Improved Techniques for GAN based Facial Inpainting

Avisek Lahiri*, Arnav Jain*, Divyasri Nadendla and Prabir Kumar Biswas, *Senior Member, IEEE*

Abstract—In this paper we present several architectural and optimization recipes for generative adversarial network(GAN) based facial semantic inpainting. Current benchmark models are susceptible to initial solutions of non-convex optimization criterion of GAN based inpainting. We present an end-to-end trainable parametric network to deterministically start from good initial solutions leading to more photo realistic reconstructions with significant optimization speed up. For the first time, we show how to efficiently extend GAN based single image inpainter models to sequences by a)learning to initialize a temporal window of solutions with a recurrent neural network and b)imposing a temporal smoothness loss(during iterative optimization) to respect the redundancy in temporal dimension of a sequence. We conduct comprehensive empirical evaluations on CelebA images and pseudo sequences followed by real life videos of VidTIMIT dataset. The proposed method significantly outperforms current GAN based state-of-the-art in terms of reconstruction quality with a simultaneous speedup of over 15×. We also show that our proposed model is better in preserving facial identity in a sequence even without explicitly using any face recognition module during training.

✦

1 INTRODUCTION

SEMANTIC inpainting is a challenging task of recovering large corrupted areas of an object based on higher level image semantics. Classical inpainting methods [1], [2], [3], [4], [5] rely on low level cues to find best matching patches from the uncorrupted sections of the same image. However, such '*copy-paste*' policy works well for background completions(sky, grass, mountains). However, the task of completing a complex object such as human face is far more challenging because the assumption of finding similar appearance patches does not always hold true. A facial image comprises of numerous unique components, which if damaged, cannot be matched with any other facial parts. An alternative is to use external reference datasets [6]. Though this paradigm enables to find similar matching patches, the low level [1] and mid level [2] features of matched patches are not sufficient to infer valid semantics of the missing regions.

Recently Yeh *et al.* [7] leveraged the recent advancement in generative modeling with Generative Adversarial Networks(GAN) [8]. Here, a trained neural network, often termed as the 'Generator', is trained to generate semantically realistic faces starting from a latent vector drawn from a known prior distribution. [7] is the current benchmark for semantic inpainting of faces. It outperforms Context Encoders [9] which was primarily designed for feature learning with inpainting. In this paper, we consider the model of Yeh *et al.* as baseline model and incorporate several architectural and optimization novelties for improving inpainting quality, optimization speed and adapting to inpaint sequences. Our application area is face inpainting. Specifically, our contributions can be summarized as follows:

- We show, for a single image inpaintng, initializing a GAN based iterative non convex optimization criterion(Eq. 2) with a learned parametric neural network(Sec. 4.1), results in more photo realistic initial reconstructions(Fig. 1a) compared to state-of-the-art GAN based single image inpainter model with random initialization.
- To our best knowledge this is the first demonstration of extending single image GAN based inpainter for sequences. For this, we design a recurrent neural network architecture(Sec. 4.2.1) for jointly initializing solutions for a group of frames. This design choice learns the scene dynamics leading to temporally more consistent initial solutions.
- In a sequence, we exploit redundancy of temporal dimension with a smoothness loss(Sec. 4.2.2) which constraints the final joint iterative solutions of a group of neighboring frames to lie close to each other in Euclidean space. The smoothness loss is not only better in enforcing temporal consistency(Sec. 5.2.2) but is also apter in preserving the facial identity(Sec. 5.2.4) of the subject compared to the baseline version.
- We present comprehensive empirical evaluations on CelebA images and pseudo sequences followed by real life facial videos from VidTIMIT dataset. In all cases, our proposed model outperforms the current benchmark baseline significantly in terms of visual reconstruction quality with an average speedup of over 15×.

