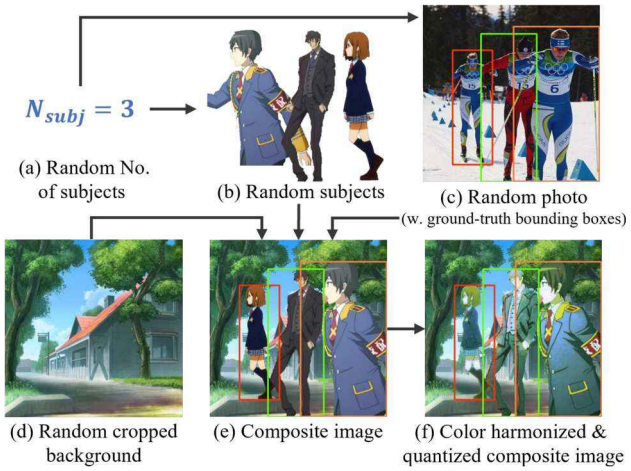

Fig. 3 Synthesis steps of our segmentation dataset

4 Anime instance extraction network

With our prepared large-scale segmentation dataset, we aim to extract high-quality instance-level subjects with resolutions up to 720^2. We propose a novel two-stage segmentation approach to achieve this. In the first stage, we perform low-resolution segmentation to locate a rough, occlusion-free mask for each subject. In the second stage, we refine this to obtain a more precise instance mask with segmentation model working natively at higher resolutions. We illustrate the pipeline in Fig. 4 and introduce our tailored model architecture and loss objective designs as follows.

4.1 Low-resolution segmentation stage

The model design of the low-resolution segmentation stage inherits from the RTMDet-Ins [17] architecture to detect subjects from the input image and obtain their coarse segmentation masks. The network model is mainly composed of three components: a detection backbone for feature extraction, a neck for feature fusion and refinement, and several detection heads to produce model output. Specifically, the detection backbone utilizes the enhanced CSPNeXt building blocks for a multi-level feature extraction. With the extracted multi-level feature, the neck performs a multi-level feature fusion [15] to further enhance the features. Finally, we apply a bounding box head to predict the bounding box, a classification head to predict the confidence score of the bounding box, and a mask head to predict the masks of the subjects. The training of this network is versatile and can be trained solely on the bounding box classification and regression or can be expanded to be supervised with instance masks with learnable dynamic convolutions. This one-pass, anchor-free design provides a lightweight solution while maintaining a good balance between computational efficiency and accu-

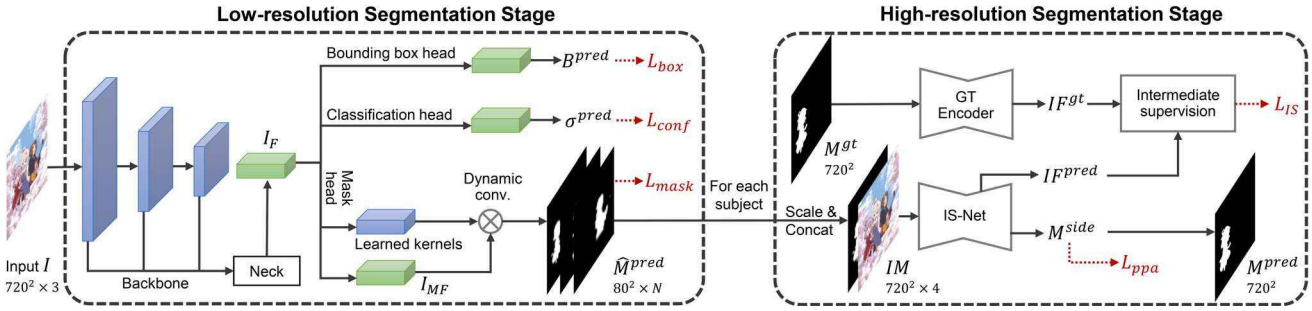

Fig. 4 Model architecture overview. The low-resolution stage extracts input features and learns to predict the bounding boxes and coarse instance masks. Subsequently, high-resolution segmentation refines the coarse masks, resulting in high-resolution subject instance extraction. Losses are highlighted in red

racy. Given an input image I at the resolution of $720^2 \times 3$, the detection backbone and neck are used to compute an enhanced multi-level feature I_F, which is further operated by two branches of heads: one branch predicts the bounding boxes of the subjects B^{pred} and their confidence values σ^{pred} with the bounding box and classification heads; the other branch predicts the instance masks $\hat{M}^{pred}$ with the mask head.

