人脸-素描合成: 一个新的挑战

范登平[1], 黄子凌†[2], 郑鹏†[3], 刘弘*[4], 秦雪彬*[3] and Luc Van Gool[1]

[1]Computer Vision Lab, ETH Zürich, Zürich, Switzerland.
[2]Information and Communication Engineering, University of Tokyo, Tokyo, Japan.
[3]Computer Vision, MBZUAI, Abu Dhabi, UAE.
[4]Digital Content and Media Sciences Research Division, NII, Tokyo, Japan.

Abstract

本文旨在对人脸素描合成（FSS）进行全面研究。然而，由于获取手绘草图数据集的高成本，缺乏一个完整的基准来评估过去十年中FSS算法的发展。本文首次构建了一个高质量的FSS数据集，名为**FS2K**，它由2,104个图像-素描对组成，并且标记了素描风格（三种类型）、图像背景、光照条件、肤色和人脸属性。FS2K在难度、多样性和可扩展性方面不同于以前的FSS数据集，因此有助于促进FSS的研究。第二，本文通过回顾**89**种经典方法从而展示了最大规模的FSS研究，其中包括**25**种基于手工特征的人脸素描合成方法、**29**种通用转换方法和**35**种图像-素描合成方法。此外，本文对现有的**19**个前沿模型进行了综合实验。第三，本文提出了一个简单的FSS基线模型，命名为**FSGAN**。它只有两个简单的组件，即人脸感知掩膜和风格向量扩张。在FS2K数据集上，FSGAN的性能大大超过了之前所有最先进的模型。最后，本文总结了过去几年的经验教训，并指出了几个未解决的挑战。代码见：https://github.com/DengPingFan/FSGAN.

Keywords: 人脸素描合成，人脸素描数据集，基准，属性，风格转换

1 引言

人脸素描合成（FSS）从人脸RGB图像产生灰度素描（图片到素描，I2S），或者相反（素描到图片，S2I）[2, 3]。FSS通常用于执法或监视，以目击者的素描为基础，协助识别和检索人脸 [2]。对于娱乐用途，素描合成可用于移动app，如TikTok和Facebook。另外，素描合成对于数字娱乐 [4]来说也是很有吸引力的话题。过去十年，FSS的研究已经取得了很大的进展。

不同于其他人脸相关的数据集，比如人脸识别 [5–7]、人脸检测 [8]、人脸关键点检测 [9]、人脸对齐 [10]和人脸合成 [11]，这些数据集不需要标注人员经过训练就可以手工标注，而人脸素描合成数据集的获取则要困难的多，因为只有一些专业的艺术家才能绘制出高质量的参考图像。由于获取专业素描数据需要很高的代价，目前人脸素描数据集 [2, 3, 12]规模较小且多样性有限。这些不足已严重限制了FSS的发展，尤其是对需要大量数据的深度学习模型。

另外，如何评价FSS模型仍然是一个有待讨论的问题。结构相似度（SSIM）[13]是评价图像质量最为广泛的评价指标之一，所以它通常也被用来评价S2I模型的性能。然而，人脸素描的特性与基于RGB的人脸图像有很大的不同，这也使得将当前的评估指标应用于I2S任务变得具有挑战性。因此，需要一种新的客观的、定量的、与人工评估高度一致的指标来评测FSS任务。

此外，由于缺少高质量数据集和合适的评价指标，不同的FSS模型（比如 [2, 3]）通常建立在不同的训练数据集上[1]，并使用不同的评

† 同等贡献；* 通讯作者。本文为 [1]的中文翻译版。由李宁翻译，郑鹏、刘弘和范登平校稿。

[1]因为他们希望学习不同风格的素描。

估方法进行测试。因此，很难提供公平且全面的比较。进一步说，许多图像-图像变换相关任务的先进变换模型也可以用到FSS任务，例如，CycleGAN [14]、UNIT [15]、Pix2pixHD [16]、SPADE [17]、DSMAP [18]、NICE-GAN [19]和DRIT++ [20]。然而，因为数据集和评价指标的不足，这些模型缺乏对于FSS任务的性能评价。因此，采用一个标准的度量指标并在一个标准的FSS数据集上对FSS相关的模型进行全面的对比和评价已是当务之急。为此，本文提出并维护一个在线文章列表（https://github.com/DengPingFan/FaceSketch-Awesome-List），目的就是为了追踪这个快速发展的领域的进展。

1.1 贡献

本文的目标是解决这些悬而未决的问题（比如，有限的数据集、度量指标和基准）并进一步为FSS社区带来新的挑战。主要的贡献如下：

1) **FSS 数据集。** 本文构建了一个新的高质量FSS数据集，名为**FS2K**。此数据集为目前最大（见表 1）公开的FSS数据集[2]，包含2,104组图像-素描对，并且搜集的人脸图像包含多种图像背景、肤色、素描风格以及光照条件。此外，本文提供额外的人脸属性，比如，性别、笑容、发型等等，目的就是为了使深度学习模型学习到更多详细的线索。
2) **FSS综述和基准。** 本文进行了大规模的FSS调研，综述了89个有代表性的方法，包括25种手工设计特征的模型、29种用于通用转换任务的模型和35种I2S转换算法。基于本文提出的FS2K，本文采用SCOOT指标 [23]，从内容和风格的角度对19个最先进的模型进行了严格的评估。
3) **FSS 基线方法。** 本文设计了基于GAN的有效的基线方法，名为**FSGAN**，其包含两个核心部分，即人脸感知掩模和风格向量扩展。前者用来修复人脸组件部分的细节，而后者被用来学习不同的人脸风格。在本文新建的FS2K数据集上，**FSGAN**作为I2S和S2I任务（图 1）统一的基准模型。本文的项目可以在此获取https://github.com/DengPingFan/FSGAN。
4) **讨论和未来发展方向。** 除了进行一个整体的性能评价，本文也进行了属性级别的评价，给出了详细的讨论，并探索一些有前景的方向。

[2]建立一个由专业艺术家绘制的FSS数据集比其他人脸数据集更有挑战，比如，人脸属性数据集 [21]。这也是为什么目前现有最大的FSS数据集（即CUFSF [22]）在过去13年中只有~1K张图像。虽然FS2K只有CUFSF大约两倍的数量，我们仍然花费了一年的时间去创建这样一个高质量的数据集。

2 相关工作

本章节首先对现有的FSS数据集进行了一个归纳和整理。然后，在第二部分，本文讨论了人脸合成的分类，并特别强调了这项任务的创新和成功的方法，包括传统的人脸合成、图像到图像转换、神经网络风格转换和深度照片素描合成。人脸素描合成的分类如图 2。关于这些模型的总结，包括他们的关键创新、数据集、代码链接和引用信息，详情见表 3 和表 4。

2.1 数据集

本文概述了4种常用于FSS任务的经典数据集（即CUFS [2]、IIIT-D [24]、CUFSF [22]和VIPSL [26]）和三个画像素描数据集[12, 27, 28]，这些是大部分FSS 模型[33]建立的基础。

CUFS [2]数据集是最早建立并广泛使用的数据集之一。它包含606组照片-素描对，其中123个样本来自AR人脸数据集[34]、188个样本来自CUHK学生数据集以及来自XM2VTS [35]数据集的295个样本。每个样本由艺术家手绘的素描图像和对应的照片组成。每张照片都是在正常的光照条件下以正面姿势拍摄的，并保持中性的表情。所有的三个子数据集都采用单色背景，比如，青色、白色和蓝色。然而，现实场景复杂多样，很难保证照片都是在这样固定的环境下获取。另外，在这个数据集中的素描图像是由同一个艺术家绘制，所以，其风格有限。

CUFSF [22]是一个用来评价FSS模型性能的常用数据集。它包含1,194张来自FERET数据集[36]的照片-素描对。一个艺术家在观察了对应的照片之后绘制所有的素描图像。CUFSF和CUFS有相同的照片收集环境，但它比CUFS更具有挑战性。因为数据集中的每张照片都具有光照变换，每张照片的前景和背景的对比度都是较低，并且每张素描都包含了夸张的形态。

VIPSL [26]数据集含200张来自FRAV2D [138]、文献[36]以及印度人脸数据集 [26]的人脸照片。不同于CUFS和CUFSF，VIPSL每个人脸有五张素描图像，由五个不同风格的艺术家绘制，同时在相同条件下观看相同的照片。

IIIT-D [24, 139]由三种形式的数据集组成，包含一个可视素描数据库，一个semi-forensic素描数据库和一个法医数据库。所有的照片都来自于CUHK学生数据库和IIIT-D素描数据库 [24]。第一个可视化数据库包含238组数字素描图像对，其中所有的素描都是由专业的艺术家在给定的照片的基础上绘制。第二个子数据库一共有140张

表1 FSS数据集对比。

数据集	年份	出版社	总数	训练	测试	属性	开源	图像对	分辨率
CUFS [2]	2009	TPAMI	606	306	300	×	√	√	200 × 250
IIIT-D [24]	2010	BTAS	231	58	173	×	×	√	-
CUFSF [22]	2011	CVPR	1,194	500	694	×	√	√	779.62±15.05 × 812.10±13.92
VIPSL [25, 26]	2011	TCSVT	1,000	100	900	×	×	√	-
DisneyPortrait [27]	2013	TOG	672	-	-	×	×	√	-
UPDG [28]	2020	CVPR	952	798	154	×	×	×	-
APDrawing [12]	2020	TPAMI	140	70	70	×	√	√	512 × 512
FS2K（本文）	2022	MIR	**2,104**	**1,058**	**1,046**	√	√	√	299.74±95.07 × 273.56±38.67

* 在[29]和[30]中，CUFS被分为268和338张图片分别进行训练和测试。对于图片的分辨率，本文分别提供宽和高为W_{avg} ± W_{std}和H_{avg} ± G_{std}。W_{avg}和W_{std}分别表示均值和标准方差。

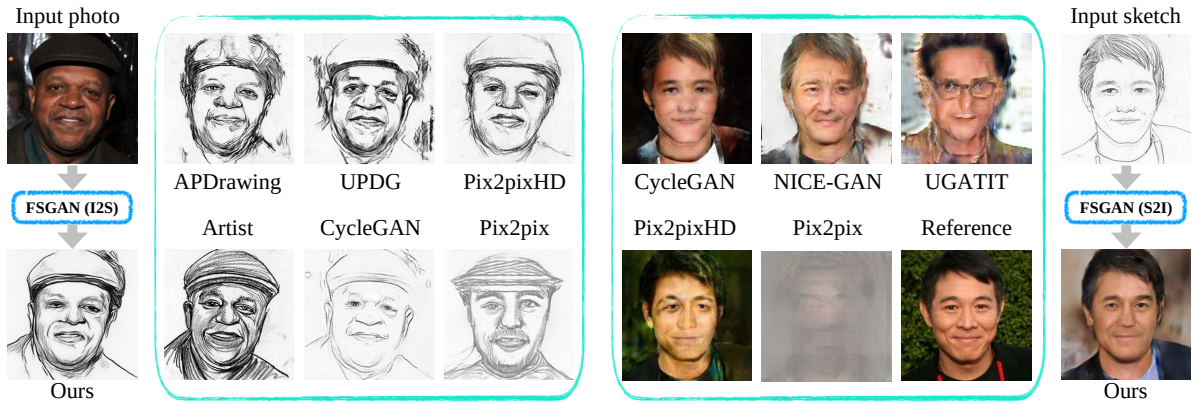


图1 **左:** 本文提出的FSGAN（I2S）从艺术家的素描中学习，并智能地将输入的照片转换为生动的人脸素描。相比之下，这五种新的风格转换方法无法获得视觉上吸引人的效果。只有UPDG [28]和Pix2PixHD [16]表现相对较好，但它们生成的内容和样式比FSGAN差。**右:** 输入一个素描图像，本文提出的FSGAN（S2I）也可以将输入转换为栩栩如生的人脸照片。同时，五种具有代表性的深度学习模型的结果要么结构受损（即CycleGan [14]、NICE-GAN [19]和UGATIT [31]）要么模糊（即Pix2pix [32]）。更多的结果可以见图 8-11。

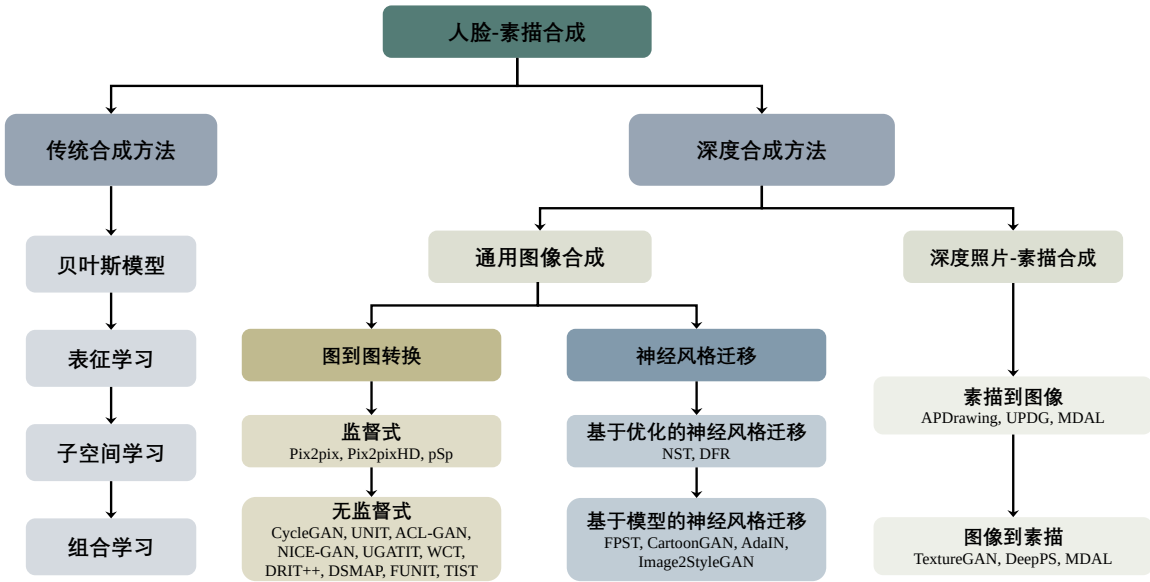


图2 人脸素描合成的分类及其代表性方法。

表2 热门相关工作总结。这些模型可以分为三个种类：传统人脸合成, 通用图像合成, 深度图像到素描合成。

#	模型	出版信息	年份	代码	核心组件	使用数据集	Assist.	引用
				传统人脸合成				
1	EFSGNS [37]	ICCV	2001	-	Active Shape Model, Non-parametric Sampling	E	-	160
2	Nonlinear [38]	CVPR	2005	-	Local Linear Preserving, Eigentransform	Y	-	398
3	E-HMM [39]	TCSVT	2008	-	Embedded Hidden Markov Model, Selective Ensemble	Y	-	165
4	HCM [40]	PAMI	2008	-	Graph, Minimum Description Length	C, D, BW, E	-	93
5	MRF [2]	PAMI	2009	Code	Multi-scale Markov Random Fields	Y	-	872
6	LPR [41]	ECCV	2010	-	Local Evidence Function, Patch Matching, Shape Prior, MRF	Y	-	120
7	LRM [42]	ICIG	2011	-	Local Regression, kNN	Y	-	19
8	MOR [43]	HCII	2011	-	Multivariate Output Regression	Y	-	22
9	MDSR [25]	ICIG	2011	-	LLE, Dictionary Learning, Sparse Representation	Y, BX	-	55
10	SVR [44]	ICIP	2011	-	Support Vector Regression	Y, BX	-	41
11	SCDL [45]	CVPR	2012	-	Sparse Coding, Semi-coupled Dictionary Learning	Y	-	613
12	MWF [46]	CVPR	2012	-	Markov Weight Fields, Cascade Decomposition	Y, E	-	173
13	SR [26]	TCSVT	2012	-	Sparse Neighbor Selection, Sparse-Representation Enhance	Y, BX	-	185
14	SAPS [27]	TOG	2013	-	Edge Detection, Shape Deformation	B	-	116
15	FESM [47]	BMVC	2013	-	Markov Random Field, Graph-cut	E	-	22
16	Transductive [48]	TNNLS	2013	-	Probabilistic graph model,Transductive Learning	Y, CU	-	167
17	CDFSL [49]	ICCV	2013	-	Coupled Dictionary and Feature Space Learning	Y	-	177
18	REB [50]	ECCV	2014	Project	kNN, Linear Estimation, Sketch Denoising	Y, D	-	124
19	RobustStyle [51]	TIP	2015	-	Sparse Representation, Multi-scale Selection	Y, E	-	49
20	SPP [52]	TCSVT	2015	Project	Superpixels, Markov Networks	Y, CU, BY	-	45
21	MR [53]	TNNLS	2016	-	Markov Networks, Edge Enhancement, Alternating Opt.	Y, BY	-	107
22	DSM [54]	IJCV	2017	Project	Perceptual Grouping, Deformable Stroke Model	A, B	-	37
23	AR [55]	NC	2017	-	Adaptive Representation, Markov Networks	Y	-	10
24	RS [56]	PR	2018	-	Offline Random Sampling, Locality Constraint	Y, CU	-	96
25	CFITT [57]	CVPR	2018	Github	PatchMatch, Guided Texture Transfer	E	Sm.	19
				通用图像合成方法				
26	NST [58, 59]	CVPR	2016	Github	Parametric Texture Mode, Representation Inversion	E	-	3853
27	FNS [60]	ECCV	2016	Github	Image Transformation and Loss Network, Perceptual Loss	F	-	7038
28	TextureNet [61]	ICML	2016	Github	Generator Network, Descriptor Network,	E	-	813
29	FPST [62]	NeurIPSW	2016	Github	CNN, Style Swap, Inverse Network	F, P	-	285
30	CIN [63]	ICLR	2017	Github	Conditional Instance Normalization	G, E	-	838
31	ITN [64]	CVPR	2017	Github	Instance Normalization, Julesz Generator Network	E	-	546
32	AdaIN [65]	ICCV	2017	Github	Adaptive Instance Normalization	F, P	-	2123
33	WCT [66]	NeurIPS	2017	Github	Multi-level Stylization, Whitening and Coloring Transforms	F, L	-	578
34	CartoonGAN [67]	CVPR	2018	Github	GAN, Semantic Content Loss, Edge-promoting Loss	E	-	227
35	I2SGAN [68]	CVPR	2019	Github	StyleGAN, Embedding	AC, BU	-	389
36	RST [69]	CVPR	2021	Github	Differentiable Renderer, Brushstrokes Parameterization	E	-	10