2 BACKGROUND ON GANS

Proposed by Goodfellow *et al.* [8], a GAN model consists of two parametrized deep neural nets, viz., generator, G_{θ_G}, and discriminator, D_{θ_D}. The task of generator is to yield an image, $x \in \mathcal{R}^{H \times W \times 3}$ with a latent vector, $z \in \mathcal{R}^d$, as input. z is sampled from a known distribution, $p_z(z)$. A common choice [8] is, $z \sim \mathcal{U}[-1, 1]^d$. The discriminator is pitted against the

- ** Denotes equal contribution. A.Lahiri and P.K. Biswas are with Dept. of E&ECE, Indian Institute of Technology, Kharagpur, India.*
  All correspondence to A. Lahiri . E-mail: avisek@ece.iitkgp.ernet.in
- *A.Jain is currently at Microsoft. Work done while graduating from IIT Kharagpur.*
- *D.Nandendla is currently at Qualcomm. Work done while graduating from IIT Kharagpur.*

generator to distinguish real samples(sampled from p_{data}) from fake/generated samples. Specifically, discriminator and generator play the following game on $V(D_{\theta_D}, G_{\theta_G})$:

$$\min_{G_{\theta_G}} \max_{D_{\theta_D}} V(D_{\theta_D}, G_{\theta_G}) = \mathbb{E}_{x \sim p_{data}(x)}[\log D_{\theta_D}(x)]$$

$$+ \ \mathbb{E}_{z \sim p_z(z)}[1 - D_{\theta_d}(G_{\theta_G}(z))]. \qquad (1)$$

With enough capacity, on convergence, G_{θ_G} fools D_{θ_D} at random [8].

3 APPROACH

3.1 GAN based semantic inpainting

We begin by reviewing the current state-of-the-art GAN based single image inpainting model of Yeh *et al.* [7], which serves as our reference baseline model. Given a damaged image, I_d, and a pre-trained GAN model, the idea is to iteratively find the 'closest' z vector(starting from $\mathcal{U}[-1, 1]^d$) which results in a reconstructed image whose semantics are similar to corrupted image. z is optimized as, 找到那个z

$$\hat{z} = \underset{z}{\mathrm{argmin}} \ \{L_{con}(z|I_d, M) + \eta L_{per}(z)\}. \qquad (2)$$

$L_{con}(\cdot)$ is contextual loss which penalizes mismatch between original and reconstructed images over the non corrupted pixels.

$$L_{con}(z|I^d, M) = ||M \odot G_{\theta_G}(z) \ - I_d||_1 , \qquad (3)$$

where $\odot$ is the Hadamard product operator. $M(x, y) = 1$ for uncorrupted pixels and 0 otherwise. η is a trade off between the two components of the loss. $L_{per}(\cdot)$ is the perceptual loss and it a measure of realism of the inpainted output. The pre-trained discriminator is leveraged for assigning this realism loss and is defined as,

$$L_{per}(z) = \log[1 - D_{\theta_D}(G_{\theta_G}(z))] \qquad (4)$$

Since, $D_{\theta_D}(\cdot)$ gives the probability of being sampled from real images, Eq. 4 drives the solution of Eq. 2 to lie near to natural image manifold. Upon convergence, the inpainted image, $\hat{I}$, is given as, $\hat{I} = M \odot I_d + (1 - M) \ \odot G_{\theta_G}(\hat{z})$. Architectures of G_{θ_G} and D_{θ_D} are provided in supplemental document.

4 SINGLE IMAGE INPAINTING

4.1 Initializing z vector for single image inpainting

The iterative optimization procedure of Yeh *et. al* [7] in Eq. 2 yields different results based on the random initialization of z; this is mainly attributed to the non convexity of the optimization space. Also, such random initialization results in longer iterations of convergence(Sec. 5.1.2) compared to a good initialization of z. The above mentioned problems can be mitigated if we learn to estimate a good z vector directly from damaged image, I_d, by feed forward mapping through a deep neural net P_{θ_z}. The parameter set, θ_z, is optimized to minimize some distance metric, $L(\cdot)$:

$$\theta_z^* = \arg\min_{\theta_z} \sum_{i=1}^{N} L(G_{\theta_G}(P_{\theta_z}(I_d^i)), I_d^i), \qquad (5)$$

where I_d^i is the ith corrupted image in dataset. Though Eq. 2 and 5 are functionally same, prediction using a learned parametric network tends to perform better than ad hoc iterative optimization. This is because, with evolution of training, the network learns to adapt parameters to map images with closely matching appearances to similar z vectors. Parameter update for a given image thus implicitly generalizes to images with similar characteristics. We formulate the loss function, $L(\cdot)$, as,

$$L(\cdot) = \frac{1}{N} \sum_{i=1}^{N} ||(I_u^i - G_{\theta_G}(P_{\theta_z}(I_d^i))||_2^2 + \lambda \log(1 - D_{\theta_d}(G_{\theta_G}(P_{\theta_z}(I_d^i)))).$$

$$(6)$$

The first component of the loss is a mean squared error(MSE) between original and inpainted images. The second component is the same as perceptual loss as defined in Eq. 4. The MSE loss helps in recovering the global low frequency components of an image while the perceptual loss helps to refine it further with incorporation of detailed high frequency texture components. The parameter, λ, strikes a balance between the two components of the loss.

4.2 Extending to series of frames

4.2.1 Initialization with a recurrent model

The naive approach of applying the formulation of [7] on sequences is to inpaint individual frames independently. However, such approach fails to leverage the temporal redundancy inherent in any sequence. In this regard, for sequences, we propose to use a Recurrent Neural Network (RNN) to jointly initialize z vectors for an entire group of frames. RNN consist of a hidden state h_t to summarize information observed upto that time step. The hidden state is updated after looking at the previous hidden state and the corrupted image, leading to more consistent reconstructions in terms of appearance.

Since, RNNs suffer from vanishing gradients problem [10] and are unable to capture long dependencies, we use Long Short Term Memory (LSTM) [11] Networks. LSTMs have produced state-of-the-art results in sequential tasks like machine translation [12], [13] and sequence generation [14], [15]. In Fig. 3, we show the LSTM based network architecture for initializing a given group of frames. Let, $V = \{I_d^1, I_d^2, ..., I_d^W\}$ be a sequence of total W corrupted successive frames. Similar to [16], each frame is passed through a weight shared CNN descriptor module. Here our CNN's architecture is same as that of P_{θ_z}. Each damaged frame, I_d^k is represented by z_d^k. The latent vector z_d^k is passed as input to LSTM module at time step k and the hidden states h^k and cell memory c^k of LSTM are updated. The hidden state h^k is used to obtain the initial z_{init}^k vector which is passed through the pre-trained(and frozen parameters) generator, G_{θ_G}, to output the initial reconstructed image, I_{init}^k. MSE loss between original image, I_u^k, and I_{init}^k is minimized w.r.t all the parameters of LSTM and the CNN descriptor network. Further details are provided in supplemental document.

4.2.2 Temporal smoothness loss(l_{sm})

The initialization method using the above mentioned recurrent model ensures that the initial solutions respect the smooth transition of scene dynamics. However, following

	64X64		128X128	
	Approach		Approach	
Mask	Yeh *et. al*	**Proposed**	Yeh *et. al*	**Proposed**
Left	20.39	24.34	16.19	22.14
Right	21.41	24.72	15.51	23.12
Top	19.72	24.03	17.07	22.83
Bottom	21.92	24.19	16.41	22.69
Central	24.32	26.32	20.22	24.12
Freehand	25.46	28.91	22.46	26.91

(a) (b)

Fig. 1: Benefit of initializing Eq. 2 with proposed learned parametric network, P_{θ_z} **(a:)** Visualization of initial solutions of Eq. 2. Row 1: original images; Row 2: corrupted images; Row 3: Initial solutions using our **proposed network**, P_{θ_z}; Row 4: Initial solutions using Yeh *et al.* [7]. Proposed outputs are more photo realistic compared to [7]. **(b:)** Average PSNR after convergence of iterative optimization. Left, right, top, bottom masks damage the respective 50% of frame. Central mask damages central 50% and freehand masks damages approximately 50% of frame with freehand drawn masks.

