Pix2NeRF: Unsupervised Conditional π-GAN for Single Image to Neural Radiance Fields Translation

Shengqu Cai
ETH Zürich

Anton Obukhov
ETH Zürich

Dengxin Dai
MPI for Informatics
ETH Zürich

Luc Van Gool
ETH Zürich
KU Leuven

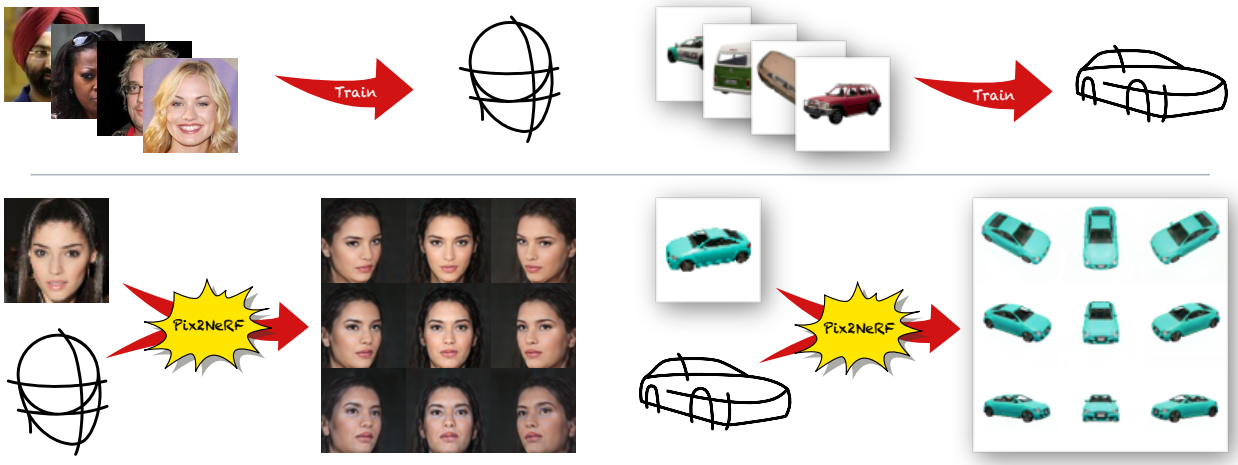

Figure 1. Overview of Pix2NeRF: We propose a method for unsupervised learning of neural representations of scenes, sharing a common pose prior. At test time, Pix2NeRF disentangles pose and content from an input image and renders novel views of the content. Top: π-GAN is trained on a dataset without pose supervision. Bottom: a trained model is conditioned on a single image to obtain pose-dependent views.

Abstract

We propose a pipeline to generate Neural Radiance Fields (NeRF) of an object or a scene of a specific class, conditioned on a single input image. This is a challenging task, as training NeRF requires multiple views of the same scene, coupled with corresponding poses, which are hard to obtain. Our method is based on π-GAN, a generative model for unconditional 3D-aware image synthesis, which maps random latent codes to radiance fields of a class of objects. We jointly optimize ***(1)*** *the π-GAN objective to utilize its high-fidelity 3D-aware generation and* ***(2)*** *a carefully designed reconstruction objective. The latter includes an encoder coupled with π-GAN generator to form an auto-encoder. Unlike previous few-shot NeRF approaches, our pipeline is unsupervised, capable of being trained with independent images without 3D, multi-view, or pose supervision. Applications of our pipeline include 3d avatar generation, object-centric novel view synthesis with a single input image, and 3d-aware super-resolution, to name a few.*

1. Introduction

Following the success of Neural Radiance Fields (NeRF) [23], encoding scenes as weights of multi-layer perceptrons (MLPs) has emerged as a promising research direction. Novel View Synthesis is an important application: given sparse sample views of a scene, the task is to synthesize novel views from unseen camera poses. NeRF addresses it by encoding color and volume density at each point of the 3D scene into a neural network and uses traditional volume rendering to compose 2D views.