表3 热门相关工作总结。这些模型可以分为三个种类：传统人脸合成, 通用图像合成, 深度图像到素描合成。

#	模型	出版信息	年份	代码	核心组件	使用数据集	Assist.	引用
					通用图像合成方法			
37	pSp [70]	CVPR	2021	Github	StyleGAN, Disentangled Latent Feature, Map2Style	AC, BU	-	194
38	Pix2pix [32]	CVPR	2017	Github	Generator with Skip, PatchGAN	A, G, Q, R, S, U, BZ	-	13244
39	CycleGAN [14]	ICCV	2017	Github	Map Functions and Discriminators, Cycle Consistency Loss	A, G, Q, R, S, U, AV, AW	-	12734
40	DualGAN [71]	ICCV	2017	Github	Trained in Closed Loop, Reconstruction Loss	R, U, Y, CU, BZ, E	-	1554
41	DiscoGAN [72]	ICML	2017	Github	GAN with a Reconstruction Loss	CI, K, I, AH, S	-	1714
42	BicycleGAN [73]	NeurIPS	2017	Github	cVAE-GAN, cLR-GAN	R, S, U, BZ	-	1114
43	UNIT [15]	NeurIPS	2017	Github	Common Latent Space, VAEs, Cycle-consistency, GAN	G, I, Q, V, W, X, BI	-	2138
44	Pix2pixHD [16]	CVPR	2018	Github	Coarse-to-fine Generator, Multi-scale Discriminator	Q, AD, AE, AF	-	2527
45	MUNIT [74]	ECCV	2018	Github	Content/Style Encoder, AdaIN, Decoder	A, S, AP, BI, E	-	1615
46	SPADE [17]	CVPR	2019	Github	Spatially-Adaptive Normalization, Pix2pixHD	F, Q, AE, AR	Sm.	1362
47	U-GAT-IT [31]	ICLR	2020	Github	Attention map, Adaptive Layer-Instance Normalization	AU, AV, AW, AX	-	248
48	CoCosNet [75]	CVPR	2020	Github	Cross-domain Correspondence, Translation Network	AE, AC, BK	-	104
49	TSIT [76]	ECCV	2020	Github	Multi-scale Feature Normalization, Two-stream Network	Q,AE, AP, AW, BH	-	34
50	DSMAP [18]	ECCV	2020	Github	Domain-specific Content Mappings	AQ, AW, AX	-	13
51	ACL-GAN [77]	ECCV	2020	Github	Adversarial Consistency Loss, MUNIT	I, AU	-	29
52	DRIT++ [20]	IJCV	2020	Github	Disentangled Representation with Cross-cycle Consistency	AP, AQ, AW, AX, I	-	218
53	CoCosNetv2 [78]	CVPR	2021	Github	ConvGRU Module, Hierarchical Strategy, PatchMatch	AE	-	32
54	SofGAN [79]	TOG	2022	Project	SOF Net, StyleGAN, Style Mixing, SPADE	AC,BU,I	Bm.,Sm.,Attri.	11

Publ.: 出版信息。**Year**: 出版年份。**Code**: 开源代码链接。**Component**: 每个模型的关键组成。**Dataset**: A = TU-Berlin Sketch Dataset [80], B = Disney Portrait Dataset [27], C = FERET [36], D = AR [34], E = Self-Collected, F = MSCOCO [81], G = ImageNet [82], I = CelebA [21], L = DTD [83], P = Wikiart [84], Q = Cityspace [85], R = CMP Facades [86], S = Edge2photo [87, 88], U = Day2night [89], V = MNIST [90], Y = CUFS [2], Z = Caltech-200 Bird [91], AC = CelebAHQ [92], AD = NYU Indoor RGBD dataset [93], AE = ADE20K [94], AG = FERET [36], AK = QMUL-Shoe-Chair-V2 [95], AL = QuickDraw dataset [96], AP = Yosemite [14], AQ = cat2dog [20], AR = Flickr Landscapes [17], AS = APDrawing Dataset [3], AT = Anime Faces of Getchu [97], AU = Selfie2anime [31], AV = hourse2zebra [14], AW = photo2vangogh [14], AX = photo2portrait [20], BH = Berkeley Deep Drive [98], BI = SYNTHIA dataset [99], BJ = UPDG [28], BK = DeepFashion [100], BU = FFHQ [101], BV = DIV2K [102], BW = LHI [103],p BX = VIPSL [25], BY = IIIT-D [24], BZ = Map2Aerial [32], CB = StandfordCars [104], CH = LSUN [105], CU = CUFSF [22]。**Assist.**: 辅助信息，例如，Bm.= 背景图，Sm.= 分割图，Fl. = 人脸特征点，Sv. = 风格向量，Cm. = 色彩图，Attri. = 人脸属性，Km. = 关键点图，Tp. = 纹理快。**Cite.**: 来自2022-05-21的谷歌引用统计。

人脸素描图像对，其中所有的素描都是艺术家在观察了对应的照片后通过记忆绘制的。第三个法医素描数据库包含190张素描，这些素描是艺术家根据目击证人对犯罪现场回忆的描述进行绘制的。IIIT-D包含多种风格的素描画像，这也使它更具有挑战性。然而获取法医素描是很困难的，因为它们通常来自执法部门。

肖像素描数据集。 Yi 等人 [12, 28]提供了两个模拟艺术人像的数据集（APDrawing）。第一个数据集 [12]包含了140对脸部图像和对应由一个画像艺术家绘制的素描画像。随后被扩展为一个更大的数据集 [28]，其中有952张人脸照片和625张画像素描。在这些收集到的照片里，有220张来自三个著名的画家，剩下的212张照片来自于摄影网站[3]。值得注意的是，在这个数据集中，照片和画像不是成对的。迪士尼研究院发布了一个画像数据集 [27]，其中包括24张来自人脸数据库 [140]的人脸图像以及672张来自七位艺术家在四个抽象层次下绘制的素描图像。另外，他们还提供了一个位图，以方便创作新的素描图像。

不同于现有的数据集，本文提供了一个更有挑战性、高质量和具有属性标注的数据集，这也是迄今为止人脸素描合成领域最大的数据集。新的数据集包括2,104组照片素描对，1,058对图像与素描用于模型训练，剩下的则用来评价。本文提出的FS2K的优势包括多种绘画风格、照片和素描之间的高度对齐、多种属性信息和复杂的背景等等。数据及详细的对比如表 1。

[3] https://vectorportal.com/

表4 热门相关著作综述。有关更详细的说明，请参阅表 3。

#	模型	出版信息	年份	代码	核心组件	使用数据集	Assist.	引用
				深度图像到素描合成				
55	FCRL [106]	ICMR	2015	-	Fully Convolutional Network	Y	-	127
56	DGFL [107]	IJCAI	2017	-	Deep CNNs, Graphic model	Y	-	34
57	Scribbler [108]	CVPR	2017	Project	Encoder-decoder with residual connections, GAN	Y, E	-	427
58	FSSC2F [109]	AAAI	2018	-	U-Net, Probabilistic Graphic Model	Y	-	11
59	TextureGAN [110]	CVPR	2018	Github	Local Texture Loss, VGG Loss, Scribbler	E, S	Bm. Tp.	221
60	SCC-GAN [111]	CVPR	2018	Code	Hybrid model, Shortcut Cycle Consistency	AK, AL	-	76
61	ContextualGAN [112]	ECCV	2018	Github	Contextual Loss, Joint Representation, GAN	I, Z, CB	-	74
62	pGAN [113]	IJCAI	2018	Github	UNet, Parametric Sigmoid, CycleGAN	Y, CU	Bm.	24
63	MRNF [114]	IJCAI	2018	-	Markov Random Neural Fields	Y	-	16
64	PS^2-MAN [115]	FG	2018	Github	Multi-Adversarial Networks, CycleGAN	Y, CU	-	98
65	DualT [116]	TIP	2018	-	Deep Features, Intra- and Inter-Domain Transfer	Y	-	51
66	MDAL [29]	TNNLS	2018	Github	Domain alignment, Interpreting by Reconstruction	Y, CU	-	45
67	FAG-GAN [117]	WACVW	2018	-	Attribute Classification, Conditional CycleGAN	I, AG	-	30
68	Geo-GAN [118]	BIOSIG	2018	Github	Geometry Discriminator, CycleGAN	CU, AG	-	17
69	PI-REC [119]	arXiv	2019	Github	Corse-to-Fine, LSGAN, VGG Loss	A, I, S, AT	Cm.	18
70	DLLRR [120]	TNNLS	2019	-	Coupled Autoencoder, Low-rank Representation	Y	-	27
71	Col-cGAN [121]	TNNLS	2019	-	Collaborative Loss, cGAN, Deep Collaborative Nets	Y, CU	-	43
72	CFSS [122]	TIP	2019	-	cGAN, VGG, Feature Selection	Y	-	14
73	KT [123]	IJCAI	2019	-	Knowledge Transfer, Teacher-Student Net	Y, CU	-	16
74	im2pencil [124]	CVPR	2019	Github	Outline and Shading Branch Networks, Pix2pix	E	Sv.	28
75	ISF [125]	ICCV	2019	Project	Shape and Appearance Generators, Two-stage	S, AC, E	-	62
76	APDrawing [3]	CVPR	2019	Github	Hierarchical GAN, DT Loss, Local Transfer Loss	AS	Fl., Bm., Sv.	82
77	APDrawing++ [12]	TPAMI	2020	Github	APDrawing, Line Continuity Loss	AS	Fl., Bm., Sv.	12
78	UPDG [28]	CVPR	2020	Github	Asymmetric CycleGAN, Cycle-consistency Loss	BJ	Fl., Bm., Sv.	22
79	WCR-GAN [126]	CVPR	2020	Github	Cartoon Representation Learning, GAN	F, BU, BV, E	-	29
80	EdgeGAN [127]	CVPR	2020	Project	SketchyCOCO, Divide-and-Conquer strategy	F	Attri.	34
81	DeepPS [128]	ECCV	2020	Github	Sketch Refinement with Dilations, Pix2pixHD	AC, I	-	25
82	DeepFaceDrawing [129]	TOG	2020	Github	Component Embedding, Feature Mapping, Image Synthesis	AC, E	Km.	41
83	CA-GAN [130]	TC	2020	Github	Composition/Appearance Encoder, P-Net, Stacked GAN	Y, CU	Fl.	44
84	IDA-CycleGAN [131]	PR	2020	-	CycleGAN, Identity Loss, Recognition Model	Y, CU	-	41
85	IPAM-GAN [132]	SPL	2020	-	Identity-preserved Adversarial Model, U-Net	Y, CU	-	12
86	MvDT [133]	TIP	2020	Github	CNN [134] Features, Hand-crafted Features	Y, E	-	10
87	MSG-SARL [135]	TIFS	2021	-	Self-attention Residual Learning, Multi-scale Gradients	Y, CU	-	6
88	GANSketching [136]	ICCV	2021	Project	Weight Adjusting, Cross-domain Fine-tuning	CH, AL	-	8
89	DoodleFormer [137]	Arxiv	2021	-	Transformer, Part Locator and Part Sketcher Networks	CK	-	1

2.2 传统人脸合成

在早期的工作中，研究人员采用启发式图像转换去交互式地或者自动地合成人脸素描 [4, 141–145]。然而这些方法往往产生不真实的、缺乏表现力的素描，导致其缺乏艺术风格。因此，最近几年，更多的注意力被放在基于学习的人脸素描合成方案，它们的分类详情见图 2。这些模型分为贝叶斯推理模型、表示学习模型和子空间学习模型。

2.2.1 贝叶斯加强模型

贝叶斯推理模型利用线索在概率模型上更新素描组件的状态，这类方法在FSS [146]中被广泛使用。在 [37]中，chen 等人首次提出一个基于样本的人脸素描合成系统，该模型通过无参抽样算法学习细微变化的素描风格。后来，嵌入式隐马尔科夫模型 [147]被用来建模照片素描对的非线性关系，然后选择集合策略产生人脸素描图 [39]。Wang和Tang [2]也遵循相似的想法，但他们考虑了不同尺度下的人脸结构，采用一个多尺度马尔可夫随机场（MRF）去构建照片素描对之间的关系。Xu 等人 [40]提出一个分层组合模型，模型考虑了人脸的规律性和结构变化。这些方法在产生素描图像方面已经取得了重要成果，但他们只考虑简单的控制条件，忽略了光线和姿势的变化。Zhang 等人 [41]通过同时考虑块匹配、强度兼容、梯度兼容和形状先验来解决这一问题，从而获得更好的视觉效果。然而，基于MRF的模型也存在两个不足：（1）在未见过的人脸信息合成方面表现不好；（2）优化问题是NP-hard。Zhou 等人 [46] 使用马尔可夫权重场和级联分解构建了一个稳健的人脸合成系统，使用候选块的线性组合来逼近新的素描块。Wang 等人 [47]建立了将照片转化为人像画的非参数模型，其中马尔可夫随机场用于增强风格参数的空间一致性，主动形状模型和图割模型用于学习人脸特征的局部信息。Wang 等人 [48]提出了一种转导式学习方法来合成人脸素描，该方法采用动态优化过程来最小化给定测试样本的损失。Peng 等人 [52]设计了一种基于马尔可夫模型的超像素方法，在不将照片分成规则的矩形块的情况下提高了灵活性。然后，他们不仅使用马尔可夫网络来建模图像块之间的关系，而且还通过多个视觉特征 [53]保留了许多视觉方面的线索（如边缘）。

2.2.2 子空间学习模型

子空间学习在FSS任务 [146]中也得到了广泛的研究，核心思想是学习嵌入在高维空间中的低维流形表征 [148]。Tang和Wang [149–151]提出一系列基于样本的线性特征变换方法。这些方法属于全局线性系统，它们不能完全解释照片-素描对之间的关系，因为这样的变换不是简单的线性关系。为此，Liu 等人 [38]用局部线性嵌入（LLE） [152]使得照片和素描在两个不同的图像空间中具有相似的局部几何形状的流型。然而，伪图像的生成和表征学习被分成两个独立的过程，导致结果次优。Huang和Wang [49]提出了一个联合学习框架，包括特定领域的词典学习和子空间学习。

2.2.3 表示学习模型

稀疏编码和字典学习，又名表示学习，被用于FSS任务 [146]。Ji 等人 [42]证明通过大部分的合成方法忽略了个性化特征。因此，一些工作 [42–44]使用不同的回归模型，例如k-NN [42]、Lasso [42]、多元输出回归[43]和支持向量回归[44]，来建立照片和素描之间的转换，目的是为了提高生成的人脸素描的质量，Wang 等人 [25, 26]使用LLE来估计初始素描或照片，然后引入了一种能够关注高频和细节信息的稀疏多字典表示模型。然而，大多数基于表示的模型假设源输入和目标输出共享相同的表示，从而限制了合成过程中特定样式的局部结构。为了放宽这一限制，Wang 等人 [45]引入了一种半耦合词典学习方法，该方法使用线性变换来弥合两个不同领域特定表示之间的差距。Gao 等人 [26]还考虑了两步算法 [44]，提出了一种选择方案来生成初始伪图像，并引入了基于稀疏表示的增强（SRE）来合成素描。

2.2.4 联合模型

最近，一些工作致力于探索不同机器学习模型组合的联合模型。如结合贝叶斯推理和子空间学习方法。Berger 等人 [27] 提出了一个模型来模拟不同艺术家的风格和抽象过程，该模型可用于人脸素描合成。Song 等人 [50] 引入了一种实时FSS方法，该方法首先使用k-NN算法来寻找top-k相似的局部块。然后使用线性组合计算相应的素描图像，最后使用图像去噪技术来提高视觉质量。然而，由于k-NN计算量大，该模型 [50]仍然很耗时，因此Wang 等人 [56]进一步利用与识别权重表示相结合的在线方案进行离线随机采样解决了这个问题。现有的大多数传统方法都完全依赖于训练数据的规模，因此Zhang 等人 [51]提出了一个在模板风格素描上训练的健壮模型。该模型包括表征学习、马尔可夫随机场和级联模型。Li 等人 [54] 提出了一种将感知分组模型和可变形笔画模型相结合的手绘素描合成方法。在工作[55]中，作者引入了一种结合表示学习和马尔

可夫网络的自适应学习方法。Men 等人 [57]提出了一种基于结构引导的交互式纹理传递的通用框架。他们的模型使用多个渠道动态地实现合成过程，包括结构提取、结构传播和引导纹理转移。

2.3 通用图像合成

深度人脸素描合成属于图像泛化的范畴。因此，一般的图像合成方法，如图像到图像的转换和神经网络风格转换，也可以用于生成人脸素描。本文将概述各种前沿转型模式。

2.3.1 图像到图像转换

图像到图像转换（I2I） [153]在机器视觉和机器学习领域属于研究热点。目标是输入图像从源域转换到目标域且保留内在的源内容并转换为外在的目标风格。当前的I2I模型通常建立在对抗生成网络（GAN） [154]。这些模型一般分为有监督模型和无监督的I2I模型。