Fig. 2: Final inpainted outputs after convergence of Eq. 2. Top Row: 64×64. Bottom Row: 128×128. For each triplet, Left: masked image, Middle: Inpainting by Yeh *et al.* [7], Right: **Proposed** inpainted output. Proposed outputs are more photo realistic. [7] specifically suffers at 128×128 resolution. More examples are provided in supplementary document.
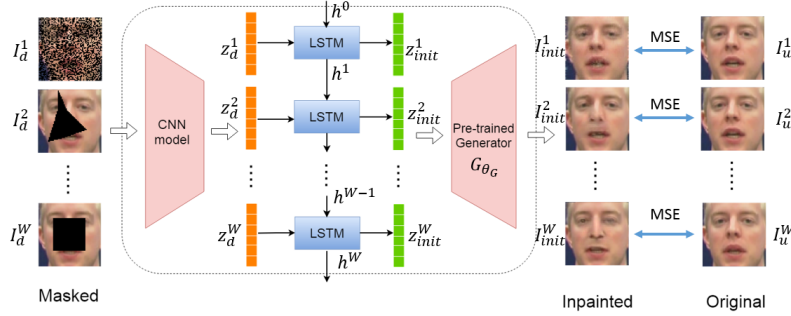

Fig. 3: Proposed LSTM based joint initialization of z vectors for a group of frames. See Sec. 4.2.1 for details of architecture.

the initialization, if we independently optimize for each frame, then the final solutions become unconstrained and manifest abrupt changes of facial appearance/expressions. To mitigate this, the idea is to jointly optimize a window of W frames to encourage the final reconstructions to respect the smooth appearance transitions. Disparity between two inpainted images, $\hat{I}^i$ and $\hat{I}^j$ can be approximated by Euclidean distance between their latent vectors $(\hat{z}^i, \hat{z}^j)$ [17]. With this approximation, we define

$$l_{sm} = \frac{1}{\binom{W}{2}} \sum_{\forall (i,j) \in W} ||\hat{z}^i - \hat{z}^j||^2. \qquad (7)$$

It can be seen as a summation of distance loss between all possible pairwise combinations of z vectors of the inpainted frames within the window of W frames. In Sec. 5.2.2 we shall see the importance of temporal smoothness loss in yielding a more consistent set of frames(along temporal dimension)

compared to the straight forward per frame application of Yeh *et al.* [7].

5 EXPERIMENTS

5.1 Single image inpainting

5.1.1 Dataset

We evaluate our method on CelebA [18] dataset which comprises of 202,599 facial images with coarse alignment. Following the protocol of [7], we used 2000 images for testing inpainting performance. Remaining images were used for training GAN. Following face detection, facial bounding boxes are central cropped to 64×64 and 128×128 resolutions.

5.1.2 Effect of initialization of z vector

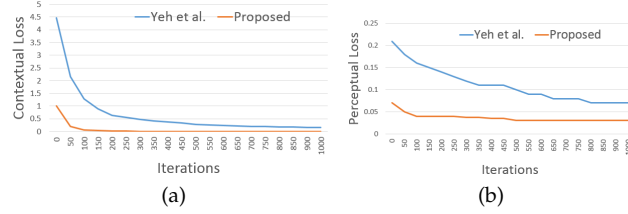In Fig. 1a we show the benefit of initializing z vector with a parametrized network as discussed in Sec. 4.1. As

Fig. 4: Convergence of **(a)** contextual loss and **(b)** perceptual loss of Eq. 2 for a batch of samples.

