MIC-GAN: Multi-view Assisted Image Completion Using Conditional Generative Adversarial Networks

Gagan Kanojia
Electrical Engineering
Indian Institute of Technology Gandhinagar
Gandhinagar, Gujarat, India
gagan.kanojia@iitgn.ac.in

Shanmuganathan Raman
Electrical Engineering
Indian Institute of Technology Gandhinagar
Gandhinagar, Gujarat, India
shanmuga@iitgn.ac.in

***Abstract*—Consider a set of images of a scene captured from multiple views with some missing regions in each image. In this work, we propose a convolutional neural network (CNN) architecture which fills the missing regions in one image using the information present in the remaining images. The network takes the set of images and their corresponding binary maps as inputs and generates an image with the completed missing regions. The binary map indicates the missing regions present in the corresponding image. The network is trained using an adversarial approach and is observed to generate sharp output images qualitatively. We evaluate the performance of the proposed approach on the dataset extracted from the standard dataset, MVS-Synth.**

***Index Terms*—Image processing, Neural networks**

I. Introduction

Often, it is desirable to capture a scene without any moving objects present. However, in public places, it is common to have people or objects moving around in the scene. To remove these objects present in the scene, the corresponding regions have to be filled appropriately. Let us assume that the regions corresponding to these objects are provided by the user in the form of a binary map in which 0 represents the regions to be filled. We call them missing regions. There are several single image completion techniques which take an image with missing regions and generate a completed image ([1]–[3]). These techniques either exploit the input image statistics or involve training on a huge amount of data to fill the missing regions. However, in such cases, it is not necessary that the missing regions would be filled by the same information which was occluded by the objects. Capturing multiple images of the same scene is a common practice. Hence, parts of the scene which are occluded in one image can be visible in some other image captured from the same or a different viewing angle. Our main idea is to exploit this observation to fill the missing regions present in the input image.

In this work, we propose a conditional generative adversarial network (GAN) which takes a set of multi-view images along with their corresponding binary maps highlighting the regions which need to be filled, as the input. Since we are dealing with multi-view images, we have to make sure that the structure is preserved while transferring information from one view to the other and the output image has to be sharp without any artifacts. Due to these reasons, we employ GAN [4]. The proposed network has two components: Image Warping Network (IWN) and Image Completion Network (ICN). IWN is responsible aligning the input set of images while ICN is jointly trained with the discriminator in an adversarial manner to obtain a sharp image as the output. We experimentally show that warping the source images to the reference images help the generator to produce better results. We show qualitatively and quantitatively that the proposed approach performs better than the single-image inpainting approaches [1], [2].

978-1-7281-5120-5/20/$31.00 © 2020 IEEE

II. Related Works

Image completion has been an active area of research in computer vision. Several methods have been proposed for single image completion in which the input is an image with some missing regions and the task is to fill them. In previous works, many algorithms have been proposed which fill the missing regions by utilizing the data from the remaining parts of the same image [3], [5]–[7]. Recently, several learning based methods based on GANs have been proposed which are trained on a huge amount of data to generate sharp and realistic images [1]–[3], [8].

Yeh *et al.* proposed a generative model based approach for semantic image inpainting [9]. Iizuka *et al.* proposed a generative adversarial network to perform the task for image completion which uses two discriminators: local and global. Global discriminator takes the complete generated image and the local discriminator takes only the generated missing region. Then, they together make the decision whether the input came from training data or it is generated by the generator [1]. Liu *et al.* utilized partial convolution to inpaint the irregular holes present in the image. They also proposed a style loss which helped in obtaining artifact-free results [3]. Recently, Yu *et al.* proposed an attention based method which utilizes the sur-

