Small facial image dataset augmentation using conditional GANs based on incomplete edge feature input

Shih-Kai Hung[1], John Q. Gan[1]

[1] School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK

Corresponding Author:
Shih-Kai Hung[1]
Wivenhoe Park, Colchester, CO4 3SQ, UK
Email address: sh19143@essex.ac.uk

Abstract

Image data collection and labelling is costly or difficult in many real applications. Generating diverse and controllable images using conditional generative adversarial networks (GANs) for data augmentation from a small dataset is promising but challenging as deep convolutional neural networks need a large training dataset to achieve reasonable performance in general. However, unlabeled and incomplete features (e.g., unintegral edges, simplified lines, hand-drawn sketches, discontinuous geometry shapes, etc.) can be conveniently obtained through pre-processing the training images and can be used for image data augmentation. This paper proposes a conditional GAN framework for facial image augmentation using a very small training dataset and incomplete or modified edge features as conditional input for diversity. The proposed method defines a new domain or space for refining interim images to prevent overfitting caused by using a very small training dataset and enhance the tolerance of distortions caused by incomplete edge features, which effectively improves the quality of facial image augmentation with diversity. Experimental results have shown that the proposed method can generate high-quality images of good diversity when the GANs are trained using very sparse edges and a small number of training samples. Compared to the state-of-the-art edge-to-image translation methods that directly convert sparse edges to images, when using a small training dataset, the proposed conditional GAN framework can generate facial images with desirable diversity and acceptable distortions for dataset augmentation and significantly outperform the existing methods in terms of the quality of synthesised images, evaluated by Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) scores.

Introduction

Deep convolutional neural networks generally require a large amount of training data to improve accuracy on unseen data or testing data (*Tan et al., 2018*). For many practical applications, it is expensive to collect a large amount of training data for deep learning. Data augmentation is often used to generate more training data. Traditional approaches for data augmentation include geometric transformations, such as translation, scaling, flip, rotation, etc., especially for image data (*Mikolajczyk & Grochowski, 2018*). However, the diversity introduced by traditional augmentation methods is limited and insufficient for many applications. The motivation of this work is to augment a small image dataset by making use of conditional edge features extracted from the available training images, and it can be expected that the synthesised images are more diverse and less distortive than those obtained from traditional methods.

In recent years, conditional generative adversarial networks (GANs), which can generate photorealistic images from conditional data, have become one of the most popular research fields in image synthesis. Conditional inputs, such as edges, mark points, masks, semantic maps, labels and so on, can be used to manipulate the images generated by GANs, making the synthesised images not only diverse but also controllable (*Isola et al., 2017*; *Torrado et al., 2020*). Image-to-image translation methods using conditional GANs directly learn the pixel mapping relationship between conditional edge features and real images (*Lin et al., 2018*; *Wang & Gupta, 2016*). Although image-to-image methods based on conditional GANs have been developed for controllable image synthesis, there are still several problems that should be resolved when applying them on a small training dataset: 1) Compared with unconditional GANs, a limitation of conditional GANs is that the output images must be generated from the corresponding conditional inputs, hence a clear mapping relationship between input and output should be correctly established. Corresponding mapping relationships are hard to be discovered, especially when only a very small training dataset is available for deep neural networks to learn. 2) With a small training dataset, the training process for image-to-image translation is easy to converge but difficult to obtain high-quality results due to the overfitting problem and insufficient information about the underlying data distribution, whether the used conditional features are of high quality or not (*Dimitrakopoulos, Sfikas & Nikou, 2020*), the discriminator will overfit the training data, and the generator would produce unexpected distortions in the generated images in the validation or application phase (*Arjovsky & Bottou, 2017*; *Gu et al., 2019*). 3) Training GANs using a small dataset must deal with the inevitable problem of mode collapse (*Zhao, Cong & Carin, 2020*), which implies that the GANs may learn the training data distribution from a limited number samples only but overlook other useful training data (Srivastava et al., 2017). Other issues, such as non-convergence and instability, would also worsen the quality of the generated images (*Salimans et al., 2016*). Clearly, it is a challenging task to synthesise photorealistic images using conditional GANs based on incomplete conditional features and a small number of training images.

Edge-based image-to-image translation using conditional GANs has the advantage of introducing diversity in image data augmentation, but it is challenging in terms of generating high-quality

80   photorealistic images. Extracted edges cannot be regarded as perfect conditional features that
81   support various advanced visual tasks and contain all visual information of potential perceptual
82   relevance (*Elder, 1999*). Since edges generally contain incomplete information, such as
83   unintegral geometry, simplified lines, discontinuous shapes, missing components, and undefined
84   contours, it is hard for image-to-image translation methods to map edges to realistic images
85   without clear conditional information.
86   To simply demonstrate the impact of the number of training images and incomplete conditional
87   edges on the quality of the images generated using conditional GANs, some preliminary
88   experimental results are shown in Figure 1. It can be observed that the level of detail in the input
89   edge features and the number of training images have considerable influence on the quality of
90   both training and inference results, which is more serious on the inference results though because
91   the input edge features used in the inference process have not been used in the training process.
92   In this paper, a new image-to-image translation framework using conditional GANs is proposed,
93   which can generate diverse photorealistic images from limited edge features after training with a
94   small number of training images. Instead of deepening the convolutional layers or increasing the
95   number of parameters, the proposed conditional GAN framework can learn additional
96   relationships between incomplete edges and corresponding images, because regional binarisation
97   and segmentation masks are used as new reference information, which can be obtained
98   automatically by image processing. In particular, the proposed method can beneficially obtain
99   extra pixel correlations between conditional edge inputs and the corresponding ground truth
100   images to mitigate the influence of the overfitting problem during training. If the conditional
101   GANs can efficiently learn from informative conditional inputs, such as colour, texture, edges,
102   labels, etc., then it would be effortless to generate corresponding photorealistic image outputs
103   (*Dosovitskiy & Brox, 2016*; *Wei et al., 2018*). A new network structure is proposed in this paper
104   to divide the image synthesis task into two stages: the first stage transforms conditional input
105   with incomplete edges into refined images as the new conditional input for the second stage,
106   whose pixel values are obtained by combining the information from segmentation masks and
107   binarised images. The second stage transforms the refined images into photorealistic image
108   outputs.
109   The experimental results have demonstrated that the proposed method can generate high-quality
110   diverse images even with a very small training dataset and very sparse edge features as
111   conditional input. In addition, the generated images do not contain large distortions from
112   incomplete or modified edge inputs for data augmentation purposes. The contributions of this
113   paper are as follows:
114   • In order to deal with the problem of distortions in the images generated by conditional
115   GANs due to using very small training data and sparse conditional input for diverse data
116   augmentation, a new conditional GAN framework has been proposed, which converts the
117   original incomplete edges into new conditional inputs in an interim domain for refining
118   images and thus alleviates distortions caused by small training data and incomplete
119   conditional edges.

- For the first time, the proposed method uses the mixture of pixel values of both binary images and segmentation masks to enhance the conditional input in an interim domain for refining images, which can integrate facial components, including eyes, nose, mouth, etc., so as to introduce diversity and enhance the quality of the images generated by conditional GANs trained using a very small training dataset.

- A facial image augmentation method using conditional GANs has been proposed, which can generate photorealistic facial images of diversity from incomplete edges or hand-drawn sketches. Compared with existing edge-to-image translation methods without ideal conditional inputs, the proposed method is tolerant to various incomplete edges as conditional inputs and able to generate diverse images of higher quality in terms of Fréchet Inception Distance (FID) (*Heusel et al., 2017*) and Kernel Inception Distance (KID) (*Binkowski et al., 2018*).

This paper is organised as follows: In section 2, the related work is reviewed, including methods for image data augmentation, image-to-image translation, challenges in training GANs using small datasets. In section 3, the proposed methods, including the proposed conditional GAN framework, image pre-processing for generating the interim domain, and training strategies, are described in detail. Section 4 describes the experiment design, including data preparation and implementation details. The experimental results are presented and the performance of the proposed conditional GAN framework is evaluated qualitatively and quantitatively in section 5. Finally, conclusions, limitations and future work suggestions are presented in section 6.

Related Work

The method proposed in this paper aims to use incomplete or modified edge features to augment small facial image datasets. The method was based on image-to-image translation using conditional GANs. This section introduces related techniques for image synthesis using conditional GANs.

Image Data Augmentation

Image data augmentation is applied in many applications to increase dataset size and data diversity. Deep neural networks are not easy to be trained well using small training datasets due to overfitting problems (*Bartlett et al., 2020*). One of the solutions to overcome overfitting is training data augmentation, which intends to boost the diversity of a small training dataset (*Shorten & Khoshgoftaar, 2019*). Traditional methods such as rotation, reflecting, translation, scaling, cropping, blurring, grey scaling and colour converting, have been commonly used for image data augmentation to reduce overfitting when training deep neural networks for image classification applications. Although these traditional techniques can produce similar images of high-quality, they rarely enlarge the feature diversity in the original images. Therefore, developing novel methods image data augumentation, which can synthesise not only diverse but also photorealistic images based on a small image dataset, is a very important but challenging task.

159   Many GANs have been proposed for image synthesis (*Gatys, Ecker & Bethge, 2016*; *Iizuka,*
160   *Simo-Serra & Ishikawa, 2017*). Since conditional GANs provide alternative methods to edit
161   images as well as manipulate generative attributes, they have been applied to generate high-
162   quality images of diverse features (*Lin et al., 2018a*). Image synthesis using conditional GANs
163   can boost the applications of data augmentation in many real applications, but it is hard to
164   generate high-quality diverse images when the training dataset is small.

165   **Image-to-image Translation and Image Synthesis Using Conditional GANs**

166   Image-to-image translation is a type of image synthesis using conditional GANs with specific
167   forms of conditions, such as videos and images (*Azadi et al., 2018*; *Szeto et al., 2021*), scenes
168   (*Ashual & Wolf, 2019*; *Johnson, Gupta & Li, 2018*), or segmentation masks (*Cherian & Sullivan,*
169   *2018*; *Park et al., 2019*). The conditional inputs can be transferred from a source domain to a
170   target domain by using supervised learning techniques. The main concept of image-to-image
171   translation is to learn the data mapping relationships (*Liu, Breuel & Kautz, 2017*). Image-to-
172   image translation methods can automatically generate images between the corresponding
173   domains (*Mo, Cho & Shin, 2019*) and discover the dependence with pairs of images to translate
174   features into realistic images. Image-to-image translation methods provide a prominent approach
175   to image synthesis with diverse results by using controllable conditional features (*Isola et al.,*
176   *2017*; *Zhu et al., 2017*).

177   With remarkable advantages in image-to-image translation, edge-to-image synthesis has
178   achieved visually pleasing performance (*Yi et al., 2017*). Compared with other conditional forms,
179   edges are one of the most easy-to-obtain and simple features in computer vision, and hand-drawn
180   sketches can be regarded as a specific form of edge features (*Chen & Hays, 2018*). Since edges
181   usually contain critical information, such as gradients, shapes, contours, profiles, boundaries and
182   so on, they directly provide simple and direct depictions for objects (*Chen et al., 2009*; *Eitz et al,*
183   *2011*). The benefit of using edge features is that they are flexible to modify for introducing
184   diversity in data augmentation. However, in contrast to labelled images, edges as conditional
185   inputs often preserve insufficient information (e.g., textures, colours, brightness, labels, etc.),
186   which makes it hard to generate desirable high-quality images using edge-to-image translation.
187   Face synthesis requires integral contours that faithfully reflect the inputs and connections to the
188   realistic context (*Li et al., 2020*). Incomplete features with missing parts or undefined
189   components will affect the quality of images generated by conditional GANs (*Jo & Park, 2019*;
190   *Karras et al., 2018*). To alleviate the influence of lacking ideal conditional inputs, this work
191   proposes to reconstruct the missed features in the conditional input and transform limited edge
192   information into refined images by introducing an interim domain to alleviate the negative effect
193   of imperfect conditional features on the quality of the generated images.