4.1.1 Bounding box classification and regression

With the multi-scale feature I_F computed, we first regress bounding boxes using three grid layouts that partition the 720^2 resolution into 80^2, 40^2, and 20^2 grids, resulting in $8,400$ candidate bounding boxes. Each bounding box is represented by four parameters *top, left, bottom, and right* relative to the cell center and is predicted through the bounding box head using a combination of convolution and normalization blocks in a regression framework. For each candidate bounding box, a confidence score is also computed by the classification head using additional convolutional blocks and logistic regression. This score indicates the likelihood that the box contains a subject. Since the number of candidate bounding boxes far exceeds the actual number of subject instances, we adopt a dynamic soft label assignment strategy [17] to select the top-N best matching boxes from the $8,400$ candidates, where N is the ground truth number of subjects.

We design two loss objectives to supervise the predicted bounding boxes and confidence scores. First, the GIoU loss [24] is adopted to supervise the overlapping with the ground truth bounding box:

$$L_{box} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \frac{C_i - |B_i^{pred} \cup B_i^{gt}|}{C_i}\right) \tag{1}$$

where i the bounding box index, B_i^{pred} and B_i^{gt} are the set of pixels in the i-th predicted bounding box and ground truth bounding box, respectively, $|\cdot|$ calculates the number of elements in a set, C_i is the area of the smallest box that contains both B_i^{pred} and B_i^{gt}, and $y_i = IoU(B_i^{pred}, B_i^{gt})$ is the IoU score between the predicted and ground truth bounding boxes.

After that, to supervise the confidence score, we use the quality focal loss [13] for an accurate and flexible optimization for the detection likelihood:

$$\begin{aligned} L_{conf} = -\textstyle\sum_{i=1}^{N} |y_i - \sigma_i|^{\beta} \left((1-y_i)\log(1-\sigma_i)\right. \\ \left. + y_i \log(\sigma_i)\right), \end{aligned} \tag{2}$$

where β is a parameter to down-weigh the easy examples achieving both good bounding box and confidence predictions. We set β to 2 in our experiments.

4.1.2 Low-resolution instance segmentation

The low-resolution instance masks are predicted after the bounding box regression. Specifically, we extract an eight-channel feature I_{MF} from I_F with four convolutional layers. With the predicted N bounding boxes obtained from the other branch, N dynamic kernels are learned to filter I_{MF} into N low-resolution instance masks $\hat{M}^{pred}$ at the dimension of $80^2 \times N$. We use the dice loss to provide balanced supervision in the precision and recall between the predicted masks and the ground truth:

$$L_{mask} = 1 - \frac{2\sum_j \left(\hat{M}_i^{gt}(j) * \hat{M}_i^{pred}(j)\right)}{\sum_j \left[\hat{M}_i^{gt}(j)\right]^2 + \sum_j \left[\hat{M}_i^{pred}(j)\right]^2 + \epsilon}, \tag{3}$$

where $\hat{M}^{gt}$ and $\hat{M}^{pred}$ denote the ground truth and predicted masks, respectively, with elements indexed by j. The parameter ϵ is a small constant included for numerical stability. Both $\hat{M}^{gt}$ and $\hat{M}^{pred}$ are rescaled to $320^2 \times N$ via bilinear interpolation for loss computation. We sum this L_{mask} for all instance detected during training. In all, we optimize the following loss objective for the first low-level segmentation stage:

$$L_{stage1} = \lambda_{box} L_{box} + \lambda_{conf} L_{conf} + \lambda_{mask} L_{mask}. \quad (4)$$

where λ_{box}, λ_{conf}, and λ_{mask} are weighting factors and set to 2, 1, and 2, respectively, in all our experiments.

