From Image to Video Face Inpainting: Spatial-Temporal Nested GAN (STN-GAN) for Usability Recovery

Yifan Wu, Vivek Singh, Ankur Kapoor

Siemens Healthineers, Digital Services, Digital Technology & Innovation,
Princeton, NJ, USA
{yifan.wu, vivek-singh, ankur.kapoor}@siemens-healthineers.com.

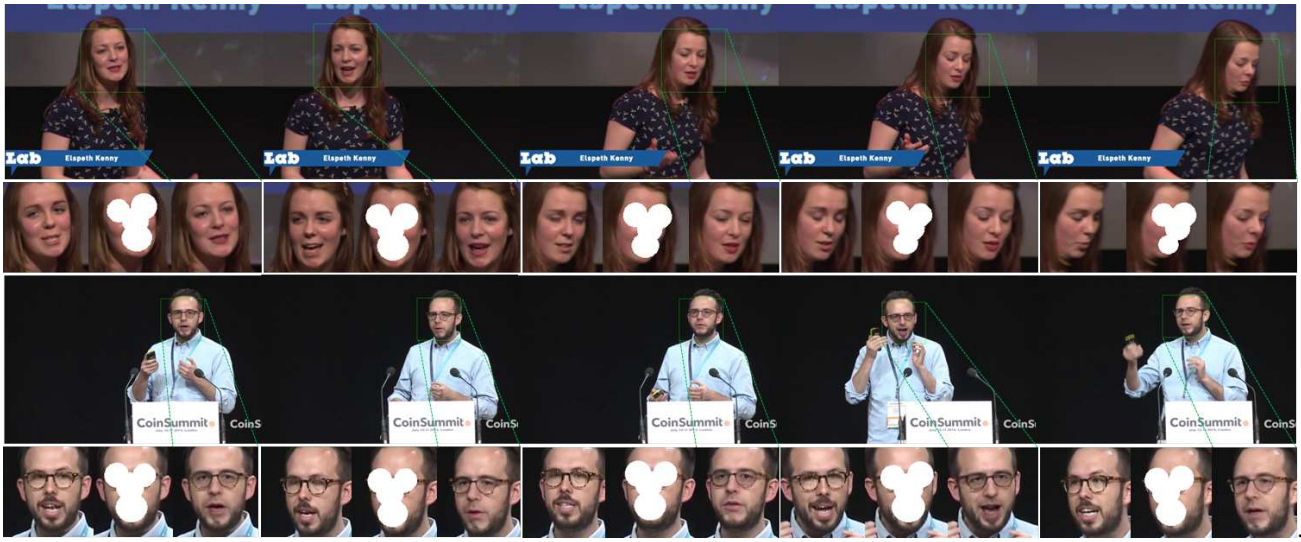

Figure 1: **Spatio-Temporal Face Inpainting.** For each sequence taken from 300-VW [7] test dataset, first row shows complete frames with faces inpainted by proposed STN-GAN from masked images, and the second row shows the corresponding groups of original, anonymized (input) and synthesized (fake) faces from left to right respectively.

Abstract

In this paper, we propose to use constrained inpainting methods to recover usability of corrupted images. Here we focus on the example of face images that are masked for privacy protection but complete images are required for further algorithm development. The task is tackled in a progressive manner: 1) the generated images should look realistic; 2) the generated images must satisfy spatial constraints, if available; 3) when applied to video data, temporal consistency should be retained. We first present a spatial inpainting framework to synthesize face images which can incorporate spatial constraints, provided as positions of facial markers and show that it outperforms state-of-the-art methods. Next, we propose Spatial-Temporal Nested GAN (STN-GAN) to adapt image inpainting framework, trained on ∼200k images, to video data by incorporating temporal information using residual blocks. Experiments on multiple public datasets show STN-GAN attains spatio-temporal consistency effectively and efficiently. Furthermore, we show that the spatial constraints can be perturbed to obtain different inpainted results from a single source. [1]

1. Introduction

Recently, image synthesis has experienced tremendous improvements following the introduction of deep generative models especially the generative adversarial networks [8], with sub-tasks like image style transfer [14], context-related image inpainting [12, 35, 18], and image super-resolution [4], which has significantly improved the possibility to inpaint partial or masked images for recovering usability of corrupted images. For example, patients' photo records are

[1] This feature is based on research and is not commercially available. Due to regulatory reasons, its future availability cannot be guaranteed.