194   **Challenges in Training GANs Using Small Training Dataset**

195   Mode collapse is a major problem in training GANs using small training data, which makes it
196   fail to switch training samples and thus the trained GANs would be over-optimised with a
197   limited number of training samples only (*Odena, Olah & Shlens, 2017*). Due to the mode
198   collapse problem, the generator is constructed with very limited training data, which makes the
199   discriminator believe that the generative outputs are real instead of fake. This is one of the key

200 causes that make it difficult for GANs to generate diverse results for data augmentation purposes
201 (*De Cao & Kipf, 2018*).
202 Another drawback in training GANs using a small training dataset is that it is hard to fine-tune
203 all parameters to discover an optimal balance between the generator and discriminator. Both the
204 generator and discriminator contain a large number of trainable parameters, which necessarily
205 need enormous training data to prevent from overfitting problems (*Gulrajani et al., 2017*;
206 *Mescheder, Geiger & Nowozin, 2018*). Ideally, training a GAN requires a large amount of
207 training data for optimising its parameters and reducing losses to generate photorealistic image
208 outputs (*Wang et al., 2018c*). Therefore, reducing the requirement for a large amount of training
209 data is a grand challenge in using GANs for image data augmentation in which available training
210 data is limited. Our previous preliminary work (*Hung & Gan, 2021*) partly addressed the above
211 challenges by proposing a new conditional GAN architecture. This paper substantially extends
212 our preliminary work via further investigation and deeper analysis of much more experimental
213 results.
214
Methods
216 Image-to-image translation methods find specific mapping relationships between source
217 distribution and target distribution. In general, a small number of paired features may not
218 comprehensively align with the source and target distributions using imperfect conditional inputs
219 such as incomplete edges and a small training image dataset. Therefore, data refining in paired
220 features can be adopted to expand mapping relationships based on a small training dataset. In this
221 section, a new method is proposed for facial image synthesis based on a very small training
222 image dataset.
223 With incomplete conditional features in the source domain and small training data in the target
224 domain, the mapping between source and target domains cannot be described by clear one-to-one
225 relationships. The method proposed in this paper transfers the source domain to an interim
226 domain for refining images with extra annotated information, in which newly defined images in
227 the interim domain need to be generated based on a small training dataset. The interim domain
228 can provide extra reference information to discover more mapping relationships between the
229 source and target domains.
230 Figure 2 shows the proposed translation method using a small training dataset. It is difficult
231 using a small training dataset to obtain a comprehensive view of correct mapping relationships
232 between the source domain and target domain without sufficient representative training samples,
233 as shown by the blue line in Figure 2. Even if changing the types of the conditional inputs, a
234 similar situation remains as it is still difficult to learn correct mapping relationships, as
235 demonstrated by the red line in Figure 2. For the purpose of comprehensively finding correct
236 mapping relationships, extending the mapping relationships in an interim domain for refining
237 images, as shown by the green dotted line in Figure 2, can reduce uncertainty caused by using a
238 small training dataset and incomplete edge features as conditional inputs for data augmentation

239 with diversity. This will be explained in more detail when introducing the proposed conditional
240 GAN framework later.
241 When training GANs using small training datasets, the following factors should be considered:
242 1) It is difficult to avoid distortions in the generated images and training imbalance with a small
243 number of training images or insufficient diverse samples. 2) Through deep convolutional neural
244 network structures, such as convolution, normalisation and downsampling, it is easy to lose
245 spatial information and impractical to completely preserve the conditional information with a
246 small number of training images (*Kinoshita & Kiya, 2020*; *Wang et al., 2018a*). If the conditional
247 inputs contain sparse, unclear, limited, discontinuous, or incomplete features, fine-tuning model
248 parameters without distortions becomes much more difficult. 3) Using a small training dataset
249 and limited conditional features will make the training easy to overfit but hard to obtain realistic
250 results. Since many parameters in a deep convolutional neural network need to be fine-tuned, it
251 is impossible to optimise all the parameter values using a small training dataset in terms of the
252 generalisation ability of the trained deep neural network. To tackle the above problems in
253 training GANs using a small number of training images, several training strategies are adopted in
254 the proposed method, which are described as follows:
255 ***Enlarging the diversity of source domain:*** The training of a GAN using a small training dataset
256 can easily converge but frequently attain unrealistic inference results, mainly because of the
257 overfitting problem. It is impossible for GANs to have a whole view of the target domain
258 through training with a limited number of training images. For the goal of achieving
259 photorealistic results, both the discriminator and generator should stay in an equilibrium balance.
260 Increasing the data diversity and widening the mapping relationship between the source domain
261 and target domain could help achieve the required balance between the generator and
262 discriminator when using a small number of training images. In the proposed method, new
263 reference information is created by image pre-processing, and the adoption of the interim domain
264 for refining images can enlarge the diversity of the source domain.
265 ***Double translation:*** Double translation strategy aims to decrease the chance of mode collapse in
266 single translation approach and reduce the impact of the uncertainty due to using incomplete
267 edges as conditional input, so as to alleviate the distortions caused by training with a small
268 number of training images. For generating additional reference information, the proposed
269 method combines binary images and segmentation masks to generate refined images as
270 conditional input in the next translation. In the first translation, refined images with annotated
271 facial components are generated from incomplete edge features. This translation is conducted
272 between the source domain and the interim domain. The second translation is conducted between
273 the interim domain and target domain, which can successively learn from the possible distortions
274 in the first stage to avoid or alleviate negative distortions in the final outputs.
275 ***Reusing the conditional information:*** Spatial information in the conditional edges can be easily
276 lost during training in the convolutional neural layers, and the relationships between the source
277 domain and target domain will become incomprehensive. In order to reduce the spatial
278 information vanishing, edge features in the source domain can be reused in each translation.

279   *Freezing weights:* Weight freezing is a strategy to overcome the gradient vanishing problem
280   during training, which often happens when using a small training dataset. If the provided training
281   data cannot give the discriminator enough information to progress the generator, the gradient will
282   become smaller and smaller when going from bottom to top layers of the network. Incomplete
283   conditions would worsen the gradient vanishing problem and make it impossible to fine-tune the
284   model parameters to obtain realistic results. Hence, freezing part of weights in separate training
285   stages allows the discriminator to acquire information from each training stage rather than tuning
286   all parameters at one time.

287   **The Proposed Conditional GAN Framework**

288   To mitigate the output distortions caused by using a small training dataset and incomplete edge
289   as conditional input, additional paired segmentation masks and regional binary images are used
290   as reference information in the proposed method, which can enrich the mapping relationships
291   between the source domain and target domain. Consequently, the proposed method creates
292   additional data distributions from the small training dataset using image pre-processing, and the
293   data in the interim domain provides more referable features than the original incomplete edges in
294   the source domain.

295   Two U-nets (*Ibtehaz & Rahman, 2020*; *Ronneberger, Fischer & Brox, 2015*) are adopted in the
296   proposed conditional GAN framework for image-to-image translation. This structure can achieve
297   better performance when training conditional GANs using small training data for two reasons: on
298   one hand, during the training process the U-nets create images based on the special
299   concatenating structure, which is beneficial to retain the matched features from limited
300   conditional features for integral perceptions in convolutional layers. On the other hand, the U-net
301   structure is simple and beneficial to generate images without using very deep convolutional
302   layers, which is critical to alleviate the gradient vanishing problem during training using small
303   training data and incomplete edges. The proposed framework also reuses the conditional input
304   information to strengthen the input features at each training stage, and freezing weights for
305   separate networks at each training stage can prevent from gradient vanishing as well. To sum up,
306   the proposed conditional GAN framework can alleviate the problems in training GANs using
307   small training data by intensifying the conditional information in the source domain. An
308   overview of the proposed conditional GAN framework is shown in Figure 3 and described as
309   follows.

310   The proposed model consists of three parts: 1) image pre-processing, 2) generators and 3)
311   discriminators. The two generators use the same convolutional structure of the U-net, both of
312   which down-sample and then up-sample to the original size of input images. All convolutional
313   layers use convolution kernels of size 3×3 (*Yu et al., 2019*), and normalisation is applied to all
314   convolutional layers except for the input and output layers (*Zhou & Yang, 2019*). In the training
315   phase, the first generator is used to create refined images based on the original sparse edges and
316   ground truth. The refined images are referred from image pre-processing, which contain features
317   related to texture, colour, shape of different facial components. The second generator is designed
318   to improve the synthetic process to generate photorealistic images from the interim domain. In

319 the inference phase, the generators use the fine-tuned parameters to generate photorealistic
320 images from conditional edges that may have not been seen during training. The two
321 discriminators have the same task of distinguishing between real and fake images: the first one is
322 to identify generated images in terms of refined images, and the other is in terms of the ground
323 truth.

324 **Image Pre-processing and Refining**

325 Image refining is essential for providing informative conditional features since incomplete edges
326 may contain much unidentical information representing the same facial component. This
327 uncertainty makes it difficult for conditional GANs to comprehensively find pixel relevance in
328 different domains. For instance, an unclear "black circle" with incomplete edges can represent
329 either nose, ear or eye, even if using a powerful network, it is difficult to learn well with a rare
330 sign of "circle" as conditional input without any other crucial information (e.g., colour, types,
331 angels, positions, textures, sub-components, brightness, layouts, shapes, etc.). A refining process
332 can be employed within an image-to-image translation method, which enhances powerful one-to-
333 one mapping by providing close to ideal conditional input. However, there is no guarantee that
334 ideal conditional inputs can be obtained in real applications, especially if the conditional inputs
335 are incomplete or sparse edges. These uncertainties can result in unexpected distortions. If the
336 interim domain for refining images can provide more specific information, the synthetic quality
337 will be improved. Therefore, enhancing conditional information is one of the important goals of
338 image pre-processing and refining.

339 *Edge Extraction:* Edges may contain incomplete features with many possible feature types,
340 including undefined density, shape or geometry (*Royer et al., 2017*). However, to achieve high
341 performance, image-to-image translation methods need clear conditions (*Lee et al., 2019*). In
342 order to generate photorealistic images from limited conditional information, extending
343 translation relationships based on proper reference images can make the mapping relationships
344 between the source domain and target domain more precise based on a small training dataset. As
345 an example, the corresponding mapping relationships among the ground truth, conditional
346 features, and refined image are shown in Figure 4. The ground truth image is responsible for
347 providing not only realistic features but also reference images to composite refined images. The
348 red boxes shown in Figure 4 indicate the eye mapping in different domains, and the new
349 relationships are expected to effectively reduce the uncertain mapping in image-to-image
350 translation.

351 *Adoption of an Interim Domain:* In contrast to directly transforming the source edges to target
352 results, the proposed conditional GAN framework converts conditional edges to refined images
353 in an interim domain first. The interim domain trends to reconstruct possible missing information
354 from incomplete conditional features using a U-net. Mode collapse problem may happen in the
355 interim domain when incomplete edges are transferred to a refined image. Nevertheless, the
356 translation at this stage is useful for facial component identification because the incomplete
357 edges in the source domain are further processed. The refined images provide clearer accessorial
358 information than the original incomplete edges, even if they are converted into simplified

359 samples when mode collapse happens. By trial and error, appropriate regional features as
360 reference images can reduce distortions and mismatch of features based on very limited edges.
361 Therefore, the refined images are constructed by combining binarised images and segmentation
362 masks, as shown in Figure 3. In short, the main function of the interim domain is to refine the
363 original data distribution and strengthen the incomplete conditional features from the source
364 domain.
365 Figure 5 (a) shows inference results of using uncertain edges to generate segmentation masks
366 from 50 paired training images. To handle the incomplete edges in the conditional input, facial
367 components can be reconstructed by a U-net in the interim domain. It is evident that the
368 proposed conditional GAN can learn from only 50 segmentation masks to generate more integral
369 face components, such as nose, eyebrow, hair and mouth. Figure 5 (b) illustrates examples where
370 incorrect eye shapes are obtained, as shown in the red boxes, which would aggravate distortions
371 in the target domain. What is worse, this situation is hard to be solved because it is difficult to
372 increase the number of diverse samples based on a small dataset as GANs generally require more
373 diverse data to be trained well. To resolve this problem, binary images with clear regional
374 information are obtained through image pre-processing, which can enhance the contours and thus
375 reduce distortions in facial components, as shown in Figure 4. In contrast to imprecisely
376 depicting facial components in segmentation masks, binary images obtained by appropriate
377 thresholding can produce more correct contours than segmentation masks and thus alleviate the
378 problem caused by very limited training data.
379 Figure 6 shows that binary images can handle uncertain edge density in the inference phase to
380 enhance crucial edge information with regional distributions. Binarised regional features can be
381 extracted by the corresponding edge distribution from a small training dataset, which can not
382 only integrate crucial contours, as shown in Figure 6 (a), but also get rid of meaningless noise
383 when various untrained edges may be unrecognisable in the inference phase, as shown in Figure
384 6 (b). It is noteworthy that the results presented in Figures 5 and 6 can be regarded as those from
385 an ablation study, which shows that removing the component of combining binarised regional
386 features in the proposed method will significantly deteriorated the performance of the proposed
387 conditional GAN.
388 **Model Training and Loss Functions**
389 ***Conditional Adversarial Loss:*** During training the proposed conditional GAN framework, it is
390 difficult to find a balance between the generator and discriminator, especially when only very
391 limited training data is available. Using an appropriate loss function is critical to ensure good
392 quality of the generated images. Firstly, to distinguish real images from fake ones, the following
393 basic loss function is used for the two convolutional neural networks, which is known as
394 conditional adversarial loss.
395 $\mathcal{L}_{adv}(D,G) = \mathbb{E}_{I,S}[log\ D(S|I)] + \mathbb{E}_{I,I'}[log\ (1 - D(I, G(I'|I)))]$

396 where $\mathbb{E}$ represents expected value, G the generator, D the discriminator, S the source image, I
397 the conditional edge feature input, and I' the generated image. In the first U-net, S should contain
398 a mixture of pixels of binary image, segmentation mask and ground truth so as to distinguish

399   between real refined image and fake generated image. In the second U-net, *S* needs to be set as
400   the ground truth only.
401   ***Feature Matching Loss:*** As in the pix2pix GAN model (*Isola et al., 2017*), the L_1 pixel loss as
402   feature matching loss in the synthesised images is adopted. Since there are paired images in the
403   training phase, the L_1 distance between the generated image (I') and source image (S) can be
404   defined as follows:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{S,I,I'}[\|S - G(I'|I)\|_1]$$