While NeRF is capable of synthesizing novel views with high fidelity, it is often impractical due to being "overfitted" to a given scene and requiring multiple views of the scene to train. Several follow-up works attempt to address these limitations via making NeRF generalize to new scenes.

Corresponding author: Shengqu Cai (shecai@ethz.ch)

Code: https://github.com/HexagonPrime/Pix2NeRF

Major progress has been made in training a general NeRF capable of encoding a scene given only one or a handful of views [5, 7, 16, 40, 41, 46]. However, these works are designed to work well only with multi-view images during either training or both training and inference.

One reason why single-shot NeRF, or in general single-shot novel view synthesis is challenging, is the incomplete content information within a single image. For example, given a frontal image of a car, there is very little information to infer a novel view from the back directly. Bringing back the traditional inverse graphics and 3D reconstruction pipelines, [44] addresses this issue by making an additional assumption on the symmetry of the scene to interpolate potentially missing geometry information within a single image. However, this technique is limited to scenes where symmetry can be introduced and does not tackle the general case.

Therefore, a natural follow-up question is how does a human brain address such a challenging task? One of the approaches we use unconsciously is learning a prior implicit model for object categories and mapping what we observe to the learned model. This line of thinking is already explored in prior works [40, 46]. An essential part missing from these works is ensuring that novel views also meet our expectation of the object class, and due to the lack of supervision from a sole image, this is normally done via imagination.

One of the closest forms of imagination developed by the machine learning community is Generative Adversarial Networks [13]. GANs have been very successful in image synthesis and transformation. Beyond 2D, studies have shown GAN's capability of synthesizing 3D content [24] from natural images. This suggests another approach to address 3D reconstruction without multi-view images via 3D GAN inversion. Such a strategy bypasses the problem of missing information within one sole image due to GAN's adversarial training. Existing works [31, 47] utilize such a method based on HoloGAN [24], StyleGAN [47], and others, but one of the drawbacks naturally from these 3D-aware generative models is their relatively weak 3D consistency.

With the rapid increase of NeRF [23] popularity, corresponding generative models are also gaining attention. GRAF [35] and π-GAN [2] follow traditional GAN settings by mapping latent codes to category-specific radiance fields. These generative models typically have high 3D consistency due to the built-in volumetric rendering design. This observation suggests the possibility of few-shot 3D reconstruction using adversarial training and radiance fields.

In this paper, we formulate the task of translating an input image of a given category to NeRF as an end-to-end pipeline termed **Pix2NeRF** (Fig. 1). The method can perform novel view synthesis given a single image, without the need of pre-training, annotation, or fine-tuning. Pix2NeRF can be trained with natural images – without explicit 3D supervision, in an end-to-end fashion. Inspired by prior works [31, 40, 46], we introduce an encoder mapping a given image to a latent space. We jointly optimize several objectives. First, we train π-GAN and the added encoder to map generated images back to the latent space. Second, we adapt the encoder coupled with π-GAN's generator to form a conditional GAN, trained with both adversarial and reconstruction loss. We show that merely doing π-GAN inversion is challenging and insufficient to complete our goal, and adaptation is important for calibrating learned representations of the encoder and generator. Our framework is able to instantiate NeRF in a single shot manner while naturally preserving the ability to synthesize novel views with high fidelity, comparable to state-of-the-art generative NeRF models.

Contributions.

- We propose Pix2NeRF, the first unsupervised single-shot NeRF model, that can learn scene radiance fields from images without 3D, multi-view, or pose supervision.
- Our pipeline is the first work on conditional GAN-based NeRF, or in general, NeRF-based GAN inversion. We expect our pipeline to become a strong baseline for future works towards these research directions.
- We demonstrate the superiority of our method compared with naive GAN inversion methods and conduct an extensive ablation studies to justify our design choices.

