Now You See Me: Deep Face Hallucination for Unviewed Sketches

Now You See Me: Deep Face Hallucination for Unviewed Sketches

Conghui Hu[1]
c.hu@qmul.ac.uk
Da Li[1]
da.li@qmul.ac.uk
Yi-Zhe Song[1]
yizhe.song@qmul.ac.uk
Timothy M. Hospedales[1, 2]
t.hospedales@ed.ac.uk

[1] SketchX Research Lab
Queen Mary University of London
London, UK

[2] School of Informatics
University of Edinburgh
Edinburgh, UK

Abstract

Face hallucination has been well studied in the last decade because of its useful applications in law enforcement and entertainment. Promising results on the problem of sketch-photo face hallucination have been achieved with classic, and increasingly deep learning-based methods. However, synthesized photos still lack the crisp fidelity of real photos. More importantly, good results have primarily been demonstrated on very constrained datasets where the style variability is very low, and crucially the sketches are perfectly align-able traces of the ground-truth photos. However, realistic applications in entertainment or law enforcement require working with more unconstrained sketches drawn from memory or description, which are not rigidly align-able. In this paper, we develop a new deep learning approach to address these settings. Our image-image regression network is trained with a combination of content and adversarial losses to generate crisp photorealistic images, and it contains an integrated spatial transformer network to deal with non-rigid alignment between the domains. We evaluate face synthesis on classic constrained, as well as unviewed, benchmarks namely CUHK, MGDB, and FSMD. The results qualitatively and quantitatively outperform existing approaches.

1 Introduction

Face hallucination addresses inferring one face image from another image in a different condition. A particularly interesting variant is that of synthesising photos based on facial sketches, which has applications in entertainment and law enforcement [13]. In the past decade, this problem has been well studied, and promising results have been achieved using both patch-based [14, 15, 19] and, more recently, deep learning-based [5] approaches.

The main limitation of existing studies is that they primarily focus on simple "*viewed*-sketch" benchmarks – those in which sketches are basically perfect traces of an underlying ground-truth photo. This assumption that sketches and photos can easily be perfectly aligned means that models can be simple because all (super)pixels are in correspondence; and easy to train because without alignment ambiguity, the mapping is a simple colour texturing process.

The standard viewed-sketch databases are also very constrained, in that there is little variability in conditions such as background, sketch style, and even subject ethnicity (CUHK). However, neither of these assumptions hold in real law or entertainment applications of sketch-photo synthesis. Here, the sketches and photos are more unconstrained, and crucially artists are drawing from their imagination, or description. This means that the sketches are affected by communication and memory imperfections [3, 13] as well as the conventional sketch-photo modality gap. So photo hallucination is now a much more complicated mapping than simple colour texturing after rigid alignment. This can be seen in the results of the few studies that test on *unviewed* forensic sketches after training on viewed benchmarks: The quality of the synthesis results in the unviewed case is much worse [5, 14].

In this paper we develop a powerful deep learning-based method for sketch-photo face hallucination that produces more crisp images than prior work while addressing the less constrained unviewed setting, that is harder but more practically relevant. We build upon a fully convolutional image-image regression network [5] that can provide a rich non-linear mapping from sketches to photos. To make this mapping learnable, given the lack of a rigid alignment between photos and sketches in the unviewed case, we integrate a modified spatial transformer network (STN) [10] into the regressor. Our STN network inputs facial geometry defined by detected facial interest points, and non-rigidly warps the sketch and photo into alignment. To enable the synthesis of high fidelity crisp photos, we first extend the image-image regression network to include two branches that process individual patches and whole images respectively, thus encoding both fine grained details and holistic structure; and train the combined model with a Markovian adversarial loss similar to that used in [11].

In summary, our main contributions are: (i) A novel sketch-photo synthesis network that specially tackles misalignment using Spatial Transformer Networks, making it more practical for law enforcement and entertainment applications. (ii) A two-branch Siamese architecture that preserves local high frequency details while maintaining overall holistic structure. (iii) Qualitative and quantitative experiments on well-aligned datasets (CUHK), and those that exhibit heavy misalignment (MGDB and FSMD), demonstrate that the proposed deep network can synthesise more realistic and crisp photos compared with prior state-of-the-art.

