A Survey of GAN-based Hand-drawn Sketch Image Translation (Postprint)

Authors: Wang Jianxin, Shi Yingjie, Liu Hao, Huang Haiqiao, Du Fang

Date: 2022-04-07T15:01:57+00:00

Abstract

Hand-drawn sketch image translation is a challenging research topic in computer vision, with significant application value in artistic design and e-commerce. Currently, GAN-based hand-drawn sketch image translation remains in its infancy. This article analyzes the challenging issues confronting sketch image translation, examines GAN-based sketch image translation research from two perspectives: uncontrolled and fine-grained controlled sketch image translation, and summarizes evaluation methods for generated images. Based on a comprehensive review of existing research, this article prospects future development trends in this field, providing valuable insights for researchers to broaden their research ideas.

Full Text

Preamble

Research on Freehand Sketch-to-Image Translation Based on Generative Adversarial Networks

Wang Jianxin[1] **, Shi Yingjie**[1] **, Liu Hao**[1] **, Huang Haiqiao**[1] **, Du Fang**[2]
(1. a. School of Arts & Sciences; b. School of Business; c. School of Fashion, Beijing Institute of Fashion Technology, Beijing 100029, China; 2. School of Information Engineering, Ningxia University, Yinchuan 750021, China)

Abstract: Freehand sketch-to-image translation is a challenging subject in computer vision with important application value in art design and e-commerce. Currently, sketch-to-image translation based on GANs remains in its infancy. This paper analyzes the challenging problems in sketch-to-image translation, summarizes GAN-based research from two perspectives—uncontrolled and finely controlled sketch-to-image translation, and reviews evaluation methods for generated images. Based on existing research, we propose possible future development trends in this field, providing researchers with clues for expanding their research directions.

Key words: freehand sketch; image-to-image translation; generative adversarial network; image synthesis; disentanglement

0 Introduction

Drawing represents one of humanity’ s earliest artistic activities, as primitive humans could depict primary prey in hunting activities through sparse sketches. Hand-drawn sketches reflect the human brain’ s visual perception of the real world, enabling anyone to express ideas and facilitate communication. Throughout history, hand-drawn sketches have remained the most direct and rapid means for visualizing objects or scenes. Consequently, sketch-related research has attracted significant attention in computer vision. Early studies focused primarily on sketch recognition, sketch-based image retrieval, and sketch-based 3D shape retrieval. With deep learning advancements, new research topics have emerged, such as synthetic sketches, deep sketch hashing, and instance-level sketch-based image retrieval. Recent breakthroughs in image translation, including style transfer and super-resolution, have sparked widespread academic and industrial interest in hand-drawn sketch-to-image translation. Image translation refers to converting one image type to another, essentially representing mutual conversion between two distinct image domains—for example, transforming winter scenes to summer scenes, semantic images to realistic images, or sketches to realistic color images. Hand-drawn sketch-to-image translation specifically converts human-drawn sketches—characterized by sparse strokes, abstraction, and certain noise—into images that remain faithful to the sketch content while achieving visual realism [1].

Traditional sketch-based image translation relied on image retrieval [2, 3]: searching for corresponding image patches from large-scale datasets based on objects and backgrounds specified in sketches, then fusing these patches. However, this approach cannot generate entirely new images. The rapid development of generative deep learning, particularly Generative Adversarial Networks (GANs) [4], has enabled sketch-to-image translation based on GANs. Due to the unique characteristics of hand-drawn sketches, GAN-based translation faces several critical challenges: First, sparse and abstract strokes require deformation correction and detail addition. Second, paired sketch-image data remains scarce, leading to insufficient training data. Third, diverse and difficult-to-imitate sketch styles cause models trained on augmented sketches to generalize poorly to real hand-drawn sketches.

Sketch-to-image translation helps users create or design novel images in practical applications, serving as an effective pathway for demonstrating creativity and communicating ideas. In design, it assists designers in rapidly visualizing products. Designers can specify color and texture for sketch regions through colored lines or incomplete color blocks within outlines, and the translation system generates realistic images with similar styles, providing powerful design references. In e-commerce, sketch translation systems convert user-drawn product sketches into realistic merchandise images, helping users search for similar online products effectively while enhancing shopping experiences and providing merchants with crucial data support for analyzing user needs, thereby boosting transaction volumes. Additionally, sketch-to-image translation demonstrates potential in other domains: generating realistic human faces from sparse sketches [5] can help eyewitnesses without drawing skills better depict criminal features, assisting law enforcement; in film production, screenwriters or directors can draw character sketches based on imagination and generate realistic facial images to select suitable actors; in image editing, sketches can edit facial contours, hair, beards, wrinkles, and combined with style transfer techniques, modify makeup and skin tones [6].

1 Challenges in Hand-drawn Sketch-to-Image Translation

Generating realistic images from sketches is not trivial. Synthesized images must remain faithful to the given sketch while maintaining realism and semantic coherence. Hand-drawn sketches depict approximate boundaries and internal contours, representing a special data domain, whereas real images precisely correspond to object boundaries with dense pixels. Thus, hand-drawn sketch-to-image translation constitutes a typical cross-modal conversion problem. GAN-based image translation is data-driven, requiring large-scale sketch and image data for training. However, collecting human-drawn sketches proves difficult and costly, resulting in limited directly usable sketch data—a problem that GAN-based hand-drawn sketch-to-image translation must address.