4.2 High-resolution segmentation stage

In the high-resolution segmentation stage, we aim to extend and refine the roughly predicted mask from the previous stage into a high-quality mask at the resolution of 720^2 to better capture the intricate details of anime-style subjects. We design approach upon the IS-Net [19] architecture. Compared to alternatives such as U^{2}Net [20] and UNet [26], IS-Net uses a multi-scale ground truth encoder network to encode the ground truth instance mask M^{gt} into a compact feature representation IF^{gt} and establishes supervision on these features. This design enables segmentation at much higher resolutions while keeping the memory footprint manageable. We design our model architecture to output and supervise two major components:

- The multi-scale intermediate features IF^{pred}. We use the IS-Net backbone to compute these implicit multi-scale features from the coarse mask and the input image.
- A set of side masks M^{side}, where intermediate features at different scales are decoded back to pixel space at the input resolution. The multi-scale decoding encourages the model to gain enough awareness in predicting accurate masks at all feature scales.

Technically, for each coarse instance mask $\hat{M}_i^{pred}$ (simplified as $\hat{M}$ in this stage for conciseness) obtained in the first stage, we upscale it to 720^2 and concatenate it with the input image I to form a $720^2 \times 4$ input IM. The network uses this input to generate the multi-scale predicted features IF^{pred} and the side masks M^{side} decoded from coarse to finest IS-Net features. In our experiment, we compute six levels of feature scales ($D = 1 \cdots 6$) for IF^{pred} and M^{side}. The final output M^{pred} is taken from the decoded results of finest level ($D = 1$) of M^{side}.

Intermediate supervision. We use the original intermediate supervision in the IS-Net architecture to align the predicted features IF^{pred} with the encoded ground truth IF^{gt}. The supervision applies MSE loss across all feature scale levels:

$$L_{IS} = \sum_{D=1}^{6} \| IF_D^{pred} - IF_D^{gt} \|_2^2, \quad (5)$$

where $\| \cdot \|_2$ is the norm-2 operator.

Pixel position-aware mask supervision. To emphasize boundary accuracy in the predicted segmentation, we introduce a tailored loss that incorporates pixel position awareness [33] into the supervision of the side masks M^{side}. This encourages the model to pay greater attention to discrepancies near object edges, which are especially crucial for the downstream processing of details such as hair, hands, and accessories in anime content. We define this pixel position-aware loss as:

$$L_{ppa} = \sum_{D=1}^{6} \frac{\lambda_D}{N} \sum_{i=1}^{N} \Big(W_i * \mathrm{BCE}\big(M_D^{side}, M^{gt}\big) + W_i * \big(1 - \mathrm{IoU}_p\big(M_D^{side}, M^{gt}\big)\big)\Big), \quad (6)$$

where N is the total number of pixels in each mask, W_i is the weighting factor for the i-th pixel, and $\mathrm{BCE}(\cdot)$ and $1 - \mathrm{IoU}_p(\cdot)$ are binary cross-entropy and pixel-wise IoU losses, focusing on pixel-wise classification accuracy and region overlap, respectively. The factor λ_D is a weight assigned to each scale D; we set $\lambda_6 = 5$ to emphasize the coarsest resolution, and $\lambda = 1$ for all other scales. The boundary weighting matrix W is introduced to stress loss on boundary pixels:

$$W = 1 + 5 \times \big|\mathrm{AvgPool2D}\big(M^{gt}\big) - M^{gt}\big|, \quad (7)$$

which assigns larger values to regions close to object edges. This highlights areas likely to contain fine structures, while still encouraging high overall mask quality.

Finally, we write the overall training loss for this stage as:

$$L_{stage2} = L_{IS} + L_{ppa} \quad (8)$$

and include training details in the supplementary.

5 Evaluations and discussions

5.1 Quantitative evaluations

We evaluate the effectiveness of our two novel components: the large-scale anime segmentation dataset and the two-stage anime instance segmentation model. We compare our solutions with existing approaches, including the classic learning-based Mask R-CNN [9], the state-of-the-art photo segmentation method Grounded Segment Anything [23], and the state-of-the-art anime segmentation method YAAS [37]. For a fair comparison, we prepare two versions of each competitor model: the vanilla model and a model fine-tuned with our proposed dataset. We perform evaluations on a separately prepared test dataset comprising 307 anime-style content samples across a wide array of Japanese-style illustrations, manga, and anime frames, with diversity in visual content

Table 1 Quantitative comparisons with existing methods. Values in bold indicate the highest performance for each metric

Model	Box AP ↑	Mask AP ↑	Boundary AP ↑
Mask R-CNN	67.1	61.4	5.9
Grounded SAM (Original)	92.2	46.7	16.6
Grounded SAM (Finetuned)	92.2	90.1	49.7
SOLOv2 (YAAS)	64.3	59.0	26.2
SOLOv2 (Finetuned)	76.8	70.8	22.8
Ours	**93.1**	**93.2**	**63.6**

and style, all manually labeled for ground truth segmentation. We verified that no subject or background in the test dataset overlaps with our training dataset.