2 Related Work

Sketch-photo Face Synthesis Sketch-photo face synthesis is now quite well studied [12, 18]. Existing studies can be categorised according to whether they use classic [14, 15, 19] or deep [5, 6] methods; and whether they process images holistically [5, 6] or patch-wise [14, 15, 19] (more common for deep and classic methods respectively).

Patch-based methods [14, 15, 19] process images by superpixel or regular grid patches. They commonly employ Markov networks to exploit connections between neighbouring image patches to improve coherence of synthesis. State-of-the-art holistic methods for image-image regression such as [5] leverage batch normalisation, residual block building, and perceptual losses to learn to predict photo from sketch images. Very recent unpublished work has further improved such deep pipelines by applying variational autoencoders [6].

We observed that both holistic and patch-based architectures have distinct shortcomings. The former typically lose some detailed information, i.e., the synthesised face may be somewhat blurry, and the latter focuses on local features without building a good entire face structure. To address these limitations, we build a two-branch generator, performing holistic and the patch-wise synthesis simultaneously, to generate faces with high fidelity.

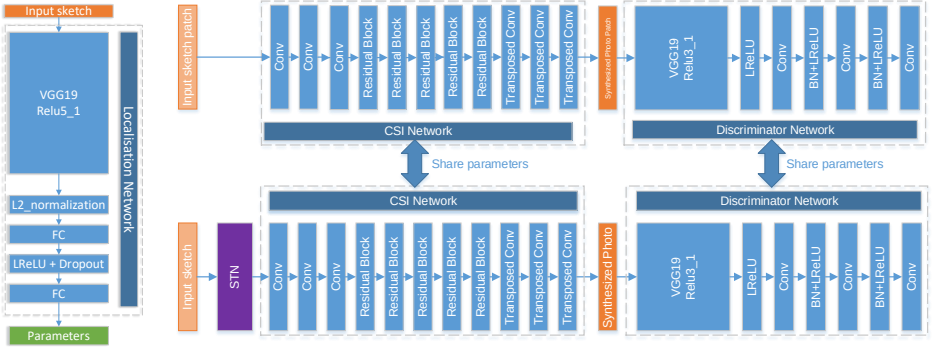

Figure 1: Architecture of the proposed network: STN subnetwork (left), overall (right).

Generative Adversarial Nets for Image Synthesis Since the Generative Adversarial Network (GAN) [4] was proposed, many extended GAN models have emerged for image synthesis. Relevant extensions include conditioning the generated image on another image, so as to achieve image-image translation. A persuasive example of this is [9], which showed it was possible to flexibly learn different mappings between the input images and the output images with a single loss function. Adversarial training was also exploited by [16] to synthesise more realistic images. The Markovian Generative Adversarial Network was proposed in [11] for texture synthesis. In our work, we leverage the flexibility of the Markovian discriminator, and adapt it to fit our two-branch generator. Our resulting model shows good performance in generating crisp images with both good high and low-frequency content.

Spatial Transformer Network Spatial Transformer Network (STN) [10] modules can help deep networks to achieve better spatial invariances to the input data by end-to-end learning to perform transformations including cropping, scaling, rotation, etc. A localization network learns to predict the appropriate transformation to be applied to any given input, which is then applied to transform the corresponding image into a more canonical form for easier subsequent processing. This technique has since been used in many applications including face pose alignment for improving detection [2] and recognition [20] with end-to-end alignment. Different from these previous applications that rigidly warp regularly structured grids, we employ the thin plate spline variant (STN-TPS) [1] to generate displacements for specific facial landmarks. By generating this displacement map, we learn end-to-end how to non-rigidly deform facial sketches so as to better align with canonical photos. This is helpful to improve photo synthesis based on unviewed sketches, which as we discussed do not have a simple rigid mapping onto a corresponding photo.

3 Methodology

Our overall approach is based on image-image regression, enhanced with adversarial training. The framework (illustrated in Fig 1) consists of two Siamese fully convolutional regression networks for patches and holistic images respectively. These are paired with Siamese discriminator networks for the adversarial training. And the holistic-image branch is enhanced with our STN-based alignment module. In the following two sections we describe each of these components, and then the losses used to train the network.