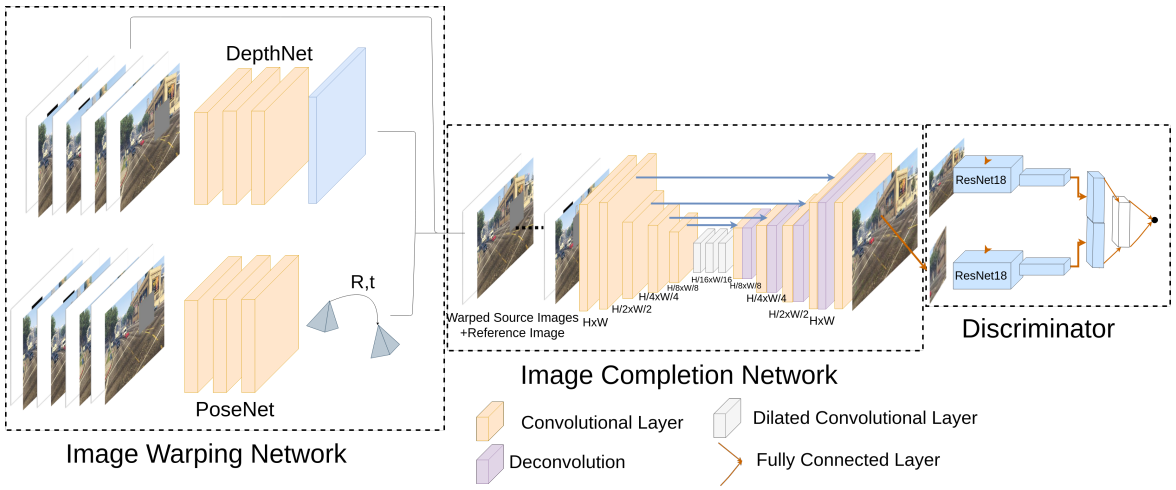

Fig. 1. **The proposed network architecture**. The proposed neural network architecture for multi-view assisted image completion. The proposed network consists of two components: Image Warping and Image Completion Network.

roundings of the missing region as reference to perform image completion [2]. Further, Ulyanov *et al.* proposed a technique which does not rely on the external dataset training and use the structure of generative models to perform inpainting [10]. These techniques rely on either the image statistics or the model learned by training over millions of images.

There are works which use multiple views of a scene for the task of inpainting [11]–[14]. Thonat *et al.* proposed a method which utilizes a set of multi-view images as input and performs multi-view inpainting such that it is consistent in all the images [11]. Their method utilizes the multi-view 3D reconstruction along with a global optimization technique. Later, Philip and Drettakis introduced a plane-based multi-view inpainting technique which exploits the local planar regions present in the scene to obtain better multi-view inpainting results [14]. Li *et al.* proposed a technique which performs multi-view inpainting using an RGB-D sequence as the input [13]. However, they are not learning-based methods.

III. Proposed Approach

The proposed network consists of two components: Image Warping Network (IWN) and Image Completion Network (ICN). Let us consider a reference image I_r and a set of n source images $\{I_s\}_{s=1}^n$ with some missing regions. I_r and $\{I_s\}_{s=1}^n$ are obtained by capturing the same scene with different (or same) viewing angles. IWN takes I_r and $\{I_s\}_{s=1}^n$ along with their corresponding binary maps as inputs. The binary maps contain two values, 0 and 1. Here, 0 represents the missing regions in the images and 1 represents the regions which are present in the images. ICN generates a completed reference image. Fig. 1 shows the visual representation of the proposed architecture.