监督I2I。 监督的I2I模型使用对齐的图像对作为源域和目标域，以学习可以将源图像转换为目标图像的变换模型。一个典型的I2I方法是Pix2Pix [32]，它将条件GAN（cGAN） [155]应用于此任务。与原始的cGAN的主要区别在于，Pix2pix中的生成器是一个U-net [156]。然而，Wang 等人 [16]观察到Pix2pix中的对抗性训练不稳定，阻碍了模型生成高分辨率图像。因此，他们对原始的Pix2pix进行了扩展，引入了新的特征匹配损失，可以生成尺寸为$2,048 \times 1,024$的高分辨率图像。Zhu 等人 [73]提出了BicycleGAN，一种包含条件VAE和条件隐含回归GAN的双环遗传算法，解决了算法的崩溃问题，提高了算法的性能。此外，为了减少Pix2pixHD模型 [16]中的语义信息损失，Park 等人 [17]引入了一种基于SPADE的生成器，该生成器将空间自适应归一化加入到Pix2pixHD [16]生成器中，以增强整个网络中的语义信息。

无监督I2I。 由于监督I2I需要大量配对数据作为支撑，而收集这些配对数据工作量沉重，且不切实际的。为了解决这个问题，研究人员开始关注并设计无监督的I2I模型。该核心思想是训练两个不同的生成网络，保证两个网络生成的样本具有循环一致性。具体而言，如果我们用第一个生成器将斑马图像转换为马图像，然后再用第二个生成器输出一张与斑马图像相近的斑马图像。反之亦然。具有代表性的工作有CycleGAN [14]、DiscoGAN [72]和DualGAN [71]。后来，Liu 等人 [15]提出了一种无监督的I2I模型（UNIT），该模型认为不同域中的图像对可以映射到同一潜在特征空间中的特征编码，进而可以使用GAN模型实现I2I任务。Kim 等人 [31]后来提出了一种新的具有归一化功能的注意力模块，将其集成到GAN模型中，以灵活地监控纹理和形状的变化。通过对标准GAN模型的重新思考，Chen 等人 [19]提出了一种NICE-GAN，其核心思想是将判别器和编码器耦合，即重用判别器参数对输入进行编码。Zhao 等人 [77]利用一种新的对抗性一致性损失而不是循环损失来强调源域和目标域之间的共性。为了提高内容表征能力，Chang 等人 [18]提出了DSMAP模型，以更好地利用内容和风格之间的关系。具体地说，该模型将内容特征从共享的领域不变特征空间映射到两个单独的领域特定特征空间。此外，DRIT++ [20]使用两个图像生成器、两个内容编码器、一个内容判别器、两个属性编码器和两个域判别器来将图像嵌入到域不变内容空间和域特定属性空间中。另外，Jiang 等人 [76]提出了一种双流I2I转换算法（TSIT），通过学习语义结构特征和文体特征，从粗到精的方式合成内容和文体的特征图。最近，Zhang 等人 [75]提出了一种基于样本的图像转换CoCosNet，它包含两个子网络。第一种是将来自不同领域的输入嵌入到依赖于语义对应的特征领域中。同时，第二种方法利用一系列反归一化块来逐步合成目标图像。Zhou 等人用全分辨率语义对应学习 [78]进一步扩展了CoCosNet，主要区别是使用了在每个语义级别迭代应用的规则和基于GRU的传播模型。最近，Chen 等人 [79] 提出了一种将人像特征分解为几何特征和纹理特征的SofGAN算法。然后，这两个功能被馈送到两个网络分支。第一个分支是用超网络表示三维空间中的语义占用场（SOF），用于将几何特征解码为SOF网络的权值。然后，使用SOF网络的输出特征通过光流投影方案来渲染分割图。第二个分支是使用GAN生成器对每个语义区域进行纹理转换，该生成器可以从纹理空间风格编码采样。最后，使用一种基于语义实例的StyleGAN模块来样式化生成的线段图并按区域输出照片级真实感肖像。

2.3.2 神经网络风格转换

神经网络风格转换（NST），旨在通过神经网络产生视觉吸引力的图像，这种技术已被引入FSS任务 [157]。具体地说，NST用于呈现不同样式的内容图像。NST方法可分为基于优化的方法和基于模型的方法。[4]

[4]请注意，一些通用的基于GAN模型的相关工作，如CartoonGAN [67]和pSp [70]，这些GAN模型可用于神经网络风格转换或图像到图像的转换。由于本文没有对广义GAN模型进行具体的回顾，所以本文将几个GAN 模型归类为神经网络风格转移任务，作为对这些方法的快速概述。

基于优化的方法。 在线NST算法迭代地更新给定的输入图像，以匹配所需的CNN特征，包括照片的内容和艺术风格信息。Gatys等人 [58, 59]对这一领域做出了第一个贡献，使用经典的CNN（如VGG [134]）来渲染具有著名画风的图像。此外，StyleGAN [101]使用潜在空间来保持图像合成结果的一致性。然而，在给定的条件下，要取得有希望的结果是具有挑战性的。最近，Abdal 等人 [68]将经典的NST [58, 59]集成到StyleGAN模型中，使用NST将输入图像投影到StyleGAN定义的潜在空间中。随后，Kotovenko 等人 [69]通过优化基于简单可微渲染机制的参数化笔划，进一步增强了经典的NST[58, 59]。

基于模型的方法。 基于优化的在线方法取得了令人满意的结果，但也存在一定的局限性。主要的缺点是计算速度慢和在线优化成本高。为了解决这些问题，一些工作引入了前馈网络来模拟风格转移的优化目标 [157]。

端到端模型可分为设计一个基本深度神经网络结构和引入新损失函数。对于基本体系结构，Johnson 等人 [60]利用神经网络和基于优化的NST模型的优点，提出了一种使用新的感知损失来训练前馈网络的方法。TextureNet [61]遵循类似的想法，但具有不同的神经网络体系结构。[60]和[61]都是实时风格转换方法。Chen和Schmidt [62]引入了风格交换操作来交换具有视觉背景和风格的块，进一步设计了一个新的优化目标，旨在学习用于任意风格转换的逆神经网络。在基于损失函数的方法方面，提出了CartoonGAN [67]将真实世界的照片转换为卡通风格的图像的方法。它由两个新的损失函数组成，旨在保留清晰的边缘信息，并处理照片和卡通之间的风格差异。

最近，一些研究人员开始使用少量的参数来表征每种风格，即改变风格转换的归一化层中的参数。Dumoulin 等人 [63]发现了归一化层可以反映不同风格的统计特性。因此，他们在保持卷积参数不变的情况下，对这些层中的参数进行了缩放和移位，以获得更好的NST。此外，他们引入了灵活的条件实例标准化，只需在线更改标准化参数即可实现样式转换。Ulyanov 等人 [64]通过简单地将归一化应用于每个图像而不是一批图像（他们称之为实例归一化），改进了他们以前的TextureNet [61]。此外，他们还证明了使用实例归一化的样式传递网络比使用批量归一化的样式传递网络收敛更快，同时获得了更好的视觉效果。后来，Huang和Belongie [65]遵循了类似的想法，在GAN模型中引入了自适应实例规范化，使内容和风格特征保持一致。Li 等人 [66] 进一步使用预先训练的VGGNet [134]的前几层来提取特征表示。然而，他们用白化和着色变换取代了AdaIN层，实现了通用的样式转换。与I2SGAN [68]类似，Richardson 等人 [70]改进了经典的StyleGAN，提出了一种新颖的编码网络，该网络学习许多输入到预先训练的生成器中的样式向量，形成扩展的$\mathcal{W}+$ 隐空间。

2.4 深度图像-素描合成

深度照片-素描合成是FSS任务最新的一个分支，其运用深度学习提升性能和质量。相关工作可以分为三个类别。第一种旨在将任意素描图像转换他们对应的RGB图像。第二种设法将任意RGB图像转换为素描图像。最后一种则主要关注人脸素描合成。

通用素描到图像。 Xian 等人 [110]提出了TextureGAN模型，在素描、颜色和纹理的监督下合成图像。TextureGAN由基本面预训练模块和外部纹理微调部分组成。然后，Lu [112] 等人引入了两阶段上下文GAN来实现素描到图像的生成。该框架在经典GAN模型的基础上，添加了一个新定义的损失函数、联合分布表示方法、以及素描与其对应图像之间的内在关系约束。受到图像绘画 [158]的启发，You 等人 [119]提出了PI-REC模型，该模型包括三个阶段：模仿阶段、生成阶段和细化阶段。PI-REC仅使用一个生成器和一个判别器进行渐进式训练。在 [125]中引入的ISF是一种基于门控的方法，它允许使用单个生成器来生成不同的类，而不需要混合特性。近期，Gao 等人 [127]提出了EdgeGAN模型，以在给定手工绘制的场景素描图像的情况下进行面相目标的图像合成。该框架包含两个顺序的模块：前景生成和背景生成。Yang 等人 [128]提出了一个深度整形修复模型来模拟人类艺术家从粗到精的绘画过程。Chen 等人 [129]提出了一个局部到全局的框架，以允许任何用户都可以生成高质量的人脸图像。他们的模型由三个模块组成：组件嵌入、特征映射和图像合成。

通用图像到素描。 Song 等人 [111]提出了第一种基于笔画特征的深度照片素描合成方法，该方法是一种混合模型，保证基于VAE的重建损失约束下的最短周期一致性。作为I2I和NST的默认设置，两者都可以合成艺术肖像绘画绘制（APD）图像。然而，由于APD图像通常具有高度抽象的风格和图形元素，因此不能满足实际需求。为此，Yi 等人 [3]提出了一种APDrawing算法 [12]，将输入的人脸图像转换为相应的图像库，并将全局网络和局部网络相结合，建立了层次化的GAN模型。接着，他们进一步提出了一个APDdrawing++，该模型用自编码强化不易察

觉的特征，同时提出一种新颖的线连续性损失从而强化了APDdrawing线的连续性。然而，所有的APDrawing模型都需要成对的数据进行训练。为了解决这个问题，Yi 等人提出一种非对称循环结构GAN 模型 [28]，模型包括了一个有松弛的前向环形一致性损失（亦称截断损失）以阻止模型从噪声中重构图片，以及一个严格的环形一致性损失以加强模型的性能。这种方法同时也采用多种局部判别器去确认人脸肖像绘制的质量。不同于肖像绘制，Wang 等人 [126]通过观察卡通画绘制的动作，合成模型需要分别考虑表面、纹理和形状等三种不同的表示。同时，他们为了更好的训练和评价模型，也开源了新的SketchyCOCO数据集。在Pix2pix的基础上，Li 等人 [124]设计了一个双分支网络结构（称为Im2Pencil）执行照片到笔画转换，模型可以模拟素描的轮廓和阴影。Wang 等人 [136]提出了一种用一个或多个素描重写GAN的方法。该方法使用正则化方法来保持原始GAN的多样性和图像质量，同时通过跨域对抗性损失将生成的素描图像与用户的需求进行匹配。Bhunia 等人 [137]引入了一种包含两个网络的新的转换器结构，以生成各种逼真的创意素描。其中，第一个是定位器，旨在通过观察局部模式之间的关系来捕捉粗略结构；第二个是基于标准GAN的素描网络，旨在合成高质量的素描图像。

图像-素描合成。 Zhang 等人 [106]是第一个使用全卷积神经网络（FCNN）来建立深度照片到素描合成模型的人。然后，文献 [87, 109, 114]将深度特征集成到概率图模型学习中，取得了比传统模型更好的性能 [2, 46]。为了使网络更加灵活，Zhang 等人 [113]借鉴CycleGAN的核心思想，提出了一种新的PGAN模型，它使用了一种特殊设计的参数化Sigmoid激活函数来减小照片先验和光照变化的影响。为了提高生成的照片/素描的质量，Wang 等人 [115]引入了一种多对抗网络综合方法（PS2MAN）。他们的模型使用两个UNET来生成从低分辨率到高分辨率的高质量图像。为了实现同样的目标，Zhang 等人 [29]在此基础上，提出了一种基于多领域对抗学习的人脸素描合成方法，克服了人脸轮廓模糊和变形的缺陷。MDAL背后的基本思想是“通过合成进行解释”的概念，它建立在两个不同的生成器之上。Kazemi 等人 [117, 118]提出了一种改进的CycleGAN算法，该算法重点考虑了人像合成过程中的人脸属性。Zhang 等人 [120, 122] 引入了自动编码器与传统的子空间学习相结合的两种方法，该方法比传统的FSS方法更有效。此外，Zhu 等人 [121] 提出了一种协同框架，该框架通过引入协同损失来利用两个相对生成器之间的交互信息。然而，由于缺乏大规模的训练数据，很难训练出一个好的模型。因此，Zhu 等人 [123]提出通过两个强教师网络，利用经典知识蒸馏学习两个定义良好的学生映射网络。最近，文献 [131, 132]中的工作引入了身份感知模型，该模型使用新的感知损失来训练更好的图像生成模型，并因此考虑下游任务，例如人脸识别，作为最终目标。Yu 等人 [130]提出了一种新的构图辅助生成性对抗网络，该网络利用人脸构图信息合成逼真的人脸素描/照片。通过利用特征之间的关系，[135] 实现了一个用于人脸照片-素描转换的多尺度自注意残差学习框架。最后，[133]提出的方法不需要来自源域的任何图像进行训练，使其能够灵活地利用深层特征（从CNN提取）和手工制作的特征。

3 提出的FS2K数据集

在本章中，我们介本文所提的FS2K数据集，图 3展示了一些示例图像。本文从数据集收集和数据注释两个关键方面来描述FS2K。总的来说，FS2K一共包含了2,104组照片-素描对，并被分为两个部分，1,058对用于模型训练，1,046对用于模型测试。完整的数据集可以在此获取：https://github.com/DengPingFan/FS2K。

3.1 数据收集

为了建立一个较好的基准，数据集构建过程中应该仔细选择数据，以涵盖来自不同视角的不同场景，如照明条件、皮肤颜色、素描样式和图像背景。为此，本文引入了FS2K，这是一个新的用于FSS任务的高质量数据集[5]。

本文的FS2K包含了来自真实场景、互联网和其他数据集的2,104张照片。FS2K多数图像来自CASIAWebFace [159]这一大规模（即500K图像）的自然环境下采集的人脸数据集。CASIA-WebFace是从IMDb6[6]网站收集的，包含了组织良好的信息，如姓名、性别和生日。得益于CASIAWebFace的丰富性和开源性，它可以用来构建本文高质量和有代表性的基准。本文从该数据集中人工选择了1,529张图片，以涵盖现实场景中面临的大范围主要挑战，例如不同的背景、发型（例如，长、短）、配饰（例如，眼镜、耳环）和皮肤信息（例如，给定人脸图像的补丁图像）。因为在CASIA-WebFace中选择的照片是同一个人的单角度拍摄，缺失了多角度人脸的拍摄。为此，我们邀请了8位演员在不同的设置下（例如，照明条件、人脸角度）拍摄了98 张

[5]此数据仅供学术交流。
[6]http://www.imdb.com

照片。此外，为了进一步增加多样性，本文还收集了一些儿童照片和一些人脸图像比例较小的照片。剩下的477张脸部照片来自其他免费的照片网站，包括Unsplash[7]、Pexels[8]、Pngimg[9]、和Google。

3.2 数据标注

在FS2K数据集中一共有四种风格的标注，包括，素描绘制、素描风格、颜色和轮廓特征标注。

3.2.1 素描绘制

参与者。 三名（两名男性和一名女性）来自四川美术学院[10]的资深艺术家受聘参与该研究。三名参与者的视力都正常或者经过矫正后正常。没有人是色盲或者色弱。参与者的年龄在20到23岁之间，有平均五年的素描绘制经验。

设备。 这三名艺术家在拷贝灯[11]的辅助下绘制所有的素描图像。图 4展示了本文使用的拷贝灯以及艺术家绘制素描的示例（图 4-d）。本文设备中的触摸开关区域支持三种级别的可调亮度，因此艺术家可以使用按钮来更改他们想要的亮度。这有助于他们根据LED板底部的照片信息定位人脸特征的轮廓。此外，该设备还有助于确保素描和对应照片之间的内容相似性和人脸对齐。同时，这些素描保留了艺术家的素描风格。

3.2.2 素描风格标注

本文提出的FS2K包括三种不同的画像风格，以保证素描风格的多样性，正如图 5所示：这可以使不同的艺术家的技能被捕获并且使得FS2K数据集比以往的数据更具有挑战性。

本文创建了一个平衡的数据集以便不同模型之间的对比。三种不同风格的图像的分布是均匀分布。具体而言，在训练数据中，风格1、风格2和风格3的图像数量分别为357、351和350张。在测试阶段，则分别为619、381和46张。

3.2.3 素描特征标注

素描是快速完成的手绘画，与原始图像相比具有更少的属性信息，例如人脸纹理、人脸表情 [160]、人脸姿势等。因此，基于单个素描图像恢复真实图像（即S2I任务）是具有挑战性的。同时，在现实世界的应用程序中，我们可以利用辅助人脸信息（如性别、饰品和发型）来缩小数据库中嫌疑人的范围。同[161]一样，我们添加了一些额外的人脸特征标签，如性别、微笑、脸部姿势、头发状况、头发颜色、耳环和皮肤纹理。我们聘请了两名数据标注员来标记所有的照片，并进行了交叉检查，以确保最终注解的准确性。总体标签可在表 5中找到，而每个标签的详细信息如下所述。

性别。 性别是传统人脸数据库中，如CelebA [21]和LFW [162]常用的高级人类属性。它在人脸检测和识别中得到了广泛的研究 [163–165]。因此，本文仔细地将FS2K中的所有照片都贴上了性别属性的标签。具体来说，训练集中有574张男性照片和484张女性照片，测试集中有632张男性照片和414张女性照片。