3.1 Network Components

The main network components are the generator, including spatial transformer and discriminator, described as follows:

Generator: The generator is a fully convolutional image-image regression network. Inspired by CSI [5], we follow their network architecture by using an encoder and an decoder architecture as our generator, and also leverage batch normalization [8] and residual blocks [7]. However, the photos synthesized by CSI are blurry and somewhat cartoon-like, due to lack of high frequency information. We address this both through a different training objective (see Sec 3.2) as well by introducing a two-branch architecture. Our generator's two Siamese branches are trained to predict patches and whole images respectively. The patch-wise branch leads the model to learn more high frequency structure (unlike CSI), while the whole-image branch keeps the holistic structure of the face.

Spatial Transformer: Localisation To address the non-rigidly misaligned data in realistic unviewed sketches, we extend the whole-image branch with the most powerful thin plate spline transformation (TPS) [1] variant of the spatial transformer network [10]. Specifically, we define the STN's localisation net by pruning the VGG-19 network [17] as shown in Fig. 1, and defining a regression network based on it. To make learning more stable, we insert a L_2 normalization after the output of the final convolutional layer. This regression network predicts the transformation parameters in the form of an interest-point displacement map, that will be used to displace facial interest points detected in the input image to generate a non-rigid warping of the input sketch.

Spatial Transformer: Warping The localization network (denoted *LOC*) inputs the raw sketch I_s and outputs $D_i = LOC(I_s)$ representing the displacements of $i = 1 \dots n$ facial landmarks $P_i = (x_i, y_i)$. After displacing the facial landmarks, the thin-plate spline transformation [1] is used to map the displaced locations to the original points on the raw input as below,

$$(C|c_0 \quad c_x \quad c_y)^T = \left[\begin{array}{c|c} K & P^* \\ \hline P^{*T} & \mathbf{0} \end{array}\right]^{-1} \left[\begin{array}{c|c} D & \mathbf{0} \end{array}\right] \tag{1}$$

$$f(x,y) = c_0 + c_x x + c_y y + \sum_{i=1}^{n} c_i U(|P_i - (x,y)|) \tag{2}$$

where the c_0, c_x, c_y and c_i are the coefficients deduced by the original landmarks P and the displaced landmarks D. P^* is the extension of P adding one column all with values one in shape $n \times 3$, and K is with shape $n \times n$, in which $k_{ij} = U(r_{ij})$ with $U(t) = t^2 \log(t^2)$ and $r_{ij} = |P_i - P_j|$. After this mapping, bilinear interpolation is applied to calculate the pixel values of the transformed input. Based on this transformation learning, raw sketches I_s are mapped to preprocessed sketches I_s^* which should be better aligned with the true photo face. These refined sketches are then passed into the generator network for synthesis as shown in Fig. 1. Training this STN end-to-end thus: (i) non-rigidly warps the train sketches and photos into alignment allowing a better cross-modal projection to be learned from roughly aligned unviewed training sketches, and (ii) learns a data-dependent alignment strategy which will also better align and thus improve the synthesis performance for unviewed testing sketches.

Discriminator: For adversarial training, we need to define a discriminator network. We employ the Markovian discriminator [11] that has been shown to work well on synthesising structured patterns and image style [9, 11]. It does so by modelling the whole image as a Markov random field, whereby pixels that are separated by more than a patch diameter are

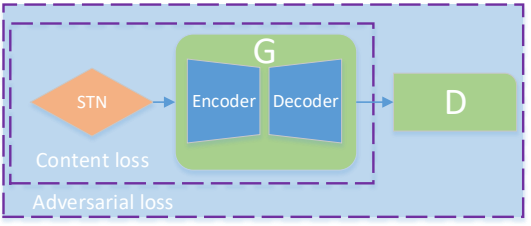

Figure 2: Optimization objectives for different modules. STN: the spatial transformer network, G: the generator, D: the discriminator.

considered statically independent. The network is trained to classify $N \times N$ image patches as real or fake. N can be any image size, and thus a single Siamese discriminator flexibly applies to both the whole image and patch-wise pathways in our two-branch model.