1.1 Abstract and Diverse Nature of Hand-drawn Sketches

Hand-drawn sketches represent a vivid data form that is concise and abstract, fundamentally different from pixel-dense natural images. First, sketches are abstract with sparse strokes and monochromatic colors; non-professionals typically depict objects with minimal strokes. Second, sketches exhibit diversity—different people possess distinct drawing styles. As shown in Figure 3, sketches of the same pair of shoes vary completely across individuals. Finally, hand-drawn sketches often contain redundant and noisy strokes, introducing certain noise.

Sketches and images belong to different data domains, making sketch-to-image translation a cross-modal conversion problem. In contrast, general image-to-image translation represents a single-modal task that incorporates hard constraints like pixel correspondence [7], ensuring strict alignment between output and input edges. Compared to general image translation, hand-drawn sketch-to-image translation possesses unique characteristics. First, sketch strokes do not precisely align with object boundaries and lack color, requiring deformation correction and colorization during conversion. Second, sketches contain minimal information about backgrounds and details, forcing generation models to insert additional information independently. Finally, sketch strokes contain detailed features that models must learn to handle, such as metal decorations on shoe surfaces depicted by sketch strokes in Figure 3 [8].

1.2 Lack of Paired Hand-drawn Sketch Data

Sketch-to-image translation constitutes cross-modal conversion, requiring both sketch and image data for model training. Table 1 summarizes datasets used in existing sketch-to-image translation research. Only Sketchy Database [9], ShoeV2 [8], and ChairV2 [8] contain both modalities; other datasets contain only real images or sketches. For data without sketches, researchers employ specific augmentation methods (Table 2). For sketch-only data, sketch-image embedding methods select images most similar to collected sketches for data augmentation.

Table 1. Datasets Used by Existing Sketch-to-Image Translation Works

Dataset	Literature	Description
Sketchy Database [9]	[1]	125 categories, 75,471 sketches of 12,500 objects
CelebA [10]	[11]	~200,000 face images
Caltech-UCSD Birds-200-2011 [12]	[11]	11.7k bird images
Stanford Cars Dataset [13]	[11]	16k car images
Flickr-Faces-HQ (FFHQ) [14]	[15], [16]	70,000 portrait images
CUFS [17]	[18]	606 face-sketch pairs
CelebAMask-HQ [19]	[5], [20]	30k high-resolution face images with segmentation masks
CelebA-HQ Dataset [21]	[22], [23]	30k portrait images
COCO Stuff [24]	[25]	91 stuff categories, 164k images with annotations
Tuberlin Dataset [26]	[28]	250 categories, 20k sketches
QuickDraw [27]	[29]	345 categories, 50M sketches
ShoeV2 [8]	[25]	1,297 sketches, 400 images
ChairV2 [8]	[25]	-
SketchyCOCO [25]	[28]	60k+ sketch-image pairs across 17 categories
Oxford-102 Dataset [30]	[31]	-

1.3 Difficulty in Imitating Human Hand-drawn Sketches

Currently, publicly available paired sketch-image datasets remain limited. Some studies employ human annotators to draw sketches [5, 32], typically achieving better translation results, yet manual sketching proves costly and unsuitable for large-scale dataset generation. Consequently, researchers have proposed various sketch augmentation methods (Table 2) for training with augmented sketches and images. These methods fall into three categories: (1) Extracting edge maps from real images as sketches using Holistically-Nested Edge Detection (HED) [33], XDoG [34] edge detectors, or FDoG [35] filters—details depend heavily on threshold values; (2) Generating sketches via image-to-sketch translation networks like Im2pencil [36] and Photosketching [37], which capture target contours well but cannot imitate sparse, abstract hand-drawn sketches; (3) Abstracting strokes to mimic hand-drawn sketches through random deformation of edge map strokes or line simplification to remove redundant, scribbled edges, making only minor modifications to original strokes. Overall, existing augmentation methods either directly extract edge maps or utilize sketch translation networks, yet these cannot simulate novice users' sparse strokes. Developing new augmentation methods or improving model generalization to hand-drawn sketches represents a key challenge.

Table 2. Sketch Augmentation Methods

Method	Literature	Advantages	Disadvantages
Extracting Real Image Edges as Sketches			
HED [33]	[1]	Complete object contour extraction	Precisely aligned with boundaries, contains excessive background
XDoG [34] Edge Detector	[1], [38]	Reduces excessive detail strokes	Cannot create deformed strokes to mimic sparse sketches
Photoshop Photocopy [39]	[11], [18]	Clear boundaries	Contains detail strokes, aligned with boundaries
FDoG [35] Filter	[5], [11], [16], [18]	Extracts relatively complete boundary information	Discontinuous strokes, contains shadow details
Semantic Map Boundaries	[11]	Obvious contour information	Contains shadow details and strokes, boundary-aligned