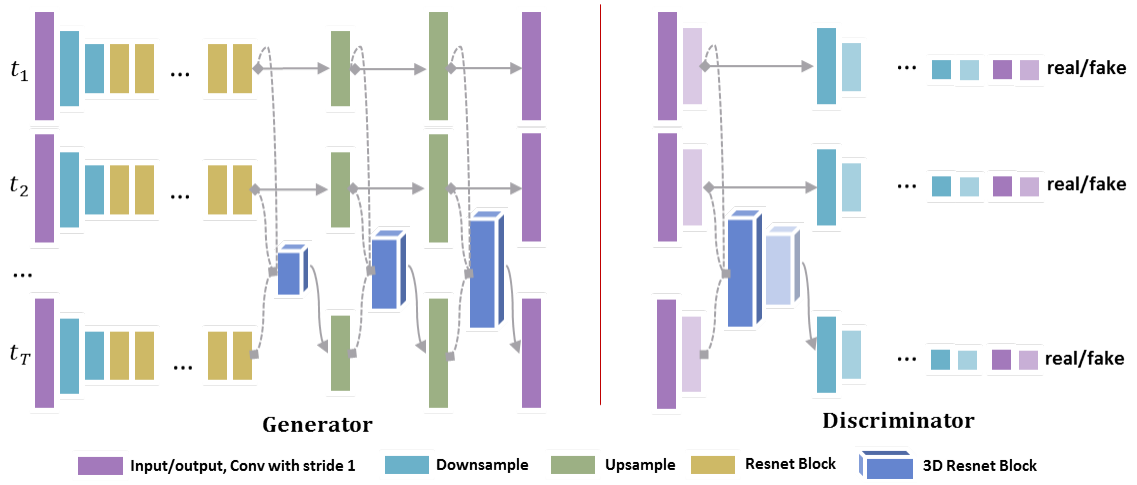

Figure 2: **Spatial-Temporal Nested GAN (STN-GAN)).** The left part shows the generator, the right part shows two discriminators: global discriminator is labeled as deep color and local discriminator is labeled as shallow color. The Generator and Discriminator for t_1 to t_T are shared.

not allowed to be released without anonymization due to the privacy protection regulation, while complete data is required for subsequent algorithm development, such as for training human pose estimation or activity recognition networks. This anonymization can be achieved by masking the biometric information [20] by covering the eyes, nose, and mouth as shown in Fig. 1. The current work addresses usability recovery of videos with such anonymized faces using STN-GAN, a novel video inpainting framework.

Recent state-of-the-art image inpainting methods [12, 18, 35, 36] aim to reconstruct corrupted images by optimizing L^1 or L^2 loss with respect to the original image. However, for anonymized face completion, such pixel reconstruction loss is not appropriate as there are many reasonable solutions beside the original image. Moreover, face inpainting must satisfy certain spatial constraints, for example, perspective consistency in pose. Furthermore, several state-of-the-art methods focus on synthesizing the whole face region rather than just the corrupted or masked areas. While this benefits the holistic context, it may lead to content and color inconsistency when stitching the face region back to the original, often much larger image (*c.f.* Fig. 1). From a usability perspective, such discrepancy is not tolerable as algorithms trained on such data may learn to exploit them as features.

When extending from image inpainting to video inpainting, the most important step is to ensure frame-to-frame consistency, i.e. the inpainted parts across frames should progress smoothly over time. Thus, a video face inpainting model should not only capture the spatial context within each frame but also the temporal context across frames. For the video face inpainting, this implies ensuring consistency both in inpainted facial structure such as eyes, nose, and facial attributes such as facial hair, eye glasses (e.g. people in speech, Fig. 1). Recent works on video inpainting [31, 32] train only using video datasets, which limits their performance due to lack of availability of large datasets. We conjecture that an effective solution should be able to learn to simultaneously enforce spatial and/or structural consistency using existing image datasets with large variations from huge amount of individuals with different attributes, and learn temporal consistency using just a few sequences.

As a solution to overcome aforementioned difficulties, we propose Spatial-Temporal Nested Generative Adversarial Network (STN-GAN), and demonstrate its ability to synthesize spatio-temporally consistent results from input videos where face is masked. In Fig. 2, we show an overview of the STN-GAN architecture. To begin with, we train a conditional generative network with perceptual loss for the single-frame inpainting. Our model takes masked images as input and learns to synthesize images that are visually and contextually consistent with the unmasked regions. When facial landmark information is available, it is used as additional input such that inpainted faces are consistent with the landmarks. This not only facilitates perspective consistency but also allows inpainting different face images with different facial markers from a single masked image. To learn temporal consistency without losing the spatial model, we link the decoding layers in the generator of the spatial model with 3D residual blocks [10], aim at learning correlation between the features in T-th frame and previous $T-1$ frames. We demonstrate the effectiveness of STN-GAN on multiple public datasets - CelebA [19], 300-VW [7] and FaceForensics [24], as well as the impact of

adding spatial and temporal constraints.

Our contributions can be summarized as:

• We propose STN-GAN, a novel generative framework that efficiently adapts models trained on image domain which usually has abundant data, to video domain where data is more expensive to acquire. By linking feature spaces using 3D residual blocks, the proposed STN-GAN learns temporal consistency effectively.

• We apply STN-GAN to inpainting masked faces in videos while strictly preserving the background. By using facial markers, different inpainting results can be obtained from one video. We demonstrate an application on recovering utility of images after face de-identification.

• We provide both qualitative and quantitative comparison on image and video face inpainting tasks, and outperform the state-of-art methods.

2. Related Work

Image inpainting. Image inpainting involves filling the missing regions of an image in a visually consistent manner. This task was traditionally approached with intensity flow inferences [3] and *PatchMatch* [33, 2] that fills the missing portions by sampling patches from surrounding image regions. They work well on texture scratches, but fail to recover large missing regions without a reference image.

Recent advances in deep generative learning, especially GANs, have enabled methods that can learn contextual as well as semantic features at different scales and thus synthesize more visually plausible images [21]. Li *et al*. introduced face parsing loss for face completion [17]. Iizuka *et al*. introduced an image inpainting network with discriminators of different scales [12]. However, these methods require post-processing steps like Poisson Blending [22]. Recent methods that do not need such post processing, such as Liu *et al*.'s partial convolution [18] and Yu *et al*.'s attention module to model long-range spatial dependency [35, 36], do not make use of adversarial or perceptual loss, and thus do not perform well when large semantic contents are missing.