3.2 Optimization Objectives

Given the network architecture outlined above, we next explain the content and adversarial optimization objectives applied to the various sub-networks, as summarised in Fig. 2.

Content Losses There are three content-based losses. These include pixel loss l_p, feature loss l_f, and total variation loss l_{tv}. The pixel loss addresses low-level image similarity with standard $L2$ loss between predicted and photo pixels. The feature loss is defined between higher level features extracted from the synthesised and target photos. Specifically, assuming $\phi()$ is the ImageNet pre-trained VGG-19 [17] feature extractor ($ReLU2\_2$), we define:

$$l_f = \frac{1}{n}\sum_{i}^{n}(\phi(I_g^i) - \phi(I_r^i))^2 \tag{3}$$

Here n is the total number of the features and I_g and I_r are the generated and real images.

To make generated skin, hair etc. more realistic, we add the total variation loss to encourage the spatial smoothness in the generated image:

$$l_{tv} = \sum_{w,h}((I_g^{w+1,h} - I_g^{w,h})^2 + (I_g^{w,h+1} - I_g^{w,h})^2)^{1.25} \tag{4}$$

where the $I_g^{w,h}$ represents the pixel of the generated image or image patches.

Adversarial Loss The adversarial hinge loss l_h is defined as:

$$l_h = \frac{1}{n \times m}\sum_{i}^{n}\sum_{j}^{m}\max(0, \Delta + y^j \hat{y}^j) \tag{5}$$

where the n is the number of samples, the m is the total number of the final Markovian neural patches [11], Δ is the margin (set to 1 in the experiments). $\hat{y}^j$ is the output of the Siamese discriminator, representing the score for the jth Markovian neural patch patch/image. $y^j \in \{+1, -1\}$ is the ground truth label of the jth true or fake (generated) neural patches respectively.

Figure 3: Illustration of varying misalignment between sketch and photo in benchmarks: (a) CUHK, (b) MGDB (viewed sketch); (c) MGDB (unviewed sketch); and (d) FSMD.

Summary Overall we train the generator, discriminator, and spatial transformer parameters $(\theta_g, \theta_d, \theta_s)$ to minimise the weighted sum of these losses:

$$\underset{\theta_g,\theta_d,\theta_s}{\text{argmin}} \ \lambda_p L_p(I_g, I_r) + \lambda_f L_f(I_g, I_r) + \lambda_{tv} L_{tv}(I_g, I_r) + \lambda_h (L_h^G(I_g) + L_h^D(I_g, I_r)) \tag{6}$$

In this way we learn to align sketches, and synthesise photos which are indistinguishable from the true photos both holistically (image branch), and per-patch (patch branch). During training we optimise this loss for the predictions of both the patch and whole-image branches. During testing we make predictions based on the whole-image branch.

4 Experiments

4.1 Datasets

Datasets: We evaluate our method on three datasets: CUHK [19], MGDB[13] and FSMD [13]. CUHK includes 188 faces and viewed sketches of students in Chinese University of Hong Kong (CUHK). MGDB contains 100 sketch-photo pairs for sketches drawn under different conditions including viewed and unviewed. FSMD contains 196 pairs of mugshot photos and corresponding unviewed forensic sketches. As shown in Fig. 3 by overlaying the sketches and photos, CUHK is possible to align near perfectly. While sketch-photo pairs in MGDB and FSMD have very obvious residual misalignment after aligning based on eye position. This illustrates the more challenging non-rigid correspondence to be overcome in these more realistic datasets.

Preprocessing: Images in each dataset are normalized to 384×288 and aligned based on the position of center of two eyes. For patch-based processing, we follow the procedure in [11] to first rotate and scale the whole image and then crop 128×128 patches with stride 64.