405
406   ***Overall Loss:*** The main purpose of using the loss function is to help the generator to synthesise
407   photorealistic images by minimising the loss value with a limited number of input images. The
408   overall loss function is defined as

$$\min_{G} \max_{D} \; \mathcal{L}_{adv}(D,G) + \alpha\mathcal{L}_{L1}(G)$$

409
410   where α is a weight parameter. A larger value of α encourages the generator to synthesise images
411   less blurring in terms of L_1 distance.
412   The second U-net uses the refined images and original sparse lines as inputs to generate
413   photorealistic images with the same loss function but different training parameters and freezing
414   weights. Another difference between these two networks is the source image *S*, which should be
415   either the refined images or the ground truth images.
416
417   ## Experiments with the Proposed Conditional GAN Framework
418   **Data Preparation**
419   A small set of randomly chosen images from CelebA-HD (*Liu et al., 2015*) formed the training
420   image dataset in our experiments. CelebA-HD includes 30,000 high-resolution celebrity facial
421   images. All the images were resized to 256×256 for our proposed model. CelebAMask-HQ
422   (*Lee et al., 2020*) is a face image dataset consisting of 30,000 high-resolution face images of size
423   512×512 and 19 classes, including skin, nose, eyes, eyebrows, ears, mouth, lip, hair, hat,
424   eyeglass, earring, necklace, neck, cloth and so on. All the images in CelebAMask-HQ were
425   selected from the CelebA-HD dataset, and each image has segmentation masks of facial
426   attributes corresponding to CelebA-HD.
427   Since different numbers of segmentation masks were used to compare the performance of
428   different methods with different numbers of training samples, the CelebAMask-HQ was used as
429   the standard segmentation masks of reference images. If a very small training dataset is used, it
430   would be fine to manually generate the segmentation masks by image pre-processing. In our
431   experiments, the segmentation masks from CelebAMask-HQ were used as the common reference
432   images of the corresponding training images.
433   **Implementation Details**
434   The hyper-parameter values were determined through trial and error in our experiments. For
435   training the proposed conditional GAN framework, the Adam optimiser was used to minimise
436   the loss function with the initial learning rate set to 0.0002 and the momentum 0.5. The weight
437   parameter in the loss function α was set to 100. All the experiments were conducted on a desktop

438   computer with *NVIDIA GeForce RTX 2080 GPU, Intel Core i7-6700* (3.4GHz) processor, and
439   16G RAM.

440   Incomplete edges or hand-drawn sketches as conditional inputs usually represent abstract
441   concepts, which are beneficial for generating diverse data augmentation results but it is difficult
442   for conditional GANs to generate photorealistic images with limited conditional inputs based on
443   small training data. In our experiments, edges were extracted by Canny edge detector (*Canny,*
444   *1986*), which can obtain simple and continuous edges using a set of intensity gradients from
445   realistic images. Two intensity gradient magnitudes are used in Canny edge detector as a
446   threshold range to control the edge density, which is determined in our experiments by a
447   threshold ratio. It is the ratio of the high threshold value to the maximum magnitude, and the low
448   threshold value is 40% of the high threshold value. The edges produced by the Canny edge
449   detector are more similar to hand-drawn sketches than those by other commonly used edge
450   detectors, as shown in Figure 7. The appropriate threshold ratio for the Canny edge detector was
451   chosen through trial and error in our experiment. The red box in Figure 7 shows the edges
452   extracted using the chosen threshold ratio, which has clear information about facial components
453   without unexpected noise and meets the requirement for good conditional inputs.

454   In the design of the interim domain, the pixel values of the refined image were set by the
455   following mixture ratios: 25% from binary image, 25% from segmentation mask, and 50% from
456   original image. Figure 8 shows the inference results of the refined images and the corresponding
457   generated image outputs. The red boxes represent blending areas in the masks, binary images and
458   texture features in the refined images, which reflect the brightness changes in the generated
459   image outputs. The overlapped regions are visually darker and gloomier compared to other
460   regions. Therefore, these blending areas from different reference images conduct transitions in
461   brightness and lightness to synthesise realistic results. With the interim domain, the proposed
462   conditional GAN can efficiently deal with both overlapped and non-overlapped mappings
463   between segmentation masks and binary regions, leading to more photorealistic image outputs.

464   Image blending with different styles is beneficial to augment training image datasets with
465   diversity. In the proposed conditional GAN framework, generated images were controlled by
466   conditional edge inputs. Exchanging or modifying edge features is an easy way to generate
467   different images that increase data diversity and expand original facial features. Figure 9 shows
468   examples of the generated images with facial features swapped on a small training dataset by
469   exchanging edge components in conditional inputs. It can be seen that the generated images can
470   preserve facial features with clear conditional edges and reconstruct the critical components in
471   incomplete or undefined areas.

472

Results and Performance Evaluation

474   For performance evaluation, the proposed conditional GAN framework was used to generate
475   images from different training images and conditional edge input settings. To demonstrate the
476   performance of the proposed method, the state-of-the-art edge-to-image translation methods

477   were compared both qualitatively by visual inspection of the quality and diversity of the

478   generated images and quantitatively in terms of FID and KID scores.

479   **Diversity in Facial Image Augmentation Using the Proposed Conditional GAN**

480   It is clear that the threshold ratio chosen for the Canny edge detector affects the density level of

481   the extracted edges, which as conditional inputs would affect the quality of the images generated

482   by the conditional GAN. It is desirable that the proposed conditional GAN can generate diverse

483   images with the change of edge density levels in the conditional input but be robust in terms of

484   the quality of the generated images. Figure 10 shows the inference results with different density

485   levels in the conditional edges, which were not included in the training phase except for those in

486   the red box. It can be seen that the generated images are slightly different with different density

487   levels in the conditional edge inputs and the distortions are small even when the GANs were

488   trained using a small dataset of 50 training images. The generated images are more photorealistic

489   when the conditional input contains less noise or unidentical edges, which correspond to those

490   generated with the edge density level chosen in the training phase, as shown in the red box.

491   Fortunately, with the change of the density levels of the conditional edge inputs, the quality of

492   the generated images is prevented from considerable deterioration because the refined images

493   can integrally represent facial features at an acceptable level based on a small dataset.

494   Consequently, as the conditional inputs to the second U-net in the second stage, they play an

495   important role in reducing distortions in the generated facial image outputs.

496   Figure 11 shows examples of facial image augmentation results using 50 training images to train

497   the proposed conditional GAN framework. Diverse new facial images can be generated from

498   each training image with extracted edges modified for desirable facial features as conditional

499   inputs. The modifications to the extracted edges include adding or deleting parts of the edges or

500   changing facial expression or direction, as shown in Figure 11. It can be seen that the image data

501   augmentation results using the proposed conditional GAN are more diverse than traditional

502   augmentation methods and the generated images are of good quality due to the use of the interim

503   domain. For data augmentation purposes, using deliberately modified edges as conditional inputs

504   to the proposed conditional GAN framework can boost the data diversity on the basis of the

505   available small set of training images.

506   Figure 12 shows some other examples of facial image data augmentation using 50 training

507   images to train the proposed conditional GAN, with edge features from multiple training images

508   swapped as conditional inputs. The red boxes in the figure indicate the original training images,

509   and the other images in a row are generated by the proposed conditional GAN, which shows

510   swapped facial features (including eyes, eyebrows, nose and mouth) or hairstyles. It can be seen

511   that by using mixed edge features from multiple training images as conditional inputs the

512   proposed conditional GAN can efficiently generate diverse facial images of good quality.

513   In general, it is difficult for image-to-image translation methods to generate high-quality images

514   with conditional inputs that are not directly corresponding to features in the training images, such

515   as hand-drawn sketches. In the previous experiments, it has been demonstrated that the interim

516   domain is helpful to generate high-quality images with various edge density levels for diversity.

517 In our experiments, hand-drawn sketches were also used as conditional inputs for the proposed
518 conditional GAN to generate facial images with specific facial expressions or features. Figure 13
519 shows the inference results with hand-drawn sketches as conditional inputs, with the proposed
520 conditional GAN trained using a dataset of 50 training images. It is obvious that, when giving
521 unidentical or incomplete facial contours in the conditional inputs, the refined images generated
522 by the first U-net in the proposed conditional GAN structure are responsible for reducing
523 distortions in the generated images while keeping the diverse facial expressions introduced by
524 the hand-drawn sketches.
525 **Qualitative Comparison**
526 To evaluate the quality of the images generated by the proposed conditional GAN framework,
527 the images generated by the proposed method were compared with those generated by the state-
528 of-the-art edge-to-image translation methods, including pix2pix (*Isola et al, 2017*) and
529 pix2pixHD (*Wang et al., 2018a*), under the same training conditions and in terms how the
530 generated images are comparable to the ground truth images. Figure 14 shows representative
531 images generated respectively by the three conditional GANs for comparison, trained using the
532 same small dataset of 50 training images. Different sparse edges as conditional inputs were
533 tested. The results in Figure 14 demonstrate that the method proposed in this paper can generate
534 more photorealistic facial images with fewer distortions than pix2pix and pix2pixHD, when the
535 GANs were trained using small training data.
536 **Quantitative Comparison**
537 Since the number of training images is small, the difference between the generated images and
538 the corresponding ground truth images is noticeable by visual inspection. In order to
539 quantitatively compare the quality of the images generated by different conditional GANs, FID
540 and KID scores were adopted to evaluate the photorealistic scales of the generated images in our
541 further experiments. FID is widely adopted to evaluate the visual quality of generated images,
542 which calculates the Wasserstein distance between the generated images and the corresponding
543 ground truth images. KID can be used similarly for image quality measurement, but KID scores
544 are based on an unbiased estimator with a cubic kernel (*Binkowski et al., 2018*). Clearly, lower
545 FID and KID scores represent a better match between the generated images and the
546 corresponding ground truth images.
547 To evaluate the effect of the interim domain adopted in the proposed conditional GAN
548 framework on the quality of the generated images, the performance of double U-nets is
549 compared with that of a single U-net in terms of FID and KID with different threshold ratios
550 used in the Canny edge detector. In the training phase, one input type with threshold ratio = 0.4
551 and three input types with threshold ratios = 0.2, 0.4 and 0.6 were considered and 50 training
552 images were used, whilst in the inference phase 11 different threshold ratios (0.01, 0.05, 0.1, 0.2,
553 0.3, 0.4, 0.5, 0.6, 0.7, 0.8 and 0.9) were tested with 1000 generated images respectively. Figure
554 15 shows the comparative results in terms of FID and KID scores. Three points can be made
555 from the experimental results presented in this figure: Firstly, double U-nets achieved lower FID
556 and KID scores than single U-net, indicating that the interim domain for refining images in the

557   proposed conditional GAN can reduce distortions caused by incomplete conditional edges and

558   small training dataset, and thus improve the quality of the generated images; Secondly, training

559   with three conditional edge density levels, compared with only one conditional type, can achieve

560   better and robust performance with various levels of conditional edge density (for better

561   diversity); Thirdly, the most photorealistic performance was achieved when the edge density

562   levels in the inference phase are close to those used in the training phase. The results presented in

563   Figure 15 can be interpreted from the perspective of ablation study because removing the interim

564   domain adopted in the proposed method will considerably deteriorate the performance of the

565   conditional GAN.

566   This paper aims to generate photorealistic facial images using conditional GANs trained with a

567   small set of training images for data augmentation. To evaluate the effect of the number of

568   training images on the quality of the images generated by the proposed conditional GAN,

569   pix2pix and pix2pixHD, different numbers (25, 50, 100, and 500) of training images were used

570   to train each of the three conditional GANs separately. Moreover, in order to demonstrate the

571   effect of different conditional edge density levels, both sparse edges (threshold ratio = 0.4) and

572   dense edges (threshold ratio = 0.2) were used to generate 1000 images by each trained

573   conditional GAN. The FID and KID scores of the images generated by the three conditional

574   GANs were calculated respectively. Figure 16 shows the changes of FID and KID scores with

575   the different numbers of training images, from which the following three points can be made: 1)

576   With the interim domain, the proposed conditional GAN achieved lower FID and KID scores

577   than pix2pix and pix2pixHD when trained with the same number of training images. 2) Dense

578   conditional edges achieved lower KID and FID scores than sparse edges, but the diversity in the

579   generated images may be constrained. 3) With the increase in the number of training images, the

580   advantage of the proposed method over the existing methods becomes less obvious. This

581   tendency indicates that the proposed conditional GAN framework is very effective when it is

582   trained with a small number training samples and its performance would approach to that of the

583   existing methods when the number of training samples becomes relatively large.