Image generation with spatial constraints. Several recent works integrate spatial constraints into image generation task. Teixeira *et al*. proposed to generate synthetic X-ray from body surface while simultaneously predicting body markers, and in turn, use predicted markers to update X-ray [28]. Bulat *et al*. integrated facial landmark detection and face super-resolution [4]. Contrary to these methods that formulate it as multi-task learning to learn a strong correlation, we only treat spatial constraints as additional input to be used only as guidance during inpainting, since input landmarks may not be sufficiently accurate.

Video generation. Unlike image inpainting which has been under active research over recent years, there are not many references on video. Wang *et al*. proposed to jointly learn temporal-spatial structure for video inpainting [31], but masks are in a fixed shape and position across all frames, which does not hold true for face inpainting where subject is in motion. Recently, a general video to video synthesis was proposed [32]; the proposed method utilizes optical flow information across frames to ensure temporal consistency and would require a large video dataset to ensure robustness to fine grained face variations.

Face de-identification with usability preservation. With the increasing privacy concerns and the need to collect larger datasets for algorithm development, face de-identification has become increasingly important. Some existing works proposed to synthesize de-identified face while ensuring structure similarity [34, 27]. Ren *et al*. trains a network to anonymize faces such that anonymization has almost not impact on action detection task [23]. However, these methods do not preserve the background well and the generated faces have visual artifacts, which limits the usefulness of the inpainted face images.

3. Method

We present a pipeline to train a video inpainting network that takes anonymized face images as input and output corresponding inpainted images, while ensuring spatio-temporal consistency. The pipeline is designed in a progressive manner. First, we present spatial inpainting method, formulated as an Image-to-Image translation [13] problem, which ensures the inpainted image looks realistic while strictly preserving the background (non-masked) regions. Next, we add sparse spatial constraints, e.g., face landmarks, to guide the structure of inpainted face. And finally, we extend the model to address video inpainting.

3.1. Spatial Inpainting

Our goal is to train an inpainting network G that learns a mapping, $G : \hat{x} \rightarrow y$ to make the generated image y indistinguishable from real image (domain of x), where x is the original image, $\hat{x}$ is a corrupted version of x. To obtain stable training process, we adopt the adversarial loss with gradient penalty (WGAN-GP) [1, 9]:

$$\begin{aligned}\mathcal{L}_{\text{adv}}(G, D) = \mathbb{E}_{x\sim P_{\text{data}}(x)}[D(x)] - \mathbb{E}_{\hat{x}\sim P_{\text{data}}(\hat{x})}[G(\hat{x})] \\ -\lambda_{gp}\mathbb{E}_{\widetilde{x}\sim P_{\text{data}}(\widetilde{x})}[(\|\nabla_{\widetilde{x}} D(\widetilde{x})\| - 1)^2],\end{aligned} \tag{1}$$

where $\widetilde{x}$ is sampled uniformly among a straight line between a pair of real and generated images. We use $\lambda_{gp} = 10$ for all experiments. To facilitate learning of the facial structure, we leverage the perceptual loss [14] that computes the L^1 distance between features obtained using ImageNet-pretrained VGG-16 [26], on original image, x and output image, $G(\hat{x})$. The loss is defined as:

$$\mathcal{L}_{\text{pct}}(G) = \sum_{i=0}^{n} \|f_n(x) - f_n(G(\hat{x}))\|_1 , \tag{2}$$

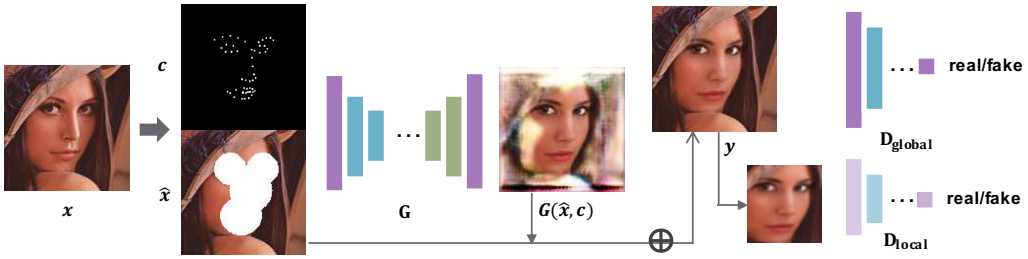

Figure 3: **Spatial inpainting framework.** Given a training sample, the face area is cropped and landmarks are detected. The masked image and facial markers are provided as input to the generator to synthesize the complete face image.

where f_n is the activation map of n^{th} selected layer, here we use pool1, pool2 and pool3 layers. Combining these losses into a single objective, the optimal G^* is obtained through the min-max procedure:

$$G^* = \arg\min_G \max_D \left(\mathcal{L}_{\text{adv}}(G, D) + \lambda \mathcal{L}_{\text{pct}}(G)\right). \quad (3)$$

where λ is the hyperparameters for multiple losses. We use $\lambda = 10$ in all of our experiments.

Background hard copy. The generator, G generates a full image instead of only the masked area. Ideally, the unmasked area in the generated image should be exactly the same as the original image. This can be represented as a hard constraint: $L^1(x \odot (1-m), y \odot (1-m)) = 0$, where m is a binary mask that takes 1 for image area to be inpainted, and 0 for the background. However, most existing methods [18, 35, 23] generate the whole images using the pixel reconstruction loss, which unavoidably causes a visual discrepancy due to color shift and/or texture in the background area. We propose to directly copy the non-masked area to ensure seamless integration. As shown in Fig. 3, we replace $y = G(\hat{x})$ by $y = \hat{x} \odot (1 - m) + G(\hat{x}) \odot m$, and calculate both $\mathcal{L}_{\text{adv}}$ and $\mathcal{L}_{\text{pct}}$ in Eq. 3.