evident, a random initialization yields a solution which lies distinctly away from natural face manifold. On the other hand, our parametrized network learns to predict the latent vector by respecting the structural and textural statistics of the uncorrupted pixels. One major advantage of good initialization is the speed up of iterative optimization of Eq. 2. In Fig. 4, we show an exemplary convergence rates of the two components of Eq. 2. With the model initialized by our method, both perceptual and contextual losses start at an order less than [7]. This leads to much faster convergence. In fact, for most of the cases, our proposed model converges after 50 iterations compared to around 700 iterations with [7]; after this the visual quality does not improve much. Moreover, our solution tends to converge at lower magnitudes of losses and thereby yielding visually more realistic solutions(See Fig. 2). This is also evident from the peak signal to noise ratio (PSNR) between original image and the final solution image reported in Fig. 1b. p-value $\leq 10^{-5}$ in all cases. It is encouraging to see difference of performance is more appreciated at higher resolution of 128×128 resolution.

5.2 Pseudo sequences

5.2.1 Motivation

Before directly applying our model on real facial sequences, we dedicate this section to analyze the benefits of our novelties on what we term as, 'pseudo sequences'. A pseudo sequence, S_W, of length W is formed by taking a single image, I, and masking it with n different/same corruptions masks. An ideal inpainter should be agnostic of the corruption masks and yield identical reconstructions for all the n frames. Since independent optimization of Eq. 2 is unconstrained, there is no explicit restriction on the $\hat{z}$ vectors to be consistent; this is an inherent drawback of GAN based framework of [7] when applied on sequences.

5.2.2 Temporal Consistency

We define temporal consistency, η_{temp}, as the mean pairwise PSNR between all possible pairs($\hat{I}^i, \hat{I}^j$) of inpainted frames within a pseudo sequence, S_W, of length, W;

$$\eta_{temp} = \frac{1}{\binom{W}{2}} \sum_{\forall(i,j)\in S_W} PSNR(\hat{I}^i, \hat{I}^j) \qquad (8)$$

Eq. 8 allows to enumerate the consistency of a generative model. Ideally we want η_{temp}=0. Please note that this evaluation is not possible on real videos because the transformation from one frame to another is not known and thus it is not possible to align the frames to a single frame of reference without incorporating interpolation noise with

motion compensator [19]. In our results, 'Smoothness Loss' refers to temporal smoothness loss (Eq. 7). 'LSTM init' refers to initializing group of 3 frames using proposed LSTM model (Sec. 4.2.1).

Benefit of initialization with LSTM: First, in Fig. 6 we show the benefit of initializing solutions for a group of pseudo frames with LSTM over per frame independent initialization with P_{θ_z}. Frames initiated with LSTM tends to be more consistent compared to initiation by P_{θ_z}. This is attributed to the recurrent structure of LSTM which learns that in pseudo sequences, the frame are static. Learning such temporal dynamics is not possible by P_{θ_z} which is curated for single image initialization.

Consistency of final solutions: In Table 1 we compare the mean temporal consistency on the 2000 pseudo sequences created over the CelebA test set with W=3. The reported mask patterns are: a) Central : randomly corrupt 40%-70% of central part of image, b)Checkboard: corrupt 50% of image with checkboard sizes drawn uniformly from the set of $\{8\times8, 16\times16, 32\times32\}$, c)Freehand: corrupt around 40% of pixels with randomly hand drawn masks. Our proposed method with Smoothness Loss (Eq. 7) fosters in a more consistent sequence of inpainting compared to the vanilla method of per frame model of Yeh *et al.* [7]. The observations are statistically significant with p value < 0.05 in all cases. Moreover if we initialize the z vectors of the pseudo sequence with a LSTM model, then the consistency of the sequence improves. This can be attributed to more consistent initialization of z vectors by LSTM followed by Smoothness Loss which maintains the similarity of z vectors. In Fig. 5 we visually show the advantage of our proposed modifications. One has to appreciate that a set of inpainted frames by [7] is a mixture of faces with neutral and smiling appearances or different levels of smiles. However, our model yields a set with consistent facial appearance/expressions. In the context of real videos, these observations would mean that there can be drastic change of facial expressions among two adjacent corrupted frames if inpainted by [7]; such abrupt change of appearance is not common in videos. However, our model has the promise to inpaint a group of neighboring frames with consistency of appearance. Also, if inpainted by [7], the stationary portions of frames would tend to show flickering effects due to different hallucinations of textural details independently on each frame.