584

Conclusions and Future Work

585

586   In this paper, a new conditional GAN framework is proposed for edge-to-image translation based

587   on a small set of training data, which can synthesise photorealistic diverse facial images using

588   incomplete edges as conditional inputs for data augmentation purposes. In order to solve the

589   problem in training conditional GANs using small training data, an interim domain for refining

590   images is introduced in the proposed conditional GAN, which can effectively reduce unexpected

591   distortions and thus improve the quality of the generated images. Experimental results have

592   demonstrated that blending segmentation masks and regional binary images as refined reference

593   images can reduce distortions in facial components of the images generated by the conditional

594   GAN trained with a small training dataset. Compared with the existing edge-to-image translation

595   methods, the proposed conditional GAN can not only automatically transfer incomplete

596   conditional edges to reference images with more facial features in the interim domain but also
597   effectively reduce unexpected distortions caused by small training data.
598   Compared to directly transferring source domain into target domain, the proposed method can
599   have a more comprehensive view to generate more photorealistic edge-to-image translation
600   results when using various incomplete conditional edges for data augmentation. More
601   informative reference images can be constructed in the interim domain from incomplete edge
602   inputs to integrate useful facial components. The proposed conditional GAN trained using a
603   small dataset can synthesise various photorealistic facial images by manipulating conditional
604   edge features or using hand-drawn facial sketches for diverse image data augmentation.
605   Compared to the existing conditional GANs for image-to-image translation, the images
606   generated by the proposed conditional GAN have less distortion and more diversity, which is
607   desirable for data augmentation purposes.
608   Due to limited GPU computing facilities available for conducting our experiments, it is hard to
609   optimise the hyperparameters of the tested models, and the performance evaluation is based on
610   the comparison with two state-of-the-art methods only. More extensive comparative study would
611   be desirable in the future research to draw more reliable conclusions.
612   The advantage of the proposed conditional GAN framework over the existing methods becomes
613   less obvious when the number of training samples are relatively large. For future work, the
614   interim domain could be improved so that the proposed conditional GAN framework would also
615   significantly outperform existing methods for image data augmentation when a reasonably large
616   number of training image dataset is available.
617

References

619   **Arjovsky M, Bottou L. 2017.** Towards principled methods for training generative adversarial
620   networks. In: *Proceedings of the international conference on learning representations*, 1–
621   6.
622   **Ashual O, Wolf L. 2019.** Specifying object attributes and relations in interactive scene
623   generation. In: *Proceedings of the IEEE international conference on computer vision*,
624   4561–4569.
625   **Azadi S, Fisher M, Kim V, Wang Z, Shechtman E, Darrell T. 2018.** Multi-content GAN for
626   few-shot font style transfer. In: *Proceedings of the IEEE conference on computer vision
627   and pattern recognition*, 7564–7573.
628   **Bartlett PL, Long PM, Lugosi G, Tsigler A. 2020.** Benign overfitting in linear regression.
629   *National Academy of Sciences* **117(48)**:30063–30070 DOI 10.1073/pnas.1907378117.
630   **Binkowski M, Sutherland D, Arbel M, Gretton A. 2018.** Demystifying MMD GANs. In:
631   *Proceedings of the international conference on learning representations*, 1–36.
632   **Canny J. 1986.** A computational approach to edge detection. *IEEE Transactions on Pattern
633   Analysis and Machine Intelligence* **8(6)**:679–698 DOI 10.1109/TPAMI.1986.4767851.
634   **Chen T, Cheng M-M, Tan P, Shamir A, Hu S-M. 2009.** Sketch2photo: Internet image
635   montage. *ACM Transactions on Graphics* **28(124)**:1–10 DOI 10.1145/1618452.1618470.

Chen W, Hays J. 2018. SketchyGAN: Towards Diverse and Realistic Sketch to Image Synthesis. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9416–9425 DOI 10.1109/CVPR.2018.00981.

Cherian A, Sullivan A. 2018. Sem-GAN: Semantically-consistent image-to-image translation. In: *Proceedings of the IEEE winter conference on applications of computer vision*, 1797–1806 DOI 10.1109/WACV.2019.00196.

De Cao N, Kipf T. 2018. MolGAN: An implicit generative model for small molecular graphs. In: *ICML 2018 workshop on theoretical foundations and applications of deep generative models* DOI 10.26434/chemrxiv.14569545.v3.

Dimitrakopoulos P, Sfikas G, Nikou C. 2020. ISING-GAN: Annotated data augmentation with a spatially constrained generative adversarial network. In: *IEEE 17th international symposium on biomedical imaging*, 1600-1603 DOI 10.1109/ISBI45749.2020.9098618.

Dosovitskiy A, Brox T. 2016. Generating images with perceptual similarity metrics based on deep networks. In: *Proceedings of the neural information processing systems conference*, 658–666.

Eitz M, Richter R, Hildebrand K, Boubekeur T, Alexa M. 2011. Photosketcher: Interactive sketch-based image synthesis. *IEEE Computer Graphics and Applications* 31(6):56–66 DOI 10.1109/MCG.2011.67.

Elder J. 1999. Are edges incomplete? *International Journal of Computer Vision* 34:97–122 DOI 10.1023/A:1008183703117.

Gatys LA, Ecker AS, Bethge M. 2016. Image style transfer using convolutional neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2414–2423 DOI 10.1109/CVPR.2016.265.

Gu S, Bao J, Yang H, Chen D, Wen F, Yuan L. 2019. Mask-guided portrait editing with conditional GANs. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3436–3445 DOI 10.1109/CVPR.2019.00355.

Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC. 2017. Improved training of Wasserstein GANs. In: *Proceedings of the neural information processing systems conference*, 5767–5777.

Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: *Proceedings of the neural information processing systems conference*, 6626–6637.

Hung SK, Gan QJ. 2021. Facial image augmentation from sparse line features using small training data. In: *Proceedings of the International Work-Conference on Artificial and Natural Neural Networks*. 547–558. DOI 10.1007/978-3-030-85030-2_45

Ibtehaz N, Rahman MS. 2020. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. *Neural Networks* 121:74–87 DOI 10.1016/j.neunet.2019.08.025.

Iizuka S, Simo-Serra E, Ishikawa H. 2017. Globally and locally consistent image completion. *ACM Transactions on Graphics* 36(4):107–110 DOI 10.1145/3072959.3073659.

676 **Isola P, Zhu J-Y, Zhou T, Efros AA. 2017.** Image-to-image translation with conditional
677   adversarial networks. In: *Proceedings of the IEEE conference on computer vision and*
678   *pattern recognition*, 1125–1134 DOI 10.1109/CVPR.2017.632.
679 **Jo Y, Park J. 2019.** SC-FEGAN: Face editing generative adversarial network with user's sketch
680   and color. In: *Proceedings of the IEEE international conference on computer vision*, 1745–
681   1753 DOI 10.1109/ICCV.2019.00183.
682 **Johnson J, Gupta A, Li F-F. 2018.** Image generation from scene graphs. In: *Proceedings of the*
683   *IEEE conference on computer vision and pattern recognition*, 1219–1228
684   DOI 10.1109/CVPR.2018.00133.
685 **Karras T, Aila T, Laine S, Lehtinen J. 2018.** Progressive growing of GANs for improved
686   quality, stability, and variation. In: *Proceedings of the international conference on learning*
687   *representations*.
688 **Kinoshita Y, Kiya H. 2020.** Fixed smooth convolutional layer for avoiding checkerboard
689   artifacts in CNNs. In: *Proceedings of IEEE international conference on acoustics, speech,*
690   *and signal processing*, 3712–3716.
691 **Lee C-H, Liu Z, Wu L, Luo P. 2020.** MaskGAN: Towards diverse and interactive facial image
692   manipulation. In: *Proceedings of the IEEE conference on computer vision and pattern*
693   *recognition*, 5549–5558.
694 **Lee H-Y, Tseng H-Y, Mao Q, Huang J-B, Lu Y-D, Singh M, Yang M-H. 2019.** DRIT++:
695   Diverse image-to-image translation via disentangled representations. *Computing Research*
696   *Repository* **1905**:1–14 DOI 10.1007/s11263-019-01284-z.
697 **Li Y, Chen X, Yang B, Chen Z, Cheng Z, Zha Z-J. 2020**. Deepfacepencil: Creating face
698   images from freehand sketches. In: *Proceedings of the ACM international conference on*
699   *multimedia*, 991– 999 DOI 10.1145/3394171.3413684.
700 **Lin J, Xia Y, Qin T, Chen Z, Liu TY. 2018.** Conditional image-to-image translation. In:
701   *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5524–
702   5532 DOI 10.1109/CVPR.2018.00579.
703 **Liu M-Y, Breuel T, Kautz J. 2017.** Unsupervised image-to-image translation networks. In:
704   *Proceedings of the neural information processing systems conference*, 700–708
705   DOI 10.1109/BigComp48618.2020.00-92.
706 **Liu Z, Luo P, Wang X, Tang X. 2015.** Deep learning face attributes in the wild. In:
707   *Proceedings of the IEEE international conference on computer vision*, 3730–3738
708   DOI 10.1109/ICCV.2015.425.
709 **Mescheder L, Geiger A, Nowozin S. 2018.** Which training methods for GANs do actually
710   converge. In: *Proceedings of the international conference on machine learning*, 3478–
711   3487.
712 **Mikolajczyk A, Grochowski M. 2018.** Data augmentation for improving deep learning in image
713   classification problem. In: *Proceedings of the international interdisciplinary PhD*
714   *workshop*, 117–122 DOI 10.1109/IIPHDW.2018.8388338.

715 **Mo S, Cho M, Shin J. 2019.** InstaGAN: Instance-aware image-to-image translation. In:
716   *Proceedings of the international conference on learning representations*.
717 **Odena A, Olah C, Shlens J. 2017.** Conditional image synthesis with auxiliary classifier GANs.
718   In: *Proceedings of the international conference on machine learning*, 2642–2651.
719 **Park T, Liu M-Y, Wang T-C, Zhu J-Y. 2019.** Semantic Image Synthesis with Spatially-
720   Adaptive Normalization. In: *Proceedings of the IEEE conference on computer vision and*
721   *pattern recognition*, 2337–2346 DOI 10.1109/CVPR.2019.00244.
722 **Ronneberger O, Fischer P, Brox T. 2015.** U-net: Convolutional networks for biomedical image
723   segmentation. In: *Proceedings of the medical image computing and computer-assisted*
724   *intervention*, 234–241 DOI 10.1007/978-3-319-24574-4_28.
725 **Royer A, Bousmalis K, Gouws S, Bertsch F, Moressi I, Cole F, Murph, K. 2017.** XGAN:
726   Unsupervised image-to-image translation for many-to-many mappings. *Computing*
727   *Research Repository* **1711**:1–19 DOI 10.1007/978-3-030-30671-7_3.
728 **Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. 2016.** Improved
729   techniques for training GANs. In: *Proceedings of the neural information processing*
730   *systems conference*, 2234–2242.
731 **Shorten C, Khoshgoftaar TM. 2019.** A survey on image data augmentation for deep learning.
732   *Journal of Big Data* **6(1)**:60 DOI 10.1186/s40537-019-0197-0.
733 **Srivastava A, Valkoz L, Russell C, Gutmann MU, Sutton C. 2017.** VEEGAN: Reducing
734   mode collapse in GANs using implicit variational learning. In: *Proceedings of the neural*
735   *information processing systems conference*, 3308–3318.
736 **Szeto R, El-Khamy M, Lee J, Corso JJ. 2021.** HyperCon: Image-to-video model transfer for
737   video-to-video translation tasks. In: *Proceedings of the IEEE winter conference on*
738   *applications of computer vision*, 3080–3089.
739 **Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C. 2018.** A survey on deep transfer learning.
740   In: *Proceedings of the 27th international conference on artificial neural networks*, 270–
741   279.
742 **Torrado RR, Khalifa A, Green MC, Justesen N, Risi S, Togelius J. 2020.** Bootstrapping
743   conditional GANs for video game level generation. In: *Proceedings of the IEEE*
744   *conference on games*, 41–48 DOI 10.1109/CoG47356.2020.9231576.
745 **Wang T-C, Liu M-Y, Zhu J-Y, Tao A, Kautz J, Catanzaro B. 2018a.** High-resolution image
746   synthesis and semantic manipulation with conditional GANs. In: *Proceedings of the IEEE*
747   *conference on computer vision and pattern recognition*, 1–13
748   DOI 10.1109/CVPR.2018.00917.
749 **Wang X, Gupta A. 2016.** Generative image modeling using style and structure adversarial
750   networks. In: *Proceedings of the European conference on computer vision*, 318–335 DOI
751   10.1007/978-3-319-46493-0_20.
752 **Wang X, Yu K, Dong C, Loy CC. 2018b.** Recovering realistic texture in image super-
753   resolution by deep spatial feature transform. In: *Proceedings of the IEEE conference on*
754   *computer vision and pattern recognition*, 606–615.