Incorporating spatial constraints. After combining the GAN and perception losses, the generator could synthesize visually reasonable image but cannot control shape and pose of the synthesized objects in the masked area (e.g. face and nose in the face mask) because this part of information is typically extremely hard to infer from the background. Thus to allow for the flexibility of manipulating the hardly-inferable information, we add landmarks as the spatial condition into the generator as soft constraints that still provides enough variation. Our goal is then to learn $G : \{\hat{x}, c\} \rightarrow y$, here c is landmarks. As shown in Fig. 3, we concat the $\hat{x}$ and c as our final input.

Architecture. The generator network is composed of two convolution layers with stride size of two for downsampling, six residual blocks [10], and two transposed convolution layers with the stride size of two for upsampling (similar to StarGAN [6]). We use instance normalization [30] for the generator but no normalization for the discriminator. We leverage PatchGANs [13] for the discriminator network, which classifies whether local image patches are real or fake. Similar to [12, 35], we utilize two discriminators D_{global} and D_{local} on both global and local scales, and concatenate the outputs of the two discriminators for computing training loss. The global discriminator is composed of six convolution layers with the stride size of two while local discriminator is composed of five convolution layers.

3.2. Spatial-Temporal Inpainting

For the video inpainting, we learn a mapping that generates corresponding frames given an anonymized sequence of a temporal window, and retain the last frame as the output during testing. To capture the temporal consistency across frames, we insert 3D residual blocks before the last M upsampling layers in the generator and after the first N layers in the discriminator (See Fig. 2 for the complete architecture). Each 3D residual block takes the outputs of its previous layers in the T frames as its input, generates an output of the same shape, and feeds its T-th frame in the output into the succeeding layer in the generator. The complete infrastructure of STN-GAN then consists of T replicas of spatial inpainting network in Sec. 3.1 with shared weights, and multiple 3D residual blocks linking the first $T-1$ instances to the T-th frame in both generator and discriminator.

3D Residual Blocks. The architecture of linking multiple frames of generators with 3D Residual Blocks benefits both efficiency and consistency. For efficiency, as the Residual Block, $R(\cdot)$ outputs $x + R(x)$ for input x, we can make full use of the pre-trained model by setting the initial weights in $R(\cdot)$ as zeros, i.e., the spatial-temporal inpainting is equivalent to spatial inpainting frame by frame if 3D Residual Blocks are 0 weighted. This would ensure that even if $R(\cdot)$ learns nothing, it would not be worse than inpainting each frame separately. For ensuring consistency, we link the feature layers in generaotr thus the feature space in the previous $T-1$ frames are then visible for the last frame. For the T-th generator, taking more temporal in-

formation constrains its variety in the sense that it needs to keep consistency with the previous $T-1$ frames. In addition, we only project the output of Residual Blocks onto the T-th frame instead of all frames. Empirically, feeding it to all T frames could potentially improve the model performance slightly. However, this is not naturally reasonable as it reverse the time-line of video while our current setting, i.e., only retain the last frame in a given time window, is extendable to an online procedure.

Algorithm Description. In this section, we provide details on how the STN-GAN is constructed and trained. First we denote the input frames as $\hat{x}_t^T = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_T\}$, output frames as $y_t^T = \{y_1, y_2, \ldots, y_T\}$. Next, the generator of image inpainting is formularized as $G = G_{head} \circ G_1 \circ G_2 \circ \cdots \circ G_M \circ G_{tail}$, where $G_1, ..., G_M$ are the layers in the generator before where the 3D-Resnet blocks $R_1^G, \ldots, R_M^G$ are inserted, G_{head} are the layers before M layers, G_{tail} are the layers after M layers, and the circle symbol represents the function composition. Similarly, the discriminator is expanded as $D = D_{head} \circ D_1 \circ D_2 \circ \cdots \circ D_N \circ D_{tail}$. To clarify the updating procedure, we further express the feature map with $\hat{x}_t$ as a_t^G and a_t^D, final output as $\mathbf{a}_t^G$ and $\mathbf{a}_t$ for $t = 1 : T$ correspondingly. In the model training, we first initialize with pretrained model from image inpainting, and then fine-tune the temporal parameters, i.e. the weights in 3D-Resnet blocks. See Algorithm1 for detailed steps.