2. Related works

Our work can be classified as a category-specific 3D-aware neural novel view synthesis method, which is strongly based on NeRF [23] and π-GAN [2].

Neural scene representations. The field of encoding a scene into neural networks has proven to be a promising research direction. This includes, but is not limited to: parameterizing the geometry of a scene via signed distance functions or occupancy [6, 22, 28, 36], encoding both geometry and appearance [18, 26, 33, 38], etc. Recently, the impressive performance of Neural Radiance Fields (NeRF) [23] has drawn extensive attention to this field. It encodes a scene as a multi-variable vector-valued function $f(x, y, z, \theta, \phi) = (r, g, b, \sigma)$ approximated by MLP, where (x, y, z) denotes spatial coordinates, (θ, ϕ) denotes viewing direction, and (r, g, b, σ) corresponds to color and volume density. This function is then called repeatedly by any of the volume rendering techniques to produce novel views. The outstanding performance of NeRF inspired follow-up works to extend it towards alternative settings, such as training from unconstrained images [20], training without poses [21, 43], etc.

NeRF-based GANs. Following the developments of GANs and NeRFs, several works tried combining them to form generative models producing NeRFs. One of the first attempts

in this direction is GRAF [35]; it performs category-specific radiance fields generation by conditioning NeRF on shape and appearance code. Following the NeRF pipeline, the generator can synthesize an image given a random code and a view direction. The generated image is passed into the discriminator together with real images, thus implementing a GAN. GRAF is an unsupervised model, since it does not require ground truth camera poses; therefore, it can be trained using "in the wild" images. This is done by introducing a *pose prior* relative to a canonical view frame of reference, e.g., Gaussian distribution to describe head pitch and yaw relative to a front face view. π-GAN [2] is similar to GRAF, but conditions on a single latent code and utilizes FiLM [10, 30] SIREN [37] layers instead of simple MLPs. More recently, several works improved synthesis quality with high resolutions [14], better 3D shapes [45], and precise control [25, 48].

Few-shot NeRF. The main property of NeRFs is the ability to bake in a 3D scene into MLP weights. However, this is also a limitation since it must be retrained for each new scene, which takes a lot of time and money. To lift this constraint, PixelNeRF [46] and GRF [40] condition MLPs on pixel-aligned features extracted by a CNN encoder. During the novel view rendering phase, 3D points along the rays are projected onto the extracted feature grid to get aligned features, then fed into an MLP with the points. More recently, CodeNeRF [16] suggested training NeRF with learnable latent codes and utilizing test-time optimization to find the best latent codes (and camera poses) given an image. However, these methods still require multi-view supervision during training, which constrains their usage in real-world settings, where multi-view datasets are challenging to collect.

Therefore, single-shot NeRF without additional supervision (e.g., 3D objects, multi-view image collections) remains an under-explored research direction. In this paper, we bridge this gap by incorporating an auto-encoder architecture into an existing π-GAN NeRF framework to obtain a conditional single-shot NeRF model, retaining the best properties of all components. We note that the concurrent work [31] shares similar ideas. The key differences are a different backbone network (HoloGAN [24]) and its lack of 3D consistency, which the authors point out. Contrary, we utilize the newly-proposed NeRF-based GAN method called π-GAN [2], which naturally provides stronger 3D consistency by design. We demonstrate that merely applying the approach of [31] is insufficient to obtain an accurate mapping from image to latent space with π-GAN as a backbone.