笑容。 微笑是一种主要的人类活动，代表着积极的情绪状态。因此，许多研究都集中在笑容检测 [166, 167]或使用笑容作为识别的属性 [168]。因此，我们也认为笑容是本文数据集中的一个关键属性。具体来说，训练集包含645名面带笑容的人和413名没有明显表情的人，而测试集包含670名面带笑容的人和376名没有表情的人。我们确保面带笑容的人在训练和测试集中的比例尽可能接近。

人脸姿态。 人脸属性可能只覆盖图像的一小部分，但照片通常受姿势 [169]的影响。此外，姿势会影响人脸识别 [170]、跟踪 [171]和合成 [172]的性能。因此，人脸姿势是有用的辅助信息。我们将头部旋转30度以内的肖像定义为正面姿势。根据这个定义，训练集有917张正面照片，而测试集有872张。其余的都是侧脸姿势。

头发的状态和颜色。 头发是头部的一个显著特征，可能会在不同的情况下发生变化。即使人脸内部特征中有足够的信息用于识别，改变头发也会损害性能 [173, 174]。此外，人脸合成和检索系统经常使用毛发作为改善生成图像质量的重要线索 [175, 176]。对于FSS，尽管素描包含头发轮廓，但缺少相应的颜色信息和头发状态（有或没有头发）。因此，在FS2K中，本文提供了头发状态的注释，包括四种可用颜色（即黑色、棕色、红色和金色）和另一种状态（即秃顶或戴帽子），如图 6所示。换句话说，对于有头发的人脸，本文直接标记颜色信息，而稀疏头发或戴帽子的情况被标记为单独的属性。该注释的统计结果可以在表 5中找到。

耳饰。 素描的简化特征导致耳环轮廓不清晰。同时，真实照片中的耳环是可见的，如图 6。因此，在FS2K中，本文提供了关于是否存

[7] http://www.unsplash.com
[8] http://www.pexels.com/
[9] http://pngimg.com/
[10] 四川美术学院是中国四大美院之一。三位资深艺术家都来自设计学院。
[11] 图 4-a 展示了一个拷贝板，它有一个LCD背光灯。需要100 ∼ 240V的高压输入和0.6A的工作电流。如图 4-b所示，它的尺寸是A4（即，300 × 200 × 3.5mm），且亮度为300 ∼ 350LM。因此，对于动画师（如图图 4-c）来讲它已经成为继铝合金拷贝板之后最受欢迎的拷贝板产品。

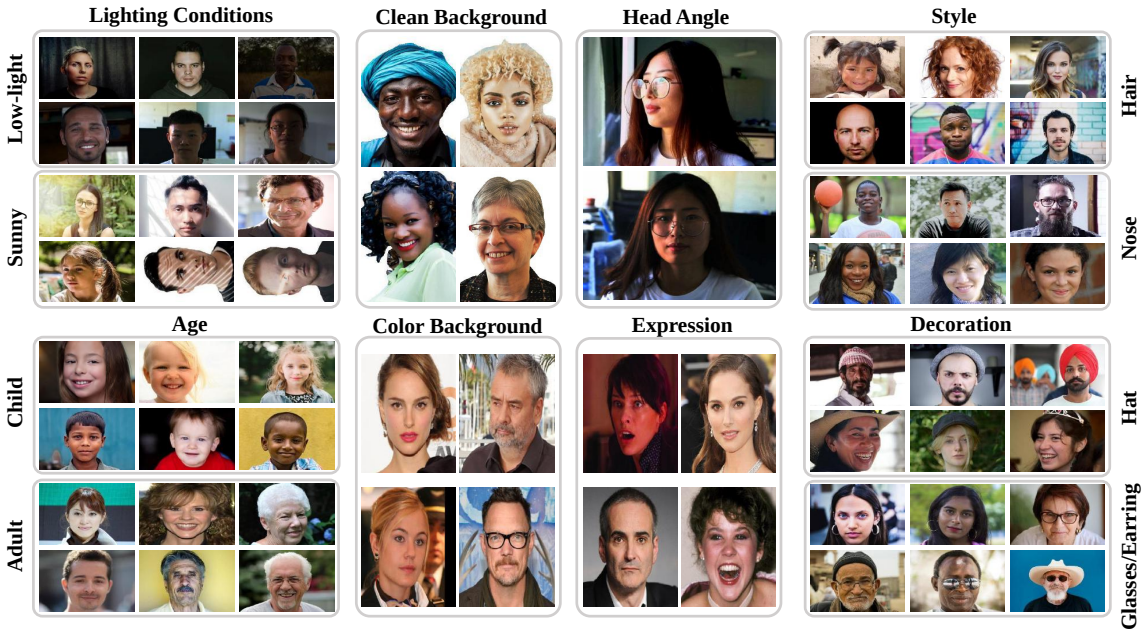


图3 有代表性的FS2K数据集图像样本。收集的图像根据不同的选择标准来描绘不同的场景，例如不同的照明条件（例如，弱光、阳光）、年龄（例如，儿童或成人）、背景（例如，干净或有色）、头部角度、人脸表情（例如，严肃、微笑和大笑）、发型（例如，黑色、金色、长发和短发）以及配饰（例如，帽子或耳环）。

（a）示意图　（b）规格　（c）虚拟示例　（d）我们的例子

图4 拷贝板的使用以及示意图。放大以获取最佳视图。有关更多详细信息，请参阅章节 3.2。

在耳环的注释，这有助于模型训练。具体来说，训练组有209人戴耳环，测试组有187人。

皮肤纹理。 皮肤纹理提供了大量详细的局部信息，是人脸识别 [177, 178]的重要特征。然而，这些关键信息在素描图像中完全消失了。因此，本文从真实的照片上剪下一小块作为皮肤纹理，如图 5所示。本文还包括了相应的嘴唇和眼球区域的平均RGB值，为进一步的研究提供了更多的信息。

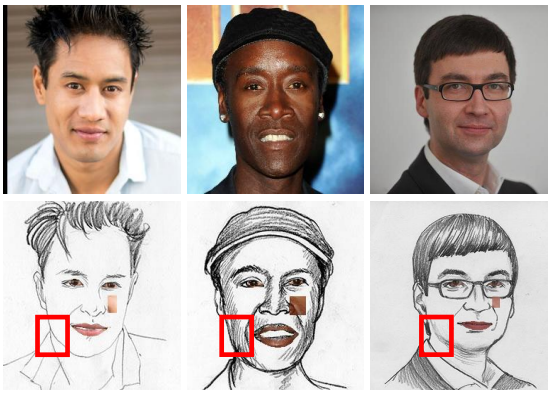
4 本文的FSGAN基准

4.1 问题定义

人脸合成（FS）模型目的是在给定的输入图像的基础上生成人脸的目标表示。这个过程可以表示为$X_o = F(X_i)$，其中X_i和X_o表示输入和输出（即RGB图像和素描），F表示合成公式。在本文中，受Pix2pixHD [16]启发，在 [3, 12]整体结构的基础上，本文设计了一个基准模型，FSGAN，可以适用于I2S任务[12]和S2I任务[13]。本文提出一种两级自底向上人脸合成结构如图 7所示，而不是采用直接的图像级的人脸合成。因此，本文提出的FSGAN由两个建立在多个生成器基础上的级联阶段模型组成（即GANs）。

第一阶段由五个并行GAN组成，它们被设计成分别合成局部人脸成分。给定输入，在第一阶段中，四个人脸区域（例如，左眼、右眼、鼻子和嘴）和其余输入被裁剪并馈送到它们对应的GAN中，用于合成关键人脸特征。随后，这些合成的人脸组件块被粘贴在一起从而获得完整的人脸表示。因为局部人脸块是单独合成，缝合的连接区域，以及它们的外观是不一致的。因此，

[12] $X_{ske} = F(X_{img}, X_{style})$，其中$X_{style}$表示输入的素描风格。
[13] $X_{img} = F(X_{ske})$。



（a）风格1　（b）风格2　（c）风格3

图5 本文提出的FS2K数据集中的三种素描风格。如脸颊区域所示，这些风格包含简单的线条（风格1）、长笔画（风格2）和重复的细小细节（风格3）。

第二个阶段通过考虑全局结构和纹理进一步优化合成结果。在这个阶段，人脸素描的风格向量用于协助素描合成。

4.2 人脸组件合成

几乎所有的人脸具有相似的全局结构。不同的只是局部人脸组件的细节，如眼睛、眉毛、鼻子和嘴。为了获取不同人脸组件更多的细节，模型的第一阶段是单独将其合成。具体来说，给定人脸输入，MTCNN [179]首先检测四个关键模式，包括左眼、右眼、鼻子和嘴巴。然后，输入X_i在检测结果的基础上被分为五个部分，$X_{parts} = \{X_{leye}, X_{reye}, X_{nose}, X_{mouth}, X_{rest}\}$。这些部分包括左眼、右眼、鼻子、嘴和剩下的部分。五个并行的GANs被用来合成这五部分对应的块。因此，问题可以表述为：$G_{parts} = \{G_{leye}, G_{reye}, G_{nose}, G_{mouth}, G_{rest}\}$和$D_{parts} = \{D_{leye}, D_{reye}, D_{nose}, D_{mouth}, D_{rest}\}$，其中$G$和$D$分别表示为生成器和判别器。

首先，合成左眼、右眼、鼻子和嘴巴的四个GAN具有相同的架构。每个GAN由一个生成器和一个判别器组成。该生成器被设计成一个编解码器，由一个编码器、一个底层连接和一个解码器组成。该编码器由三个卷积块组成，每个卷积块由卷积层（核大小为3，步长为2）、批归一化层和RELU激活层组成。同时，第二个底部连接由9个类似于 [180]的颈部残差块组成。最后，解码器建立在三个去卷积块上：去卷积层、批归一化层和ReLU激活层。注意，用于合成X_{rest}的GAN类似于前面描述的GAN。然而，编码器包含四个卷积块，而解码器包括四个去卷积块以获得更大的感受野。

上述5个GAN的判别器的结构是一致的。每一个判别器经过都包括三个级联卷积层（卷积核尺寸为3，步长为2）然后是全局平均池化。接着，一个1×1的卷积层和一个sigmoid函数则被用来预测生成的结果为真或假的概率。

基于以上的设计，在第一个阶段，FSGAN无论在I2S还是S2I任务中都可以保留人脸组件部分的细节。在这一阶段的最后，合成的图像块被缝合在一起以恢复完整的人脸素描结果X_{intact}。因为合成的素描块是由不同的生成器生成，整体是不连续的，在缝合的结果中尤其明显。为此，缝合的结果被送入下一个阶段，以调整和完善整体结构和外观。

4.3 人脸-素描合成

为了解决第一阶段输出不连续的问题，本文引入了第二阶段，受Pix2pixHD [16]的启发，第二阶段被设计为另一个GAN模型，目的是为了进行局部细节加强和全局结构调整。

在这阶段，本文使用和Pix2pixHD [16]一样的多尺度判别器D_{fs}和粗糙到精细的生成器G_{fs}。具体来讲，生成器G_{fs}包括两个子网络$G1$和$G2$，两个子网络都保持编码-解码结构，如图 7中的右图所示。本文以50%的采样率对第一阶段的输出进行下采样。最新的采样图像为$X_{intact}^{1/2}$ $(height/2, width/2)$，随后，采样图像被传入第一个用来获取全局特征的子网络$G1$。$G2$则是用来获取局部细节，并以第一阶段的输出为输入。本文同时使用串联和逐个元素的加法运算来合成风格、局部和全局信息。具体地说，拼接将风格特征图和$G1$的输出组合在一起，生成新的合成特征图。然后，利用元素相加将新的特征映射与$G2$的编码部分的潜在特征相结合。最后，本文使用$G2$的译码部分来生成最终的输出X_o。值得注意的是，风格向量可以控制生成的素描的风格，这有助于提高它们的质量和多样性。此外，真实照片的风格往往是固定的、且独立于艺术家的风格。因此，本文在I2S任务中引入风格信息，而在S2I中摒弃这种风格信息。

4.4 损失函数

本文联合使用几个损失函数进行模型训练。分别把X和Y表示为输入和其对应的参考图。简单来说，本文将$G(X)$定义为输入X和其对应的第k个判别器的预测概率为$D_k(X, Y)$的生成的输出。然后，本文将判别器D_k的第i层特征提取器记为D_k^i，其中k为判别器的索引。

对抗损失。 本文采用对抗损失 [154]去使得生成的图像更具有视觉吸引力。对抗损失定义为：

表5 在训练和测试数据集中每个属性的图像数目。

FS2K（Ours）	w/ H	w/o H	H（b）	H（bl）	H（r）	H（g）	M	F	w/ E	w/o E	w/ S	w/o S	w/ F	w/o F	S1	S2	S3
Train	1010	48	288	423	60	239	574	484	209	849	645	413	917	141	357	351	350
Test	994	52	290	418	44	242	632	414	187	859	670	376	872	174	619	381	46

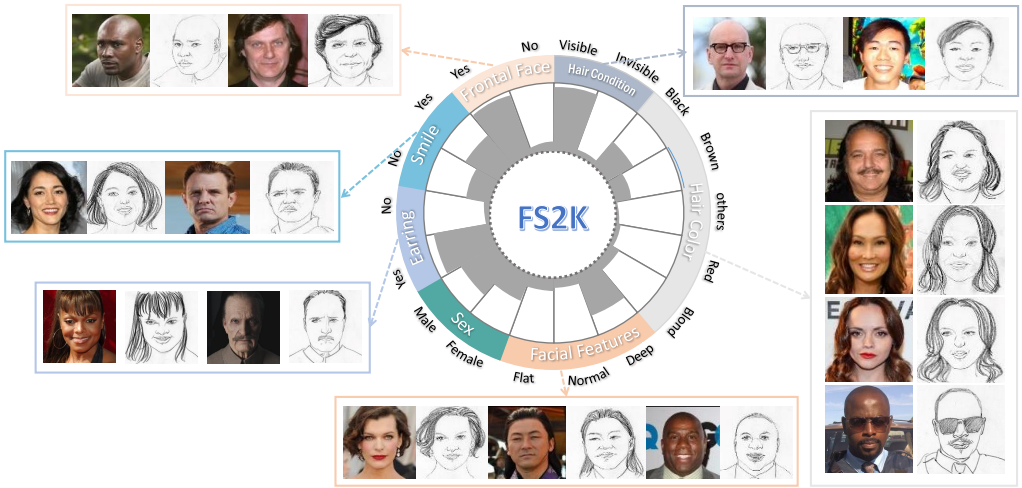


图6 本文提出的FS2K数据集的统计信息和样例。详情参考章节 3。

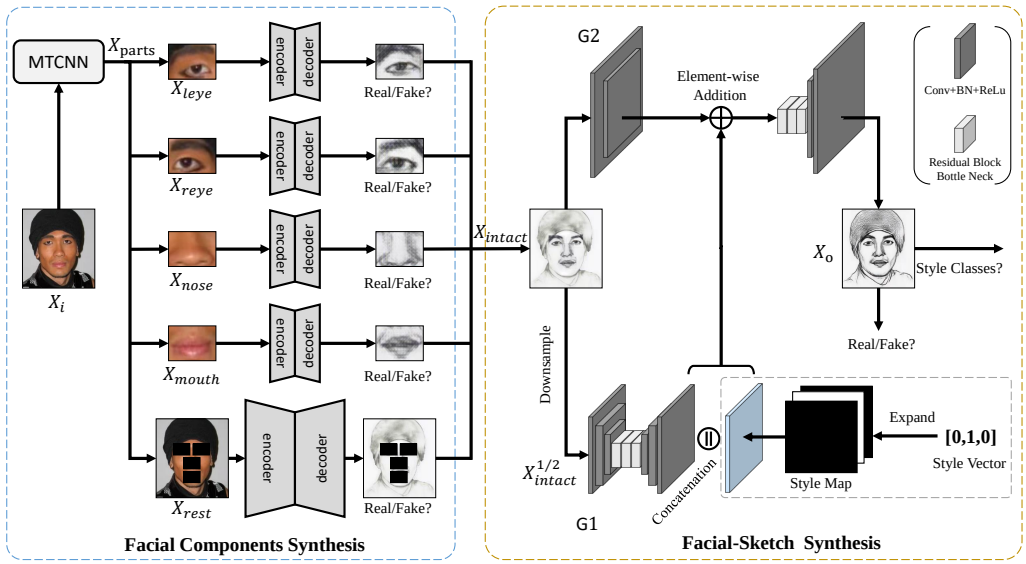


图7 本文提出的用于I2S任务的FSGAN基准线的流程。包括两个阶段：1）人脸组件合成和，2）人脸素描合成。详情参考章节 4.2和章节 4.3。

$$L_{\text{adv}}(G,D) = \mathbb{E}_{X,Y}[\log D(X,Y)] + \mathbb{E}_X[1-\log D(X,G(X))]. \quad (1)$$

特征匹配损失。 与 [16]类似，本文采用特征匹配损失在第k个判别器的基础上去提高对抗损失。特征匹配损失定义如下：

$$L_{\text{fm}}(G, D_k) = \mathbb{E}_{X,Y} \sum_{i=0}^{T} \frac{1}{N_i} \left[\left\| D_k^i(X, Y) - D_k^i(X, G(X)) \right\|_1 \right], \quad (2)$$

其中T表示每个判别器的总层数，N_i表示i-th层的特征图个数。此损失被用来匹配真实的和合成的图像的中间特征图，使得生成器产生多尺度统计信息。另外，这也使得训练过程更稳定以及恢复出高保真的结果。