Method	Literature	Advantages	Disadvantages
Sketch Master [40]	[20]	Not precisely boundary-aligned	Only contains semantic map boundary lines
Discrete H-channel Edge Detection	[41]	Relatively complete object information	Excessive shadow details, boundary-aligned
Network-Generated Sketches			
Im2pencil [36]	[38]	Extracts apparel patterns and accessory edges	Precisely boundary-aligned
Photosketching [37]	[42]	Relatively complete object information	Cannot imitate sparse, deformed sketches
Unsupervised Sketch Generation Network	[18]	Sparse strokes with some deformation	Partially discontinuous, incoherent strokes
Stroke Abstraction			
TOM Stroke Deformation [43]	[23]	Generates 10 different sketch styles	Limited contour deformation, some incoherent lines
Stroke Simplification [44]	[20]	Mimics human sketches not boundary-aligned	Style gap with real hand-drawn sketches
Moving Least Squares Contour Deformation	[11], [16], [18], [22], [42]	Clearer boundaries, removes shadow details	Partially incoherent strokes
Integral Vector Fields [45] Simulating Strokes	[31]	Achieves deformed hand-drawn sketch style	Cannot imitate excessively exaggerated deformations

	[46]	Generates hand-drawn style strokes	Dense strokes, single style

2 GAN-based Sketch-to-Image Translation Methods

Sketch-to-image translation aims to learn cross-domain mapping from sketches to images. Based on control granularity over generated images, existing research divides into two categories: (1) Uncontrolled sketch-to-image translation, where most methods employ Conditional GANs (CGAN) [47] with paired or unpaired data; (2) Finely controlled sketch-to-image translation, where researchers propose using attributes and strokes to control output, as the mapping from sketch to image is inherently multimodal.

2.1 Uncontrolled Sketch-to-Image Translation

Sketch-to-image translation seeks to learn conversion between two distinct image domains. By training approach, methods categorize into supervised and unsupervised (Table 3). Supervised methods use paired sketches and images for one-to-one mapping with conditional GANs. Unsupervised methods employ dual GANs to map images from source to target domain and back, enabling training with unpaired data.

2.1.1 Supervised Methods Pix2pix [7] serves as a general image translation framework and common baseline. However, it is not sketch-specific, producing reasonable results only with professional realistic sketches or edge maps. Its translation process infers missing texture or shading information between strokes, failing with sparse hand-drawn sketches. Image-to-image models generally cannot generate sketches due to large domain gaps preventing direct pixel-wise alignment in visual space. Pix2pixHD [48] also performs image-to-image translation, generating 2048$×$1024 resolution images, but cannot handle hand-drawn sketches.

Hand-drawn sketches, as a universal expression, depict diverse content. Methods can be categorized by generated object type: multi-category images, hairstyle/faces, and scene-level images.

1) Multi-category Image Generation In 2018, James Hays et al. proposed SketchyGAN [1], training an encoder-decoder model conditioned on class labels from sketch-image pairs. This end-to-end GAN-based multimodal synthesis method generates 50 object categories including horses, sofas, and motorcycles. It replaces convolutional layers with Masked Residual Unit (MRU) blocks in generator and discriminator, extracting features from image pyramids at different scales via masked inputs. To encourage diversity, the authors propose a diversity loss maximizing L1 distance between outputs of two identical input sketches with different noise vectors. Yongyi Lu et al. concurrently proposed

ContextualGAN [11], reframing sketch-to-image conversion as an image completion problem with sketches as weak contextual constraints. Using joint images to learn joint distributions of sketches and corresponding images avoids complex cross-domain learning issues, also enabling image-to-sketch generation. Reference [29] proposes a two-stage sketch-to-edge-to-image model, where inter-feature correlation learning enables category-consistent generation without class labels. To assist novice users, Arnab Ghosh et al. developed iSketchNFill [42], an interactive GAN-based system introducing a gated class-conditioning method to generate 10 image categories (basketballs, chicken, cookies, cupcakes, etc.) from a single generator network. When users draw sketches, the system automatically recommends stroke feedback and performs texture filling based on class conditions. Comprising a shape completion stage [49] and class-conditioned appearance transfer stage based on MUNIT [50], it generates 256$×$256 resolution images.

Overall, multi-category generation requires substantial training data. These methods propose different augmentation techniques, but their augmented sketches resemble edge maps, often failing to produce reasonable results with sparse, abstract human sketches (Figure 4 [51]). Additionally, generated image resolution remains low—SketchyGAN [1] only produces 64$×$64 images.

2) Hairstyle and Face Generation Hair simulation represents a challenging computer graphics problem, requiring simulation of hundreds of thousands of hairs while considering motion characteristics and inter-hair collisions. With generative deep learning, researchers have focused on GAN-based hair generation. HIS [32] proposes a two-stage GAN model for sketch-to-hairstyle conversion, constructing 640 high-resolution hair sketch-image pairs with hair areas limited to 512$×$512. The model inputs hairstyle sketches or low-resolution hair images to produce realistic hair images. Specifically, it first applies the Pix2pix [7] framework to generate coarse hairstyle images, then feeds these into a regeneration network with self-enhancement capability to produce high-quality results. This self-enhancement uses a structure extraction layer to extract texture and orientation maps from hair images, generating finer textures and hair strands.