Algorithm 1: Training of STN-GAN

Input: Anonymous sequence with window size T: $\hat{x}_t^T = \{\hat{x}_1, \hat{x}_2 ..., \hat{x}_T\}$
Output: Synthesized sequence $y_t^T = \{y_1, y_2, ..., y_T\}$. During testing stage, we only retain the last frame y_T as final result in each time window.
Initialize the generator G and discriminator D with weights trained on image dataset only, initialize 3D-Resnet modules R^G and R^D with weight 0;
while *not converged* **do**
  for $t = 1 : T$ **do**
    Compute feature maps $a_t \leftarrow G_{head}(\hat{x}^t)$;
  end
  for $m = 1 : M$ **do**
    $a^T \leftarrow$ concat $\{a_1, a_2, ...a_T\}$ along time axis;
    $a^{T'} \leftarrow R_m^G(a^T)$;
    Decompose $a^{T'}$ to $\{a_{1'}, a_{2'}, ..., a_{T'}\}$ along time axis;
    Remain $a_1, a_2, ..., a_{T-1}$ unchanged, update $a_T \leftarrow a_{T'}$;
  end
  for $t = 1 : T$ **do**
    $\mathbf{a_t} \leftarrow G_{tail}(a_t)$;
  end
  Repeat the above operations for D;
  Compute $\mathcal{L}(G, D)$ regarding to eq. 3 w.r.t $\mathbf{a}_t^G$ and $\mathbf{a}_t^D$ for $t = 1 : T$, average loss along time axis ;
  Back propagate the error and update parameters of G, D, R^G, R^D;
end

4. Experiments

In this section, we first introduce the image and video datasets used in the experiments. Next, we show that the proposed spatial inpainting method performs better than state-of-the-art methods; we present several experiments studying the impact of background copy as well as incorporating landmarks as spatial constraints. We further evaluate our method on video inpainting task on multiple public datasets, both qualitatively and quantitatively. Lastly, we present quantitative experiments to show that even when proposed inpainting method utilizes landmarks as constraints, it does not recover the original face or look; we further demonstrate that perturbing the landmarks produces different inpainting results from a single masked image.

4.1. Datasets

CelebA. The CelebFaces Attributes (CelebA) dataset [19] contains 202,599 face images of celebrities. We first apply open source toolkit *dlib* [15] to detect face bounding box on initial 178×218 unaligned images and localize facial landmarks, followed by padding each face bounding box with ratio 0.2 and resizing to 128×128. In total, there are 197,036 faces detected. We use the first 98 identities as testing set, which contains 2353 images.

300-VW. The 300 Videos in the Wild (300-VW) dataset [7, 29, 25] contains 114 videos,with annotation of 68 landmarks for each frame. Designed for face landmarks detection task, the face images in 300-VW have large variations in pose, expression, illumination, background, occlusion, and image quality. Each video has duration around 1-2 minutes (at 25-30 fps). We use 70 for training, 10 for testing, and omit the remaining videos that are in low quality, black and white, or with part of faces out of view. Faces are cropped and resized in the same way as CelebA.

FaceForensics. FaceForensics [24] is a large-scale video dataset which contains 704 videos for training and 150 videos for testing. Unlike 300-VW dataset, actors are generally in the same position across the frames with only small head motions. For each video, we use the first 30 frames.

4.2. Spatial Inpainting

Experimental setting. For spatial inpainting, we trained our models only on CelebA dataset. To study the influence of landmarks in our model, we trained 2 models, *SI-lm* and

Table 1: **Spatial Inpainting testing results.** We compare 4 different algorithms: 1) Regression based Partial Convolution [18]; 2) ours without hardly copying background (*(SI-lm) w/o BGCP*), similar to other two GAN based methods with extra perceptual loss [12, 35], and ours proposed *SI* with and without landmarks.

CelebA dataset	PSNR	MS-SSIM	TV	FID
PartialConv[18]	27.38	0.979	12.82	41.23
ours w landmarks (SI-lm) w/o BGCP	27.30	0.979	**12.66**	37.10
ours w/o landmarks (SI)	27.06	0.976	12.78	38.52
ours w landmarks (SI-lm)	**28.61**	**0.982**	12.75	**36.21**
300-VW	PSNR	MS-SSIM	TV	FID
PartialConv[18]	27.26	0.970	9.40	13.92
ours w landmarks (SI-lm) w/o BGCP	26.29	0.971	9.48	13.66
ours w/o landmarks (SI)	27.17	0.970	9.34	12.86
ours w landmarks (SI-lm)	**28.67**	**0.976**	**9.29**	**11.13**

SI, with and without landmarks respectively.

All models are trained using Adam [16] with $\beta_1 = 0.5$ and $\beta_2 = 0.99$. For data augmentation, we flip the images horizontally with a probability of 0.5. To reduce the influence of errors in landmark detectors as well as avoid overfitting to the mask shape, we perturb the shape of the mask by randomly changing the radius of circular masks by 3 pixels. The batch size is set to 16 and the generator is updated once for every 5 discriminator updates; learning rate decay is set as in [6]. Training takes $\sim$15 hours for 200k iterations on a single NVIDIA TITAN X.

We test on both CelebA and 300-VW dataset. For 300-VW dataset, we sample frames from 70 video sequences at interval of 100, resulting in 828 test images. The identities in 300-VW are completely unseen during the training phase. We compare with PartialConv [18], a state-of-the-art solution for irregular hole inpainting. We use the same experiment setting, including data augmentation etc. and train for 500k iterations, which took $\sim$47 hours.

Quantitative results. As mentioned in [35], image generation tasks lack good quantitative evaluation metrics. Nevertheless, we report traditional measurements of image quality including peak-signal-to-noise ratio (PSNR), total variation (TV) loss, and Multi-scale structural similarity index (MS-SSIM). Additionally, we report Fréchet Inception Distance (FID) [11], a widely used metric for implicit generative models to measure similarity between two datasets of images, correlating well with human judgment of visual quality. Here we compute FID score on the features computed using Inception network.