755 **Wang Y, Wu C, Herranz L, van de Weijer J, Gonzalez-Garcia A, Raducanu B. 2018c.**
756   Transferring GANs: Generating images from limited data. In: *Proceedings of the European*
757   *conference on computer vision*, 218–234.
758 **Wei X, Gong B, Liu Z, Lu W, Wang L. 2018.** Improving the improved training of Wasserstein
759   GANs: A consistency term and its dual effect. In: *Proceedings of the international*
760   *conference on learning representations*.
761 **Yi Z, Zhang H, Tan P, Gong M. 2017.** DualGAN: Unsupervised dual learning for image-to-
762   image translation. In: *Proceedings of the IEEE conference on computer vision and pattern*
763   *recognition*, 2849–2857 DOI 10.1109/ICCV.2017.310.
764 **Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS. 2019.** Freeform image inpainting with gated
765   convolution. In: *Proceedings of the IEEE international conference on computer vision*,
766   4471–4480 DOI 10.1109/ICCV.2019.00457.
767 **Zhao M, Cong Y, Carin L. 2020.** On leveraging pretrained GANs for generation with limited
768   data. In: *Proceedings of the international conference on machine learning* **119**:11340–
769   11351.
770 **Zhou X-Y, Yang G-Z. 2019.** Normalization in training U-Net for 2D biomedical semantic
771   segmentation. *IEEE Robotics and Automation Letters* **4(2)**:1792-1799
772   DOI 10.1109/LRA.2019.2896518.
773 **Zhu J, Shen Y, Zhao D, Zhou B. 2020.** In-domain GAN inversion for real image editing. In:
774   *Proceedings of the European conference on computer vision*, 592-608.
775 **Zhu J-Y, Park T, Isola P, Efros AA. 2017.** Unpaired image-to-image translation using cycle-
776   consistent adversarial networks. In: *Proceedings of the IEEE conference on computer*
777   *vision and pattern recognition*, 2223–2232 DOI 10.1109/ICCV.2017.244.

Figure 1

Examples of training results (left side) and inference results (right side) with a different number of training images and different conditional edge inputs by using the same parameter setting for the image-to-image translation method.
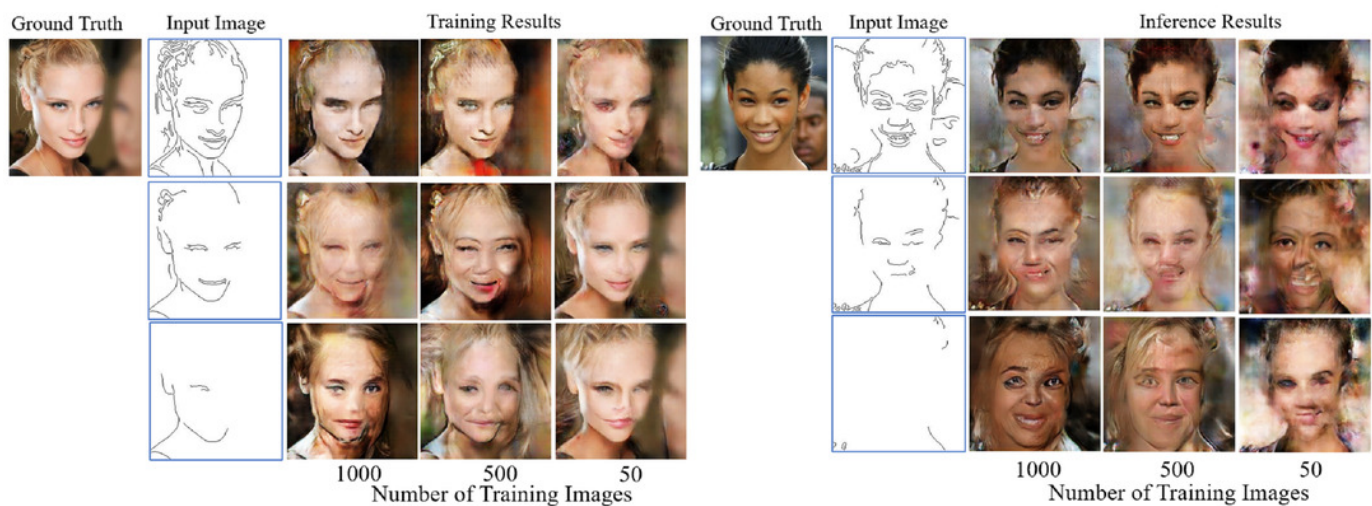
Figure 2

The proposed translation method by defining an interim domain for refining images based on a small training dataset.

Image pre-processing is adopted in the proposed conditional GAN to enhance the mapping relationship from source domain to target domain.
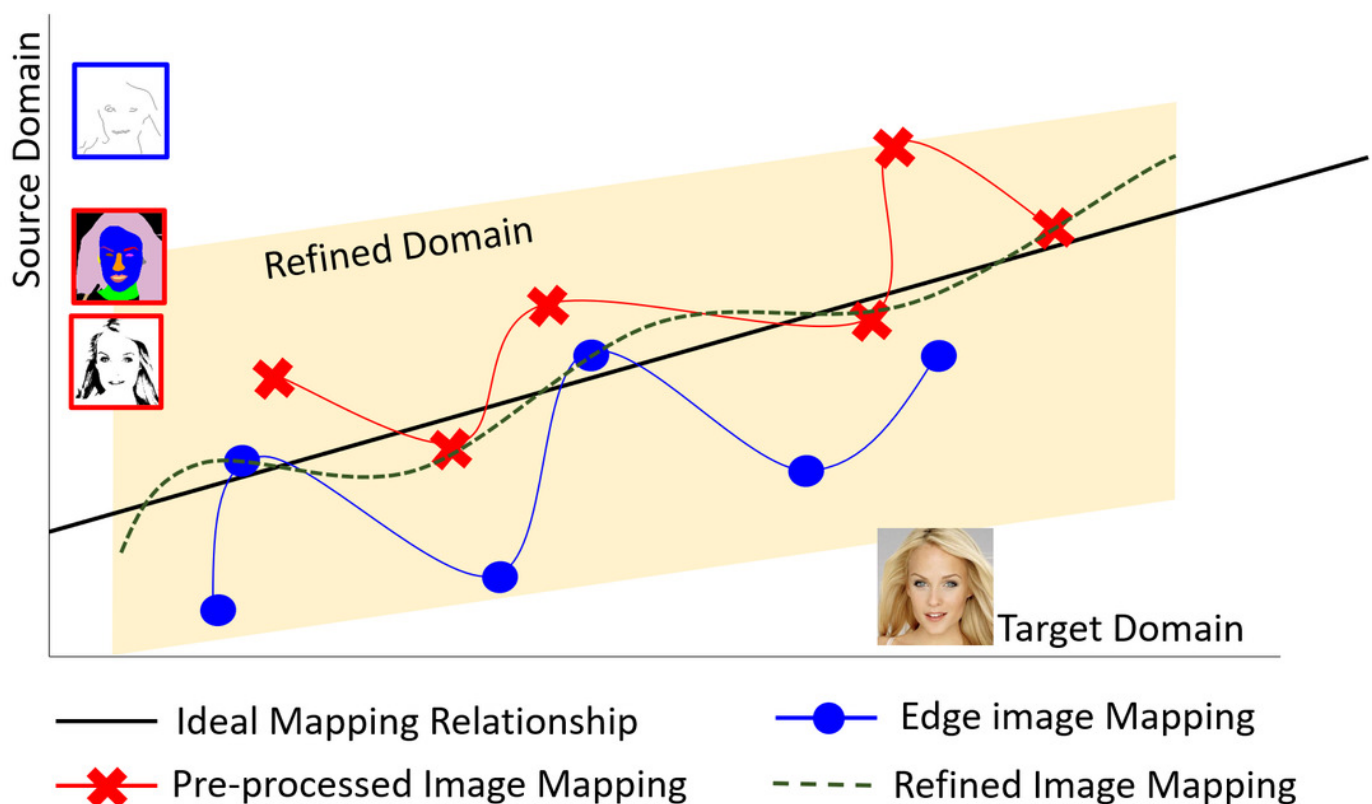
Figure 3

Overview of the proposed conditional GAN for translating edges to photorealistic images using two U-nets.
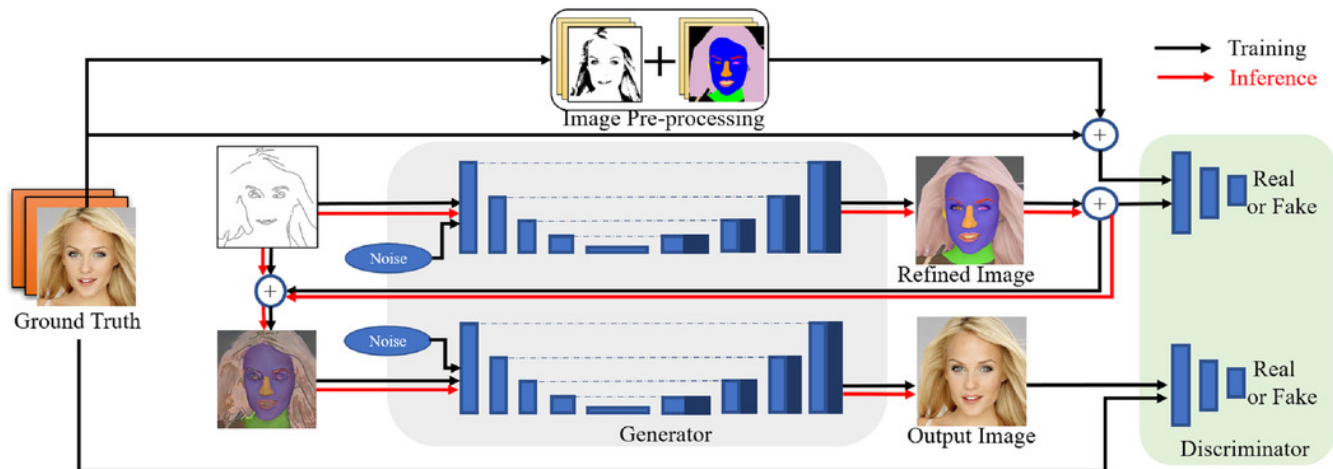
Figure 4

Corresponding mapping relationships among the conditional inputs, refined image and ground truth.
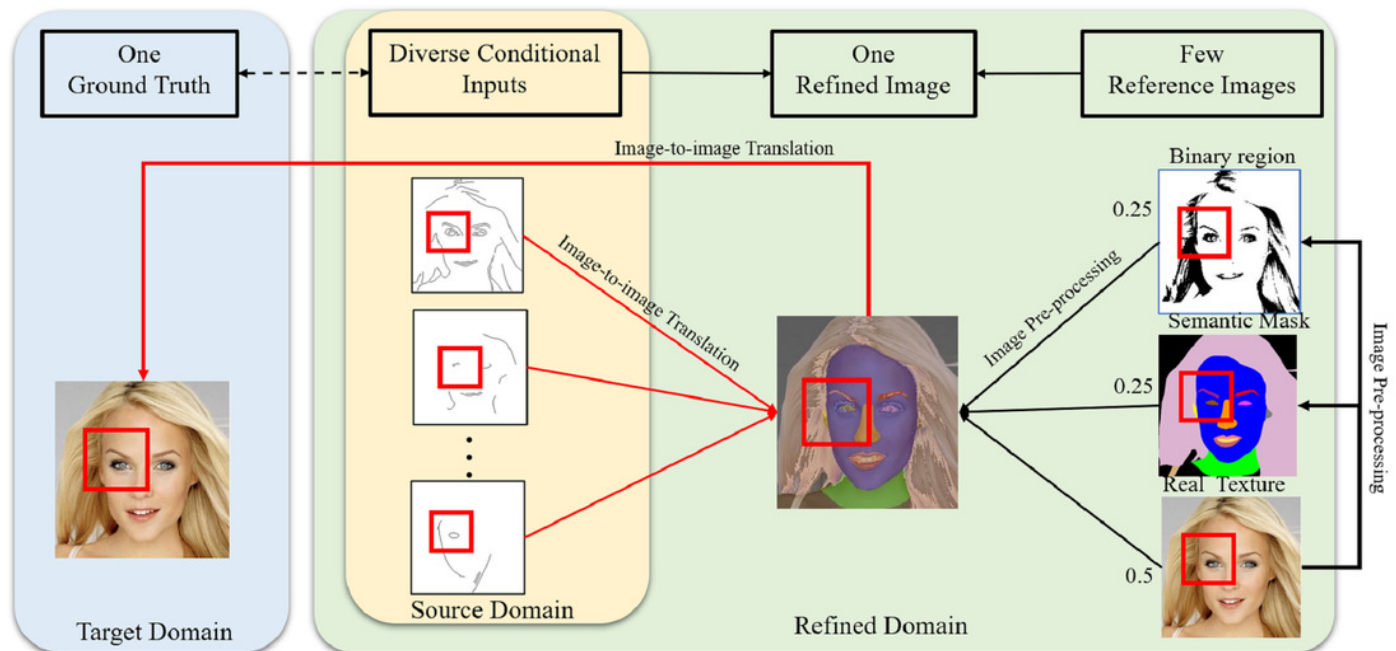
Figure 5

Inference results in translating sparse edges to labelled segmentation masks with 50 training images.