We conduct the quantitative evaluation with three metrics: Box AP, Mask AP, and Boundary AP [6]. Box AP computes the precision of predicted bounding boxes against their ground truths; Mask AP measures the precision of instance masks; and Boundary AP computes precision over instance boundary areas to highlight boundary fitting accuracy. The boundary IoU computes the intersection-over-union only at the pixels within the distance of 2% of the image diagonal length from their exterior mask contours, to highlight the boundary extraction precision. We reveal the evaluation statistics in Table 1. Our evaluation first confirms that fine-tuning on our large-scale dataset significantly enhances the performance of all existing models. The distinctive features of anime-style content in our dataset enable learning-based solutions to handle delicate styles more proficiently. Additionally, our ablation study comparing our full model with one trained on only 1/8 of the dataset (Table 2) demonstrates that a substantial anime-specific dataset is indispensable and cannot be effectively replaced by a few-shot alternative.

We afterward examine the model performances. Among the competitors, Mask R-CNN shows the lowest performance due to numerous unidentified subjects, reflected in its low Box AP scores. The model is significantly limited by its architecture design, where high-resolution features are not fully utilized during inference. The transformer-based Grounded SAM achieves better results with our dataset, but still falls short of our model's performance. Additionally, the model features a lower mask output resolution compared to ours (256^2 vs. 720^2), leading to a significantly weaker ability to extract precise boundaries. Furthermore, their method is not fully automatic and requires a manual stage of bounding box extraction from the text prompt. In contrast, our proposed model is designed to work fully automatic on anime-style subject instances. Meanwhile, the SOLOv2 mode cannot also surpass our performance with any of the datasets. We believe this is due to their relatively smaller model capacity and the lack of flexible image feature extraction and understanding. The model also works poorly at boundary pixels.

Table 2 Ablation study reports on model and dataset. Values in bold indicate the highest performance for each metric

Method	Box AP ↑	Mask AP ↑	Boundary AP ↑
1/8 dataset	89.4	83.3	44.2
W/o 2nd stage	**93.1**	84.6	52.1
W/o L_{ppa}	**93.1**	92.9	63.1
Full model	**93.1**	**93.2**	**63.6**

We also evaluated the computational efficiency of our proposed model against the SAM model, focusing on anime-style frame inputs at 1024px resolution. Our model demonstrated a significant advantage, extracting masks at 25.6 images per second compared to Grounded SAM's 3.1 on the same test dataset. This performance gain is achieved while delivering a much higher mask output resolution, highlighting our model's optimization for domain-specific tasks where speed and resolution are critical.

5.2 Qualitative evaluations

We present our qualitative visual comparison in Fig. 5. In this evaluation we do not compare with Mask R-CNN due to its large margin to the state of the art in performance. As shown in Fig. 5b, the SOLOv2-based segmentation model, YAAS, fails to extract complete instances in most sample cases and produces checkerboard artifacts. We also find our dataset helps the SOLOv2 model to better deal with the parsing of crowded multiple instance images, yet its performance is still suboptimal. In Fig. 5c, the fine-tuned Grounded SAM model extracts all instances but struggles to distinguish characters without clear boundaries (top row) and creates tear-like artifacts (2nd and 3rd rows). Moreover, the SAM model cannot extract very precise instance boundaries, especially when the instance features delicate drawings such as long hair, fancy clothes, or accessories. Additionally, the model does not generalize well to abstract line drawings and manga images (3rd show). In contrast, our method, supported by a large-scale anime-style segmentation dataset and a high-resolution segmentation model tailored for anime-style images, successfully identifies all subjects with