Face-related problems have long been central to computer vision, including face recognition and detection. Similarly, face synthesis remains a hot topic in generative deep learning. Weihao Xia et al. proposed Cali-Sketch [18], a two-stage network for sketch-based portrait synthesis. Stage one calibrates sparse input sketches into detailed, calibrated edge-map-like sketches. Stage two synthesizes realistic portrait images from refined sketches. Reference [41] uses latent codes for multimodal face image output, but resolution is limited to 64$×64.*Toaddressoverfitting*, *LinGaoetal.proposedDeepFaceDrawing*[5], *generatingrealistic*512×$512 images. Using high-resolution face datasets, they augmented sketches via Photoshop photocopy [39] plus stroke simplification [44]. To generate high-quality faces from rough, sparse, or incomplete sketches, they treat augmented sketches as soft constraints, adopting a local-to-global approach. The face is divided into five key components (left eye, right eye, nose, mouth, remaining face), learning feature embeddings for each component. A deep neural network maps embedded component features to realistic images, using manifold projection to improve generation quality and robustness for hand-drawn sketches. Yuhang Li et al. proposed DeepFacePencil [20], employing a Spatial Attention Pooling (SAP) module to adaptively balance spatially-varying trade-offs between realism and sketch consistency. Using a dual-generator framework, SAP identifies locally unrealistic strokes and corrects synthesized facial regions from imperfect sketches to realistic domains. pSp [22] is a general framework combining encoders with StyleGAN2 [52] decoders for sketch-to-image conversion, enabling diverse outputs beyond frontal faces. However, sketch geometry is encoded in latent codes, so pSp-generated faces often do not faithfully respect input sketches, and its style-mixing operations adversely affect geometric realism.

Currently, most work focuses on frontal face generation, leveraging fixed facial structures to produce high-quality images. Future challenges include exploring other attributes like head pose and lighting, and overcoming sketch semantic ambiguity to generate accurate boundaries for hair, backgrounds, and necks.

3) Scene-level Image Generation Unlike single-object images, scene-level images involve complex structures with multiple objects and intricate background relationships. Chengying Gao et al. proposed SketchyCOCO [25], focusing on generating entire scene images from hand-drawn sketches. Accounting for varying sketch roughness, it sequentially generates foreground and background. Foreground includes animals (deer, zebra, elephant) from the dataset, while background comprises grass, blue sky, and trees. Foreground generation aims to match user requirements, while background aligns with sketches. For abstract, diverse foreground sketches, the authors designed EdgeGAN, requiring no paired hand-drawn sketches and images during training—only images and corresponding edge maps. The method learns a shared attribute vector representation for images and edge maps, then maps sketch attribute vectors to corresponding images. Background generation uses the Pix2pix [7] architecture, feeding generated foreground and background sketches to produce 128$×128*and*256×$256 scene-level images.

Scene-level synthesis remains limited, with low-resolution outputs. Dataset construction depends on advanced sketch segmentation techniques for abstract sketches. Reference [31] uses dual-level cascaded GANs to generate higher-resolution, richly textured images for cats and flowers, proposing moving least squares to deform extracted edge contours and simulate hand-drawn styles. Reference [38] focuses on Chinese ethnic costume sketch-to-image translation, designing specialized contour extraction and edge processing methods to mimic sketch styles. However, both methods produce insufficiently realistic images and cannot handle sketches with dense strokes or exaggerated lines.

2.1.2 Unsupervised Methods Due to high cost and difficulty acquiring paired data, researchers have developed unsupervised methods. In general image translation, CycleGAN [53] pioneered unsupervised approaches. MUNIT [50] decomposes images into content and style components, sampling from different spaces for reconstruction to achieve many-to-many domain mapping. U-GAT-IT [54] proposes an attention module to distinguish source and target domains, with AdaLIN functions flexibly controlling shape and texture changes. These methods are not sketch-specific and cannot effectively handle sparse, geometrically deformed human sketches.

US2P [28] is a two-stage unsupervised model using unpaired sketch-image data, generating diverse realistic images. It first converts input sketches to grayscale images via cycle-consistency loss [53] supervision, then performs sample-based colorization using a separate GAN model.

The first stage performs shape translation to handle spatial deformations, including abstract lines and varied drawing styles. Using unpaired sketches and grayscale images, it includes sketch-to-grayscale and grayscale-to-sketch mappings supervised by cycle-consistency loss, similar to CycleGAN [53]. For sketch particularities—dense useless strokes or detail noise—it introduces self-supervision and attention modules. The self-supervision module restores noisy sketches to clean originals (Figure 5 [28]). Given large blank sketch regions, the attention module re-weights attention maps to suppress activation in dense stroke areas, ignoring noise interference (Figure 6 [28]). The second stage, content enrichment, generates detailed color images from grayscale maps. Using paired grayscale and color images provides reference style guidance, following AdaIN [55] to diversify outputs via feature adjustment.

Since the shape translation network is bidirectional, US2P [28] can also convert images to sketches and apply to unsupervised sketch-based retrieval. However, US2P focuses only on shoes and sofas with limited sketch data, generating 128$×$128 resolution images. As paired hand-drawn sketches and images are difficult to obtain, future work should break cycle-consistency loss bottlenecks and explore more advanced unsupervised methods.

2.2 Finely Controlled Sketch-to-Image Translation

While some methods support multimodal generation, attributes and styles remain uncontrollable. To enable fine-grained user control, researchers have proposed methods controlling image attributes and strokes (Table 4).