A. Image warping

Given I_r and $\{I_s\}_{s=1}^n$, we synthesize a new image for each I_s by warping them in such a way that they align with I_r. The warping is achieved using the depth map $\mathcal{D} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ of I_r and the pose $P_{r \rightarrow s}$ of the source image I_s's view with reference to I_r's view. Here, $P_{r \rightarrow s}$ is a 4×4 camera transformation matrix (3-D rotation and 3-D translation). Zhou *et al.* show that this process can be achieved in a fully differentiable manner using CNNs [15]. They have proposed a differentiable depth-based renderer which reconstructs the reference image by sampling the pixel values of the source image using $\mathcal{D}$ and $P_{r \rightarrow s}$. They obtain the corresponding location x_s of x_r in I_s as shown in Eq. 1.

$$x_s \sim K \hat{P}_{r \rightarrow s} \hat{\mathcal{D}}(x_r) K^{-1} x_r \tag{1}$$

Here, x_r and x_s are the homogeneous coordinates of the locations in I_r and I_s, respectively. K is the intrinsic camera matrix, $\hat{\mathcal{D}}$ is the predicted depth map of I_r, and $\hat{P}_{r \rightarrow s}$ is the predicted camera transformation matrix. To obtain $\hat{I}_s(x_r)$ using the values $I_s(x_s)$, a differentiable bilinear sampling method is used [15]. Here, $\hat{I}_s$ is the source image warped to the coordinate frame of I_r.

B. Image Warping Network (IWN)

The Image Warping Network consists of two components: DepthNet and PoseNet. Let B_r and $\{B_s\}_{s=1}^n$ be the binary maps corresponding to I_r and $\{I_s\}_{s=1}^n$, respectively. Let I_r^b and $\{I_s^b\}_{s=1}^n$ be the volumes obtained by concatenating I_r and $\{I_s\}_{s=1}^n$ with B_r and $\{B_s\}_{s=1}^n$ as their last channel, respectively. DepthNet and PoseNet take I_r^b and $\{I_s^b\}_{s=1}^n$ as inputs and output the depth map $\mathcal{D}$ corresponding to I_r and $\{q_s\}_{s=1}^n$, respectively. Here, q_s is a 6-D vector corresponding to I_s. q_s comprises three euler angles which are used to compute the rotation matrix and the 3-D translational vector for $\hat{P}_{r \rightarrow s}$. Then, the warped source images $\{\hat{I}_s\}_{s=1}^n$ and their corresponding warped binary maps $\{\hat{B}_s\}_{s=1}^n$ are obtained as explained in III-A.

Architecture. The DepthNet has an encoder-decoder architecture [16]. The encoder adapts the Resnet (50 layers) architecture [17]. The depth is predicted in a multi-scale fashion [18]. PoseNet is a convolutional neural network which consists of

TABLE I
THE BUILDING BLOCKS USED IN DEPTHNET, POSENET AND ICN.

Layer Name	Layer Type	Kernel	Stride/ Dilation	In/Out
dconv	conv	3×3	1/1	In/Out
	conv	3×3	1/1	Out/Out
upconv	conv	3×3	1/1	In/Out
	upsample	$\frac{1}{2} \times \frac{1}{2}$	1/-	Out/Out
pred	conv	3×3	1/1	In/Out
	Sigmoid	-	-	Out/Out
oconv	conv	7×7	2/1	$In/32$
	conv	3×3	1/1	$32/32$
	conv	3×3	1/1	$32/Out$

TABLE II
THE DETAILED ARCHITECTURE OF THE POSENET. EACH CONVOLUTION LAYER, EXCEPT CONV8 , IS FOLLOWED BY A RELU LAYER AND BATCH NORMALIZATION LAYER.

Name	Type	Kernel	Stride/ Dilation	In/Out	Input
oconv0	oconv	-	-/-	4/16	I_r
oconv1	oconv	-	-/-	4/16	I_1
oconv2	oconv	-	-/-	4/16	I_2
oconv3	oconv	-	-/-	4/16	I_3
conv2	conv	3×3	2/1	64/64	oconv0+oconv1 +oconv2+oconv3
conv3	conv	3×3	2/1	64/64	conv2
conv4	conv	3×3	2/1	64/128	conv3
conv5	conv	3×3	2/1	128/256	conv4
conv6	conv	3×3	2/1	256/256	conv5
conv7	conv	3×3	2/1	256/256	conv6
conv8	conv	3×3	2/1	256/6*3	conv7
gpool1	global avgpool	-	-	6*3/6*3	conv6