(a) The outputs can roughly resume the missing facial components from incomplete layouts when given abstract inputs. (b) The red boxes indicate the corresponding indefinite contours in the original inputs and generated masks.
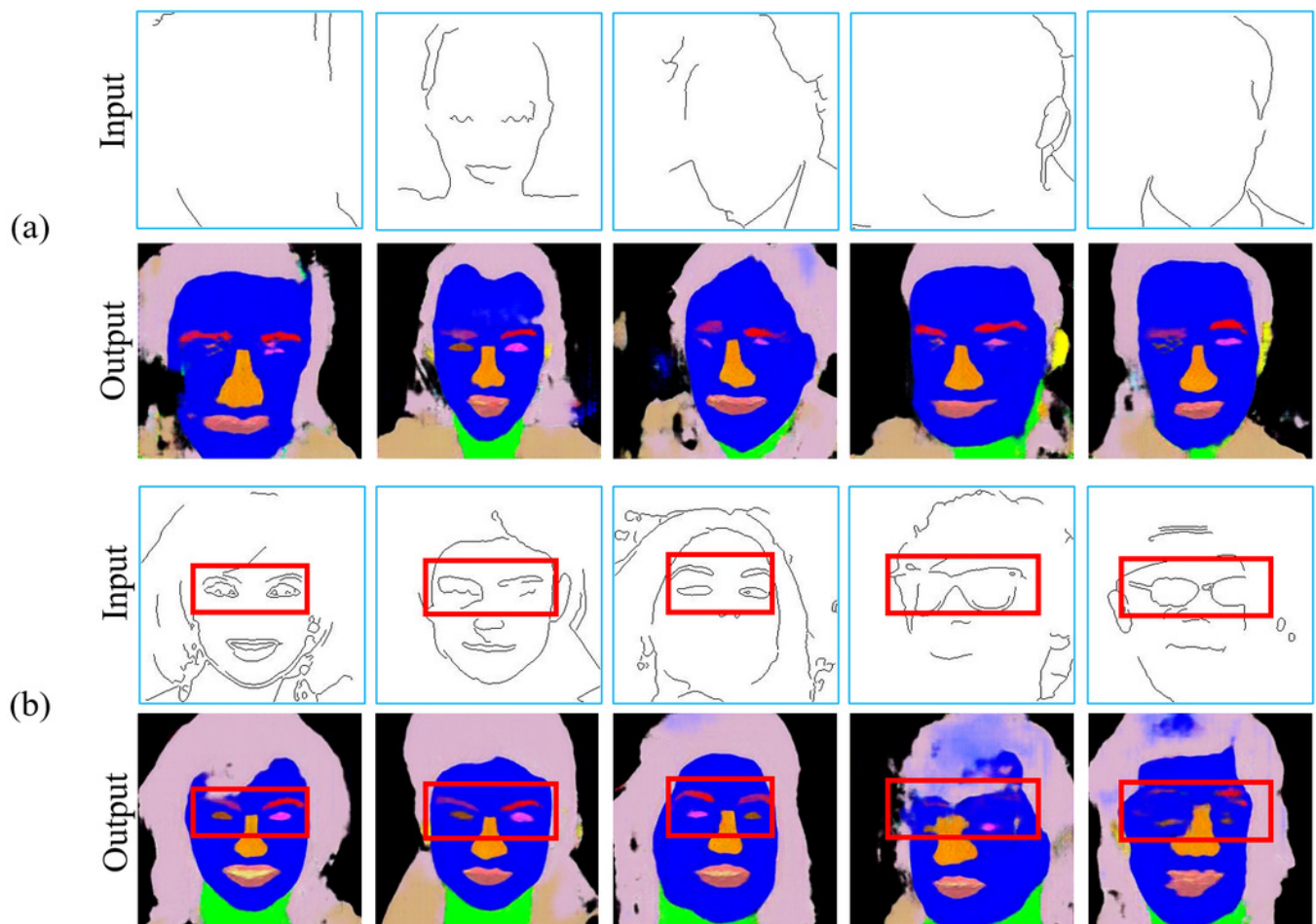
Figure 6

Inference results in translating sparse edges to binary regional images with 50 training images.

(a) The outputs can integrate discontinued contours when given sparse inputs. (b) The outputs can get rid of 'bogus' edges when given very dense inputs.
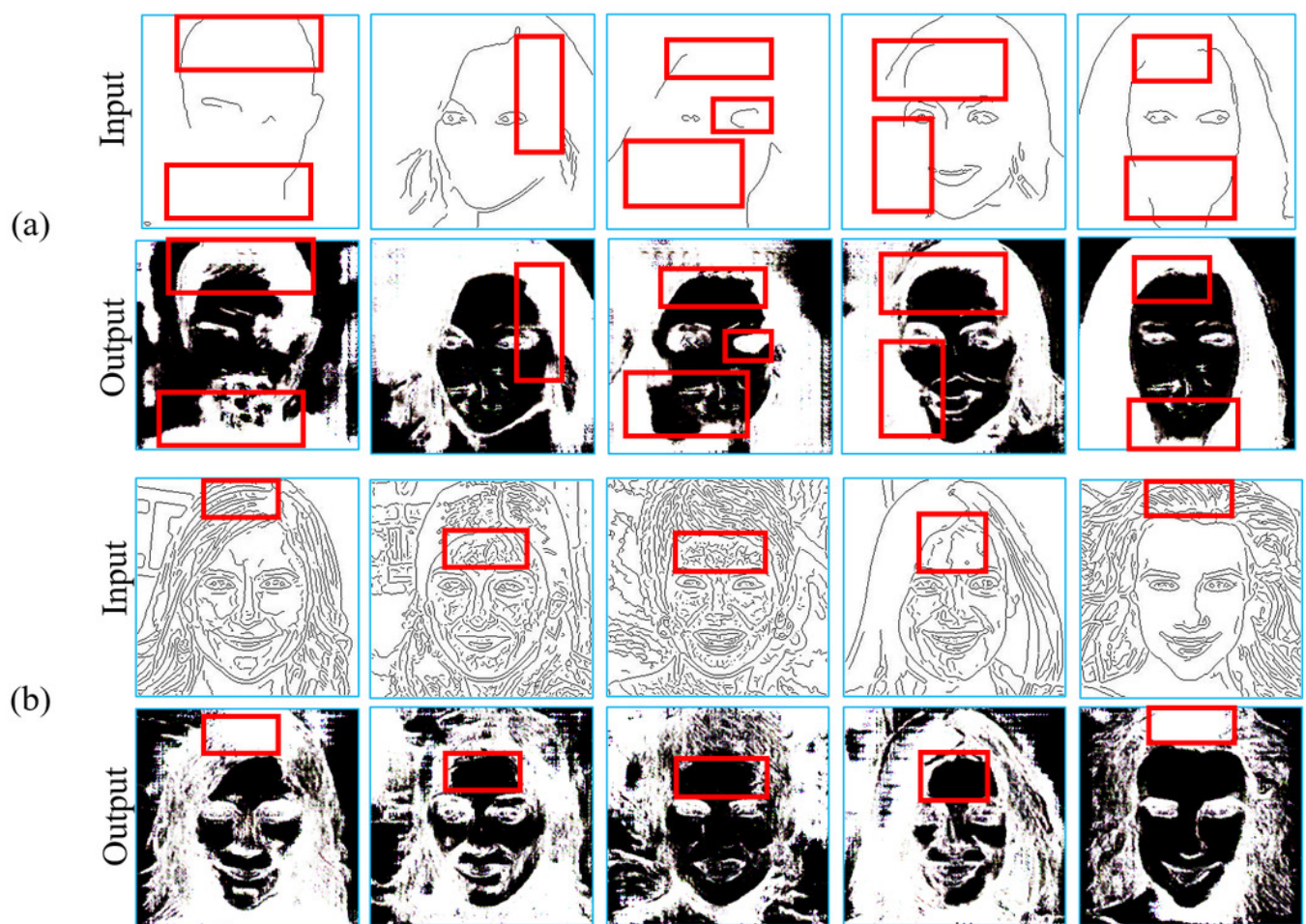
Figure 7

Comparison of different edge detectors: (a) results of Canny. (b) results of Sobel. (c) results of Laplace. (d) results of Gradient.
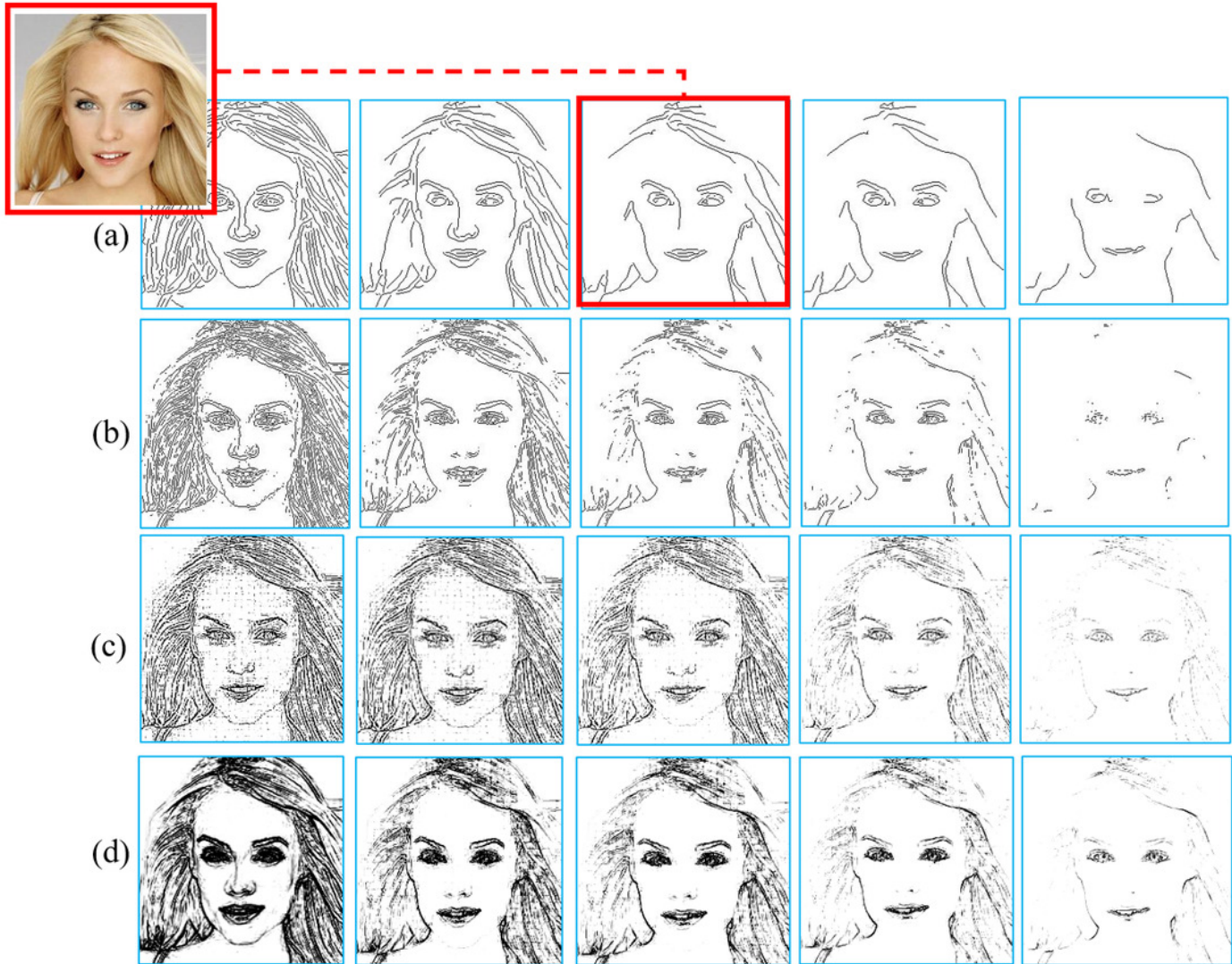
Figure 8

Inference results with examples of refined images and final outputs.