Table 4. Methods of Finely Controlled Sketch-to-Image Translation

Method	Subject	Resolution	Interactive	Control Mechanism
BHS [46]	Hair/Beard	512$×512|*Yes*|*Vectorfieldintegration,decoupledstructureandcolorattributes*||*Micha* *completion*||*DeepFaceEditing*[16]|*Face*|512×512|*Yes*|*Sketcheditingwithreferencein* *supervisedAE*[57]*forcontent−styledisentanglement,momentum*[58]*mutualinformationminimization*||*SketchYou*	No	Fine-tunes pretrained models with few sketches to match shape and pose

2.2.1 Image Attribute Control Image attribute control decomposes target images into visual attributes, designing corresponding modules for each. Texture and style better help users specify desired targets, prompting research on exemplar-based translation. Exemplar-based methods translate images (e.g., semantic maps, skeletons, edge maps) according to reference images with desired style (color, texture). Networks receive both source and target exemplar images with similar semantics, learning to output images matching the specified style. CoCosNet [61] establishes dense semantic correspondence between input and exemplar images to locate corresponding color and texture information, enabling style-consistent generation for image editing and facial makeup. RBNet [62] colors sketches or edge maps using reference images. Reference [59] proposes an artistic style exemplar-based method generating 512$×$512 images, using SketchyGAN [1] augmentation and demonstrating human body generation. However, these are not designed for human hand-drawn sketches.

For controllable hair manipulation, MichiGAN [15] proposes interactive portrait hair generation conditioned on disentangled attributes (shape, structure, appearance, background), enabling local and detailed editing via reference portraits or paintings. Reference [46] also interactively synthesizes hair and beards. DeepFaceEditing [16] is Lin Gao et al.' s latest work—a structural disentanglement framework for face images enabling generation and editing through geometry-appearance disentanglement. It decomposes local component images into geometry and appearance representations, then globally fuses them for high-quality results. By extracting geometry from sketches, it supports face editing via sketches. SSS2I [23] is an exemplar-based synthesis method with hand-drawn sketches. To address paired data scarcity, it proposes TOM, an unsupervised domain-transfer GAN model treating sketch synthesis as mapping from RGB domain R to line sketch domain S, synthesizing multiple sketches per image via online feature matching. The sketch-to-image generation comprises two stages: first converting sketches to color images, then refining details, resolution, and quality via adversarial networks. Using synthetic paired data, a self-supervised autoencoder (AE) [57] disentangles content and style features, with a style classifier further separating them before generation.

2.2.2 Stroke Control For hair generation, researchers believe colored strokes provide attribute guidance. BHS [46] uses sketch-like "guide strokes" to describe local hair shape and color for easier interaction. Editing a vector field extracted from hair information adjusts overall hairstyle structure with minimal user input, enabling subtle local changes through synthesized guide strokes for editing, adding, or deleting individual strokes. Hongbo Fu et al. argue colored hair sketches implicitly contain target shape and appearance information, proposing SketchHairSalon [6]—a novel framework synthesizing realistic 512$×$512 hair images directly from colored strokes. It comprises S2M-Net (sketch-to-matte) and S2I-Net (sketch-to-image) generators with self-attention modules. For training, they constructed a new dataset with thousands of manually annotated hair sketch-image pairs and corresponding hair masks. Its interface

(Figure 2) includes hair structure customization, shape optimization, appearance customization, and auto-completion. Since training high-quality generative models requires large datasets and high-performance computing, Reference [60] proposes customizing generative models with few sketch examples. Leveraging pretrained models on large-scale data, it adjusts model weight subsets to match user sketches via cross-domain fine-tuning, creating images similar to user sketches while preserving color, texture, and detail from the pretrained model.

Current work primarily focuses on fine control for hair and face tasks, with strong algorithm specificity unsuitable for other tasks. In art design, fine-grained control over generated images or editing can assist designers, offering significant commercial value and remaining highly challenging—a promising future direction.

3 Result Evaluation

Evaluating generative model performance is complex. Since quantitative metrics often lack consistency with human perception [63], many studies rely on qualitative human evaluation. For specific tasks, evaluation should consider not only final image quality but also input matching and intended application suitability. GAN-based hand-drawn sketch-to-image translation evaluation comprises qualitative and quantitative approaches (Table 5).

Table 5. Evaluation Metrics for Sketch-to-Image Translation

Type	Method	Description	Literature
Qualitative	Perception Study	Untrained participants evaluate generated images via online questionnaires	[1], [5], [16]
	Usability Study	Users test systems and evaluate usefulness/effectiveness	[1], [5], [16]
	Generalization Comparison	Test with sparse/exaggerated sketches from novices	[6]
	Ablation Study	Analyze component effectiveness	[5], [6]
	Comparison with SOTA	Compare against advanced models	[1], [16], [20]
Quantitative	FID [64]	Measures distribution similarity	[11], [25]
	LPIPS [65]	Perceptual similarity via deep features	[23]

Type	Method	Description	Literature
	IS [66]	KL divergence of class distributions	[23]
	SR [61]	Style correlation for color/texture consistency	[25]
	L2 Gabor Feature [67]	Shape similarity	[42]
	SAD [68]	Sum of absolute differences for matte accuracy	[5]
	SSIM [69]	Structural similarity	[5]

Qualitative Evaluation includes perception studies, usability studies, generalization comparisons, ablation studies, and comparisons with state-of-the-art models. Perception studies invite non-experts to evaluate generated images via questionnaires and scoring. Usability studies have users experience systems directly. Generalization testing uses sparse or exaggerated novice sketches, as most training data comprises edge maps or professional sketches. These qualitative methods are most direct and effective, truly reflecting generation quality.