The evaluation results are presented in Table. 1. The proposed spatial inpainting method, both with and without landmarks, significantly outperform the PartialConv on both datasets on FID score (Fig. 4). While PartialConv works well when mask is complex but narrow, it produces artifacts when large regions with semantic information is missing, as in the case in our experiments. To study the impact of background copy, we trained a network that generates the whole image (without background copy) using GAN based framework with perceptual loss, similar to [12, 35]. As can be observed in Fig. 5, synthesizing the whole image causes artifacts and inconsistencies in the background area, e.g., the skin and hair color.

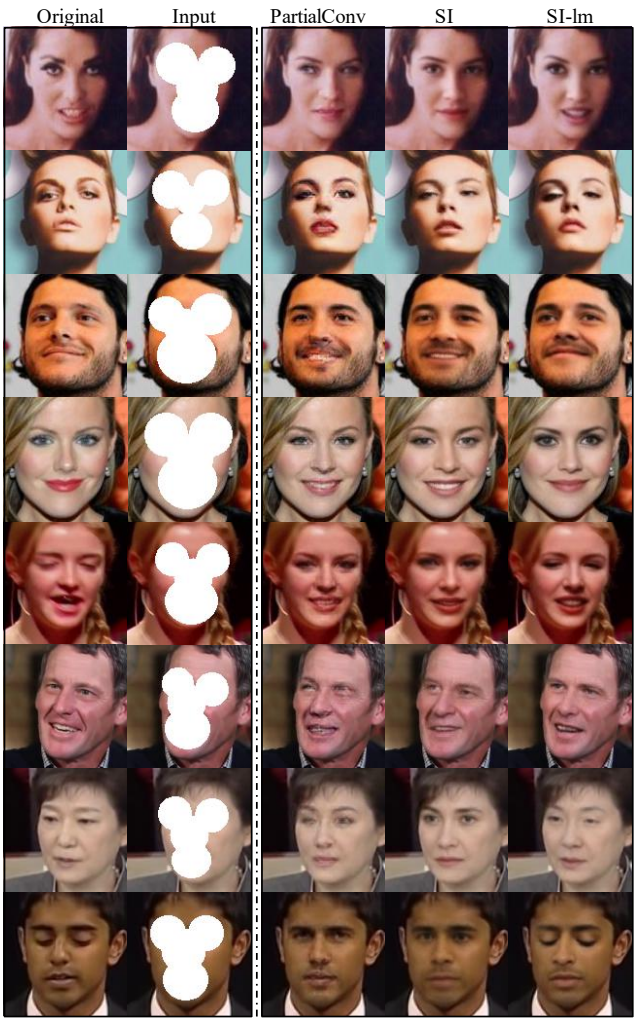

Figure 4: **Testing results of spatial inpainting trained on CelebA.** From left to right are original, anonymized (input), inpainted results of PartialConv [18], ours w/o landmarks (*SI*) and ours w landmarks (*SI-lm*). The first 4 rows are examples in CelebA, the bottom 4 rows are from 300-VW.

Qualitative results. Fig. 4 shows our spatial inpainting testing results. Although most inpainted images of our *SI* (the 4^{th} column) looks realistic without landmarks, by adding landmarks, not only the inpainted images are controllable and manipulable, e.g., the opened and closed months or eyes are retained in the row 1, 3, 5, 8, the eyes sizes are controlled in row 4, 7, but also significantly ben-

Figure 5: **Impact of background copy.** We present two testing examples with visual bias in the color space here, the left one is from 300-VW, the right one is from CelebA.

Datasets	SI	SI-lm	STN	STN-lm	Vid2Vid [32]
300-VW	8.89	6.95	7.14	**6.02**	39.61
Face Forensics	8.08	6.34	6.82	**5.28**	35.17

(a) FID Scores on 2 different testing sets.

Methods	Human Perference Score
SI / SI-lm / STN / STN-lm	1.47% / 1.96% / 12.25% / 84.31%

(b) Human preference Score on 300-VW testing set.

Table 2: **Spatial-Temporal inpainting evaluation**.

efit the visual quality, especially for the images with non-frontal poses, e.g., the proper months' angles in the row 6. Although the landmarks constraint is not regularized in the objective, it benefits the training and control the shapes and locations of eyes, nose, and mouth.

4.3. Spatial-Temporal Inpainting

Experimental Setting. For spatio-temporal inpainting, we set $T = 3$, i.e., each training sample contains a set of 3 consecutive frames. We report evaluation results on both 300-VW and FaceForensics datasets. From the 70 training and 10 videos in 300-VW, we have 126,273 training and 20,755 testing samples. For FaceForensics, we use the standard train-test split with 19,690 samples for training and 4,348 for testing. We fine-tune the pre-trained spatial models (Sec. 4.2) with a batch size of 12 and do not utilize any data augmentation. We train 2 separate models, with and without landmarks, referred as *STN* and *STN-lm* respectively. We only need to train for 5 epochs to achieve visually satisfying results. Training only takes $\sim$8 hours for 300-VW and $\sim$1.5 hours for FaceForensics on TITAN X.

Quantitative results. We compute the FID scores on both 300-VW and FaceForensics datasets. Table. 2a shows a comparison of STN-GAN with spatial inpainting methods *(SI, SI-lm)* as well as Vid2Vid [32], a general video to video synthesis approach. For Vid2Vid, we use the pre-trained model provided by the authors and compute the FID score only on cropped faces. Even on cropped faces, proposed methods significantly outperform Vid2Vid in terms of FID score. Furthermore, the proposed networks trains in matter of hours, compared to the Vid2Vid, a computationally expensive model, that takes several days to train [32].