Pipeline: The patch branch takes cropped sketch patches as well as their XY channels as input. The discriminator is then invoked to produce corresponding neural patch scores for both generated patch and real patch. For the holistic image branch, normalised sketches and manually labelled N facial landmarks are first fed into a STN. The localisation network then generates a $2N$-D vector, which represents facial landmark displacements in X and Y coordinates for each sketch. Afterwards, the TPS transformation module utilises the displacement vectors to wrap the original input sketches to mitigate misalignment. [1]

Settings: For our model, we set $\lambda_p = 1.0$, $\lambda_f = 0.5$, $\lambda_{tv} = 1e-4$, $\lambda_h = 1.0$. And we implement our framework in Tensorflow, using Adam optimizer, with batch size 4 for holistic-image branch and 64 for patch branch. For CUHK we use 35 facial interest points obtained from [19]. For experiments conducted on MGDB and FSMD, we manually labelled 36 facial landmarks, by removing those from CUHK that are not clearly exhibited in MGDB and

[1] Full implementation will be made available at the SketchX website: http://sketchx.eecs.qmul.ac.uk/downloads/

Table 1: Quantitative comparison against state of the art.

Benchmark	Method	PSNR	SSIM	R
CUHK	MrFSPS_SP [14]	-	0.633	-
	CSI [5]	17.295 ± 0.203	0.774 ± 0.006	0.921 ± 0.002
	Scribbler [16]	17.540 ± 0.175	0.785 ± 0.004	0.915 ± 0.003
	Proposed	**17.683** ± **0.167**	**0.791** ± **0.004**	**0.922** ± **0.002**
MGDB-Viewed	CSI [5]	**16.397** ± **0.421**	**0.699** ± **0.003**	0.695 ± 0.029
	Scribbler [16]	15.894 ± 0.375	0.660 ± 0.023	0.700 ± 0.036
	Proposed	16.067 ± 0.497	0.675 ± 0.028	**0.707** ± **0.032**
MGDB-Unviewed	CSI [5]	15.283 ± 0.297	**0.682** ± **0.028**	0.619 ± 0.041
	Scribbler [16]	15.241 ± 0.376	0.648 ± 0.014	0.637 ± 0.054
	Proposed	**15.340** ± **0.489**	0.663 ± 0.021	**0.646** ± **0.062**
FSMD	CSI [5]	**13.465** ± **0.384**	0.476 ± 0.013	0.461 ± 0.025
	Scribbler [16]	11.961 ± 0.247	0.371 ± 0.010	0.440 ± 0.020
	Proposed	13.328 ± 0.297	**0.505** ± **0.017**	**0.467** ± **0.021**

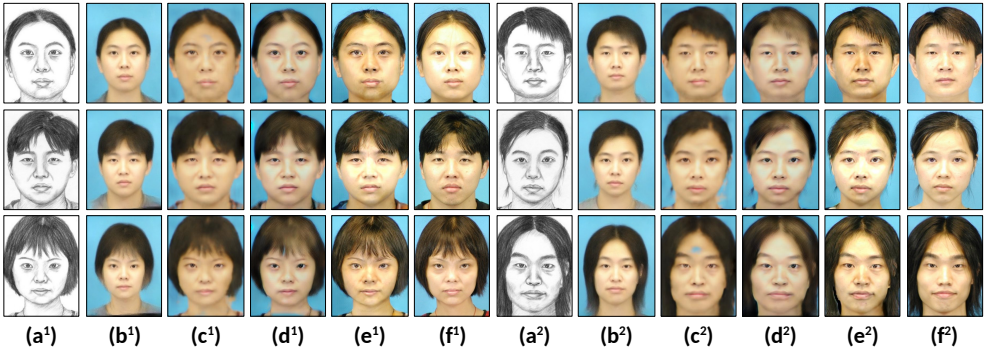

Figure 4: Qualitative comparison against state of the art on CUHK. (a) Input sketch (b) MrFSPS_SP [14] (c) CSI [5] (d) Scribbler [16] (e) Proposed method and (f) Ground-truth.

FSMD (e.g., in the hair region), and re-introducing a few in the inner face region to allow for more non-rigidity.

Pretraining: We follow the procedure of [5, 6] to define a pre-training strategy for our model. Specifically, we exploit the CelebA dataset [21] by automatically generating sketches via edge extraction, and using these as input to predict the corresponding photos. In this way we pre-train on CelebA before fine-tuning on the target datasets CUHK/MGDB/FSMD.