Fig. 5 Qualitative comparisons with existing methods

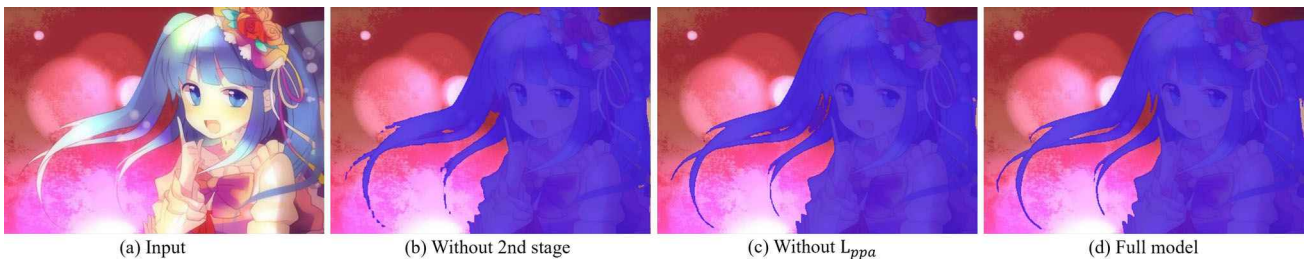

Fig. 6 Qualitative ablation studies. **b** Without the high-resolution segmentation stage, the model produces inaccurate masks. **c** Without loss of L_{ppa} in the high-resolution stage, mask boundary accuracy is compromised. **d** Our full model output

high-quality masks across all test images. Particularly, we emphasize our method's effectiveness in challenging scenarios, such as large areas of occlusion, intricate structures at the detail level, and adaptability to various styles of anime-style content. In addition, our method provides smoother and more adhesive boundaries in extracting these instances. Refer to the supplemental material for additional qualitative comparisons.

5.3 Ablation studies

We conducted an ablation study for both our large-scale dataset and building blocks in our model design. Besides the studies on the constrained dataset discussed in Sect. 5.1, we also validate our model design by removing the second stage and the pixel position-aware loss L_{ppa}, respectively. Figure 6 presents the visual comparisons. Observably, without the second high-resolution segmentation stage, the model

Fig. 7 3D Ken Burns synthesis with instance awareness. Our extraction of instances provides consistent geometry and texture in the synthesized 3D scene due to the additional prior provided from the instances

fails to produce accurate masks for intricate structures such as hairs. Incorporating the high-resolution segmentation stage substantially improves the quality of the segmented instance masks. Without the pixel position-aware loss L_{ppa}, the model still struggles to capture precise structures and mask boundaries. The qualitative statistics in Table 2 also support these specific model designs.

6 Applications

The high-quality instance-level extraction of anime-style subjects enables various anime editing applications. These were previously considered challenging and time-consuming due to the significant reliance on manual segmentation of scenes and characters. In this section, we introduce three major applications essential for the real-world production of anime and cartoons. Refer to the supplementary material for details of implementation and evaluations of these applications.

3D Ken Burns. The 3D Ken Burns effect [18], widely used in low-cost cartoons, creates dynamic visuals from static frames by simulating camera movement based on a roughly estimated depth of the scene. However, when dealing with complex scenes, the original implementation of depth estimation and inpainting needed for a coherent scene can be problematic, often resulting in visual artifacts and inconsistent rendering, as in Fig. 7e. To address this, we leverage the instance awareness gathered from our proposed instance segmentation model, for a refined scene depth estimation and inpainting for occluded areas. In our implementation, we first obtain the initial depth using a refined depth estimator [35]. After that, we apply an average pooling filter on the initial depth of each subject instance. As in Fig. 7c, our depth estimation better differentiates between the foreground and the background. For more accurate inpainting of the background textures, we first isolate and remove the foreground subjects. This ensures that the inpainting models do not utilize any material from the subject areas. After that, we utilize a customized latent diffusion [25] model to achieve high-quality and consistent inpainting on foreground instances and occluded pixels (Fig. 7d). We show the final rendered effects in Fig. 7f. Compared to the original pipeline, we can achieve more engaging and immersive visual experiences, thanks to instance-aware optimization steps for both the geometry and texture during rendering.