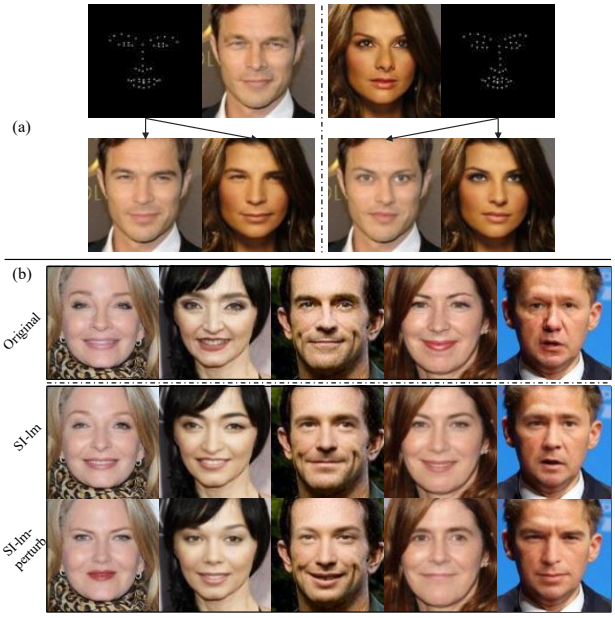

Figure 6: **Examples of perturbing landmarks.** (a) represents landmark switching results, the top row shows two original face images with detected landmarks on two sides, the bottom row represents corresponding inpainted images with original and switched landmarks. (b) shows landmark perturbation results, The top row shows the original face images, the second and third row show inpainted results of *SI-lm* with and without landmark perturbation.

We also performed a user-study for evaluating the visual quality of synthesized videos on 300-VW dataset. Each user was shown 4 synchronized videos, synthesized by proposed algorithms *(SI, SI-lm, STN, STN-lm)* and asked which one looks most realistic. We gathered responses from 20 different subjects, where each subject went over 10 videos of 20s each. We report the results of user study in Table. 2b. The proposed *STN-lm* consistently performs better than others.

Qualitative results. Sample test results obtained using different methods are shown in Fig. 8. Notice that the image based methods *(SI, SI-lm)* achieve visually consistent face inpainting results within each frame but fails to ensure a temporal consistency of the facial attributes, which is properly addressed using spatio-temporal methods *(STN, STN-lm)*. Furthermore, the use of landmarks as spatial constraints clearly improves the consistency of the landmarks with head pose. *STN-lm*, which incorporates all information, clearly produces the best overall results.

4.4. ID Distance Test and Landmark Perturbation

In this section we examine whether the proposed inpainting method can restore the original identity. To this end,

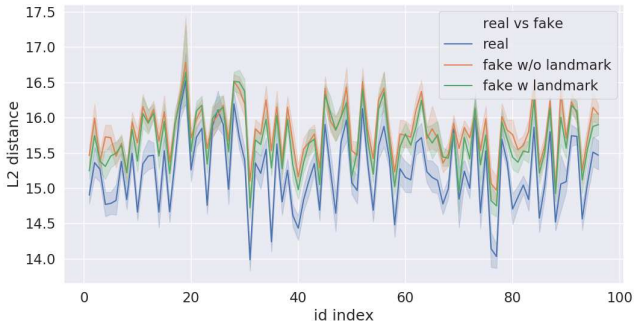

Figure 7: $L2$ **distance on VGGFace2 representation.** x-axis represents 98 IDs, y-axis represents ID distance (mean and std) between real / inpainted from *SI* / inpainted from *SI-lm* and remaining images that belong to the same person.

we propose to compare a pair of face images using the L^2 distance on VGGFace2 representation [5], a state-of-the-art face recognizer with ResNet-50 architecture [10]. For each (original) image of each subject, we compute the mean L^2 distance between the image and the remaining images of the same subject and compare it with the distance with the inpainted image (with and without landmarks). Fig.7 shows the mean score and deviation for all 98 identities (2353 images) in CelebA testing set. Note that inpainted images consistently hold larger distance, which confirms that the inpainted image is sufficiently different from the original regarding the identity.

Furthermore, the spatial constraints (landmarks) can be perturbed to obtain different inpainted results from a single source image. Fig. 6 (a) shows results by switching facial landmarks between 2 source images. Notice that inpainted face images for 2 subjects with different landmarks are clearly different. Fig. 6 (b) shows different inpainting results obtained from the masked original image by using the original landmarks as well as after a small perturbation. Notice the clear visual difference between the inpainted faces generated from different landmark configurations.

5. Conclusion

We presented the STN-GAN framework to approach for spatio-temporal inpainting and demonstrated its effectiveness to inpaint masked facial areas, often generated during privacy preserving or anonymization techniques. The proposed framework adapts image based solution, trained on datasets with thousands of individuals, to video based solution for which only tens of videos are available. We demonstrate the effectiveness of the approach on 2 public benchmark datasets. We further go on to show that the proposed framework can also be used for parameterized inpainting, by adjusting the facial markers.