a series of convolution layers. Each convolution layer has a stride of two. Each convolution layer is followed by the ReLU activation function and a batch normalization layer, except the final layer. Table I shows the building blocks used in IWN, i.e., DepthNet and PoseNet, and ICN. Table II shows the detailed architecture of PoseNet. Table III shows the detailed architecture of DepthNet. The terms $conv$, $avgpool$, $Sigmoid$, $deconv$, and $global\ avgpool$ stand for convolution, average pooling, sigmoid, deconvolution, and global average pooling layer, respectively. $res01$, $res02$, $res03$ and $res04$ are the four residual blocks used in ResNet (50 layers) architecture [17]. $upred3$, $upred4$, $upred5$, and $upred6$ are the disparity maps estimated at different scales. The reciprocal of $upred3$, $upred4$, $upred5$, and $upred6$ are the depth maps at different scales. The reciprocal of $upred6$ is the depth map $\hat{\mathcal{D}}$ at the scale of reference image. The output of $gpool1$ is $\{q_s\}_{s=1}^n$.

C. Image Completion Network (ICN)

Let $\{\hat{I}_s^b\}_{s=1}^n$ be the volumes obtained by concatenating the warped source images $\{\hat{I}_s\}_{s=1}^n$ with the corresponding warped binary maps $\{\hat{B}_s\}_{s=1}^n$ along their last channel. ICN takes I_r^b and $\{\hat{I}_s^b\}_{s=1}^n$ as inputs. We concatenate I_r^b and $\{\hat{I}_s^b\}_{s=1}^n$ along the channels and pass them through ICN. The task of ICN is to output the reference image with missing regions completed. The regions are filled by utilizing the information present in the warped source images.

Architecture. ICN is a fully convolutional neural network. It has an encoder-decoder architecture [16]. The encoder comprises a series of two 3×3 convolutions which are followed by a ReLU, a batch normalization layer, and an average pooling layer for downsampling. Then, we perform dilated convolutions on it. The dilated convolutions help the network to consider more spatial area to produce the output [1]. In the decoder, the resolution of the feature maps is increased through deconvolution. Table IV shows the detailed architecture of ICN. $TanH$ stands for the hyperbolic tangent layer. In all our experiments, we have used three source images, i.e., $n = 3$.

TABLE III
THE DETAILED ARCHITECTURE OF THE DEPTHNET. EACH CONVOLUTION LAYER, EXCEPT IN $upred3$, $upred4$, $upred5$, AND $upred6$, IS FOLLOWED BY A RELU LAYER AND BATCH NORMALIZATION LAYER.

Name	Type	Kernel	Stride /Dilation	In/Out	Input
oconv0	oconv	-	-/-	4/16	I_r
oconv1	oconv	-	-/-	4/16	I_1
oconv2	oconv	-	-/-	4/16	I_2
oconv3	oconv	-	-/-	4/16	I_3
maxpool1	maxpool	3×3	2/-	64/64	oconv0+oconv1+ oconv2+oconv3
res01	ResBlock	-	-/-	64/64	maxpool1
res02	ResBlock	-	-/-	64/128	res01
res03	ResBlock	-	-/-	128/256	res02
res04	ResBlock	-	-/-	256/512	res03
upconv1	upconv	-/-	-/-	512/512	res04
conv1	conv	3×3	1/1	768/512	res03+ upconv1
upconv2	upconv	-/-	-/-	512/256	conv1
conv2	conv	3×3	1/1	384/256	res02+ upconv2
upconv3	upconv	-/-	-/-	256/128	conv2
conv3	conv	3×3	1/1	192/128	res01+ upconv3
upred3	pred	-/-	-/-	128/1	conv3
upconv4	upconv	-/-	-/-	128/64	conv3
conv4	conv	3×3	1/1	129/64	maxpool1+ upconv4+ upred3
upred4	pred	-/-	-/-	64/1	conv4
upconv5	upconv	-/-	-/-	64/32	conv4
conv5	conv	3×3	1/1	33/32	upconv5 +upred4
upred5	pred	-/-	-/-	32/1	conv5
upconv6	upconv	-/-	-/-	32/16	conv5
conv6	conv	3×3	1/1	17/16	upconv6 +upred5
upred6	pred	-/-	-/-	16/1	conv6