5.2.3 Disparity between converged $\hat{z}$ vectors

To bolster the finding in the above section we also study the disparity of the converged $\hat{z}$ vectors. Ideally, for a given pseudo sequence, the converged $\hat{z}$ vectors should be identical. We can quantify the disparity using the temporal

Fig. 5: Visualization of consistency of inpainting pseudo sequences. A pseudo sequence is created by masking a given image with different corruption patterns. Ideally we want an inpainter to yield exactly same outputs for a given subject's pseudo sequence.; Top: Masked original pseudo sequence. Middle: Inpainted sequence with Yeh *et al.* [7]. Bottom: **Proposed** inpainted sequence. Proposed method yields more consistent sequence w.r.t facial appearances.

TABLE 1: Mean consistency (Eq. 8) on CelebA test set measured in terms of PSNR(in dB). A sequence was randomly perturbed by Central, Freehand or Checkboard masks. Higher consistency is better.

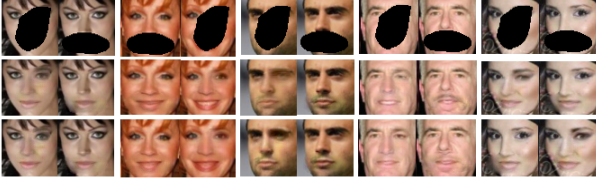|	Resolution @ 64X64			Resolution @ 128X128		
	Central	Freehand	Checkboard	Central	Freehand	Checkboard
Yeh *et al.* [7]	22.43	22.87	20.71	22.15	20.19	19.81
Proposed(Smoothness Loss)	27.14	28.95	25.12	25.11	25.40	23.75
Proposed(LSTM init + Smoothness Loss)	28.01	29.15	25.73	26.01	26.10	25.09

Fig. 6: Benefit of LSTM for initialization of sequences. Top Row: A pseudo sequence with same image masked twice differently; Middle Row: Initial solutions by independently predicting z vectors by P_{θ_z}; Bottom Row: Initial solutions by jointly initializing each pair of z vectors with LSTM. Solutions with LSTM are more consistent (similarity near mouth, eye regions).

smoothness loss of Eq. 7. In Fig. 7a we show an exemplary plot of decay of smoothness loss for a pseudo sequence. The proposed method of implicitly minimizing Eq. 7 results in near identical solutions for Eq. 2 over the sequence. However, the converged $\hat{z}$ vectors using [7] shows more variation. Also, the latter method is slower in convergence.

5.2.4 Identity preservation

It is important that a sequence of inpainted frames not only appears visually realistic but also maintains the facial identity of the subject. To evaluate this we use FaceNet embeddings [20]. FaceNet learns a parametrized network, F, to represent a given facial image into a 128-D real vector; $F : I^{H \times W \times 3} \rightarrow \mathcal{R}^{128 \times 1}$. Images of same subject yields similar embeddings and is enumerated by the L_2 distance between the embeddings. For a given sequence, S_W, identity loss, l_{seq}, is,

$$l_{seq} = \frac{1}{W} \sum_{\forall i \in S_W} ||F(\hat{I}^i) - F(I_u)||_2^2. \qquad (9)$$

$\hat{I}_i$ is the ith inpainted frame within the pseudo sequence and I_u is the original uncorrupted image. In Fig. 7b we report

the mean identity loss over the 2000 pseudo sequences. Our proposed method(LSTM init + Smoothness Loss), retains the identity of a person over a sequence more veraciously than [7]. In our initial experiments, we explicitly included l_{seq} for optimizing Eq. 2. However, we get similar sequence identity preservation prowess with the Smoothness Loss constraint alone. Authors in [21] showed that a z vector can be semantically decomposed to z_I, the identity component and z_o, the appearance component. Since our proposed Smoothness Loss enforces similarity of converged $\hat{z}$ vectors, the task of identity preservation is implicitly incorporated in the process.