Quantitative Metrics include Fréchet Inception Distance (FID) [64] for distribution similarity, Learned Perceptual Image Patch Similarity (LPIPS) [65] using deep features, Inception Score (IS) [66] via ImageNet-pretrained models, Style Correlation (SR) [61] for color/texture consistency, L2 Gabor features [67] and Structural Similarity Index (SSIM) [69] for shape similarity. Reference [5] uses Sum of Absolute Differences (SAD) [68] for hair matte accuracy and Intersection over Union (IoU) for boundary region evaluation.

While some metrics demonstrate effectiveness, different methods suit different models. IS [66] has limitations and does not reflect realism. FID [64] evaluates non-ImageNet data but cannot indicate overfitting. SSIM [69] performs well for denoising and similarity assessment, serving as a widely-used quality metric.

4 Conclusion

GAN-based hand-drawn sketch-to-image translation enables controllable image generation through sketches. This paper analyzed challenges, summarized related work and evaluation metrics, and identified future trends. Despite existing research, the field remains nascent. Human sketches are complex and varied, with many valuable problems awaiting solutions.

a) Hand-drawn Sketch Data Augmentation. The lack of large-scale sketch-image datasets and the time-consuming nature of sketch collection hinder progress. Different translation tasks require different datasets, necessitating large-scale training data. Existing augmentation methods—global transformations (rotation, shifting), stroke deformation, or thickening—fail to mimic authentic drawing styles [70]. Reference [23] explores unsupervised sketch synthesis, but results resemble professional realistic styles. As shown in Figure 7 [51], models trained on synthetic sketches cannot generalize to real sketches. Synthesizing sketches that imitate diverse human drawing styles and bridging the domain gap represents a key future challenge.

b) Fine-grained Control over Generated Images. While multimodal translation is supported, controlling specific texture, color, and material features remains difficult. Exemplar-based methods enable style control via single reference images. Future work referencing multiple style exemplars or using colored strokes holds commercial value, potentially reducing repetitive work in animation, film, and game storyboarding. In art design, representing material properties beyond color to better assist designers represents a valuable direction.

c) Sketch-to-Artistic-Style Image Generation. Most research focuses on synthesizing realistic natural photos from sketches. Artistic images differ in style diversity, affecting how sketches synthesize into full-color textured images. Reference [59] studies artistic style synthesis (e.g., impressionism, realism), though some style features prove difficult to learn, challenging model balancing between sketch semantics and style references. Converting sketches to artistic paintings can advance deep networks' ability to capture and translate various artistic styles. Future work could serve both entertainment—allowing users to experience artistic creation—and professional applications, synthesizing images from multiple styles to assist creative artists.

References

[1] Chen Wengling, Hays J. Sketchygan: towards diverse and realistic sketch to image synthesis [C]// Proc of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 9190-9200.

[2] Chen Tao, Cheng Mingming, Tan Ping, et al. Sketch2photo: internet image montage [J]. ACM Transactions on Graphics, 2009, 28 (5): 1-10.

[3] Eitz M, Richter R, Hildebrand K, et al. Photosketcher: interactive sketch-based image synthesis [J]. IEEE Computer Graphics and Applications, 2011, 31 (6): 56-66.

[4] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C]// Proc of the 28th Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680.

[5] Chen Shuyyu, Su Wanchao, Gao Lin, et al. Deepfacedrawing: deep generation of face images from sketches [J]. ACM Trans on Graphics, 2020, 39 (4): 72: 1-72: 16.

[6] Xiao Chufeng, Yyu Deng, Han Xiaoguang, et al. Sketchhairsalon: deep sketch-based hair image synthesis [J]. ACM Trans on Graphics, 2021, 40 (6):

216: 1-216: 16.

[7] Isola P, Zhu Junyan, Zhou Tinghui, et al. Image-to-image translation with conditional adversarial networks [C]// Proc of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 1125-1134.

[8] Yyu Qian, Liu Feng, Song Yizhe, et al. Sketch me that shoe [C]// Proc of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 799-807.

[9] Sangkloy P, Burnell N, Ham C, et al. The sketchy database: learning to retrieve badly drawn bunnies [J]. ACM Trans on Graphics, 2016, 35 (4): 119: 1-119: 12.

[10] Liu Ziwei, Luo Ping, Wang Xiaogang, et al. Deep learning face attributes in the wild [C]// Proc of the 2015 IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 3730-3738.

[11] Lu Yongyi, Wu Shangzhe, Tai Y, et al. Image generation from sketch constraint using contextual gan [C]// Proc of the 2018 European Conference on Computer Vision. Berlin, Springer Press, 2018: 205-220.

[12] Wah C, Branson S, Welinder P, et al. The Caltech-UCSD Birds-200-2011 dataset [EB/OL]. (2011) [2022-01-01]. citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.372.852&rep=rep1&

[13] Krause J, Stark M, Deng Jia, et al. 3D object representations for fine-grained categorization [C]// Proc of the 2013 IEEE International Conference on Computer Vision Workshops. Washington: IEEE Computer Society Press, 2013: 554-561.

[14] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks [C]// Proc of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2019: 4401-4410.