3. Method

Pix2NeRF consists of three neural networks, a Generator G, a Discriminator D, together forming a Generative Adversarial Network, and an Encoder E forming an auto-encoder together with G. The generator is conditioned on the output view pose d and a latent code z, broadly describing content variations, such as color or shape. It employs 3D-volume rendering techniques and outputs a single parameterized scene view as RGB image I. The discriminator D is a CNN, which simultaneously predicts distribution origin of the input RGB image via logit l (*real* – "in the wild", or *fake* – generated by G), and the corresponding scene pose d. The encoder E is a CNN tasked to map an input image onto the latent manifold, learned by G, and at the same time predict the input's pose:

$$\begin{aligned} G &: z, d \rightarrow I \\ D &: I \rightarrow l, d \\ E &: I \rightarrow z, d. \end{aligned} \tag{1}$$

Functionally, Pix2NeRF extends π-GAN [2] with the encoder E trained jointly with the GAN to allow mapping images back to the latent manifold. Because the encoder E disentangles the content z and the pose d of the input I, content can be further used to condition the π-GAN generator G and obtain novel views by varying the rendered pose d.

Having defined network modules, we turn to specifying the inputs and outputs of the modules. The latent code z comes from a simple prior distribution p_z (multivariate uniform in our case) – it makes sampling random codes z_{rand} easy and lets us design E such that it can encode any input image I into some z_{pred} within the support of p_z. Following prior art [2, 35], the unsupervised setting we operate in assumes we have access to the prior distribution of poses p_d of real images $I_{\text{real}} \sim p_{\text{real}}$ used for training. Depending on the dataset and choice of pose coordinates, it can be multivariate Gaussian with diagonal covariance (for images of faces) or uniform on a (hemi-)sphere (for images of cars). Parameters of this distribution must be known to allow easy sampling random poses d_{rand} for the generator, and that p_d is representative of poses of real images I_{real}.

Simply training the encoder E to map an image I into GAN latent space (as in Stage 1 of [31]) simultaneously with training GAN is challenging. This is because the encoder needs to correctly map images of the same scene from different views to a single latent code. This is especially hard when these views contain variations of fine details due to occlusions. As seen from Eq. 1 and the design Fig. 2, our method disentangles latent representation of image mapped by the encoder and generator input into content z and pose d, which undergo separate treatment.

Given an input image, Pix2NeRF disentangles pose and content and produces a radiance field of the content, which is (1) consistent with the input under the disentangled pose and (2) consistent and realistic under different poses from p_d. To achieve these properties, we devise several training objectives for (1) generator, (2) discriminator, (3) GAN in-

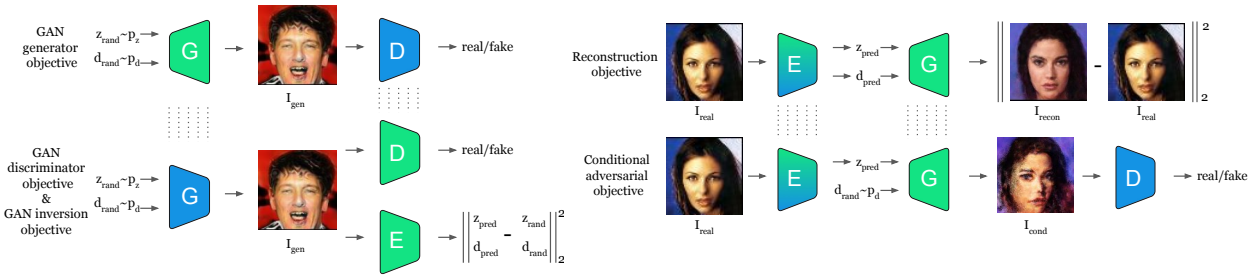

Figure 2. Overview of building blocks and objectives, used in Pix2NeRF. GAN objectives follow π-GAN [2] and ensure that NeRF outputs match the distribution of real images p_{real} under the latent prior p_z and pose prior p_d. Reconstruction and GAN inversion objectives ensure calibrated latent representations, such that E and G can operate as an auto-encoder, similar to [31]. The conditional adversarial objective enables learning better representations without explicit pose supervision. Legend: green - trained module, blue - frozen, gradient - warm-up.

version, (4) reconstruction, and (5) conditional adversarial training.