Baselines: We compare the performance of our full model (Two branch generator, Markov discriminator, Spatial transformer) with state-of-the-art competitors: **MrFSPS_SP** [14]: combines the multiple features from face images processed using multiple filters and exploits Markov networks of to model the relationships between neighbouring image patches. **CSI** [5]: A deep image-image regression network that leverages batch normalization, residual blocks, and three content losses for sketch inversion. **Scribbler** [16]: A deep adversarial network for general sketch-photo synthesis but not specifically designed for facial sketches (our best re-implementation based on the descriptions provided in the paper). CSI, Scribbler and our method are pre-trained on CelebA, but MrFSPS_SP is not.

4.2 Results

CUHK Benchmark: For the CUHK database, we follow the same train/test split in [19], taking 88 sketch-photo pairs for training the remaining 100 pairs for testing. This sparse

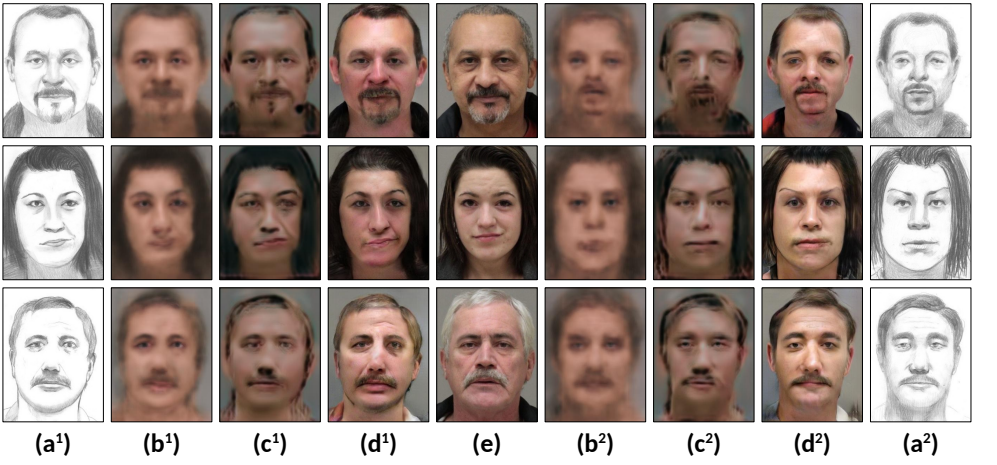

Figure 5: Qualitative results for synthesis in MGDB viewed (left) and unviewed (right). (a) Input sketch; (b) CSI [5] (c) Scribbler [16] (d) Proposed method and (e) Ground truth.

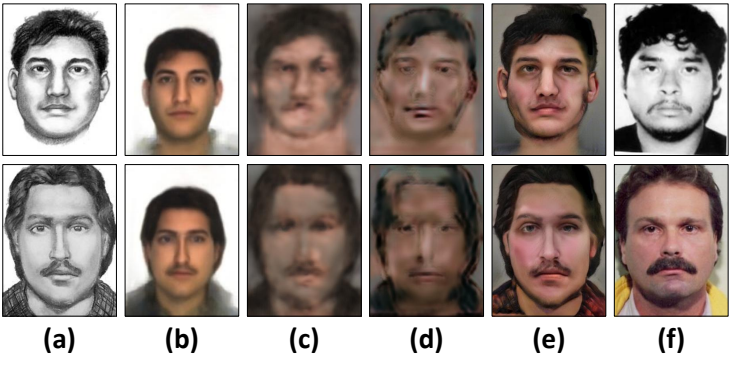

Figure 6: Qualitative results for synthesis in FSMD using model trained on MGDB: (a) Input sketch, (b) MrFSPS_SP [14] (c) CSI [5] (d) Scribbler [16] (e) Proposed method and (f) Ground truth.

training data approach is disadvantageous to our deep method, but we stick with this for direct comparison to previous work. The quality of our synthesized photos outperforms competitors qualitatively (Fig 4) and quantitatively in terms of three commonly used measures [5, 14]: peak signal to noise ratio (PSNR), structural similarity (SSIM) and Pearson product-moment correlation coefficient R (Table 1). With our two branch model, both the overall structure and the local details of the synthesized image are well preserved. In contrast, all alternatives produce less crisp images appearing blurry and lacking high frequency detail.