D. Discriminator

The discriminator network is trained to identify whether the image is from the original distribution or it is generated by ICN. It enables ICN to produce sharp images.

Architecture. The discriminator network has two parts: local and global [1]. The global part takes the complete image and the local part takes the image patch corresponding to the filled region. For both the parts, we adapt ResNet (18 layers) architecture [17]. We remove their classification layer and concatenate their outputs which is further passed through a fully connected layer to make a prediction. The image and the patch are both resized to 224×224 before passing them through the discriminator.

TABLE IV
THE DETAILED ARCHITECTURE OF THE ICN. EACH CONVOLUTION LAYER, EXCEPT CONV4, IS FOLLOWED BY A LEAKY RELU LAYER AND BATCH NORMALIZATION LAYER.

Name	Type	Kernel	Stride/Dilation	In/Out	Input
dconv0	dconv	-	-/-	16/64	Images
down1	down	-	-/-	64/128	dconv0
down2	down	-	-/-	128/256	down1
down3	down	-	-/-	256/512	down2
down4	down	-	-/-	512/512	down3
conv1	conv	3×3	1/2	512/512	down4
conv2	conv	3×3	1/3	512/512	conv1
conv3	conv	3×3	1/4	512/512	conv2
deconv1	deconv	2×2	$\frac{1}{2}$/1	512/512	conv3
dconv1	dconv	-	-/-	1024/256	down3+deconv1
deconv2	deconv	2×2	$\frac{1}{2}$/1	256/256	dconv1
dconv2	dconv	-	-/-	512/128	down2+deconv2
deconv3	deconv	2×2	$\frac{1}{2}$/1	128/128	dconv2
dconv3	dconv	-	-/-	256/64	down1+deconv3
deconv4	deconv	2×2	$\frac{1}{2}$/1	64/64	dconv3
dconv4	dconv	-	-/-	128/64	dconv0+deconv4
conv4	conv	1×1	1/1	64/3	dconv4
tanh1	Tanh	-	-/-	3/3	conv4

E. Training

We jointly train DepthNet and PoseNet, i.e., IWN, in an unsupervised manner independent of ICN to obtain the depth map $\mathcal{D}$ corresponding to the reference image I_r and the poses $P_{t\rightarrow s}$ corresponding to each source image I_s [15]. We train IWN to minimize the photometric loss [19] shown in Eq. 2 in a multi-scale fashion [18].

$$\begin{aligned}\mathcal{L}_w =& \lambda_1 \sum_{s=1}^{n} ||\hat{B}_s \odot (I_o - \hat{I}_s)||_1 + \lambda_2 \mathcal{L}_{smooth} \\ &+ \lambda_3 \sum_{s=1}^{n} ||\hat{B}_s \odot (1 - SSIM(I_o, \hat{I}_s))||_1\end{aligned} \tag{2}$$

Here, $\odot$ is the Hadamard product, $||.||_1$ is the ℓ_1 norm, $SSIM$ is the structural similarity index [20], and I_o is the ground truth image for I_r (i.e., the image we want to achieve as output). $\hat{B}_s$ is multiplied to ignore the missing regions present in warped source images $\hat{I}_s$. $\mathcal{L}_{smooth}$ is the ℓ_1 norm of the second-order gradients for $\hat{\mathcal{D}}$ [15]. λ_1, λ_2, and λ_3 are real constants. We used Adam for the weight update with $\beta_1 = 0.9$, $\beta_2 = 0.99$, and learning rate = 0.0001 [21]. After training IWN, we trained ICN with the loss function $\mathcal{L}_g$ shown in Eq. 3.