5.3 Experiments on VidTIMIT dataset

The experiments with pseudo sequences taught us two lessons, viz., a)LSTM based group initialization is better than independent initialization with P_{θ_z} and b) Temporal smoothness loss is essential in imposing temporal consistency. With these understandings we proceed to test our model on real life facial video sequences. To our best knowledge, this is the first attempt towards GAN based inpainting on real videos. For this, we selected the VidTIMIT dataset [22][1] which consists of video recordings of 43 subjects each narrating 10 different sentences. Images of CelebA dataset are of superior resolution than those of VidTIMIT. Due to this intrinsic difference of data distribution we finetuned our pretrained(trained on CelebA) GAN models on randomly selected 33 subjects of VidTIMIT. Remaining videos of 10 subjects were kept for testing inpainting performances. In total there are total 9600 frames for testing. We follow the same procedure of Sec. 5.1.1 for cropping faces and Sec. 5.2.2 for creating random masks. In Table 2 we compare PSNR of different inpainting approaches for each subject (each subject has 10 videos). Here again we observe that our proposed models perform superior compared to [7]. From Table 3 it is evident that our proposed model yields more visually

1. Availabe at: http://conradsanderson.id.au/vidtimit/#examples

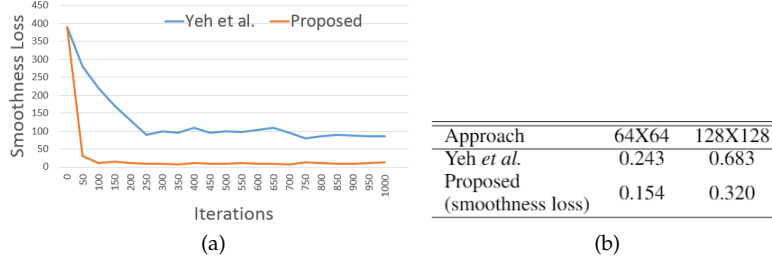| Approach	64X64	128X128
Yeh *et al.*	0.243	0.683
Proposed (smoothness loss)	0.154	0.320

|(a)| |(b)|

Fig. 7: (a): Convergence of temporal smoothness loss(Eq. 7) on a batch of CelebA pseudo sequence. (b): Mean FaceNet loss on CelebA pseudo sequences.

Fig. 8: Inpainting on sample snippets of VidTIMIT video dataset. Top Row: Damaged sequence. Middle Row: Inpainting by Yeh. *et al.* [7]. Bottom Row: **Proposed method**. It is evident that proposed framework yields visually better samples. More examples are in supplemental document.

TABLE 2: Inpainting PSNR (in dB) on test sequences of VidTIMIT dataset.

Approach	Resolution @ 64X64									
	Subject Name									
	mrj001	mwbt0	mtmr0	mtas1	mreb0	mrgg0	mdbb0	mjsw0	fjre0	fjas0
Yeh *et al.* [7]	24.32	25.32	23.61	26.11	25.12	26.01	25.98	26.09	25.31	25.81
Proposed (Smoothness Loss)	26.12	27.01	25.11	27.11	26.91	26.98	27.11	27.21	27.00	27.71
Proposed (LSTM init + Smoothness Loss)	27.02	28.07	27.11	28.87	28.87	28.78	29.21	29.12	28.21	29.01
	Resolution @ 128X128									
	Subject Name									
	mrj001	mwbt0	mtmr0	mtas1	mreb0	mrgg0	mdbb0	mjsw0	fjre0	fjas0
Yeh *et al.*	22.22	23.09	21.11	23.98	23.11	24.12	23.65	25.45	24.09	23.11
Proposed (Smoothness Loss)	24.15	25.51	23.78	25.23	25.08	25.18	25.32	25.36	25.78	25.98
Proposed (LSTM init + Smoothness Loss)	25.01	27.12	25.98	27.02	26.81	27.11	27.62	27.32	27.34	27.78