Instance-aware style editing. Image-to-image style editing and translation for cartoons are rapidly evolving, with existing solutions demonstrating potential in style translation between Japanese anime and Western comics [34] or from photos to anime styles [5]. However, they struggle with arbitrary style translation due to a lack of high-level semantic understanding and the ability to synthesize high-quality textures and shadings. Recent trends favor diffusion models [10, 29] for their superior generation quality and flexibility, using texts to condition the generation process. However, without instance-level guidance, these models often fail to preserve the identity and visual semantics of instances. As shown in Fig. 8b, instances may lose original visual properties, such as hair colors, background semantics, and face expressions, and fail to fully adhere to the given style prompts (e.g., *3D pixar style*). This issue arises from the indiscriminate nature of CLIP embedding injection during the denoising step, leading to confusion in applying proper prompts to specific visual elements. A uniform style prompt also cannot provide sufficient semantic guidance, resulting in decreased

Fig. 8 Instance-aware style editing with global prompt *3D pixar style* (top), *artist_cg, cg* (bottom). Our solution respects the global style prompt while maintaining visual properties and semantics (e.g., hair color) for all instances

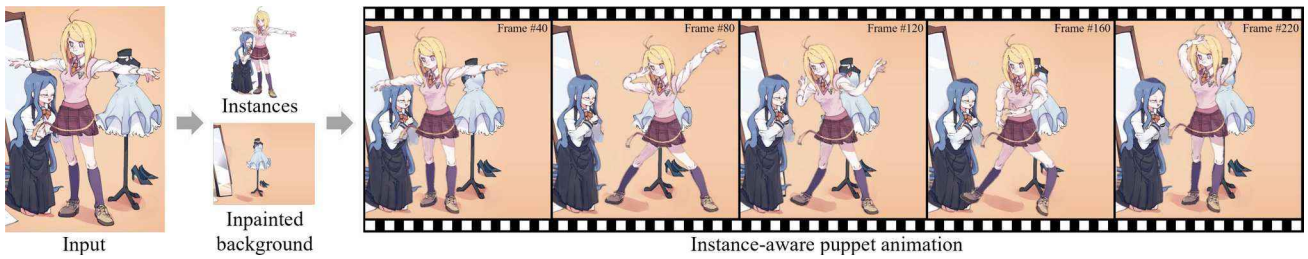

Fig. 9 Our instance-level segmentation and inpainting streamlines the making of puppet animation from still illustrations

visual quality, such as blurry edges and halo artifacts. We address these visual discrepancy issues with an instance-level diffusion-based solution for consistent style translation. For each extracted instance, we compute its visual tags using the SwinV2 tagger [27] as a supplement to the global visual tags for style editing. We setup the style editing with the same noise application and denoising schedule for all instances and denoise each instance with its corresponding global and instance-specific visual tags. Additionally, we use the lineart ControlNet [36] as an additional structure constraint for the instances. Note that we also inpaint the background in the subject areas and treated it as a unique instance to maintain overall visual consistency after style editing. Our method maintains style consistency, achieving clearer and sharper results, as shown in Fig. 8c.

Puppet Animation from Illustration and Manga. The extraction of high-resolution subject instances enables creating puppet animation from single illustrations or manga images. Without individual instance masks, only basic deformation can be applied to the overall image, often resulting in distorted subjects or backgrounds. We propose a conceptual approach for generating puppet animation, as depicted in Fig. 9. The process begins with the extraction of individual instances, followed by the application of techniques presented in Animated Drawings [28] to create high-quality, time-varying deformation frames of the subject. To manage any missing pixels due to warping, we use the same inpainting technique employed in the 3D Ken Burns application to avoid creating additional subjects. Our proposed method allows for the creation of visually coherent puppet animation, either from a reference animation or through manual manipulation of image keypoints. Crucially, the movement of each subject operates independently, without any impact on other subjects or the background.

7 Discussion and conclusion

Generalization to comics and cartoons. We demonstrate one qualitative result for western comics in Fig. 5. Our method reliably identifies subjects and produces accurate boundaries, particularly in challenging regions such as hands. While some overlap errors may occur (e.g., a hand assigned to both the seated woman and the standing man), our model

Fig. 10 Limitations of our approach. The segmentation model may be confused by too crowded contents, or major occlusion

still provides clearer and more precise instance boundaries than competing approaches, demonstrating generalization to other hand-drawn style contents. Additionally, we also tested the performances qualitatively and found that our model achieves state-of-the-art performance on manga109, and on-par performance with Grounded SAM on western comics. In general, our approach generalizes to color comics but performs less effectively on old black-and-white comics and animations or fine-art domains like oil paintings. Refer to the supplementary for more details.