$$\mathcal{L}_g = ||I_o - I_g||_1 + 0.5 * \mathcal{L}_{perceptual} \tag{3}$$

Here, $\mathcal{L}_{perceptual}$ is the perceptual loss used in [3].

$$\mathcal{L}_{perceptual} = \sum_{n=0}^{N-1} ||\phi(I_o) - \phi(I_g)||_1 + \sum_{n=0}^{N-1} ||\phi(I_{og}) - \phi(I_g)||_1 \tag{4}$$

Here, I_o is the ground truth image for the reference image I_r, $I_g = G(I_r^b, \{\hat{I}_s^b\}_{s=1}^n)$ is the image generated by ICN, and G is a function representing ICN. I_{og} is the generated image I_r with the missing region filled with the corresponding region of the ground truth image. $\phi(I)$ is the feature vector obtained by passing the image I through the vgg19 network [22]. We found that the perceptual loss helps in stabilizing the training when trained with the adversarial loss. Training ICN using $\mathcal{L}_g$ produced images completed with information which had significant blur. To produce sharper images, we train ICN jointly with the discriminator using the min-max approach proposed in [23] as shown in Eq. 5.

$$\min_G \max_D \log(D(I_o, p_{I_o}) + \log(1 - D(I_g, p_{I_g})) \tag{5}$$

Here, $I_g = G(I_r^b, \{\hat{I}_s^b\}_{s=1}^n)$, p_{I_g} is the region in I_g corresponding to the missing region in I_r, and p_{I_o} is a patch extracted from I_o. D is a function representing the discriminator. To produce sharp outputs, $\mathcal{L}_g$ is combined with Eq. 5 as shown in Eq. 6 [1].

$$\min_G \max_D \mathcal{L}_g + \lambda_4 (log(D(I_o, p_{I_o}) + \log(1 - D(I_g, p_{I_g}))) \tag{6}$$

Here, λ_4 is a constant. The discriminator D is trained to identify whether the input came from the training data or it is generated by the ICN. The generator G is trained to minimize $\log(1 - D(I_g, p_{I_g}))$. However, $\log(1 - D(I_g, p_{I_g}))$ does not provide sufficient gradient for the generator and it saturates in the early stage. Hence, the training of the generator becomes poor. As suggested in [23], instead of training the generator to minimize $\log(1 - D(I_g, p_{I_g}))$, we train it to maximize $\log(D(I_g, p_{I_g}))$. For the experiments, the values used for λ_1, λ_2, λ_3 and λ_4 are 0.15, 0.1, 0.85 and 0.2. For the weight update of ICN, we used Adam with $\beta_1 = 0.5$, $\beta_2 = 0.99$, and an initial learning rate = 0.0001 [21]. For the weight update of the discriminator, we used stochastic gradient descent with a learning rate of 0.01 and momentum of 0.9.

IV. EXPERIMENTS

Dataset. We have used MVS-Synth dataset used in [24]. It consists of 120 sequences of static urban scenes. Each sequence consists of 100 images. We split the sequences into two parts. We used the first 80% sequences for training and the remaining for testing. In each split, we created sets of four images by clubbing each image in each sequence with its three consecutive images. During training, one of the images in the set is used as I_r and the other images are used as the source images. We randomly insert missing regions in the input images and compute their corresponding binary maps. The width and height of the missing regions vary between 70-120 pixels. The images and their corresponding binary maps are resized such that the number of rows is equal to 256.