[15] Tan Zhentao, Chai Menglei, Chen Dongdong, et al. Michigan: multi-input-conditioned hair image generation for portrait editing [J]. ACM Trans on Graphics, 2020, 39 (4): 95: 1-95: 13.

[16] Chen Shuyyu, Liu Fenglin, Lai Yyukun, et al. Deepfaceediting: deep face generation and editing with disentangled geometry and appearance control [J]. ACM Trans on Graphics, 2021, 40 (4): 90: 1-90: 15.

[17] Wang Xiaogang, Tang Xiaoou. Face photo-sketch synthesis and recognition [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2009, 31 (11): 1955-1967.

[18] Xia Weihao, Yang Yyujiu, Xue Jinghao. Cali-sketch: stroke calibration and completion for high-quality face image generation from poorly-drawn sketches [EB/OL]. (2019-11-01) [2022-1-13]. https://doi.org/10.48550/arXiv.1911.00426.

[19] Lee C, Liu Ziwei, Wu Lingyun, et al. Maskgan: towards diverse and interactive facial image manipulation [C]// Proc of the 2020 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 5548-5557.

[20] Li Yyuhang, Chen Xuejin, Yang Binxin, et al. Deepfacepencil: creating face images from freehand sketches [C]// Proc of the 28th International Conference on Multimedia. New York: ACM Press, 2020: 991-999.

[21] Karras T, Aila T, Laine S, et al. Progressive growing of gans for improved quality, stability, and variation [C/OL]/ Proc of the 6th International Conference on Learning Representations. (2018) [2022-01-01]. https://arxiv.org/abs/1710.10196.

[22] Richardson E, Alaluf Y, Patashnik O, et al. Encoding in style: a stylegan encoder for image-to-image translation [C]// Proc of the 2021 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2021: 2287-2296.

[23] Liu Bingchen, Zhu Yizhe, Song Kunpeng, et al. Self-supervised sketch-to-image synthesis [C]// Proc of the 35nd AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2021: 2073-2081.

[24] Caesar H, Uijlings J, Ferrari V. Coco-stuff: thing and stuff classes in context [C]// Proc of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 1209-1218.

[25] Gao Chengying, Liu Qi, Xu Qi, et al. Sketchycoco: image generation from freehand scene sketches [C]// Proc of the 2020 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 5174-5183.

[26] Eitz M, Hays J, Alexa M. How do humans sketch objects? [J]. ACM Trans on Graphics, 2012, 31 (4): 44: 1-44: 10.

[27] Ha D, Eck D. A neural representation of sketch drawings [C/OL]/ Proc of the 6th International Conference on Learning Representations. (2017) [2022-01-01]. https://arxiv.org/abs/1704.03477.

[28] Liu Runtao, Yyu Qian, Yu S. Unsupervised sketch-to-photo synthesis [C]// Proc of the 16th European Conference on Computer Vision. Berlin, Springer Press: 2020: 36-52.

[29] Zong Yyujia. A two-stage method and application implementation for image generation from sketch [D]. Dalian: Dalian University of Technology, 2021.

[30] Nilsback M, Zisserman A. Automated flower classification over a large number of classes [C]// Proc of the 6th Indian Conference on Computer Vision, Graphics&Image Processing. Washington: IEEE Computer Society Press, 2008: 722-729.

[31] Cai Yyuting, Chen Zhaojiong, Ye Dongyi. Bi-level cascading GAN-based heterogeneous conversion of sketch-to-realistic images [J]. Pattern Recognition

and Artificial Intelligence, 2018, 31 (10): 877-886.

[32] Qiu Haonan, Wang Chuan, Zhu Hang, et al. Two-phase hair image synthesis by self-enhancing generative model [J]. Computer Graphics Forum, 2019, 38 (7): 403–412.

[33] Xie Saining, Tu Zhuowen. Holistically-nested edge detection [C]// Proc of the 2015 IEEE International Conference on Computer Vision. Washington: IEEE Computer Society Press, 2015: 1395–1403.

[34] Winnem H, Kyprianidis, J E, Olsen S. Xdog: an extended difference-of-gaussians compendium including advanced image stylization [J]. Computers & Graphics, 2012, 36 (6): 740 -753.

[35] Kang H, Lee S, Chui C. Coherent line drawing [C]// Proc of the 5th International Symposium on Non-Photorealistic Animation and Rendering. New York: ACM Press, 2007: 43–50.

[36] Li Yijun, Chen Fang, Hertzmann A, et al. Im2pencil: controllable pencil illustration from photographs [C]// Proc of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2019: 1525–1534.

[37] Li Mengtian, Lin Zhe, Mech R, et al. Photosketching: inferring contour drawings from images [C]// Proc of the 2019 IEEE Winter Conference on Applications of Computer Vision. Piscataway, NJ: IEEE Press, 2019: 201-210.

[38] Liu Bo. Automatic coloring method for national costume sketches [D]. Yunnan: Yunnan Normal University, 2020.

[39] Photocopy. [2022-01-01]. Create filter gallery photocopy effect with single photoshop. https://www.youtube.com/watch?v=QNmniB_{5Nz0}.

[40] Sketch master. [2022-01-01]. http://www.ouyaoxiazai.com/soft/txtx/108/8389.htm1.

[41] Wang Pengcheng. Research on gan translation from sketch to real image based on perceptual attention and latent space [D]. Anhui: Anhui University, 2020.