These objectives are used to compute gradients for parameters of G, D, and E within a single optimization process. However, certain parts remain "frozen" during optimizer updates (such as G during D updates and vice-versa); we denote them with an asterisk in equations (e.g., G^*) and blue color in Fig. 2. We empirically find that training encoder from the start has a detrimental effect on the whole pipeline and employ a warm-up strategy (denoted with green-blue transitions), explained further.

3.1. GAN generator objective

The generator is trained to "fool" the discriminator by serving it progressively realistic images. Pix2NeRF follows the same procedure of training the generator as π-GAN: it samples latent codes $z_{\text{rand}} \sim p_z$ and random poses $d_{\text{rand}} \sim p_d$ in pairs, which are then passed through the generator to obtain fake generated images:

$$I_{\text{gen}} = G(z_{\text{rand}}, d_{\text{rand}}), \tag{2}$$

which are further fed into the frozen discriminator:

$$l_{\text{gen}}, d_{\text{gen}} = D^*(I_{\text{gen}}). \tag{3}$$

Following [2], another component helpful to the stability and performance of GAN training is MSE supervision of predicted poses d_{gen} of images generated with d_{rand}. It penalizes the generator if the image pose recovered by the discriminator does not correspond to the sampled pose, thus setting the goal of learning a "canonical" 3D space. This is especially helpful if the pose distribution of real data is noisy, such as seen in CelebA [19].

$$\mathcal{L}_{\text{GAN}}(G) = \mathop{\mathbb{E}}_{\substack{z_{\text{rand}}\sim p_z \\ d_{\text{rand}}\sim p_d}} \Big[\text{softplus}\left(-l_{\text{gen}}\right) + \lambda_{\text{pos}} \left\|d_{\text{rand}} - d_{\text{gen}}\right\|_2^2\Big], \tag{4}$$

where λ_{pos} is a tuned weighting factor.

3.2. GAN discriminator objective

The discriminator is trained to distinguish between the generated fake samples and real data sampled from the dataset. Pix2NeRF follows the exact procedure of training the discriminator in π-GAN: it samples latent codes $z_{\text{rand}} \sim p_z$ and random poses $d_{\text{rand}} \sim p_d$ in pairs, which are then passed through the frozen generator to obtain fake generated images:

$$I_{\text{gen}} = G^*(z_{\text{rand}}, d_{\text{rand}}). \tag{5}$$

The discriminator is then trained using these generated images I_{gen} and real images $I_{\text{real}} \sim p_{\text{real}}$:

$$\begin{aligned} l_{\text{real}}, d_{\text{real}} &= D(I_{\text{real}}), \\ l_{\text{gen}}, d_{\text{gen}} &= D(I_{\text{gen}}). \end{aligned} \tag{6}$$

The discriminator objective modified to take into account MSE supervision over the known pose can then be formulated as follows:

$$\begin{aligned} \mathcal{L}_{\text{GAN}}(D) = &\mathop{\mathbb{E}}_{I_{\text{real}}\sim p_{\text{real}}} \left[\text{softplus}\left(-l_{\text{real}}\right)\right] + \\ &\mathop{\mathbb{E}}_{\substack{z_{\text{rand}}\sim p_z \\ d_{\text{rand}}\sim p_d}} \Big[\text{softplus}\left(l_{\text{gen}}\right) + \\ &\lambda_{\text{pos}} \left\|d_{\text{rand}} - d_{\text{gen}}\right\|_2^2\Big], \end{aligned} \tag{7}$$

where λ_{pos} is a tuned weighting factor.