MGDB and FSMD Benchmarks: These images are more challenging than CUHK due to being more varied photos and sketches under more realistic uncontrolled conditions. We explore both viewed and unviewed images in MGDB, and true forensic sketches in FSMD. For the experiments on MGDB, we use 90 sketch-photo pairs for training and the rest 10 pairs to test. From the quantitative results in Table 1 we can see that as expected the quality

Table 2: Analysis on the contribution of each component of our method. TB: Two branch. MD: Markov Discriminator. STN: Spatial Transformer. PT: Pretraining.

Benchmark	Methods	PSNR	SSIM	R
MGDB-Viewed	TB+MD	15.772 ± 0.549	0.674 ± 0.027	0.694 ± 0.032
	TB+MD+STN	15.833 ± 0.633	0.669 ± 0.028	0.703 ± 0.025
	TB+MD+PT	15.873 ± 0.475	0.669 ± 0.032	0.704 ± 0.028
	TB+MD+STN+PT	**16.067 ± 0.497**	**0.675 ± 0.028**	**0.707 ± 0.032**
MGDB-Unviewed	TB+MD	15.288 ± 0.395	0.649 ± 0.021	0.641 ± 0.058
	TB+MD+STN	15.315 ± 0.356	0.656 ± 0.026	0.642 ± 0.040
	TB+MD+PT	15.300 ± 0.525	0.659 ± 0.020	0.641 ± 0.062
	TB+MD+STN+PT	**15.340 ± 0.489**	**0.663 ± 0.021**	**0.646 ± 0.062**

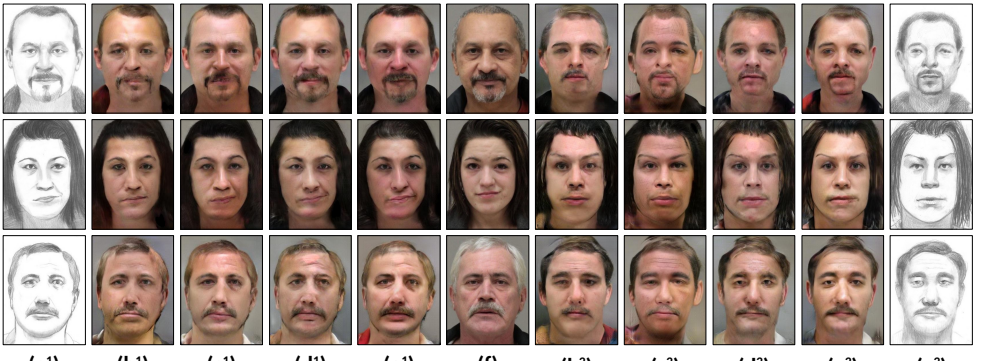

Figure 7: Ablative analysis on MGDB viewed (left) and unviewed (right). (a) Input sketch (b) TB+MD (c) TB+MD+STN (d) TB+MD+PT (e) TB+MD+STN+PT and (f) Ground-truth.

of the synthesised images is better for viewed than unviewed sketches, and both of these more challenging unconstrained sets are lower quality than easier CUHK sketch synthesis. Qualitative results for MGDB are shown in Figure 5 for viewed (left) and unviewed (right) sketches. For FSMD, we applied the model trained on MGDB to synthesise photos, mimicking the practical setting in law enforcement where one needs to synthesise a realistic photo from a forensic sketch to show to the public [3]. The corresponding qualitative results are shown in Figure 6. We can see that the synthesis quality is much lower across the board in this more challenging setting. However in each case our approach produces more crisp realistic results. We also quantitatively evaluated the synthesised faces using 147 sketch-photo pairs from FSMD as training data and the rest for testing. Results are summarised in Table 1, where our results compare favourably against state-of-the-art alternatives.