Results. In all the experiments, we have used three source images, i.e., $n = 3$. We assume that the intrinsic parameters of the camera are known. Fig. 2 shows the image completion achieved using the proposed approach. It shows the results on multi-view images of three different scenes from the test split of the dataset when the corresponding image is taken as the reference image and others as the source images. The generated images are not only sharp but also have the same information as that of the original scene. The previous image completion works used a single image as the input ([1], [2]). Since the proposed approach requires multiple

Fig. 2. **Multi-view assisted image completion.** The figure shows the results on multi-view image sets corresponding to three different scenes (two columns each). In each set, first column shows the input multi-view image set with the missing regions and the second column shows the image completion obtained by the proposed network when the corresponding image is taken as the reference image and others as the source images.

Fig. 3. **Comparisons with single image inpainting approach.** The figure shows the image completion results obtained using the proposed approach and their comparison with Lizuka *et al.* [1]. The first row shows the reference images of four different multi-view image sets. The second row shows the results obtained using Lizuka *et al.* [1]. The third and fourth rows show the results obtained using the proposed approach and the corresponding ground truth images, respectively. It can be seen that using the proposed approach we are able to fill the missing region with the similar information as in the actual scene which is not the case with [1].

views, we cannot train our network on datasets like Places2 [25] used in those works. Hence, a fair comparison is not plausible. However, we have provided comparison with the recent work by [1] in Fig. 3. In Fig. 3, we show the image completion results obtained using the proposed approach and their comparison with [1] trained on Places2 dataset [25]. It can be seen that using the proposed approach we are able to fill the missing region with the similar information as in the actual scene which is not the case with [1]. Also, just for reference, we have compared our results in terms of mean ℓ_1 norm, mean ℓ_2 norm and PSNR with the state-of-the-art single image completion method by [2] in Table V. The gpu and cpu runtimes of the proposed model for an input image set of size four are 0.75s and 7.3s, respectively. We implemented the proposed network using PyTorch and performed all the experiments on a system with Intel i7-5960X processor and an Nvidia Titan X GPU.

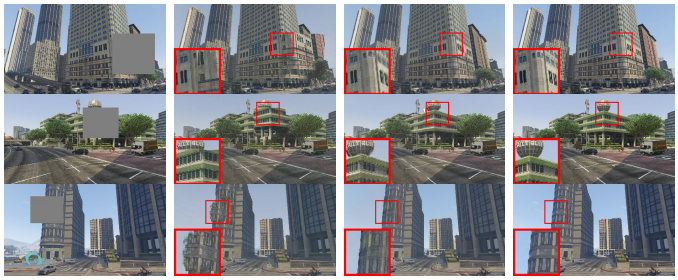

Fig. 4. **Without IWN.** The figure shows the comparison of the results obtained using the proposed approach with the results obtained without using IWN, i.e., by feeding the reference and the source images directly to the ICN. The first and second columns show the reference images and the results obtained without using IWN, respectively. The third and fourth columns show the results obtained using the proposed approach and ground truth images, respectively. It can be seen that the information filled in the missing region is quite distorted when we do not use IWN in comparison to the results obtained using the proposed approach.

Fig. 5. **Without ICN.** The figure shows comparison of the results obtained using the proposed approach with the results obtained using only IWN. The first column shows the reference images. The second column shows the results obtained using only IWN. The third and the fourth columns show the results obtained using the proposed approach and the ground truth images, respectively. It can be seen that the information filled in the missing region is quite blurry when we only use IWN in comparison to the proposed method.

Ablation Study. We performed the following ablation studies.

1. Without IWN. In this study, we check the usefulness of IWN. Instead of passing the concatenated I_r^b and $\{\hat{I}_s^b\}_{s=1}^n$ through ICN, we trained ICN jointly with the discriminator

TABLE V
COMPARISON OF THE RESULTS OBTAINED IN [2] WITH THE RESULTS OBTAINED USING THE PROPOSED APPROACH.