3.3. GAN inversion objective

The encoder E is jointly optimized with the discriminator D and reuses I_{gen} computed for GAN discriminator objective Eq. (5):

$$z_{\text{pred}}, d_{\text{pred}} = E(I_{\text{gen}}). \tag{8}$$

This objective aims to ensure consistency between the sampled content and pose and those extracted from the generated

image by the encoder. This is done using the MSE loss:

$$\mathcal{L}_{\text{GAN}^{-1}}(E) = \mathop{\mathbb{E}}_{\substack{z_{\text{rand}}\sim p_z \\ d_{\text{rand}}\sim p_d}} \big[\|z_{\text{pred}} - z_{\text{rand}}\|_2^2 + \|d_{\text{pred}} - d_{\text{rand}}\|_2^2 \big]. \tag{9}$$

Up until now, the objectives only ensured a generative mapping from the latent space to radiance fields and some basic form of consistency to learn auto-encoder. However, our experiments show that optimizing just these three objectives does not produce a reasonable mapping. Therefore, Pix2NeRF adds two more objectives to address reconstruction quality and 3D consistency in the unsupervised setting.

3.4. Reconstruction objective

While the GAN inversion objective promotes consistency in latent space, nothing so far directly promotes consistency in the image space. To this end, we condition the generator G on a real image by extracting its latent code and pose prediction using the encoder, and then render its view using the predicted pose:

$$\begin{aligned} z_{\text{pred}}, d_{\text{pred}} &= E\left(I_{\text{real}}\right) \\ I_{\text{recon}} &= G\left(z_{\text{pred}}, d_{\text{pred}}\right). \end{aligned} \tag{10}$$

Ideally, we expect to get back the original image. However, using MSE loss alone in the image space is known to promote structural inconsistencies and blur. In line with [31], we employ Structural Similarity Index Measure loss (SSIM [42]) with weighting factor λ_{ssim} and a perceptual loss (VGG [44]) with weighting factor λ_{vgg}. We can therefore aggregate the reconstruction loss as follows:

$$\begin{aligned} \mathcal{L}_{\text{recon}}(G, E) = \mathop{\mathbb{E}}_{I_{\text{real}}\sim p_{\text{real}}} \Big[&\|I_{\text{recon}} - I_{\text{real}}\|_2^2 + \\ &\lambda_{\text{ssim}}\mathcal{L}_{\text{ssim}}\left(I_{\text{recon}}, I_{\text{real}}\right) + \\ &\lambda_{\text{vgg}}\mathcal{L}_{\text{vgg}}\left(I_{\text{recon}}, I_{\text{real}}\right) \Big]. \end{aligned} \tag{11}$$

3.5. Conditional adversarial objective

The reconstruction objective promotes good reconstruction quality for just one view extracted by the encoder E. This may push the combination of networks towards either predicting trivial poses or unrealistic reconstructions for other poses from p_d. To alleviate that, we further apply an adversarial objective while conditioning the generator on an image I_{real} when it is rendered from random poses. Reusing results from Eq. (10),

$$\begin{aligned} l_{\text{cond}}, d_{\text{cond}} &= D^*(G\left(z_{\text{pred}}, d_{\text{rand}}\right)) \\ \mathcal{L}_{\text{cond}}(G, E) &= \mathop{\mathbb{E}}_{\substack{I_{\text{real}}\sim p_{\text{real}} \\ d_{\text{rand}}\sim p_d}} \left[\text{softplus}\left(-l_{\text{cond}}\right)\right]. \end{aligned} \tag{12}$$

3.6. Encoder warm-up

As pointed out in [31], reconstruction loss may easily dominate and cause the model overfitting towards input views while losing its ability to represent 3D. We, therefore, introduce a simple "warm-up" strategy to counter this issue. For the first half iterations of the training protocol, we freeze the encoder while optimizing reconstruction and conditional adversarial loss and optimize only the generator for these two objectives. This serves as a warm-up for the generator to roughly learn the correspondence between encoder outputs and encoded images. The encoder is then unfrozen, enabling further distillation of its learned representations.

After the warm-up stage, the encoder and generator directly form a pre-trained auto-encoder capable of producing 3D representations close to ground truth, bypassing the cumbersome early-stage reconstruction objective, which is extremely hard to balance with GAN objectives. We show the necessity of this strategy and comparison with merely assigning a smaller weight for reconstruction loss in the ablation studies.