The red boxes represent blending areas in the refined region, which can be reflected by the brightness in the generated image outputs.
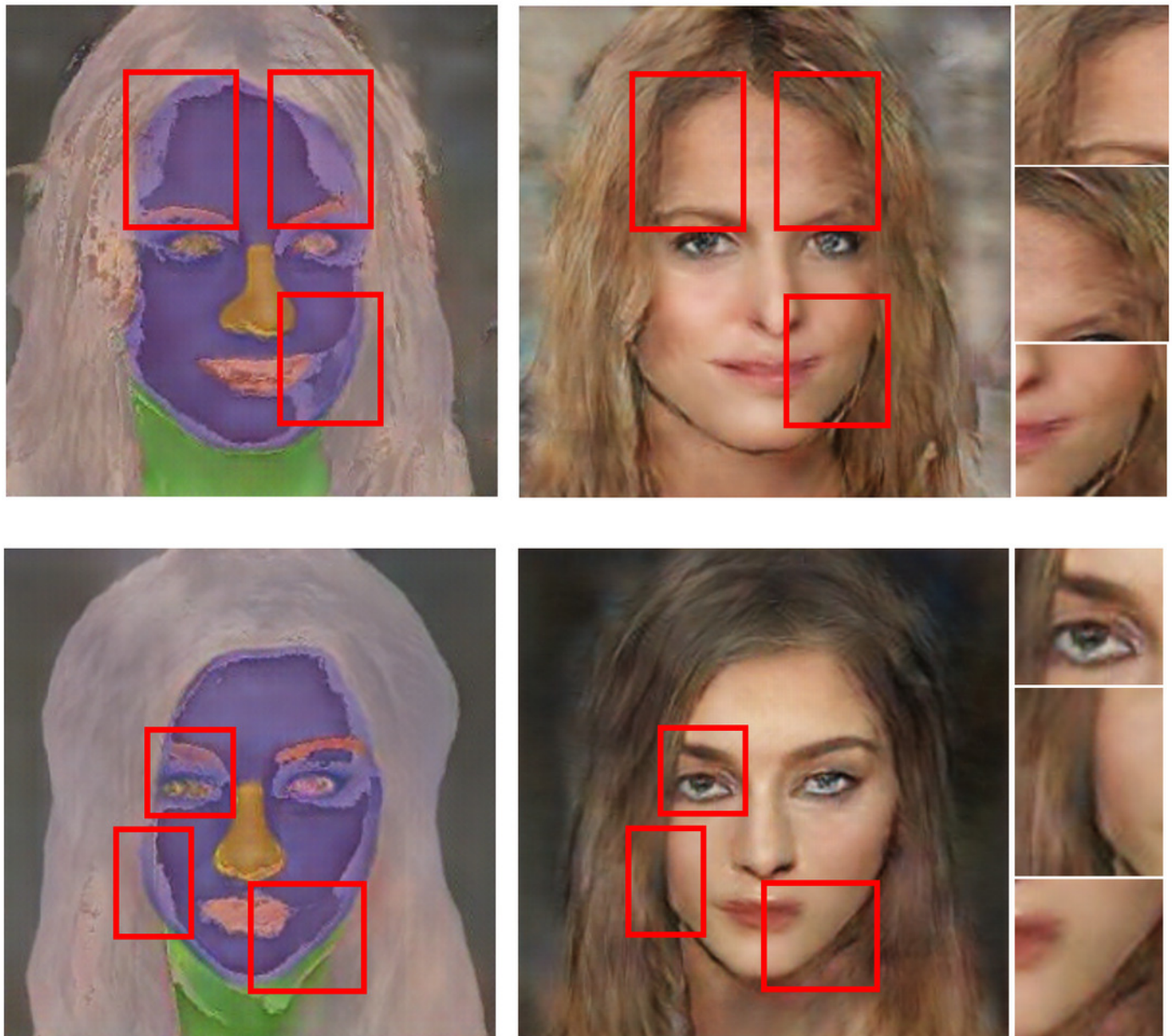
Figure 9

Results from exchanging conditional facial edges to generate diverse styles of facial images.
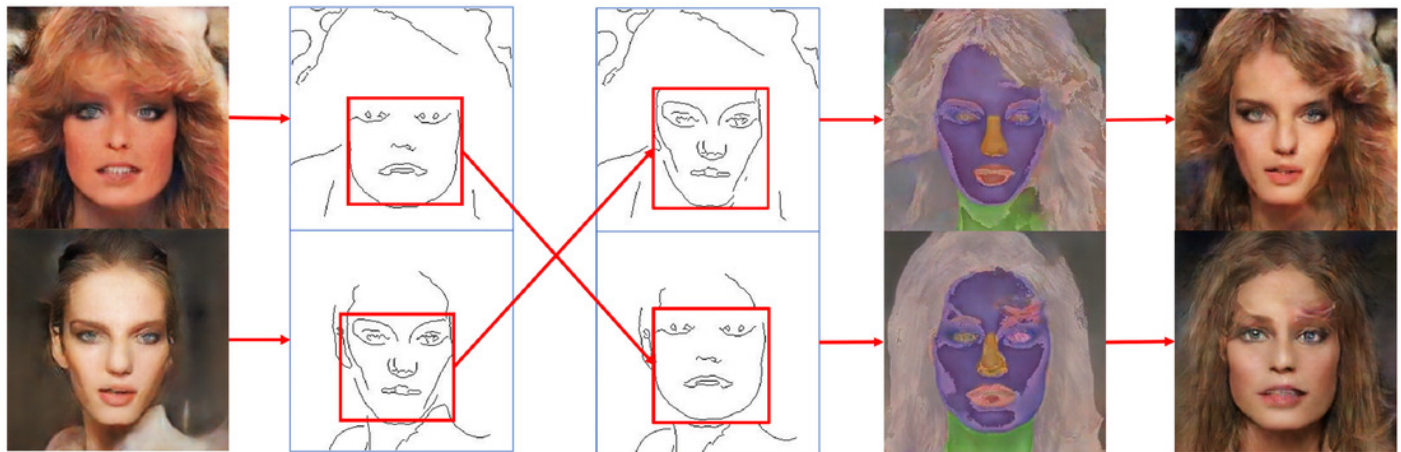
Figure 10

Inference results with example images in the source, interim, and target domains respectively.

The various density levels in the conditional inputs were not included in the training phase except the one in the red box generated by Canny edge detector with threshold ratio = 0.4. The results are from GANs trained using 50 images only.
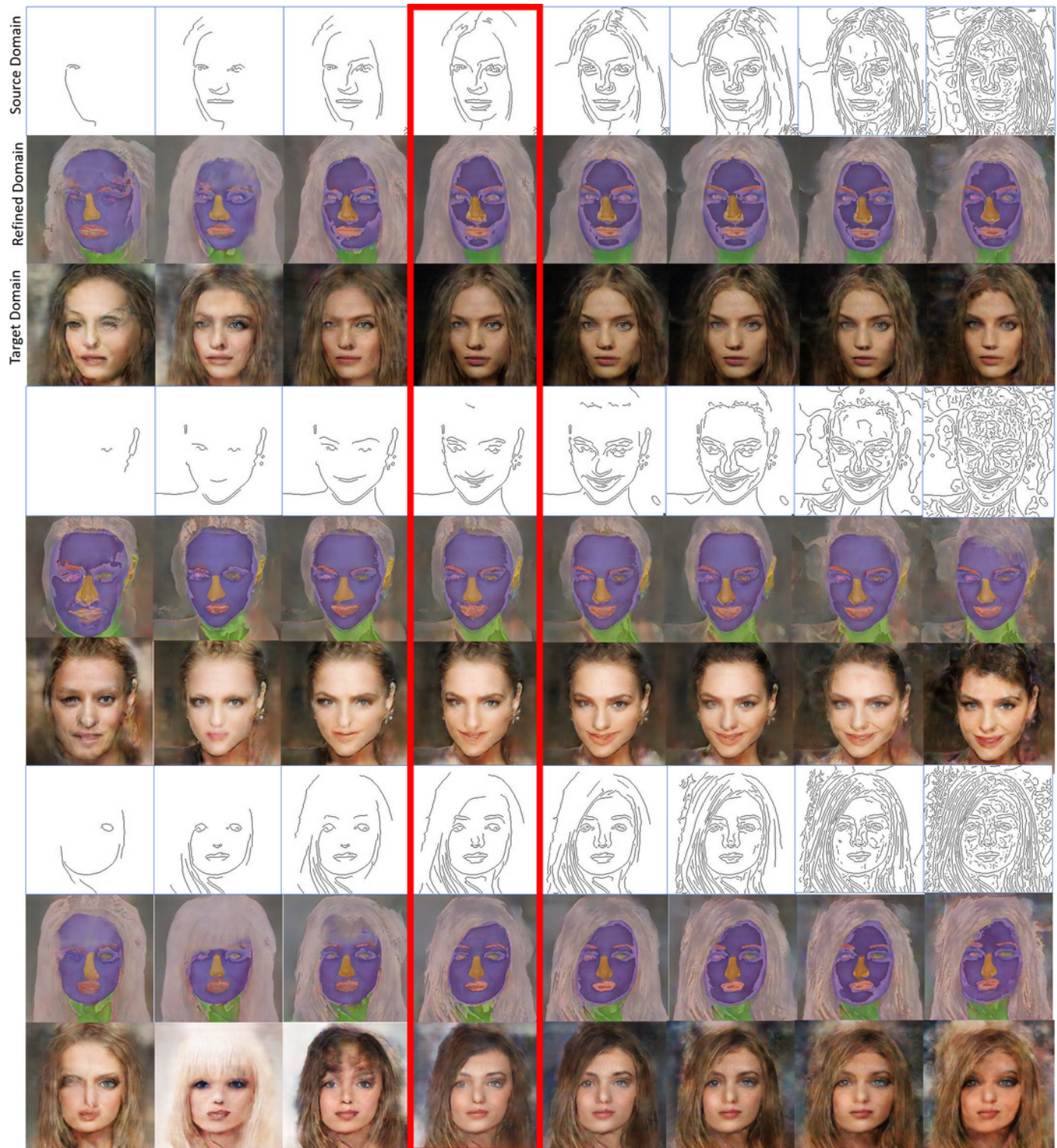
Figure 11

Examples of facial image augmentation results using 50 training images, with parts of input edges modified for introducing diversity so as to augment each training image with desirable facial features.
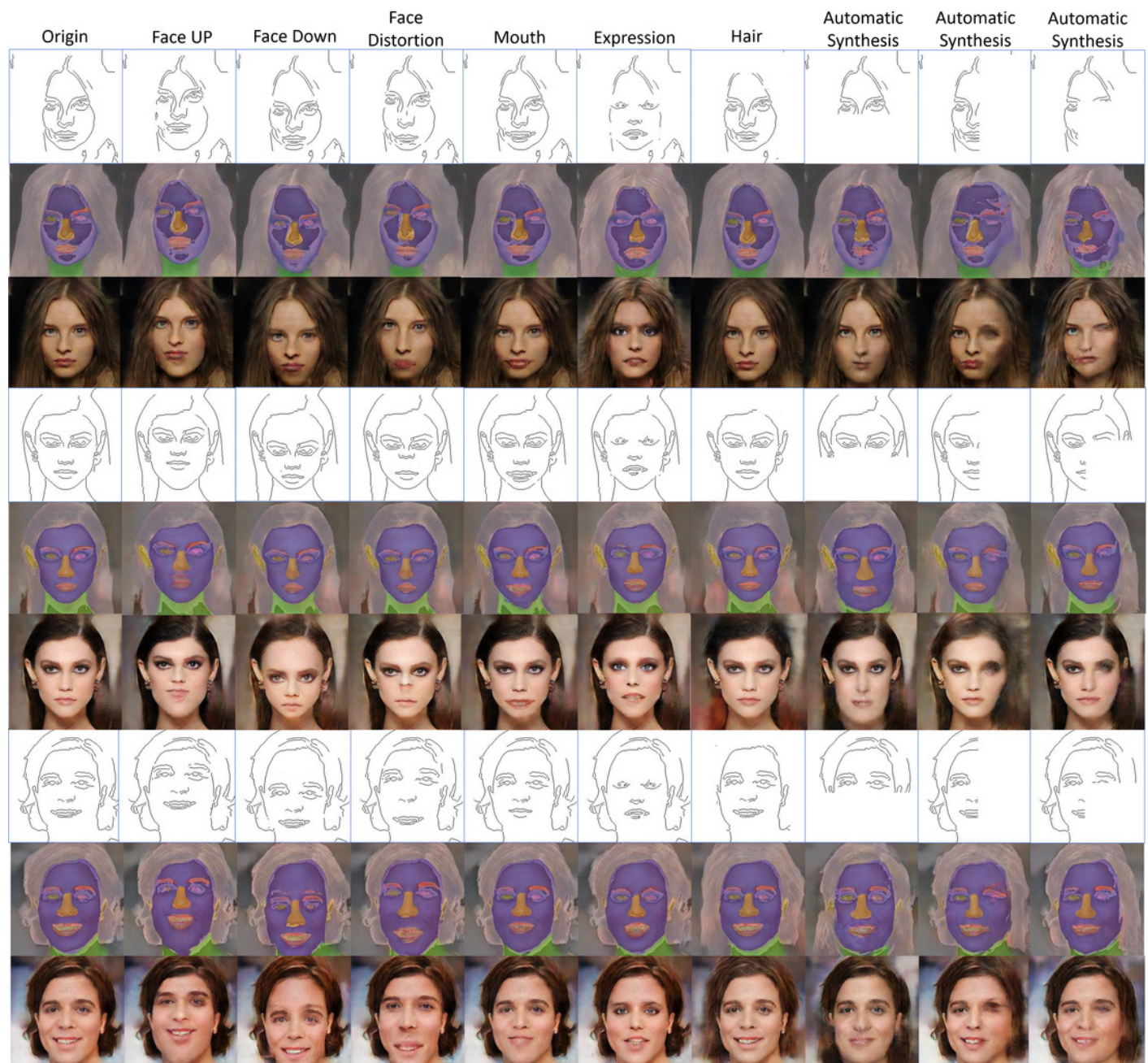
Figure 12

Examples of facial image augmentation results using 50 training images, with face components and hairstyles in different training images swapped in the edges as conditional inputs so as to generate diverse facial images.