感知损失。 为了保持感知和语义统一，本文采用感知损失 [60]去测量原始图像和对应的合成素描图像的差异。本文从预训练模型VGGNet [134]的第i层的激活层中提取感知特征，这被表示为$\phi_i(\cdot)$。感知损失的定义为：

$$L_{\text{per}}\big(G(X), Y\big) = \mathbb{E}_{G(X),Y} \sum_{i=0}^{t} \left\| \phi_i(Y) - \phi_i(G(X)) \right\|_1. \quad (3)$$

像素级损失。 L_1为生成的图像和参考图像之间的距离，Y则是像素级别损失，它被定义为：

$$L_1\big(G(X), Y\big) = \frac{1}{h \times w} \sum_{(i,j)=(0,0)}^{(h,w)} \left\| Y(i,j) - G(X)_{(i,j)} \right\|_1, \quad (4)$$

其中(i, j)和(h, w)分别是像素坐标以及结果图像的尺寸（即图像的高与宽）。

风格分类损失。 与 [181, 182]类似，本文定义了一个辅助分类器去预测生成图像的素描风格。对于任意生成的图像$G(X)$，风格分类器损失为：

$$L_{\text{sty}}(G, S, c) = \mathbb{E}_{X,c}\big[l_{\text{ce}}\big(S(G(X)), c\big)\big], \quad (5)$$

其中$l_{\text{ce}}(\cdot, \cdot)$是交叉熵损失，$S(\cdot)$是一个输出不同风格概率的CNN，$c$是给定的艺术家的风格标签。需要注意是，只有在I2S任务的第二个阶段使用了风格分类器损失。

整体损失。 最后，多尺度判别器的整体损失函数如下：

$$L_{D\sim(D_{parts}, D_{fs})} = \sum_{i}^{K} -L_{\text{adv}} + \lambda_{\text{fm}} L_{\text{fm}}, \quad (6)$$

对于生成器，总体的损失函数为：

$$L_{G\sim(G_{parts}, G_{fs})} = L_{\text{adv}} + \lambda_{\text{fm}} L_{\text{fm}} + \lambda_1 L_1 + \lambda_{\text{per}} L_{\text{per}} + \lambda_{\text{sty}} L_{\text{sty}}, \quad (7)$$

其中λ_{fm}、λ_1、λ_{per}和λ_{sty}分别为控制特征匹配损失重要性的超参数、像素级别损失、感知损失和风格分类器损失。

4.5 实施细节

我们使用PyTorch [183]实现本文的基准模型FSGAN。实验在一块NVIDIA V100S上进行。

对于I2S任务，在人脸组件合成阶段，本文设置参数$\lambda_{fm} = 25.0$、$\lambda_1 = 25.0$和$\lambda_{per} = 12.5$去训练模型，设置参数$\lambda_{fm} = 100.0$、$\lambda_1 = 100.0$、$\lambda_{per} = 50.0$和$\lambda_{sty} = 100.0$用于人脸合成。整个网络的训练采用Adam优化器 [184]。生成器和判别器的初始学习率分别是$2e{-}4$和$1e{-}5$。优化器的其他超参采用PyTorch推荐的默认参数。本文设置epochs为50。所有的生成器和判别器交替训练。

对于S2I任务，本文设置$\lambda_{fm} = 50.0$、$\lambda_1 = 50.0$和$\lambda_{per} = 0.2$训练神经网络的人脸组件合成阶段，设置$\lambda_{fm} = 100.0$、$\lambda_1 = 100.0$和$\lambda_{per} = 0.2$训练人脸合成。本文使用Adam优化器，以初始学习率$2e-4$训练全部的生成器和判别器。训练策略与I2S任务几乎相同。而本文设置epoch为400[14]，在250 个epoch之后冻结人脸组件合成模块的权重并在后面的epoch中训练人脸合成模块。

5 基准

本章节提供现有模型在FS2K数据集上关于I2S和S2I任务方面的综合对比和分析。

5.1 实验设置

5.1.1 评价度量

对于I2S任务，最常用的人脸素描度量是结构相似度度量指标（SSIM） [13, 146]。然而，它忽视了预测图像和参考图像的感知相似度。因此，本文采用了近期提出的结构共生纹理度量指标（SCOOT） [23]。对于S2I任务，本文仍然采用最广泛应用的SSIM度量去评价合成的人脸。本文所采用评价工具包可以在此获取 https://github.com/DengPingFan/FS2KToolbox。

5.1.2 对比模型

为了评估I2S和S2I任务的性能，本文展示了19个有代表性的方法以及FSGAN评测的经验结果。

[14] S2I任务需要保留RGB图像的细节信息，需要更多的训练次数。

表6 I2S任务的热门模型的定量结果。“↑” 表示更高，更好。

#	模型 出版信息	SCOOT↑	SSIM↑
1	DualGAN [71] Yi 等人 ICCV	0.261	0.324
2	FPST [62] Chen 等人 NeurIPSW	0.271	0.460
3	NST [58, 59] Gatys 等人 CVPR	0.273	0.326
4	Pix2pix [32] Isola 等人 CVPR	0.275	0.438
5	ACL-GAN [77] Zhao 等人 ECCV	0.278	0.404
6	WCT [66] Li 等人 NeurIPS	0.282	0.369
7	AdaIN [65] Huang 等人 ICCV	0.303	0.365
8	UNIT [15] Liu 等人 NeurIPS	0.304	0.504
9	TSIT [76] Jiang 等人 ECCV	0.307	0.441
10	DRIT++ [20] Lee 等人 IJCV	0.308	0.492
11	CartoonGAN [67] Chen 等人 CVPR	0.319	0.400
12	UGATIT [31] Kim 等人 ICLR	0.323	0.457
13	NICE-GAN [19] Chen 等人 CVPR	0.327	0.473
14	CycleGAN [14] Zhu 等人 ICCV	0.348	0.435
15	MDAL [29] Zhang 等人 TNNLS	0.355	0.466
16	UPDG [28] Yi 等人 CVPR	0.364	0.471
17	Pix2pixHD [16] Wang 等人 CVPR	0.374	0.492
18	APDrawing [3] Yi 等人 CVPR	0.375	0.464
19	DSMAP [18] Chang 等人 ECCV	0.378	0.493
20	**FSGAN** Fan 等人 MIR	**0.405**	**0.510**

表7 S2I任务热门模型的定量结果。“↑” 意味这更高，更好。

#	模型	出版信息	SSIM↑
1	DualGAN [71]	Yi 等人 ICCV	0.241
2	WCT [66]	Li 等人 NeurIPS	0.311
3	ACL-GAN [77]	Zhao 等人 ECCV	0.314
4	TSIT [76]	Jiang 等人 ECCV	0.316
5	UGATIT [31]	Kim 等人 ICLR	0.317
6	NST [58, 59]	Gatys 等人 CVPR	0.335
7	CycleGAN [14]	Zhu 等人 ICCV	0.339
8	Pix2pix [32]	Isola 等人 CVPR	0.346
9	SPADE [17]	Park 等人 CVPR	0.361
10	UNIT [15]	Liu 等人 NeurIPS	0.362
11	AdaIN [65]	Huang 等人 ICCV	0.373
12	DRIT++ [20]	Lee 等人 IJCV	0.381
13	FNS [60]	Johnson 等人 ECCV	0.391
14	NICE-GAN [19]	Chen 等人 CVPR	0.397
15	FPST [62]	Chen 等人 NeurIPSW	0.400
16	pSp [70]	Richardson 等人 CVPR	0.428
17	Pix2pixHD [16]	Wang 等人 CVPR	0.433
18	DSMAP [18]	Chang 等人 ECCV	0.471
19	DeepPS [128]	Yang 等人 ECCV	0.487
20	**FSGAN**	Fan 等人 MIR	**0.503**

5.1.3 训练/测试协议

所有对比的方法的选择基于以下标准：a）公认的技术、b）开源代码、以及c）最好的性能。所有的模型都是在FS2K数据集上按照文章中设定的图像尺寸进行训练和测试。如果文章中没有提供图像的尺寸，则默认设置为512 × 512。

5.2 整体结果和分析

5.2.1 I2S任务

本文首先展示在I2S任务上，SCOOT和SSIM分数的性能总结。定量结果和定性的对比分别见表 6和图 8-10。实验观察证明FSGAN基准获得了更好的结果。为了进一步分析，本文根据SCOOT分数将所有的对比模型分为三个类别：

- 得分<= 0.3;
- 0.3 < 得分<= 0.35;
- 0.35 < 得分。

分析。 第一组模型为SCOOT分数低于0.3的。这些模型包括DualGAN [71]、FPST [62]、ST [58, 59]、Pix2pix [32]、ACL-GAN [77]和WCT [66]。如图 8所示，DualGAN、NST和WCT产生了结构变形，许多局部人脸细节丢失。DualGAN产生的图像很差，要检测出它们的人脸组件是很有挑战的。这也解释了为什么SSIM和SCOOT分数比较低。另外和其他结果相比，Pix2Pix和FPST产生了模糊的结果。ACL-GAN在视觉审美上似乎取得了较理想的结果，得到了较高的SSIM分数。然而，ACL-GAN几乎完全恢复了原始人脸结构，缺乏艺术风格。

第二组模型包括AdaIN [65]、UNIT [15]、TSIT [76]、DRIT++ [20]、CartoonGAN [67]、UGATIT [31]、NICE-GAN [19]和CycleGAN [14]，它们的SCOOT分数在0.3和0.35之间。正如图 9所示，与第一组模型相比，合成的素描图像在结构保留方面取得了更好的结果。然而，除了AdaIN，所有的模型都因复杂的背景而缺失信息（详见第二行的头发区域）。另外，CartoonGAN的结果似乎改变了输入图像的颜色，导致了较低的SSIM分数。

MDAL [29]、UPDG [28]、Pix2pixHD [16]、APDrawing [3]、DSMAP [18] 和FSGAN基准模型则被分为第三组，这些模型可以生成没有变形且全局细节丢失不多的素描。

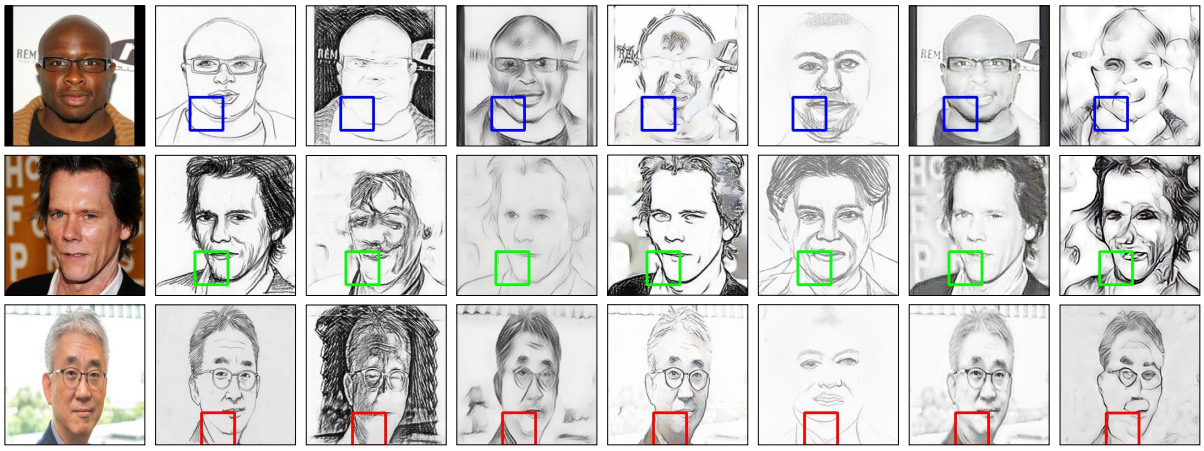


图8 从左到右依次: 输入人脸图像, 参考图像, DualGAN [71]、FPST [62]、NST [58, 59]、Pix2pix [32]、ACL-GAN [77]和WCT [66]。本文为每个结果使用蓝色、绿色和红色框标记这三种样式。可放大以便查看细节。

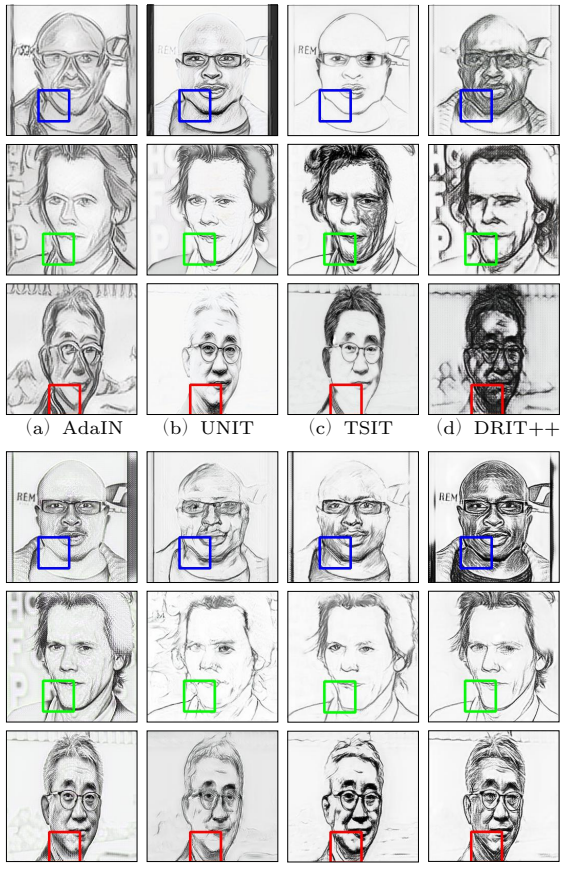


图9 AdaIN [65]、UNIT [15]、TSIT [76]、DRIT++ [20]、CartoonGAN [67]、UGATIT [31]、NICE-GAN [19]和CycleGAN [14]的对比。他们的输入和参考图像见图 8。

然而，UPDG和APDrawing在头发区域丢失了一些细节，从而导致了较差的视觉效果。APDrawing引入了一些多余的笔画，尤其是对于第一种素描风格。另外，APDrawing经常会导致局部结构的变形或者缺失，如头发区域。与此同时，UPDG生成的素描通常有更好的风格元素，但是模型不能处理有复杂背景的情况。Pix2PixHD可以生成一个全局结构相对较好，且背景干净的素描，但是它没能生成最好的人脸组件。比如，图 10-e，眼睛周围的区域不太清晰，许多细节都丢失了。以第三行为例，眼镜有一部分丢失，而眼球则为纯黑色。我们继续观察DSMAP和MDAL，这两个模型取得了最好的素描结果，但是在局部人脸信息方面有变形。最后，基准模型可以合成出高质量素描，其重点关注全局结构和局部细节，同时考虑不同的风格。此外，正如高亮框中显示（绿色、蓝色和红色），我们发现同其他先进的模型相比，FSGAN的结果与参考图像更相似。

5.2.2 S2I任务

本文在表 7和图 11中报告了实验结果。相较于其他现有的先进模型，我们发现FSGAN在本文提出的极具挑战的数据集FS2K上得到了最好的结果。

分析。 正如图 11所示，我们观察发现对比的模型都不能成功精确地恢复图像，这也揭示了S2I任务比I2S任务更复杂。我们认为这是因为素描图像的高度抽象，有价值信息的丢失，从而导致神经网络无法恢复出原始图像。我们也观察到高分辨率模型，比如Pix2PixHD和FSGAN可以生成视觉效果更好的结果。

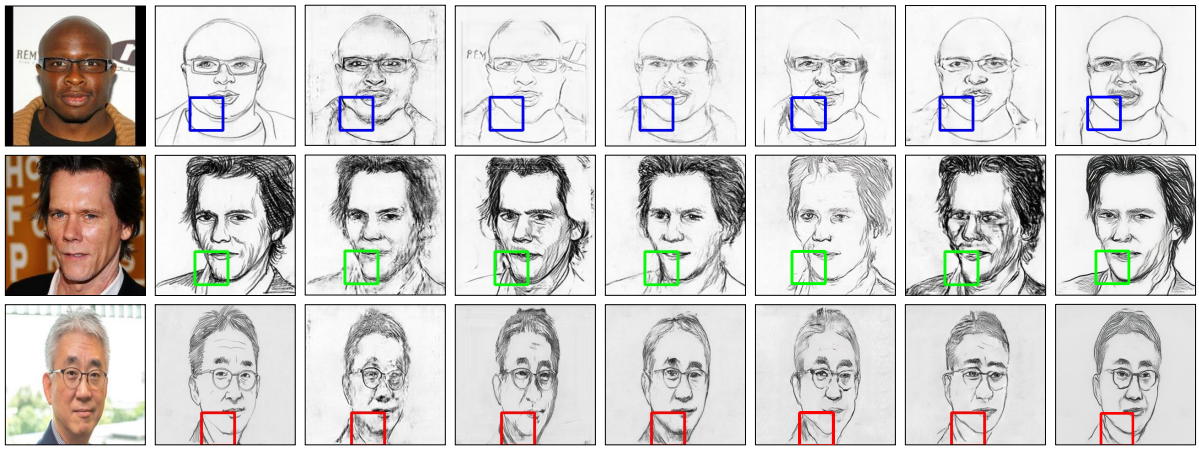


图10 同MDAL [29]、UPDG [28]、Pix2pixHD [16]、APDrawing [3]和DSMAP [18]一起的比较结果。

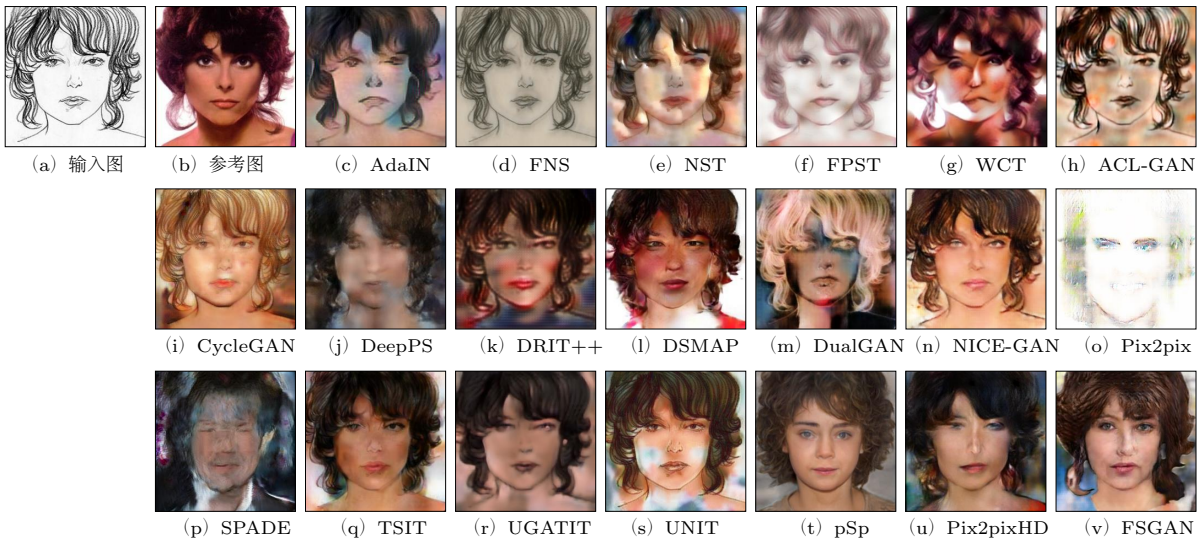


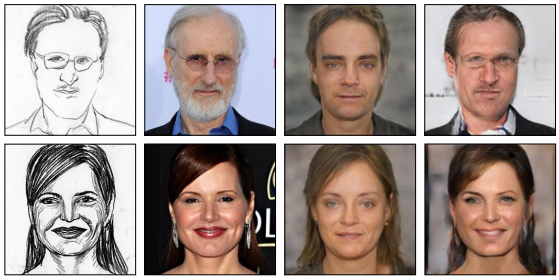
图11 本文选取了19个经典模型，包括AdaIN [65]、FNS [60]、FPST [62]、WCT [66]，ACL-GAN [77]、CycleGAN [14]、DeepPS [128]、DRIT++ [20]、DSMAP [18]、DualGAN [71]、NICE-GAN [19]、Pix2pix [32]、SPADE [17]、TSIT [76]、UGATIT [31]、UNIT [15]、pSp [70]和Pix2pixHD [16]，用于定性比较。