TABLE 3: Mean contextual, perceptual and FaceNet losses on VidTIMIT test videos

Approach	Resolution @ 64X64			Resolution @ 128X128		
	Contextual	Perceptual	FaceNet	Contextual	Perceptual	FaceNet
Yeh *et al.* [7]	0.25	0.13	0.28	0.41	0.20	0.68
Proposed(LSTM init + Smoothness Loss)	0.09	0.02	0.11	0.23	0.11	0.23

realistic solutions(lower perceptual loss) while retaining the appearance of the non corrupted pixels(lower contextual loss) and facial identity(lower FaceNet Loss).

6 DISCUSSION

In this paper we proposed several innovations for better optimization of the GAN based inpainting cost function. The study on pseudo sequences enabled us to do ablation studies to appreciate the benefits of each component of our proposals. Since the generator was same for both the comparing models, the improvements are solely due to our contributions. Finally, we bolstered our understandings with experiments on real videos. However, the performance of inpainting strongly relies on the generative model and the GAN training procedure. An immediate extension would be to improve the generative model itself to generate photo realistic samples at higher resolutions. The recent works of Stacked GAN [23] and progressive stagewise training of GANs [24] show promise towards this end. It would be interesting to integrate the innovations of this paper in such high resolution generative model pipelines.

REFERENCES

[1] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patch-match: A randomized correspondence algorithm for structural image editing," *ACM Transactions on Graphics-TOG*, vol. 28, no. 3, p. 24, 2009.

[2] J.-B. Huang, S. B. Kang, N. Ahuja, and J. Kopf, "Image completion using planar structure guidance," *ACM Transactions on graphics (TOG)*, vol. 33, no. 4, p. 129, 2014.

[3] M. V. Afonso, J. M. Bioucas-Dias, and M. A. Figueiredo, "An augmented lagrangian approach to the constrained optimization formulation of imaging inverse problems," *IEEE Transactions on Image Processing*, vol. 20, no. 3, pp. 681–695, 2011.

[4] A. A. Efros and T. K. Leung, "Texture synthesis by non-parametric sampling," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, vol. 2. IEEE, 1999, pp. 1033–1038.

[5] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "Patch-match: A randomized correspondence algorithm for structural image editing," *ACM Transactions on Graphics-TOG*, vol. 28, no. 3, p. 24, 2009.

[6] J. Hays and A. A. Efros, "Scene completion using millions of photographs," in *ACM Transactions on Graphics (TOG)*, vol. 26, no. 3. ACM, 2007, p. 4.

[7] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *CVPR*, 2017, pp. 5485–5493.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.

[9] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016, pp. 2536–2544.

[10] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[12] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[13] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[14] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[15] A. Kumar Jain, A. Agarwalla, K. Krishna Agrawal, and P. Mitra, "Recurrent memory addressing for describing videos," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

[16] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. IEEE, 2015, pp. 3156–3164.

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

[18] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3730–3738.

[19] J. Caballero, C. Ledig, A. Aitken, A. Acosta, J. Totz, Z. Wang, and W. Shi, "Real-time video super-resolution with spatio-temporal networks and motion compensation," *CVPR*, 2016.

[20] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[21] C. Donahue, A. Balsubramani, J. McAuley, and Z. C. Lipton, "Semantically decomposing the latent spaces of generative adversarial networks," *arXiv preprint arXiv:1705.07904*, 2017.

[22] C. Sanderson and B. C. Lovell, "Multi-region probabilistic histograms for robust and scalable identity inference," in *International Conference on Biometrics*. Springer, 2009, pp. 199–208.

[23] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, and S. Belongie, "Stacked generative adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2017, p. 4.

[24] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.