3.7. Training and Inference

The objectives mentioned above can be trained jointly; however, we optimize them in alternative iterations due to GPU memory constraints. The discriminator and GAN inversion objectives are optimized upon every iteration; the GAN generator objective is optimized on even iterations; reconstruction and conditional adversarial objectives are optimized jointly during odd iterations with weighting factor λ_{recon}:

$$\mathcal{L}_{\text{odd}} = \mathcal{L}_{\text{cond}} + \lambda_{\text{recon}}\mathcal{L}_{\text{recon}}. \tag{13}$$

During the inference stage, Pix2NeRF only requires a single input image, which can be fed into the encoder E and then generator G, coupled with arbitrarily selected poses for novel view synthesis. At the same time, instead of obtaining the latent code z from the encoder, it is possible to sample it from the prior distribution p_z, to make the model synthesize novel samples like a π-GAN.

4. Experiments

4.1. Evaluation

Datasets. We train and evaluate our pipeline on several 3D datasets listed below. CelebA [19] is a dataset of over 200k images of celebrity faces. We use its "aligned" version and apply center cropping to keep the face area roughly. We hold out 8k images as the test set. CARLA [8] contains 10k images of 16 car models rendered with Carla driving simulator with random textures. ShapeNet-SRN is a dataset hosted by the authors of SRN [38], from which we use the "chairs"

Figure 3. Reconstructions and novel views on CARLA [8], CelebA [19], and ShapeNet-SRN [4, 38] chairs. See Appendix for more results.

split for the comparison with prior multi-view methods. The dataset contains 50 rendered views from ShapeNet [4] with Archimedean spiral camera poses for each of the 6591 instances. As the ShapeNet-SRN dataset does not include the lower hemisphere in its validation and test sets, we filter the training set to contain only the upper hemisphere as well.

Evaluation metrics. Pix2NeRF is evaluated in two modes: unconditional, which assumes sampling directly from p_z and p_d, and conditional, which corresponds to using $z = E(I_{\text{real}}),\ I_{\text{real}} \sim p_{\text{real}}$, while still sampling from p_d. For "in the wild" datasets, as we do not possess multi-view ground truth images, we resort to reporting generative metrics: Inception Score (IS) [34], Frechet Inception Distance (FID) [15], and Kernel Inception Distance (KID) [1] with scaling factor $\times 100$ following the steps of prior works [2, 35] using the implementation [27]. To compare with multi-view-based novel view synthesis methods on Shapenet-SRN, we follow the evaluation protocols in pixelNeRF and CodeNeRF and report PSNR (Peak Signal to Noise Ratio) and SSIM (Structural Similarity Index Measure) [42].

Technical details. We choose the latent code prior distribution p_z as a multivariate uniform on $[-1, 1]$. We build our model on top of the π-GAN implementation in PyTorch [29], re-using its released generator and discriminator architectures. We also use the discriminator architecture as the backbone of our encoder, where we add a $tanh$ at the end of the latent code head. All models are optimized with Adam [17] optimizer for 300k iterations, which is approximately the same computational cost to obtain a π-GAN model. CelebA [19] models are trained with batch size 48

Method	64 × 64 FID ↓	KID ↓	IS ↑	128 × 128 FID ↓	KID ↓	IS ↑
HoloGAN [24]	-	2.87	-	39.7	2.91	1.89
GRAF [35]	-	-	-	41.1	2.29	2.34
π-GAN [2]	**5.15**	**0.09**	2.28	**14.7**	**0.39**	**2.62**
Pix2NeRF unconditional	6.25	0.16	**2.29**	14.82	0.91	2.47
Pix2NeRF conditional	24.64	1.93	2.24	30.98	2.29	2.20

Table 1. Quantitative results on CelebA [19].