图 11展示的结果表明FNS和FPST不能将素描图像转换为彩色图像。SPADE和Pix2Pix生成了较差的人脸轮廓（如Pix2Pix）或者黑色背景（如SPADE）。五种模型（如，NST、WCT、DeepPS、DSMAP和UNIT）在显著区域产生了噪声块，破坏了整个人脸结构。与此同时，AdaIN、ACL-GAN、Dual-GAN和UGATIT则得到比上述模型更好的结果，生成了不真实的卡通图像。只有CycleGAN、NICE-GAN、TSIT、pSp和Pix2pixHD克服了各种挑战在人脸完整性方面取得了较好的结果。尤其，Pix2pixHD [16]和pSp [70]生成的眼部区域要比其他模型更好。然而和FSGAN相比，Pix2pixHD的人脸特征相对较差，因为它采用像素级别而不是块级别策略进行学习。虽然PSP可以生成高质量的结果，但是和FSGAN相比，它的结果缺乏多样性。比如，pSp [70]在两个不同素描风格中生成了相似的人脸表情，而本文基准模型可以合成出不同的内容，如图 12所示。

表8 19种最先进的模型在I2S任务上基于属性的性能比较。

模型	SCOOT↑																
	w/ H	w/o H	H（b）	H（bl）	H（r）	H（g）	M	F	w/ E	w/o E	w/ S	w/o S	w/ F	w/o F	S1	S2	S3
DualGAN [71]	0.260	0.279	0.250	0.267	0.216	0.279	0.275	0.240	0.239	0.266	0.255	0.271	0.261	0.262	0.298	0.194	0.319
FPST [62]	0.269	0.304	0.254	0.294	0.214	0.304	0.288	0.245	0.246	0.276	0.262	0.286	0.269	0.278	0.329	0.168	0.332
NST [58, 59]	0.272	0.283	0.268	0.287	0.236	0.283	0.280	0.262	0.258	0.276	0.268	0.282	0.272	0.276	0.310	0.205	0.332
Pix2pix [32]	0.272	0.335	0.255	0.300	0.217	0.335	0.298	0.240	0.250	0.281	0.267	0.290	0.276	0.272	0.333	0.178	0.302
ACL-GAN [77]	0.276	0.309	0.265	0.298	0.226	0.309	0.292	0.256	0.254	0.283	0.270	0.291	0.276	0.284	0.330	0.183	0.355
WCT [66]	0.281	0.315	0.271	0.302	0.229	0.315	0.296	0.261	0.262	0.287	0.277	0.292	0.281	0.290	0.332	0.195	0.346
AdaIN [65]	0.303	0.295	0.307	0.317	0.258	0.295	0.306	0.298	0.283	0.307	0.298	0.310	0.300	0.314	0.348	0.215	0.419
UNIT [15]	0.301	0.364	0.292	0.328	0.225	0.364	0.330	0.265	0.261	0.313	0.293	0.324	0.301	0.319	0.376	0.175	0.411
TSIT [76]	0.307	0.307	0.308	0.320	0.259	0.307	0.320	0.288	0.283	0.313	0.300	0.320	0.306	0.316	0.359	0.208	0.432
DRIT++ [20]	0.305	0.348	0.291	0.336	0.248	0.348	0.329	0.276	0.279	0.314	0.299	0.323	0.305	0.323	0.380	0.181	0.378
CartoonGAN [67]	0.319	0.318	0.320	0.337	0.262	0.318	0.329	0.304	0.291	0.325	0.314	0.329	0.317	0.332	0.382	0.204	0.428
UGATIT [31]	0.321	0.365	0.315	0.347	0.265	0.365	0.339	0.298	0.298	0.328	0.314	0.338	0.322	0.325	0.391	0.204	0.400
NICE-GAN [19]	0.325	0.355	0.320	0.357	0.262	0.355	0.342	0.303	0.302	0.332	0.317	0.343	0.325	0.333	0.398	0.201	0.401
CycleGAN [14]	0.348	0.343	0.358	0.362	0.287	0.343	0.351	0.343	0.326	0.353	0.341	0.360	0.346	0.357	0.397	0.252	0.483
MDAL [29]	0.354	0.363	0.348	0.380	0.292	0.363	0.369	0.333	0.329	0.360	0.345	0.372	0.352	0.365	0.436	0.211	0.446
UPDG [28]	0.362	0.411	0.349	0.390	0.290	0.411	0.390	0.325	0.336	0.371	0.356	0.379	0.363	0.370	0.423	0.259	0.448
APDrawing [3]	0.374	0.395	0.372	0.399	0.322	0.395	0.380	0.369	0.356	0.380	0.370	0.385	0.373	0.390	0.456	0.227	0.524
Pix2pixHD [16]	0.374	0.392	0.365	0.403	0.307	0.385	0.392	0.351	0.343	0.378	0.371	0.392	0.371	0.381	0.462	0.212	0.508
DSMAP [18]	0.375	0.431	0.357	0.405	0.322	0.431	0.400	0.343	0.354	0.383	0.369	0.393	0.377	0.381	0.437	**0.276**	0.423
FSGAN	**0.403**	**0.435**	**0.389**	**0.435**	**0.335**	**0.435**	**0.423**	**0.377**	**0.381**	**0.410**	**0.395**	**0.422**	**0.403**	**0.414**	**0.481**	0.268	**0.509**

这里，w/ H = 头发可见，w/o H = 头发不可见，H（b）= 棕色头发，H（bl）= 黑色头发，H（r）= 红色头发，H（g）= 金色头发，M = 男性，F = 女性，w/ E = 有耳环，w/o E = = 无耳环，w/ S = 有笑容，w/o S = 无笑容，w/ F = 正脸，w/o F = 非正脸，S1 = 风格1，S2 = 风格2和S3 = 风格3。



（a）输入图　（b）参考图　（c）pSp　（d）FSGAN

图12 S2I任务生成的数据的视觉多样性。

5.3 基于属性分析

5.3.1 SCOOT度量结果

为了深入了解模型，本文在表 8列出了基于属性性能评价。

分析。 头发是头部的主要特征之一。表 8中，我们发现除AdaIN、CartoonGAN和CycleGAN外，大部分模型在无毛发的图像上的性能略好于有头发的图像。同时，我们发现红色和黑色的头发也分别是最具挑战性和最容易检测/重建的。我们认为，这是因为红色和黑色头发的图像分别占所有数据的最低和最大比例（>40%）。因此，模型对这些属性并不熟悉/熟悉。

此外，我们还注意到，对于几乎所有的模型来说，女性（F）比男性（M）更具挑战性，因为女性通常有不同的配饰和发型。例如，模型在戴耳环（W/E）的图像上的表现比不戴耳环的差。此外，有微笑的人脸图像比没有微笑的人脸图像更具挑战性。有趣的是，现有模型取得了不同的性能，而与头发的颜色无关（例如，H（b）、H（bl）、H（r）和H（g））。最后，与风格1（简单的线条）和风格3（即重复的细小细节）相比，我们看到风格2（长笔划）对所有模型来说都是最具挑战性的。

5.3.2 SSIM评测结果

除了SCOOT度量，对于I2S任务，本文在表 9中提供了SSIM度量。

分析。 我们发现，在头发、性别、配饰和发型等几个关键属性上，总体表现与SCOOT指标结果相似。我们注意到，在“w/ F”上的表现低于在“w/o F”上的表现，如表 9所示。一个可能的原因是正脸比非正脸保留了更多的结构特征。因此，在I2S任务中，具有“w/ F”等属性的图像比具有“w/o F”属性的图像更具挑战性。

5.4 消融实验

本节展示了FSGAN在本文FS2K数据集上的详细分析。与大多数现有的人脸合成模型不同 [16]，本文的模型具有用于I2S和S2I任务的两阶段GAN架构。此外，还引入了素描风格向量，以支持I2S任务第二阶段的多样化风格合成。因此，对I2S任务的消融研究主要集中在以下两个关键部分：（1）人脸成分合成阶段和（2）风格向量辅助生成阶段。请注意，在消融实验中本文采用了与章节 4.5中相同的超参数。

表 10列出了I2S任务的消融结果。我们发现，人脸成分合成阶段SCOOT和SSIM得分分别提高

表9 19种最先进的模型在I2S任务上基于属性的性能比较。

模型	SSIM↑																	
	w/ H	w/o H	H (b)	H (bl)	H (r)	H (g)	M	F	w/ E	w/o E	w/ S	w/o S	w/ F	w/o F	S1	S2	S3
DualGAN [71]	0.320	0.393	0.310	0.342	0.276	0.393	0.352	0.282	0.292	0.331	0.313	0.343	0.318	0.354	0.364	0.247	0.424
FPST [62]	0.459	0.481	0.442	0.492	0.383	0.481	0.492	0.411	0.416	0.469	0.448	0.481	0.455	0.486	0.517	0.351	0.597
NST [58, 59]	0.325	0.347	0.317	0.349	0.256	0.347	0.339	0.306	0.305	0.330	0.316	0.344	0.324	0.338	0.372	0.241	0.417
Pix2pix [32]	0.434	0.526	0.410	0.470	0.332	0.526	0.478	0.377	0.391	0.449	0.425	0.461	0.438	0.439	0.503	0.319	0.558
ACL-GAN [77]	0.402	0.432	0.392	0.430	0.334	0.432	0.427	0.369	0.363	0.413	0.393	0.423	0.398	0.434	0.445	0.316	0.583
WCT [66]	0.368	0.389	0.368	0.387	0.316	0.389	0.389	0.339	0.334	0.377	0.362	0.381	0.367	0.380	0.407	0.297	0.461
AdaIN [65]	0.364	0.367	0.364	0.382	0.319	0.367	0.378	0.343	0.340	0.370	0.359	0.375	0.362	0.379	0.399	0.297	0.460
UNIT [15]	0.501	0.556	0.488	0.528	0.421	0.556	0.539	0.450	0.460	0.514	0.492	0.526	0.498	0.532	0.563	0.395	0.616
TSIT [76]	0.439	0.465	0.430	0.461	0.371	0.465	0.465	0.404	0.408	0.448	0.431	0.458	0.435	0.468	0.485	0.351	0.587
DRIT++ [20]	0.490	0.534	0.479	0.519	0.411	0.534	0.524	0.444	0.451	0.501	0.480	0.512	0.487	0.515	0.547	0.387	0.617
CartoonGAN [67]	0.399	0.420	0.397	0.421	0.345	0.420	0.419	0.372	0.368	0.407	0.392	0.416	0.395	0.425	0.438	0.321	0.552
UGATIT [31]	0.455	0.497	0.445	0.476	0.386	0.497	0.489	0.409	0.416	0.466	0.447	0.476	0.451	0.491	0.499	0.373	0.593
NICE-GAN [19]	0.472	0.497	0.463	0.492	0.398	0.497	0.505	0.424	0.429	0.483	0.464	0.490	0.468	0.498	0.518	0.384	0.603
CycleGAN [14]	0.433	0.461	0.429	0.455	0.374	0.461	0.460	0.395	0.401	0.442	0.425	0.452	0.429	0.463	0.471	0.358	0.580
MDAL [29]	0.465	0.487	0.457	0.491	0.399	0.487	0.496	0.420	0.426	0.475	0.458	0.481	0.462	0.488	0.506	0.386	0.593
UPDG [28]	0.468	0.507	0.456	0.500	0.391	0.507	0.501	0.424	0.431	0.479	0.459	0.493	0.465	0.501	0.534	0.355	0.584
APDrawing [3]	0.461	0.522	0.441	0.497	0.373	0.522	0.504	0.402	0.419	0.473	0.452	0.484	0.458	0.492	0.512	0.371	0.582
Pix2pixHD [16]	0.492	0.552	0.473	0.523	0.419	0.546	0.531	0.431	0.457	0.505	0.481	0.513	0.488	0.524	0.537	0.402	0.618
DSMAP [18]	0.490	0.551	0.472	0.527	0.405	0.551	0.532	0.433	0.447	0.503	0.481	0.515	0.488	0.518	0.557	0.373	0.622
FSGAN	**0.507**	**0.565**	**0.491**	**0.539**	**0.424**	**0.565**	**0.549**	**0.451**	**0.466**	**0.520**	**0.498**	**0.531**	**0.505**	**0.534**	**0.568**	**0.403**	**0.629**

表10 FSGAN在I2S任务上的消融实验。

设置	多块	风格向量	SCOOT↑	SSIM↑
Baseline			0.381	0.487
	✓		0.386 (+1.31%)	0.500 (+2.67%)
FSGAN	✓	✓	**0.405** (+6.30%)	**0.510** (+4.72%)

了1.31%（相对）和2.67%，而风格向量则分别提高了6.30%和4.72%。如图 13所示，在没有多块策略的情况下，合成嘴唇中的线条往往缺少结构细节。同时，随着多块阶段的进行，线条变得更加平滑。此外，合成的图形在没有风格向量分量的情况下更加混乱，并且可能在嘴唇区域中引入阴影。对于S2I任务，进行了一项消融研究，以验证人脸组件合成阶段的有效性，如表 11所示。与I2S任务类似，多贴片组成在基线模型上取得了显著的性能提升（即, 3.3%）。图 14提供了本文的模型和没有人脸成分合成阶段的模型产生的结果的示例。正如我们所看到的，本文的人脸成分合成模型捕捉到了更多的细节，并确保了更真实的整体外观（见图 14-c）。

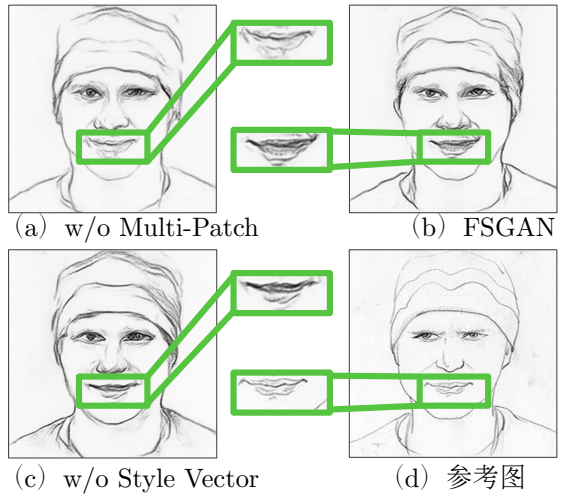


图13 I2S任务的消融实验。

6 讨论

虽然FSS已经取得了巨大进展，但仍有很大的改进空间。这一部分总结未来可能的研究方向。

（1）数据集。 由于专业素描艺术家的相对缺乏，获取大量的图像仍然是一个悬而未决的问题，这也阻碍了FSS的发展。此外，需要更多样化的素描（或绘图）风格来构建更吸引人的模型，并获得更好的合成效果。为了解决这些问题，本文认为为FSS设计新的数据增强技术 [95, 185, 186]和迁移学习策略[187–189]是很有前景的研究方向。

（2）模型。 目前，大多数最先进的模型都是用大量的成对图像和素描 [16, 28]进行训练，以克服数据短缺的问题。然而，可以更多地关注少样本学习 [190]、半监督 [191]、弱监督 [192]、自监督 [193]和非成对无监督 [79]学习等技术，以在有限的数据集下实现风格迁移。此外，开发新颖的、人在回路（Human in the loop）[194]的深度模型是另一个很有前景的方向，它将为用户提供更多的互动选择，以生成和编辑个性化的风格。利用本文FS2K中的属性的交互式模型也可为专业艺术家提供绘画工具，以促进素描和其他绘画风格的创作。此外，自然场景的FSS仍然具有挑战性，因为图像质量，包括分辨率、噪声和背景，变化很大。除了上述技术外，还可以重点研究基础模型单元，以设计新的策略。例如，大