Limitations. One potential issue arises when handling images with a high degree of subject crowding, particularly if the subjects share similar colors and styles. As shown in the left image of Fig. 10, our method can struggle to identify the belonging of the dark pixels in the lower middle part. Furthermore, occlusions can cause discontinuity in the subject area, which can cause small detached areas to be misidentified as part of the occluding subject. As shown in the right image of Fig. 10, the right hand of the girl in the back is incorrectly identified as part of the girl in the front. In rare cases, our model may misrecognize thin elements as hairs to produce false results, such as in the first row of Fig. 5. Our model may struggle when the hairs are extremely thin as 1–2 pixels. In such situations, matting techniques may be more suitable. Besides, our model primarily identifies human characters due to biases of the data set, but can also extract non-human subjects such as cats and robots. When full automation is not feasible, semiautomatic extraction remains possible by manually providing bounding box prompts. We demonstrate one example in the supplementary.

*Conclusion.*In this work, we propose to improve anime editing pipelines with a large-scale segmentation dataset and a novel two-stage segmentation framework, producing high-quality instance masks. Our dataset is prepared and synthesized using a reverse engineering approach based on chroma-keying videos and illustrations. Our method shows significant improvements over existing techniques in qualitative and quantitative evaluations, enabling high-quality downstream applications for modifying anime and cartoons, with more uses and applications to be explored in future.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00371-025-03970-1.

Acknowledgements This work was fully supported by the Research Grants Council of the Hong Kong SAR, China (Project No. UGC/FDS11/E02/23).

Author Contributions J.L. was primarily responsible for designing and conducting the experiments, as well as performing the quantitative and qualitative evaluations presented in the manuscript. C.L. supervised the project, providing overarching guidance, project management, and critical discussions that shaped the research direction. X.L. contributed by offering essential feedback during all stages of the work, assisting with the development of core ideas and providing insights that improved the quality and rigor of the project. Z.G. participated in the experimental evaluation and was solely responsible for programming and developing all demonstration applications related to this project, ensuring the practical utility of the proposed methods. All authors contributed to discussions of the results and collaboratively reviewed and approved the final manuscript.

Data Availability Codes and datasets avaiable at: https://github.com/CartoonSegmentation/CartoonSegmentation

Competing interests The authors declare no competing interests.

References

1. Anonymous, community, D., Branwen, G.: Danbooru2021: A large-scale crowdsourced and tagged anime illustration dataset. https://gwern.net/danbooru2021 (2022). Accessed 07 Mar 2024
2. Bhattacharjee, D., Süsstrunk, S., Salzmann, M.: Dense multitask learning to reconfigure comics. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 5646–5655 (2023). https://doi.org/10.1109/CVPRW59228.2023.00598
3. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
4. Chen, S., Zwicker, M.: Transfer learning for pose estimation of illustrated characters. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2022)
5. Chen, Y., Lai, Y.K., Liu, Y.J.: Cartoongan: Generative adversarial networks for photo cartoonization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9465–9474 (2018)
6. Cheng, B., Girshick, R., Dollár, P., Berg, A.C., Kirillov, A.: Boundary iou: improving object-centric image segmentation evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15334–15342 (2021)
7. Fujimoto, A., Ogawa, T., Yamamoto, K., Matsui, Y., Yamasaki, T., Aizawa, K.: Manga109 dataset and creation of metadata. In: Proceedings of the 1st International Workshop on CoMics ANalysis, Processing and Understanding, MANPU '16. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/3011549.3011551
8. Grönquist, P., Bhattacharjee, D., Aydemir, B., Ozaydin, B., Zhang, T., Salzmann, M., Süsstrunk, S.: Unlocking comics: the ai4va dataset for visual understanding. Workingpaper, arXiv ECCV 2024 Workshop Proceedings (2024)
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988 (2017). https://doi.org/10.1109/ICCV.2017.322