Facial Components

Style
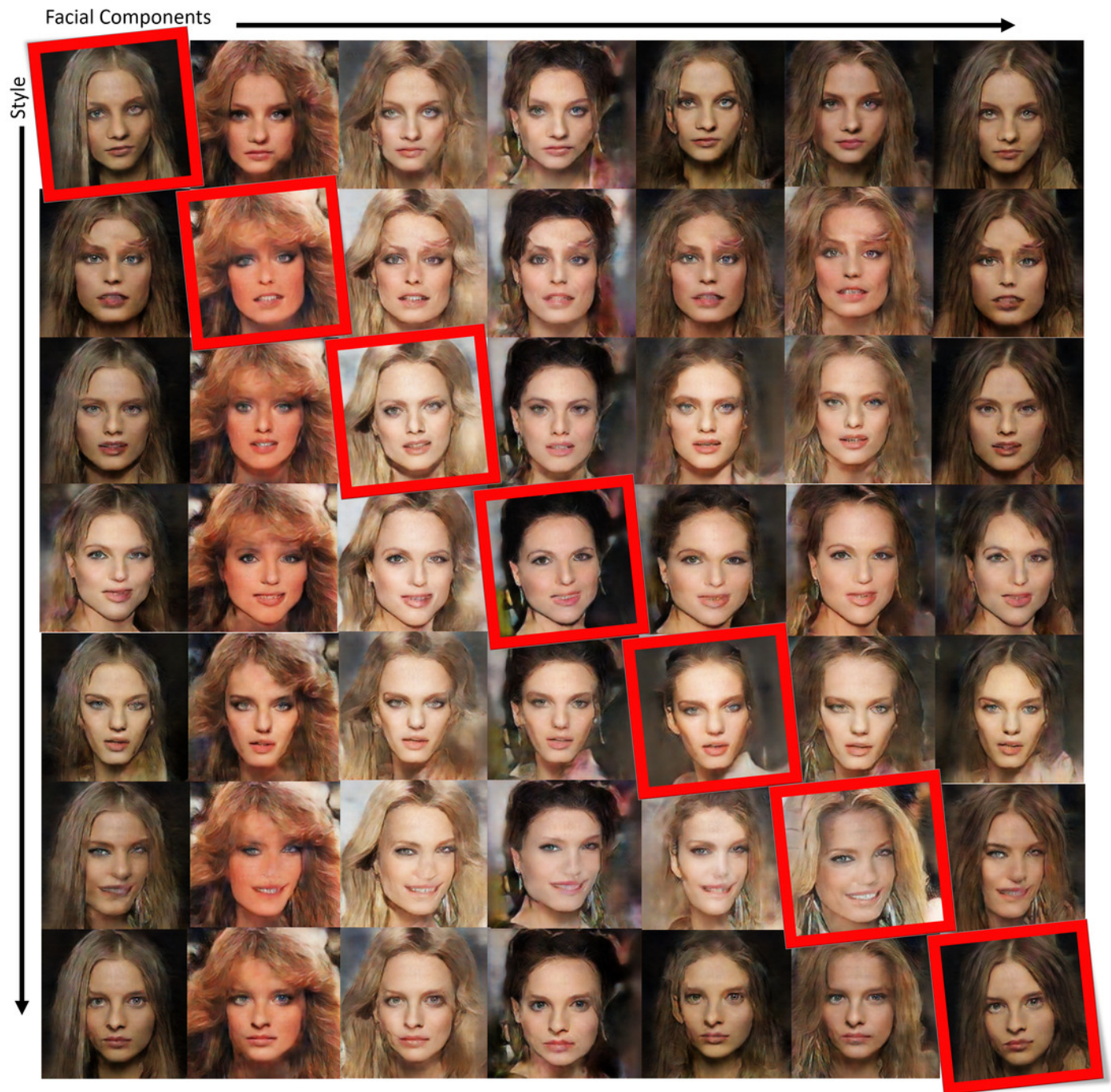
Figure 13

Inference results with example images in the source, interim, and target domains respectively.

The conditional inputs are hand-drawn sketches showing different facial expressions. The proposed conditional GAN was trained using 50 training images.
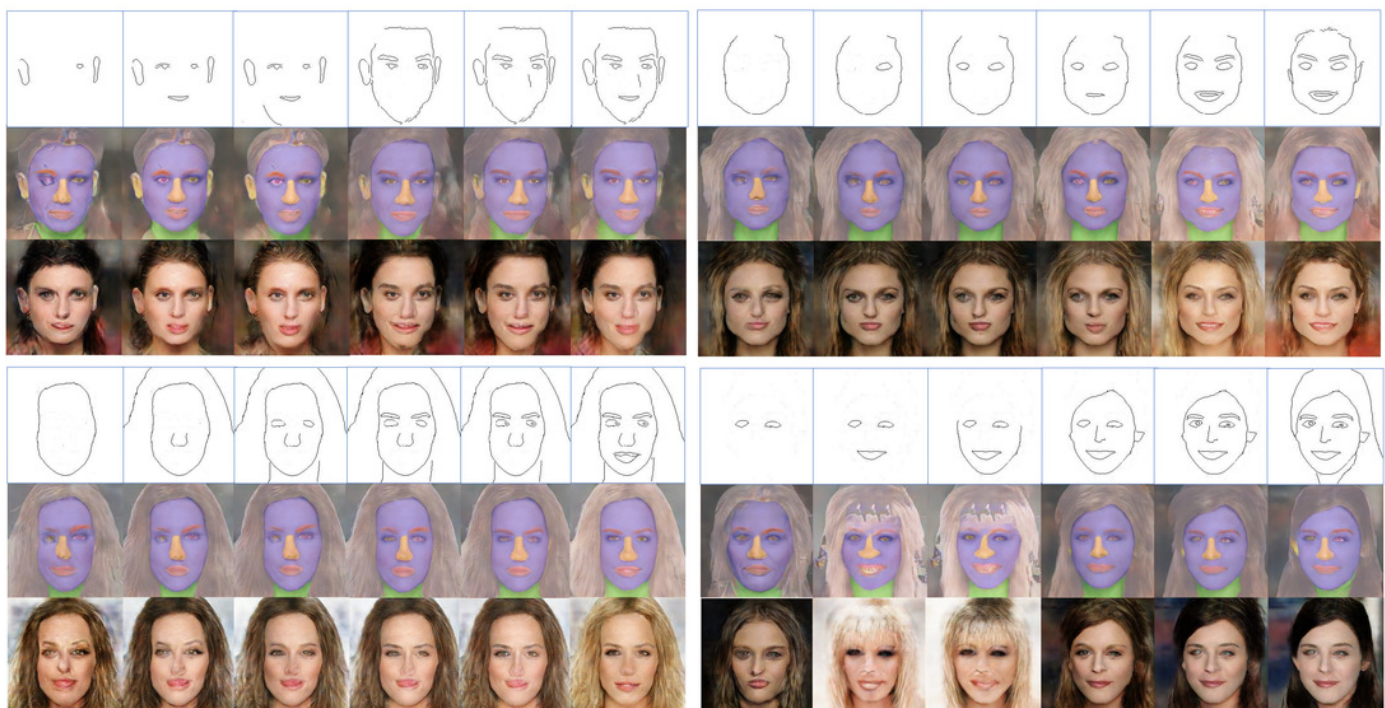
Figure 14

Inference results generated with sparse edge inputs (the first row), in comparison with those obtained from the state-of-the-art conditional GANs.

The images were generated respectively by the three GANs for comparison, trained using the same small dataset of 50 training images.
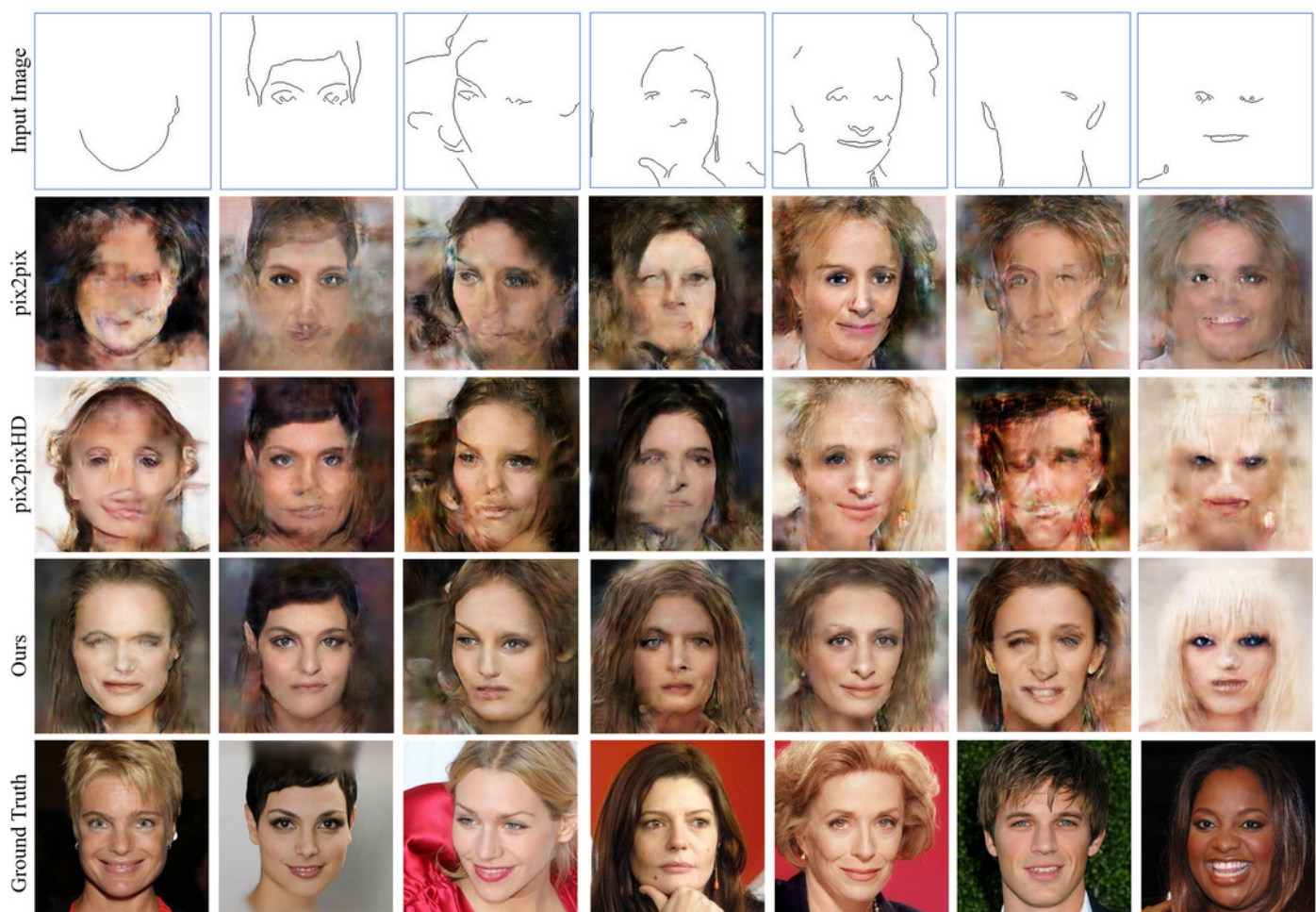
Figure 15

FID and KID scores of double U-nets with an interim domain and single U-net with different levels of input edge density respectively.

One input type (threshold ratio = 0.4) and three input types (threshold ratios = 0.2, 0.4, 0.6) in the source domain were used respectively during training with a small training dataset of 50 images. The FID and KID scores were calculated based on the same 1000 inference images at different edge density levels.
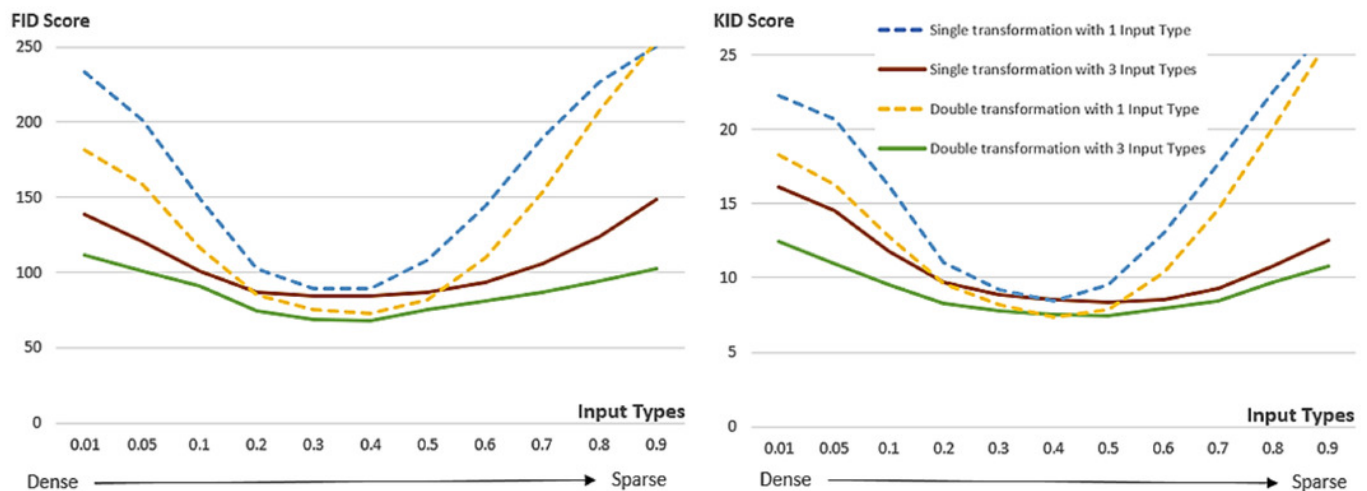
Figure 16

Changes of FID scores (first row) and KID scores (second row) with different number of training images.

Comparison among three edge-to-image translation methods with sparse and dense edge inputs respectively: pix2pix, pix2pixHD and ours.