表11 FSGAN在S2I任务上的消融实验。

设置	多块	SSIM↑
Baseline		0.487
FSGAN	✓	**0.503** （+3.3%）

（a）输入 （b）无多块策略 （c）FSGAN

图14 S2I任务的消融实验。

多数当前的模型都是建立在CNN [195]单元之上的。因此，还可以对其他框架进行更多的探索，如MLP [196]和Transformers [197, 198]。

（3）测评。 评测指标对于新模型的开发和现有模型的基准测试至关重要。目前，使用的主要为几种定量评估指标 [13, 199]和人眼视觉排序方法 [50]。然而，由于这些指标的目的是在所有模型之间提供相对客观和公平的比较，因此没有考虑FSS在不同应用场景的情况。这可能会导致对具体任务的评估有偏见或不可靠。因此，更多针对任务的评估指标和方法可能是未来研究的另一个重要方向。

（4）应用。 目前，FSS（I2S和S2I）的唯一直接应用是娱乐和执法 [2, 146]。随着FSS技术的发展，FSS研究也可以隐含或显式地促进许多其他有前景的应用，如艺术设计、动画制作等。除了这些工业应用，我们相信FSS的方法和思想也可以惠及其他研究领域。例如，素描可用于辅助调整图像大小 [200]、超分辨率 [201]等。此外，素描通常包含图像的最显著信息，因此可被认为是RGB图像的压缩版本 [202]。这一特性使得素描对于图像压缩任务非常有用。此外，S2I任务可以被认为是广义上的图像超分辨率的特定情况，因为这两个任务的目的都是从给定的输入重建详细的RGB图像。不同的是，S2I的输入是高频信息，而标准超分辨率任务的输入是原始图像的低频信息。

7 结论

本文对人脸素描合成问题进行了全面的综述。据本文所知，这是第一次在素描到图像和图像到素描任务中对FSS进行深层的系统研究。为了实现这一点，本文建立了一个新的具有挑战性的数据集，名为FS2K。本文还引入了拷贝灯，以解决艺术家绘制的素描与原始图像之间的对齐问题。提出的简单基准FSGAN通过两级架构取得了最先进的性能。最后，作为最广泛的概述（即89种文献方法）和基准（即19种前沿模型），本文揭示了该领域的发展仍处于起步阶段。因此，这篇论文的主要目的是激发新奇的想法，而不是对所有标杆作品进行排名。由于该领域发展繁荣，要对所有现有模型进行基准测试并非易事。本文希望这次概述能引起社会的关注，并产生令人兴奋的后续方向，如用音乐生成生动的小品，由小品开发卡通、合成小品视频、以及假脸检测 [203]等。

References

[1] D.-P. Fan, Z. Huang, P. Zheng, H. Liu, X. Qin, and L. Van Gool, "Facial-sketch synthesis: A new challenge," *Machine Intelligence Research*, 2022.

[2] X. Wang and X. Tang, "Face photo-sketch synthesis and recognition," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 31, no. 11, pp. 1955–1967, 2008.

[3] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "APDrawingGAN: Generating artistic portrait drawings from face photos with hierarchical GANs," in *Conference on computer vision and pattern recognition*. IEEE, 2019, pp. 10 743–10 752.

[4] H. Koshimizu, M. Tominaga, T. Fujiwara, and K. Murakami, "On kansei facial image processing for computerized facial caricaturing system picasso," in *International Conference on Systems, Man, and Cybernetics*. IEEE, 1999, pp. 294–299.

[5] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and simile classifiers for face verification," in *International conference on computer vision*. IEEE, 2009, pp. 365–372.

[6] H.-S. Du, Q.-P. Hu, D.-F. Qiao, and I. Pitas, "Robust face recognition via low-rank sparse representation-based classification," *International Journal of Automation and Computing*, vol. 12, no. 6, pp. 579–587, 2015.

[7] Y.-Z. Lu, "A novel face recognition algorithm for distinguishing faces with various angles," *International Journal of*

Automation and Computing, vol. 5, no. 2, pp. 193–197, 2008.

[8] V. Jain and E. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," UMass Amherst technical report, Tech. Rep., 2010.

[9] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *European conference on computer vision*. Springer, 2014, pp. 94–108.

[10] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks)," in *International conference on computer vision*. IEEE, 2017, pp. 1021–1030.

[11] J. Sun, Q. Li, W. Wang, J. Zhao, and Z. Sun, "Multi-caption text-to-face synthesis: Dataset and algorithm," in *International conference on Multimedia*. ACM, 2021, pp. 2290–2298.

[12] R. Yi, M. Xia, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "Line drawings for face portraits from photos using global and local structure based GANs," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3462–3475, 2020.

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

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

[15] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in neural information processing systems*. Curran Associates, Inc., 2017.

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

[17] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Conference on computer vision and pattern recognition*. IEEE, 2019, pp. 2337–2346.

[18] H.-Y. Chang, Z. Wang, and Y.-Y. Chuang, "Domain-specific mappings for generative adversarial style transfer," in *European conference on computer vision*. Springer, 2020, pp. 573–589.

[19] R. Chen, W. Huang, B. Huang, F. Sun, and B. Fang, "Reusing discriminators for encoding: Towards unsupervised image-to-image translation," in *Conference on computer vision and pattern recognition*. IEEE, 2020, pp. 8168–8177.

[20] H.-Y. Lee, H.-Y. Tseng, Q. Mao, J.-B. Huang, Y.-D. Lu, M. Singh, and M.-H. Yang, "Drit++: Diverse image-to-image translation via disentangled representations," *International journal of computer vision*, vol. 128, no. 10, pp. 2402–2417, 2020.

[21] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *International conference on computer vision*. IEEE, 2015, pp. 3730–3738.

[22] W. Zhang, X. Wang, and X. Tang, "Coupled information-theoretic encoding for face photo-sketch recognition," in *Conference on computer vision and pattern recognition*. IEEE, 2011, pp. 513–520.

[23] D.-P. Fan, S. Zhang, Y.-H. Wu, Y. Liu, M.-M. Cheng, B. Ren, P. L. Rosin, and R. Ji, "Scoot: A perceptual metric for facial sketches," in *International conference on computer vision*. IEEE, 2019, pp. 5612–5622.

[24] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "On matching sketches with digital face images," in *International Conference on Biometrics: Theory, Applications and Systems*. IEEE, 2010, pp. 1–7.

[25] N. Wang, X. Gao, D. Tao, and X. Li, "Face sketch-photo synthesis under multi-dictionary sparse representation framework," in *International Conference on Image and Graphics*. IEEE, 2011, pp. 82–87.

[26] X. Gao, N. Wang, D. Tao, and X. Li, "Face sketch-photo synthesis and retrieval using sparse representation," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 8, pp. 1213–1226, 2012.

[27] I. Berger, A. Shamir, M. Mahler, E. Carter, and J. Hodgins, "Style and abstraction in portrait sketching," *ACM Transactions on graphics*, vol. 32, no. 4, pp. 1–12, 2013.

[28] R. Yi, Y.-J. Liu, Y.-K. Lai, and P. L. Rosin, "Unpaired portrait drawing generation via asymmetric cycle mapping," in *Conference on computer vision and pattern recognition*. IEEE, 2020, pp. 8217–8225.

[29] S. Zhang, R. Ji, J. Hu, X. Lu, and X. Li, "Face sketch synthesis by multidomain adversarial learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 5, pp. 1419–1428, 2018.

[30] M. Zhu, J. Li, N. Wang, and X. Gao, "Knowledge distillation for face photo-sketch synthesis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 893–906, 2022.

[31] J. Kim, M. Kim, H. Kang, and K. Lee, "U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation," in *International Conference on Learning Representations*. OpenReview.net, 2020.

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

[33] C. Peng, X. Gao, N. Wang, and J. Li, "Face recognition from multiple stylistic sketches: Scenarios, datasets, and evaluation," *Pattern Recognition*, vol. 84, pp. 262–272, 2018.

[34] A. M. Martinez, "The ar face database," *CVC Technical Report24*, 1998.

[35] K. Messer, J. Matas, J. Kittler, J. Luettin, G. Maitre *et al.*, "XM2VTSDB: The extended m2vts database," in *International conference on audio and video-based biometric person authentication*, vol. 964. Springer, 1999, pp. 965–966.

[36] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[37] H. Chen, Y.-Q. Xu, H.-Y. Shum, S.-C. Zhu, and N.-N. Zheng, "Example-based facial sketch generation with non-parametric sampling," in *International conference on computer vision*. IEEE, 2001, pp. 433–438.

[38] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, "A nonlinear approach for face sketch synthesis and recognition," in *Conference on computer vision and pattern recognition*. IEEE, 2005, pp. 1005–1010.

[39] X. Gao, J. Zhong, J. Li, and C. Tian, "Face sketch synthesis algorithm based on e-hmm and selective ensemble," *IEEE Transactions on circuits and systems for video technology*, vol. 18, no. 4, pp. 487–496, 2008.

[40] Z. Xu, H. Chen, S.-C. Zhu, and J. Luo, "A hierarchical compositional model for face representation and sketching," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 6, pp.

955–969, 2008.

[41] W. Zhang, X. Wang, and X. Tang, "Lighting and pose robust face sketch synthesis," in *European conference on computer vision*. Springer, 2010, pp. 420–433.

[42] N. Ji, X. Chai, S. Shan, and X. Chen, "Local regression model for automatic face sketch generation," in *International Conference on Image and Graphics*. IEEE, 2011, pp. 412–417.

[43] L. Chang, M. Zhou, X. Deng, Z. Wu, and Y. Han, "Face sketch synthesis via multivariate output regression," in *International conference on human-computer interaction*. Springer, 2011, pp. 555–561.

[44] J. Zhang, N. Wang, X. Gao, D. Tao, and X. Li, "Face sketch-photo synthesis based on support vector regression," in *International Conference on Image Processing*. IEEE, 2011, pp. 1125–1128.

[45] S. Wang, L. Zhang, Y. Liang, and Q. Pan, "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in *Conference on computer vision and pattern recognition*. IEEE, 2012, pp. 2216–2223.

[46] H. Zhou, Z. Kuang, and K.-Y. K. Wong, "Markov weight fields for face sketch synthesis," in *Conference on computer vision and pattern recognition*. IEEE, 2012, pp. 1091–1097.

[47] T. Wang, J. P. Collomosse, A. Hunter, and D. Greig, "Learnable stroke models for example-based portrait painting," in *British Machine Vision Conference*. BMVA Press, 2013.

[48] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "Transductive face sketch-photo synthesis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24, no. 9, pp. 1364–1376, 2013.

[49] D.-A. Huang and Y.-C. F. Wang, "Coupled dictionary and feature space learning with applications to cross-domain image synthesis and recognition," in *International conference on computer vision*. IEEE, 2013, pp. 2496–2503.

[50] Y. Song, L. Bao, Q. Yang, and M.-H. Yang, "Real-time exemplar-based face sketch synthesis," in *European conference on computer vision*. Springer, 2014, pp. 800–813.

[51] S. Zhang, X. Gao, N. Wang, and J. Li, "Robust face sketch style synthesis," *IEEE Transactions on image processing*, vol. 25, no. 1, pp. 220–232, 2015.

[52] C. Peng, X. Gao, N. Wang, and J. Li, "Superpixel-based face sketch-photo synthesis," *IEEE Transactions on circuits and systems for video technology*, vol. 27, no. 2, pp. 288–299, 2015.

[53] C. Peng, X. Gao, N. Wang, D. Tao, X. Li, and J. Li, "Multiple representations-based face sketch-photo synthesis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 11, pp. 2201–2215, 2015.

[54] Y. Li, Y.-Z. Song, T. M. Hospedales, and S. Gong, "Free-hand sketch synthesis with deformable stroke models," *International journal of computer vision*, vol. 122, no. 1, pp. 169–190, 2017.

[55] J. Li, X. Yu, C. Peng, and N. Wang, "Adaptive representation-based face sketch-photo synthesis," *Neurocomputing*, vol. 269, pp. 152–159, 2017.

[56] N. Wang, X. Gao, and J. Li, "Random sampling for fast face sketch synthesis," *Pattern Recognition*, vol. 76, pp. 215–227, 2018.

[57] Y. Men, Z. Lian, Y. Tang, and J. Xiao, "A common framework for interactive texture transfer," in *Conference on computer vision and pattern recognition*. IEEE, 2018, pp. 6353–6362.

[58] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," *arXiv preprint arXiv:1508.06576*, 2015.

[59] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Conference on computer vision and pattern recognition*. IEEE, 2016, pp. 2414–2423.

[60] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.

[61] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. S. Lempitsky, "Texture networks: Feed-forward synthesis of textures and stylized images," in *International conference on machine learning*. PMLR, 2016, p. 1349–1357.

[62] T. Q. Chen and M. Schmidt, "Fast patch-based style transfer of arbitrary style," in *Advances in neural information processing systems workshops*. Curran Associates, Inc., 2016.

[63] V. Dumoulin, J. Shlens, and M. Kudlur, "A learned representation for artistic style," in *International Conference on Learning Representations*. OpenReview.net, 2017.

[64] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Conference on computer vision and pattern recognition*. IEEE, 2017, pp. 6924–6932.

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

[66] Y. Li, C. Fang, J. Yang, Z. Wang, X. Lu, and M.-H. Yang, "Universal style transfer via feature transforms," in *Advances in neural information processing systems*. Curran Associates, Inc., 2017.

[67] Y. Chen, Y.-K. Lai, and Y.-J. Liu, "Cartoongan: Generative adversarial networks for photo cartoonization," in *Conference on computer vision and pattern recognition*. IEEE, 2018, pp. 9465–9474.

[68] R. Abdal, Y. Qin, and P. Wonka, "Image2styleGAN: How to embed images into the stylegan latent space?" in *International conference on computer vision*. IEEE, 2019, pp. 4432–4441.

[69] D. Kotovenko, M. Wright, A. Heimbrecht, and B. Ommer, "Rethinking style transfer: From pixels to parameterized brushstrokes," in *Conference on computer vision and pattern recognition*. IEEE, 2021, pp. 12 196–12 205.

[70] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," in *Conference on computer vision and pattern recognition*. IEEE, 2021, pp. 2287–2296.

[71] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *International conference on computer vision*. IEEE, 2017, pp. 2849–2857.

[72] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *International conference on machine learning*. PMLR, 2017, pp. 1857–1865.

[73] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Advances in neural information processing systems*. Curran Associates, Inc., 2017.

[74] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *European conference on computer vision*. Springer, 2018, pp. 172–189.

[75] P. Zhang, B. Zhang, D. Chen, L. Yuan, and F. Wen, "Cross-domain correspondence learning for exemplar-based image translation," in *Conference on computer vision and pattern recognition*. IEEE, 2020, pp. 5143–5153.

[76] L. Jiang, C. Zhang, M. Huang, C. Liu, J. Shi, and C. C. Loy, "TSIT: A simple and versatile framework for image-to-image translation," in *European conference on computer vision*. Springer, 2020, pp. 206–222.

[77] Y. Zhao, R. Wu, and H. Dong, "Unpaired image-to-image translation using adversarial consistency loss," in *European conference on computer vision*. Springer, 2020, pp. 800–815.

[78] X. Zhou, B. Zhang, T. Zhang, P. Zhang, J. Bao, D. Chen, Z. Zhang, and F. Wen, "CoCosNet v2: Full-resolution correspondence learning for image translation," in *Conference on computer vision and pattern recognition*. IEEE, 2021, pp. 11 465–11 475.

[79] A. Chen, R. Liu, L. Xie, Z. Chen, H. Su, and J. Yu, "Sofgan: A portrait image generator with dynamic styling," *ACM Transactions on graphics*, vol. 41, no. 1, pp. 1–26, 2022.

[80] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *ACM Transactions on graphics*, vol. 31, no. 4, pp. 1–10, 2012.

[81] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[82] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[83] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *Conference on computer vision and pattern recognition*. IEEE, 2014, pp. 3606–3613.

[84] S. Y. Duck, "Painter by numbers, wikiart.org," https://www.kaggle.com/c/painter-by-numbers, 2016.

[85] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Conference on computer vision and pattern recognition*. IEEE, 2016, pp. 3213–3223.

[86] R. Tyleček and R. Šára, "Spatial pattern templates for recognition of objects with regular structure," in *German conference on pattern recognition*. Springer, 2013, pp. 364–374.

[87] J.-Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *European conference on computer vision*. Springer, 2016, pp. 597–613.

[88] A. Yu and K. Grauman, "Fine-grained visual comparisons with local learning," in *Conference on computer vision and pattern recognition*. IEEE, 2014, pp. 192–199.

[89] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays, "Transient attributes for high-level understanding and editing of outdoor scenes," *ACM Transactions on graphics*, vol. 33, no. 4, pp. 1–11, 2014.

[90] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[91] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," 2011.

[92] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *International Conference on Learning Representations*. OpenReview.net, 2018.

[93] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*. Springer, 2012, pp. 746–760.

[94] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Conference on computer vision and pattern recognition*. IEEE, 2017, pp. 633–641.

[95] Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Sketchx!-shoe/chair fine-grained SBIR dataset," 2017.

[96] D. Ha and D. Eck, "A neural representation of sketch drawings," in *International Conference on Learning Representations*. OpenReview.net, 2018.