10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle H., Ranzato M., Hadsell R., Balcan M., Lin H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, Virtual (2020)
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
12. Li, J., Shahjahan, T.: Aniseg. https://github.com/jerryli27/AniSeg (2020)
13. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection. Adv. Neural. Inf. Process. Syst. **33**, 21002–21012 (2020)
14. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision - ECCV 2014, pp. 740–755. Springer, Cham (2014)
15. Liu, S., Qi, L., Qin, H., Shi, J., Jia, J.: Path aggregation network for instance segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8759–8768 (2018)
16. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J. et al.: Grounding dino: marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)
17. Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S., Chen, K.: Rtmdet: an empirical study of designing real-time object detectors (2022)
18. Niklaus, S., Mai, L., Yang, J., Liu, F.: 3d ken burns effect from a single image. ACM Trans. Graph. **38**(6), 184:1-184:15 (2019)
19. Qin, X., Dai, H., Hu, X., Fan, D.P., Shao, L., Gool, L.V.: Highly accurate dichotomous image segmentation. In: ECCV (2022)
20. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O., Jagersand, M.: U2-net: Going Deeper with Nested u-Structure for Salient Object Detection, p. 107404 (2020)
21. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
22. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Cortes C., Lawrence N.D., Lee DD, Sugiyama M., Garnett R. (eds.) Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pp. 91–99 (2015)
23. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded sam: Assembling open-world models for diverse visual tasks (2024)
24. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 658–666 (2019)
25. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10,674–10,685. IEEE Computer Society, Los Alamitos, CA, USA (2022). https://doi.org/10.1109/CVPR52688.2022.01042
26. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18, pp. 234–241. Springer, New York (2015)
27. SmilingWolf: Sw-cv-modelzoo. https://github.com/SmilingWolf/SW-CV-ModelZoo (2023). Accessed 07 Mar 2024
28. Smith, H.J., Zheng, Q., Li, Y., Jain, S., Hodgins, J.K.: A method for animating children's drawings of the human figure. ACM Trans. Graph. (2023). https://doi.org/10.1145/3592788
29. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv:2010.02502 (2020)
30. Sun, Y., Tang, C.K., Tai, Y.W.: Human instance matting via mutual guidance and multi-instance refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2647–2656 (2022)
31. Tian, Z., Zhang, B., Chen, H., Shen, C.: Instance and panoptic segmentation using conditional convolutions. IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI) **45**(1), 669–680 (2022)
32. Wang, X., Zhang, R., Kong, T., Li, L., Shen, C.: SOLOv2: Dynamic and fast instance segmentation. In: Proc. Advances in Neural Information Processing Systems (NeurIPS) (2020)
33. Wei, J., Wang, S., Huang, Q.: F^3net: fusion, feedback and focus for salient object detection. Proceedings of the AAAI Conference on Artificial Intelligence **34**, 12,321-12,328 (2020)
34. Xie, M., Li, C., Liu, X., Wong, T.T.: Manga filling style conversion with screentone variational autoencoder. ACM Trans. Graph. (TOG) **39**(6), 1–15 (2020)
35. Yin, W., Zhang, J., Wang, O., Niklaus, S., Chen, S., Liu, Y., Shen, C.: Towards accurate reconstruction of 3d scene shape from a single monocular image. IEEE Trans. Pattern Anal. Mach. Intell. **45**(5), 6480–6494 (2022)
36. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023)
37. zymk9: Yet-Another-anime-segmenter. https://github.com/zymk9/Yet-Another-Anime-Segmenter (2020)

Jian Lin received his B.Eng. degree in Computer Science and Engineering from Chongqing University of Posts and Telecommunications in 2021. He is currently a research assistant at the School of Computing and Information Sciences, Saint Francis University. His research interests include computer vision and computer graphics.

Chengze Li received their B.Eng. degree from University of Science and Technology of China in 2013 and PhD degree in Computer Science and Engineering from the Chinese University of Hong Kong in 2020. They are currently an Assistant Professor in the School of Computing and Information Sciences, Saint Francis University. Their research interests include 2D non-photorealistic media analysis and processing, computational photography, and computer graphics.

Xueting Liu received her B.Eng. degree in Computer Science and Technology from Tsinghua University and PhD degree in Computer Science from The Chinese University of Hong Kong in 2009 and 2014, respectively. She is currently an Assistant Professor in the School of Computing and Information Sciences, Saint Francis University. Her research interests include computer graphics, computer vision, and computational/intelligent art.

Zhongping Ge (a.k.a. *Francis Simpson*) is currently pursuing a Master of Computing (Computer Science) at Curtin University. He received his B.Eng. degree Shenyang Agricultural University. During his tenure as a Research Assistant at Caritas Institute of Higher Education (now Saint Francis University), he contributed to full development cycle of Python-based graphical user interfaces supporting AI research projects. He also provided contribution to tooling developments to facilitate research projects in the institute.