[42] Ghosh A, Zhang R, Dokania P, et al. Interactive sketch&fill: multi-class sketch-to-image translation [C]// Proc of the 2019 IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2019: 7775-7784.

[43] AutoTrace. [2022-01-01]. http://autotrace.sourceforge.net/.

[44] Simo-Serra E, Iizuka S, Sasaki K, et al. Learning to simplify: fully convolutional networks for rough sketch cleanup [J]. ACM Trans on Graphics, 2016, 35 (4): 121: 1–121: 11.

[45] Kyprianidis J E, Kang H. Image and video abstraction by coherence-enhancing filtering [J]. Computer Graphics Forum, 2011, 30 (2): 593-602.

[46] Olszewski K, Ceylan D, Xing Jun, et al. Intuitive, interactive beard and hair synthesis with generative models [C]// Proc of the 2020 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 7446-7456.

[47] Mirza M, Osindero S. Conditional generative adversarial nets [EB/OL]. (2014) [2022-01-01]. https://arxiv.org/abs/1411.1784.

[48] Wang Tingchun, Liu Mingyyu, Zhu Junyan, et al. High-resolution image synthesis and semantic manipulation with conditional gans [C]// Proc of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 8798-8807.

[49] Mescheder L, Geiger A, Nowozin S. Which training methods for gans do actually converge? [C]// Proc of the 35th Annual International Conference on Machine Learning. New York: ACM Press, 2018: 3478-3487.

[50] Huang Xun, Liu Mingyu, Belongie S, et al. Multimodal unsupervised image-to-image translation [C]// Proc of the 15th European Conference on Computer Vision. Berlin, Springer Press, 2018: 172-189.

[51] Xiang Xiaoyyu, Liu Ding, Yang Xiao, et al. Adversarial open domain adaption for sketch-to-photo synthesis [C/OL]/ Proc of the 2022 IEEE Conference on Applications of Computer Vision. (2021) [2022-01-01]. https://arxiv.org/abs/2104.05703.

[52] Karras T, Laine S, Aittala M. Analyzing and improving the image quality of stylegan [C]// Proc of the 2020 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 8107-8116.

[53] Zhu Junyan, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks [C]// Proc of the 2017 IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2017: 2242-2251.

[54] Kim J, Kim M, Kang H, et al. U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation [C/OL]/ Proc of the 8th International Conference on Learning Representations. (2020-04-08) [2022-01-01]. https://arxiv.org/abs/1907.10830.

[55] Huang Xun, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization [C]// Proc of the 2017 IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2017: 1510-1519.

[56] Fu Jun, Liu Jing, Tian Haijie, et al. Dual attention network for scene segmentation [C]// Proc of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2019: 3146-3154.

[57] Kingma D P, Welling M. Auto-encoding variational bayes [C/OL]/ Proc of the 2nd International Conference on Learning Representations. (2013-12-20) [2022-01-01]. https://arxiv.org/abs/1312.6114.

[58] He Kaiming, Fan Haoqi, Wu Yyuxin, et al. Momentum contrast for unsupervised visual representation learning [C]// Proc of the 2020 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 9729-9738.

[59] Liu Bingchen, Song Kunpeng, Zhu Yizhe, et al. Sketch-to-art: synthesizing stylized art images from sketches [C]// Proc of the 15th Asian Conference on Computer Vision. Berlin, Springer Press, 2020: 315-330.

[60] Wang Shengyyu, Bau D, Zhu Junyan. Sketch your own gan [C]// Proc of the 2021 IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2021: 14030-14040.

[61] Zhang Pan, Zhang Bo, Chen Dong, et al. Cross-domain correspondence learning for exemplar-based image translation [C]// Proc of the 2020 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 5143-5153.

[62] Lee J, Kim E, Lee Y, et al. Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence [C]// Proc of the 2020 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2020: 5800-5809.

[63] Lucic M, Kurach K, Michalski M, et al. Are gans created equal? a large-scale study [C]// Proc of the 2018 Annual Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2018: 698-707.

[64] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local Nash equilibrium [C]// Proc of the 2017 Annual Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2017: 6626-6637.

[65] Zhang R, Isola P, Efros A, et al. The unreasonable effectiveness of deep features as a perceptual metric [C]// Proc of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 586-595.

[66] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training gans [C]// Proc of the 2016 Annual Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2016: 2226-2234.

[67] Eitz M, Richter R, Boubekeur T, et al. Sketch-based shape retrieval [J]. ACM Trans on Graphics. 2012, 31 (4): 31: 1-31: 10.

[68] Li Yaoyi, Lu Hongtao. Natural image matting via guided contextual attention [C]// Proc of the 34th Conference on American Association for Artificial Intelligence. New York, AAAI Press, 2020: 11450-11457.

[69] Wang Zhou, Bovik A, Sheikh H, et al. Image quality assessment: from error visibility to structural similarity [J]. IEEE trans on image processing. 2004, 13 (4): 600-612.

[70] Xu Peng, Hospedales T, Yin Qiyue. Deep learning for free-hand sketch: a survey [J/OL]. IEEE Trans on Pattern Analysis and Machine Intelligence. (2020-06-01) [2022-01-01]. http://doi.org/10.1109/TPAMI.2020.2997469.