[97] Y. Jin, J. Zhang, M. Li, Y. Tian, H. Zhu, and Z. Fang, "Towards the automatic anime characters creation with generative adversarial networks," in *Advances in neural information processing systems workshops*. Curran Associates, Inc., 2017.

[98] H. Xu, Y. Gao, F. Yu, and T. Darrell, "End-to-end learning of driving models from large-scale video datasets," in *Conference on computer vision and pattern recognition*. IEEE, 2017, pp. 2174–2182.

[99] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Conference on computer vision and pattern recognition*. IEEE, 2016, pp. 3234–3243.

[100] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Conference on computer vision and pattern recognition*. IEEE, 2016, pp. 1096–1104.

[101] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Conference on computer vision and pattern recognition*. IEEE, 2019, pp. 4401–4410.

[102] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *Conference on computer vision and pattern recognition workshops*. IEEE, 2017, pp. 126–135.

[103] B. Yao, X. Yang, and S.-C. Zhu, "Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks," in *CVPRW*. IEEE, 2007, pp. 169–183.

[104] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *International conference on computer vision workshops*. IEEE, 2013, pp. 554–561.

[105] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

[106] L. Zhang, L. Lin, X. Wu, S. Ding, and L. Zhang, "End-to-end photo-sketch generation via fully convolutional representation learning," in *International Conference on Multimedia Retrieval*. ACM, 2015, pp. 627–634.

[107] M. Zhu, N. Wang, X. Gao, and J. Li, "Deep graphical feature learning for face sketch synthesis," in *International Joint Conference on Artificial Intelligence*. IJCAI, 2017, pp. 3574–3580.

[108] P. Sangkloy, J. Lu, C. Fang, F. Yu, and J. Hays, "Scribbler: Controlling deep image synthesis with sketch and color," in

Conference on computer vision and pattern recognition. IEEE, 2017, pp. 5400–5409.

[109] M. Zhang, N. Wang, Y. Li, R. Wang, and X. Gao, "Face sketch synthesis from coarse to fine," in *AAAI Conference on Artificial Intelligence*. AAAI Press, 2018, pp. 7558–7565.

[110] W. Xian, P. Sangkloy, V. Agrawal, A. Raj, J. Lu, C. Fang, F. Yu, and J. Hays, "TextureGAN: Controlling deep image synthesis with texture patches," in *Conference on computer vision and pattern recognition*. IEEE, 2018, pp. 8456–8465.

[111] J. Song, K. Pang, Y.-Z. Song, T. Xiang, and T. M. Hospedales, "Learning to sketch with shortcut cycle consistency," in *Conference on computer vision and pattern recognition*. IEEE, 2018, pp. 801–810.

[112] Y. Lu, S. Wu, Y.-W. Tai, and C.-K. Tang, "Image generation from sketch constraint using contextual GAN," in *European conference on computer vision*. Springer, 2018, pp. 205–220.

[113] S. Zhang, R. Ji, J. Hu, Y. Gao, and C.-W. Lin, "Robust face sketch synthesis via generative adversarial fusion of priors and parametric sigmoid," in *International Joint Conference on Artificial Intelligence*. IJCAI, 2018, pp. 1163–1169.

[114] M. Zhang, N. Wang, X. Gao, and Y. Li, "Markov random neural fields for face sketch synthesis," in *International Joint Conference on Artificial Intelligence*. IJCAI, 2018, pp. 1142–1148.

[115] L. Wang, V. Sindagi, and V. Patel, "High-quality facial photo-sketch synthesis using multi-adversarial networks," in *International conference on automatic face & gesture recognition*. IEEE, 2018, pp. 83–90.

[116] M. Zhang, R. Wang, X. Gao, J. Li, and D. Tao, "Dual-transfer face sketch-photo synthesis," *IEEE Transactions on image processing*, vol. 28, no. 2, pp. 642–657, 2018.

[117] H. Kazemi, M. Iranmanesh, A. Dabouei, S. Soleymani, and N. M. Nasrabadi, "Facial attributes guided deep sketch-to-photo synthesis," in *Winter Applications of Computer Vision Workshops*. IEEE, 2018, pp. 1–8.

[118] H. Kazemi, F. Taherkhani, and N. M. Nasrabadi, "Unsupervised facial geometry learning for sketch to photo synthesis," in *International Conference of the Biometrics Special Interest Group*. IEEE, 2018, pp. 1–5.

[119] S. You, N. You, and M. Pan, "Pi-rec: Progressive image reconstruction network with edge and color domain," *arXiv preprint arXiv:1903.10146*, 2019.

[120] M. Zhang, N. Wang, Y. Li, and X. Gao, "Deep latent low-rank representation for face sketch synthesis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 10, pp. 3109–3123, 2019.

[121] M. Zhu, J. Li, N. Wang, and X. Gao, "A deep collaborative framework for face photo-sketch synthesis," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 10, pp. 3096–3108, 2019.

[122] M. Zhang, Y. Li, N. Wang, Y. Chi, and X. Gao, "Cascaded face sketch synthesis under various illuminations," *IEEE Transactions on image processing*, vol. 29, pp. 1507–1521, 2019.

[123] M. Zhu, N. Wang, X. Gao, J. Li, and Z. Li, "Face photo-sketch synthesis via knowledge transfer," in *International Joint Conference on Artificial Intelligence*. IJCAI, 2019, pp. 1048–1054.

[124] Y. Li, C. Fang, A. Hertzmann, E. Shechtman, and M.-H. Yang, "Im2pencil: Controllable pencil illustration from photographs," in *Conference on computer vision and pattern recognition*. IEEE, 2019, pp. 1525–1534.

[125] A. Ghosh, R. Zhang, P. K. Dokania, O. Wang, A. A. Efros, P. H. Torr,

and E. Shechtman, "Interactive sketch & fill: Multiclass sketch-to-image translation," in *International conference on computer vision*. IEEE, 2019, pp. 1171–1180.

[126] X. Wang and J. Yu, "Learning to cartoonize using white-box cartoon representations," in *Conference on computer vision and pattern recognition*. IEEE, 2020, pp. 8090–8099.

[127] C. Gao, Q. Liu, Q. Xu, L. Wang, J. Liu, and C. Zou, "SketchyCOCO: image generation from freehand scene sketches," in *Conference on computer vision and pattern recognition*. IEEE, 2020, pp. 5174–5183.

[128] S. Yang, Z. Wang, J. Liu, and Z. Guo, "Deep plastic surgery: Robust and controllable image editing with human-drawn sketches," in *European conference on computer vision*. Springer, 2020, pp. 601–617.

[129] S.-Y. Chen, W. Su, L. Gao, S. Xia, and H. Fu, "DeepFaceDrawing: Deep generation of face images from sketches," *ACM Transactions on graphics*, vol. 39, no. 4, pp. 72–1, 2020.

[130] J. Yu, X. Xu, F. Gao, S. Shi, M. Wang, D. Tao, and Q. Huang, "Toward realistic face photo-sketch synthesis via composition-aided gans," *IEEE Transactions on cybernetics*, vol. 51, no. 9, pp. 4350–4362, 2020.

[131] Y. Fang, W. Deng, J. Du, and J. Hu, "Identity-aware CycleGAN for face photo-sketch synthesis and recognition," *Pattern Recognition*, vol. 102, p. 107249, 2020.

[132] Y. Lin, S. Ling, K. Fu, and P. Cheng, "An identity-preserved model for face sketch-photo synthesis," *IEEE Signal Processing Letters*, vol. 27, pp. 1095–1099, 2020.

[133] C. Peng, N. Wang, J. Li, and X. Gao, "Universal face photo-sketch style transfer via multiview domain translation," *IEEE Transactions on image processing*, vol. 29, pp. 8519–8534, 2020.

[134] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*. OpenReview.net, 2015.

[135] S. Duan, Z. Chen, Q. J. Wu, L. Cai, and D. Lu, "Multi-scale gradients self-attention residual learning for face photo-sketch transformation," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 1218–1230, 2020.

[136] S.-Y. Wang, D. Bau, and J.-Y. Zhu, "Sketch your own GAN," in *International conference on computer vision*. IEEE, 2021, pp. 14 050–14 060.

[137] A. K. Bhunia, S. Khan, H. Cholakkal, R. M. Anwer, F. S. Khan, J. Laaksonen, and M. Felsberg, "Doodleformer: Creative sketch drawing with transformers," *arXiv preprint arXiv:2112.03258*, 2021.

[138] Á. Serrano, I. M. de Diego, C. Conde, E. Cabello, L. Shen, and L. Bai, "Influence of wavelet frequency and orientation in an SVM-based parallel gabor PCA face verification system," in *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, 2007, pp. 219–228.

[139] H. S. Bhatt, S. Bharadwaj, R. Singh, and M. Vatsa, "Memetically optimized MCWLD for matching sketches with digital face images," *Transactions on Information Forensics and Security*, vol. 7, no. 5, pp. 1522–1535, 2012.

[140] M. Minear and D. C. Park, "A lifespan database of adult facial stimuli," *Behavior research methods, instruments, & computers*, vol. 36, no. 4, pp. 630–633, 2004.

[141] J. Nishino, T. Kamyama, H. Shira, T. Odaka, and H. Ogura, "Linguistic knowledge acquisition system on facial caricature drawing system," in *International Fuzzy Systems*. IEEE, 1999, pp. 1591–1596.

[142] S. Iwashita, Y. Takeda, and T. Onisawa, "Expressive facial caricature drawing," in *International Fuzzy Systems*. IEEE, 1999, pp. 1597–1602.

[143] Y. Li and H. Kobatake, "Extraction of facial sketch image based on morphological processing," in *International Conference on Image Processing*. IEEE, 1997, pp. 316–319.

[144] M. Tominaga, S. Fukuoka, K. Murakami, and H. Koshimizu, "Facial caricaturing with motion caricaturing in PICASSO system," in *International Conference on Advanced Intelligent Mechatronics*. IEEE, 1997, p. 30.

[145] S. E. Brennan, "Caricature generator," Ph.D. dissertation, Massachusetts Institute of Technology, 1982.

[146] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "A comprehensive survey to face hallucination," *International journal of computer vision*, vol. 106, no. 1, pp. 9–30, 2014.

[147] A. V. Nefian and M. H. Hayes III, "Face recognition using an embedded hmm," in *Conference on Audio and Video-based Biometric Person Authentication*. IEEE, 1999.

[148] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[149] X. Tang and X. Wang, "Face photo recognition using sketch," in *International Conference on Image Processing*. IEEE, 2002, pp. I–I.

[150] X. Tang and X. Wang, "Face sketch synthesis and recognition," in *International conference on computer vision*. IEEE, 2003, pp. 687–694.

[151] ——, "Face sketch recognition," *IEEE Transactions on circuits and systems for video technology*, vol. 14, no. 1, pp. 50–57, 2004.

[152] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[153] S. Saxena and M. N. Teli, "Comparison and analysis of image-to-image generative adversarial networks: A survey," *arXiv preprint arXiv:2112.12625*, 2021.

[154] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Advances in neural information processing systems*. Curran Associates, Inc., 2014.

[155] M. Mirza and S. Osindero, "Conditional generative adversarial nets," in *Advances in neural information processing systems workshops*. Curran Associates, Inc., 2014.

[156] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer Assisted Intervention*. Springer, 2015, pp. 234–241.

[157] Y. Jing, Y. Yang, Z. Feng, J. Ye, Y. Yu, and M. Song, "Neural style transfer: A review," *IEEE Transactions on visualization and computer graphics*, vol. 26, no. 11, pp. 3365–3385, 2019.

[158] Y. Song, C. Yang, Y. Shen, P. Wang, Q. Huang, and C.-C. J. Kuo, "Spg-net: Segmentation prediction and guidance network for image inpainting," in *British Machine Vision Conference*. BMVA Press, 2018, p. 97.

[159] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[160] L. Wang, R.-F. Li, K. Wang, and J. Chen, "Feature representation for facial expression recognition based on facs and lbp," *International Journal of Automation and Computing*, vol. 11, no. 5, pp. 459–468, 2014.

[161] X. Zheng, Y. Guo, H. Huang, Y. Li, and R. He, "A survey of deep facial attribute analysis," *International journal of computer vision*, vol. 128, no. 8, pp. 2002–2034, 2020.

[162] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[163] R. Ranjan, V. M. Patel, and R. Chellappa, "Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 121–135, 2017.

[164] E. M. Hand and R. Chellappa, "Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification," in *AAAI Conference on Artificial Intelligence*. AAAI Press, 2017, pp. 4068–4074.

[165] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen, "Heterogeneous face attribute estimation: A deep multi-task learning approach," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 40, no. 11, pp. 2597–2609, 2017.

[166] Y. Jang, H. Gunes, and I. Patras, "Smilenet: registration-free smiling face detection in the wild," in *International conference on computer vision workshops*. IEEE, 2017, pp. 1581–1589.

[167] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *International conference on automatic face & gesture recognition*. IEEE, 2017, pp. 17–24.

[168] S. Li and W. Deng, "Deep facial expression recognition: A survey," *Transactions on affective computing*, 2020.

[169] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "Panda: Pose aligned networks for deep attribute modeling," in *Conference on computer vision and pattern recognition*. IEEE, 2014, pp. 1637–1644.

[170] M. Kan, S. Shan, H. Chang, and X. Chen, "Stacked progressive auto-encoders (spae) for face recognition across poses," in *Conference on computer vision and pattern recognition*. IEEE, 2014, pp. 1883–1890.

[171] Y. Wu, Z. Wang, and Q. Ji, "Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines," in *Conference on computer vision and pattern recognition*. IEEE, 2013, pp. 3452–3459.

[172] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Conference on computer vision and pattern recognition*. IEEE, 2017, pp. 1415–1424.

[173] U. Toseeb, D. R. Keeble, and E. J. Bryant, "The significance of hair for face recognition," *PloS one*, vol. 7, no. 3, p. e34144, 2012.

[174] S. J. Bartel, K. Toews, L. Gronhovd, and S. L. Prime, "Do i know you? altering hairstyle affects facial recognition," *Visual Cognition*, vol. 26, no. 3, pp. 149–155, 2018.

[175] N. Kumar, P. Belhumeur, and S. Nayar, "Facetracer: A search engine for large collections of images with faces," in *European conference on computer vision*. Springer, 2008, pp. 340–353.

[176] L. Huai-Yu, D. Wei-Ming, and B.-G. Hu, "Facial image attributes transformation via conditional recycle generative adversarial networks," *Journal of Computer Science and Technology*, vol. 33, no. 3, pp. 511–521, 2018.

[177] J.-S. Pierrard and T. Vetter, "Skin detail analysis for face recognition," in *Conference on computer vision and pattern recognition*.

IEEE, 2007, pp. 1–8.

[178] S. Z. Li, *Encyclopedia of Biometrics: I-Z*. Springer Science & Business Media, 2009, vol. 2.

[179] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[180] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conference on computer vision and pattern recognition*. IEEE, 2016, pp. 770–778.

[181] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Conference on computer vision and pattern recognition*. IEEE, 2018, pp. 8789–8797.

[182] B. Zhao, B. Chang, Z. Jie, and L. Sigal, "Modular generative adversarial networks," in *European conference on computer vision*. Springer, 2018, pp. 150–165.

[183] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Advances in neural information processing systems workshops*. Curran Associates, Inc., 2017.

[184] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*. OpenReview.net, 2014.

[185] Q. Yu, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, and C.-C. Loy, "Sketch me that shoe," in *Conference on computer vision and pattern recognition*. IEEE, 2016, pp. 799–807.

[186] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[187] Y. Wang, C. Wu, L. Herranz, J. van de Weijer, A. Gonzalez-Garcia, and B. Raducanu, "Transferring gans: generating images from limited data," in *ECCV*. Springer, 2018, pp. 218–234.

[188] y. wang, L. Yu, and J. van de Weijer, "Deepi2i: Enabling deep hierarchical image-to-image translation by transferring from gans," in *Advances in neural information processing systems*. Curran Associates, Inc., 2020, pp. 11 803–11 815.

[189] A. Shocher, Y. Gandelsman, I. Mosseri, M. Yarom, M. Irani, W. T. Freeman, and T. Dekel, "Semantic pyramid for image generation," in *International conference on computer vision*. IEEE, 2020, pp. 7457–7466.

[190] S. Ravi and H. Larochelle, "Optimization as a model for few-shot learning," in *International Conference on Learning Representations*. OpenReview.net, 2017.

[191] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[192] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? - weakly-supervised learning with convolutional neural networks," in *Conference on computer vision and pattern recognition*. IEEE, 2015, pp. 685–694.

[193] X. Wang, K. He, and A. Gupta, "Transitive invariance for self-supervised visual representation learning," in *International conference on computer vision*. IEEE, 2017, pp. 1329–1338.

[194] R. Pinto, T. Mettler, and M. Taisch, "Managing supplier delivery reliability risk under limited information: Foundations for a human-in-the-loop DSS," *Decision support systems*, vol. 54, no. 2, pp. 1076–1084, 2013.

[195] Y. LeCun *et al.*, "Generalization and network design strategies," *Connectionism*

in perspective, vol. 19, no. 143-155, p. 18, 1989.

[196] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit *et al.*, "Mlp-mixer: An all-mlp architecture for vision," in *Advances in neural information processing systems*. Curran Associates, Inc., 2021.

[197] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*. Curran Associates, Inc., 2017.

[198] K. Lee, H. Chang, L. Jiang, H. Zhang, Z. Tu, and C. Liu, "ViTGAN: Training GANs with vision transformers," in *International Conference on Learning Representations*. OpenReview.net, 2022.

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

[200] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *ACM Transactions on graphics*, vol. 26, no. 3, p. 10–es, 2007.

[201] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[202] Y. Hu, S. Yang, W. Yang, L.-Y. Duan, and J. Liu, "Towards coding for human and machine vision: A scalable image coding approach," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2020, pp. 1–6.

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