Method	Dataset	ℓ_1 norm	ℓ_2 norm	PSNR	SSIM
Yu *et al.* [2]	Places2	8.6%	2.1%	18.91	-
Ours	MVS-Synth	4.44%	1.01%	26.92	0.93

using I_r^b and $\{I_s^b\}_{s=1}^n$ concatenated along the third dimension. We found that the results were worse than using ICN along with IWN. Fig. 4 shows the comparison of the results obtained using the proposed approach with the results obtained without using IWN. It can be seen that the completion without using IWN is really poor (second column) in comparison to the results obtained using the proposed approach (third column).

2. Without ICN. In this study, we checked the requirement of ICN. Using the outputs of IWN and the warping technique used in [15], we obtained $\{\hat{I}_s\}_{s=1}^n$. Then using Eq. 7, we obtained the completed image.

$$I_o(p) = \frac{1}{\sum_{s=1}^{n} \hat{B}_s(p)} \sum_{s=1}^{n} \hat{B}_s(p)\hat{I}_s(p) \tag{7}$$

Here, p represents the pixel locations where $B_r(p) = 0$. For other pixel locations with $B_r(p) = 1$, $I_o(p) = I_r(p)$. We found that the results were blurry in the filled region. Fig. 5 shows the comparison of the results obtained using the proposed approach with the results obtained using only IWN. It can be seen that the information filled in the missing region is blurrier (second column) in comparison to the results obtained using the proposed approach (third column).

V. CONCLUSION

We propose a novel convolutional neural network architecture for the task of image completion using multiple views. Through ablation studies, we verified the importance of both IWN and ICN. The image completion results achieved using the proposed approach are sharp and artifact-free. They are also consistent with the ground truth images. If some part of the missing region of the reference image is not present in any of the source images then for such regions, single image inpainting techniques can be used to fill them ([1]–[3]). As a future work, we will explore the possibilities to incorporate the task of single image inpainting in the proposed approach.

ACKNOWLEDGMENT

Gagan Kanojia was supported by TCS Research Fellowships. Shanmuganathan Raman was supported by SERB Core Research Grant and Imprint 2 Grant.

REFERENCES

[1] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics*, vol. 36, no. 4, p. 107, 2017.

[2] 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*. IEEE, 2018.

[3] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," *arXiv preprint arXiv:1804.07723*, 2018.

[4] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[5] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1200–1212, 2004.

[6] I. Drori, D. Cohen-Or, and H. Yeshurun, "Fragment-based image completion," vol. 22. ACM, 2003, pp. 303–312.

[7] A. Criminisi, P. Perez, and K. Toyama, "Object removal by exemplar-based inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2003, pp. II–II.

[8] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3911–3919.

[9] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5485–5493.

[10] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Deep image prior," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9446–9454.

[11] T. Thonat, E. Shechtman, S. Paris, and G. Drettakis, "Multi-view inpainting for image-based scene editing and rendering," in *International Conference on 3D Vision (3DV)*, 2016.

[12] S.-H. Baek, I. Choi, and M. H. Kim, "Multiview image completion with space structure propagation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 488–496.

[13] F. Li, G. A. G. Ricardez, J. Takamatsu, and T. Ogasawara, "Multi-view inpainting for rgb-d sequence," in *International Conference on 3D Vision (3DV)*. IEEE, 2018, pp. 464–473.

[14] J. Philip and G. Drettakis, "Plane-based multi-view inpainting for image-based rendering in large scenes," in *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*. ACM, 2018, p. 6.

[15] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 6612–6619.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[17] 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*. IEEE, 2016, pp. 770–778.

[18] Z. Yin and J. Shi, "Geonet: Unsupervised learning of dense depth, optical flow and camera pose," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1983–1992.

[19] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 270–279.

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

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[22] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[23] 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*, 2014, pp. 2672–2680.

[24] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang, "Deepmvs: Learning multi-view stereopsis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 2821–2830.

[25] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452–1464, 2018.