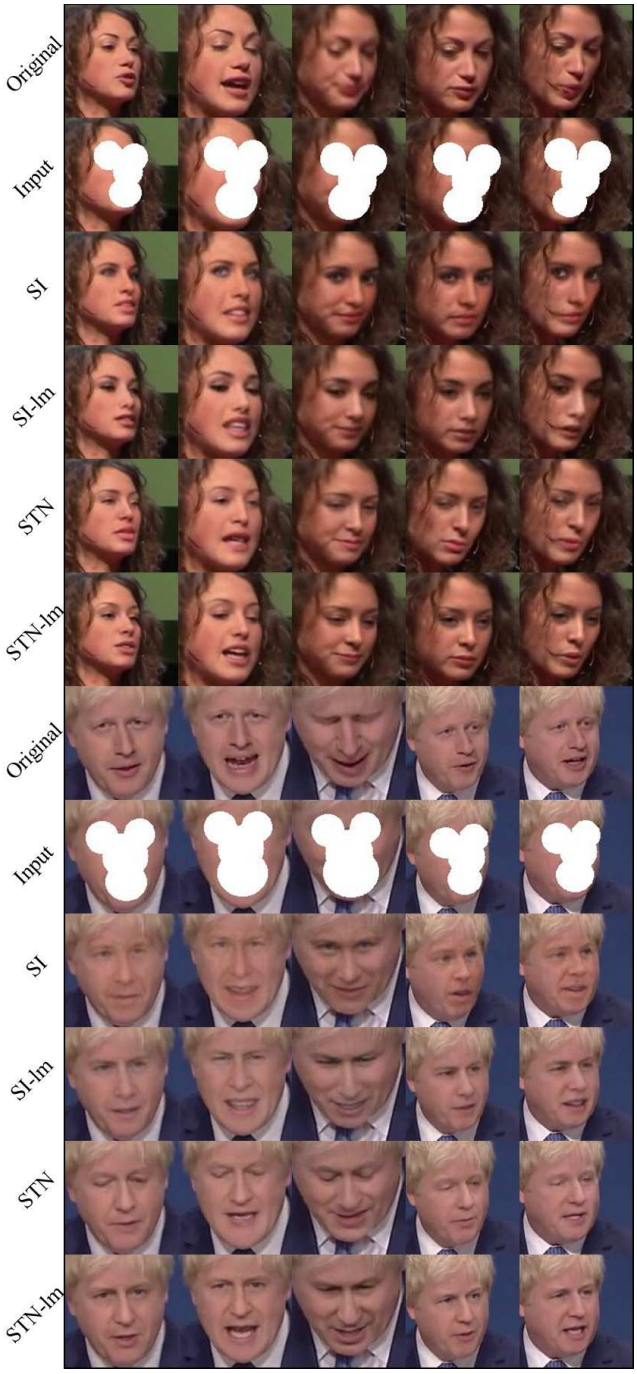

Figure 8: **Testing results of spatial-temporal inpainting on 300-VW.** For each example shows here, from top to bottom are original (real), anonymous (input), results (fake) of *SI*, *SI-lm*, *STN*, *STN-lm* respectively. From left to right is temporal continuously images with sampling interval 10 of original 25fps videos.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, 2017.

[2] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (ToG)*, 28(3):24, 2009.

[3] M. Bertalmio, A. L. Bertozzi, and G. Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Computer Vision and Pattern Recognition (CVPR)*, 2001.

[4] A. Bulat and G. Tzimiropoulos. Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[5] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition*, pages 67–74. IEEE, 2018.

[6] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] G. G. Chrysos, E. Antonakos, S. Zafeiriou, and P. Snape. Offline deformable face tracking in arbitrary videos. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (CVPR-W)*, 2015.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems (NIPS)*, pages 2672–2680, 2014.

[9] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016.

[11] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[12] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (TOG)*, 36(4):107, 2017.

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[14] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, 2016.

[15] D. E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(Jul):1755–1758, 2009.

[16] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Y. Li, S. Liu, J. Yang, and M.-H. Yang. Generative face completion. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 3, 2017.

[18] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. In *European Conference on Computer Vision (ECCV)*, 2018.

[19] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[20] E. M. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005.

[21] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[22] P. Pérez, M. Gangnet, and A. Blake. Poisson image editing. *ACM Transactions on graphics (TOG)*, 22(3):313–318, 2003.

[23] Z. Ren, Y. J. Lee, and M. S. Ryoo. Learning to anonymize faces for privacy preserving action detection. In *European Conference on Computer Vision (ECCV)*, 2018.

[24] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv*, 2018.

[25] J. Shen, S. Zafeiriou, G. G. Chrysos, J. Kossaifi, G. Tzimiropoulos, and M. Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2015.

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

[27] Q. Sun, L. Ma, S. Joon Oh, L. Van Gool, B. Schiele, and M. Fritz. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[28] B. Teixeira, V. Singh, T. Chen, K. Ma, B. Tamersoy, Y. Wu, E. Balashova, and D. Comaniciu. Generating synthetic x-ray images of a person from the surface geometry. *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2018.

[29] G. Tzimiropoulos. Project-out cascaded regression with an application to face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[30] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *CoRR*, abs/1607.08022, 2016.

[31] C. Wang, H. Huang, X. Han, and J. Wang. Video inpainting by jointly learning temporal structure and spatial details. *arXiv preprint arXiv:1806.08482*, 2018.

[32] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, and B. Catanzaro. Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NIPS)*, 2018.

[33] Y. Wexler, E. Shechtman, and M. Irani. Space-time completion of video. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (3):463–476, 2007.

[34] Y. Wu, F. Yang, and H. Ling. Privacy-protective-gan for face de-identification. *arXiv preprint arXiv:1806.08906*, 2018.

[35] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[36] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019.