4.3 Further Analysis

Our full model contains multiple contributions (two-branch generator, Markov discriminator, spatial transformer network, CelebA pre-training), so we denote it TB+MD+STN+PT. To understand the contribution of each component, we perform an ablation study comparing four different settings: TB+MD with/without STN and with/without pre-training (PT). From the results in Fig. 7 and Table 2, we can see that adding STN improves performance qualitatively and is comparable to the pre-trained model without STN. Adding both STN and pre-training results in the best quality synthesis. At last, to offer insights as towards

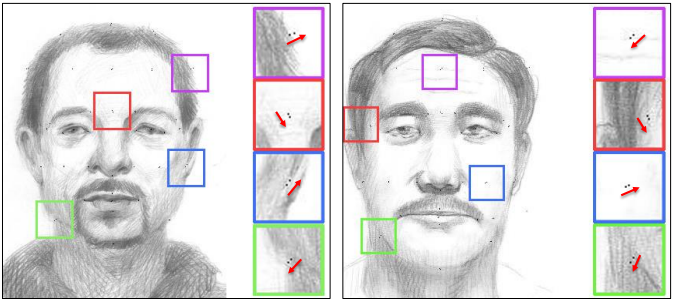

Figure 8: Visualization for the displacement transformation of the landmarks. Red arrow indicates direction of travel.

what STN had learned, Figure 8 offers illustrations for the predicted displacements of facial landmarks on two unviewed sketches.

5 Conclusion

We address the unviewed sketch-photo hallucination problem by proposing a new deep image-image regression approach. Our two-branch network models both local features and whole image structure. The non-rigid misalignment that occurs in unviewed or forensic sketches is dealt with by integrating a spatial transformer network into the generator, and cleaner images are synthesised by performing adversarial as well as content-based training. Overall the network produces higher quality photos than recent alternatives in both the conventional viewed sketch benchmarks, and the more challenging unviewed sketch setting.

References

[1] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE TPAMI*, 1989.

[2] Dong Chen, Gang Hua, Fang Wen, and Jian Sun. Supervised transformer network for efficient face detection. In *ECCV*, 2016.

[3] Charlie D. Frowd, William B. Erickson, James M. Lampinen, Faye C. Skelton, Alex H. McIntyre, and Peter J.B. Hancock. A decade of evolving composites: regression- and meta-analysis. *JFP*, 2015.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[5] Yağmur Güçlütürk, Umut Güçlü, Rob van Lier, and Marcel AJ van Gerven. Convolutional sketch inversion. In *ECCV*, 2016.

[6] Qi Guo, Ce Zhu, Zhiqiang Xia, Zhengtao Wang, and Yipeng Liu. Attribute-controlled face photo synthesis from simple line drawing. *arXiv preprint arXiv:1702.02805*, 2017.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[8] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015.

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[10] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *NIPS*, 2015.

[11] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *ECCV*, 2016.

[12] Shuxin Ouyang, Timothy Hospedales, Yi-Zhe Song, Xueming Li, Chen Change Loy, and Xiaogang Wang. A survey on heterogeneous face recognition: Sketch, infra-red, 3d and low-resolution. *IVC*, 2016.

[13] Shuxin Ouyang, Timothy M Hospedales, Yi-Zhe Song, and Xueming Li. Forgetmenot: memory-aware forensic facial sketch matching. In *CVPR*, 2016.

[14] Chunlei Peng, Xinbo Gao, Nannan Wang, Dacheng Tao, Xuelong Li, and Jie Li. Multiple representations-based face sketch–photo synthesis. *IEEE TNNLS*, 2016.

[15] Chunlei Peng, Xinbo Gao, Nannan Wang, and Jie Li. Superpixel-based face sketch–photo synthesis. *IEEE TCSVT*, 2017.

[16] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *CVPR*, 2017.

[17] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[18] Nannan Wang, Dacheng Tao, Xinbo Gao, Xuelong Li, and Jie Li. A comprehensive survey to face hallucination. *IJCV*, 106(1):9–30, 2014.

[19] Xiaogang Wang and Xiaoou Tang. Face photo-sketch synthesis and recognition. *IEEE TPAMI*, 2009.

[20] Yuanyi Zhong, Jiansheng Chen, and Bo Huang. Towards end-to-end face recognition through alignment learning. *CoRR*, abs/1701.07174, 2017.

[21] Xiaogang Wang Ziwei Liu, Ping Luo and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.