Method	64 × 64 FID ↓	KID ↓	IS ↑	128 × 128 FID ↓	KID ↓	IS ↑
HoloGAN [24]	134	9.70	-	67.5	3.95	3.52
GRAF [35]	30	0.91	-	41.7	2.43	3.70
π-GAN [2]	13.59	**0.34**	3.85	29.2	**1.36**	4.27
Pix2NeRF unconditional	**10.54**	0.37	**3.95**	**27.23**	1.43	**4.38**
Pix2NeRF conditional	12.06	0.44	3.81	38.51	2.37	3.89

Table 2. Quantitative results on CARLA [8].

Method	PSNR ↑	SSIM ↑
GRF* [40]	21.25	0.86
TCO* [39]	21.27	0.88
dGQN* [12]	21.59	0.87
ENR* [11]	22.83	-
SRN** [38]	22.89	0.89
PixelNeRF* [46]	**23.72**	**0.91**
CodeNeRF** [16]	22.39	0.87
Pix2NeRF conditional	18.14	0.84

Method	FID ↓	KID ↓	IS ↑
HoloGAN [24]	-	1.54	-
π-GAN [2]	15.47	0.55	**4.62**
Pix2NeRF unconditional	**14.31**	**0.51**	**4.62**
Pix2NeRF conditional	17.55	0.59	4.36

Table 3. Quantitative results on ShapeNet-SRN [4, 38] chairs. Top: reconstruction metrics (128 × 128). Bottom: generative metrics (64 × 64). Legend: * – requires multi-view training data; ** – requires multi-view training data and test time optimization.

on resolution 64×64, where we sample 24 points per ray. We use learning rates of 2e-4, 6e-5, and 2e-4 for discriminator, generator, and encoder, respectively. For all other models, we utilized π-GAN [2]'s progressive training strategy, starting with training on resolution 32×32 with learning rates 4e-5, 4e-4, and 4e-4 for generator, discriminator, and encoder, respectively, with 96 sampled points per ray. We increase to resolution 64×64 with learning rates 2e-5, 2e-4, and 2e-4 for generator, discriminator, and encoder, respectively, and sample 72 points per ray after 50k iterations. We empirically set $\lambda_{\text{recon}} = 5$, $\lambda_{\text{ssim}} = 1$ and $\lambda_{\text{vgg}} = 1$ for all datasets. For CelebA [19], we follow [2] and set $\lambda_{\text{pos}} = 15$. For CARLA [8] and ShapeNet-SRN [4, 38], we set $\lambda_{\text{pos}} = 0$ as we do not observe significant difference. We use $|z| = 512$ for CelebA [19] and $|z| = 256$ for CARLA [8] and Shapenet-SRN [4, 38].

Quantitative results. We show the evaluation on CelebA [19] and CARLA [8] in Tables 1 and 2 respectively. We also show evaluation with the same generative metrics on ShapeNet-SRN in Table 3 (bottom). We observe that even though our model's conditional synthesis is not as good as our backbone π-GAN (especially on CelebA), it is on par with other prior 3D view generation methods [24, 35].

Since we do not explicitly enforce prior distribution p_z on the encoded samples $E(I_{\text{real}})$ from p_{real}, the image of p_{real} resulting from the encoder mapping may occupy a small portion in p_z. Thus, conditioning on p_{real} naturally leads to a smaller variation in samples from p_z, and hence, smaller diversity of NeRF outputs. For this reason, directly sampling randomly from p_z (unconditionally) achieves better performance as measured by the generative metrics. Additionally, our generator outperforms π-GAN on most metrics on CARLA [8] and ShapeNet-SRN [4, 38]. Results on CelebA [19] are less consistent due to dataset noise (background, geometry, pose noise, artifacts, *etc.*), encouraging GANs to converge towards the mean as a trade-off to variations. These observations can be related to manifold learning [9], where we enforce the existence of a latent code for each real image in the train set.

We compare our method with other single-image 3D inference methods in Table 3 on ShapeNet-SRN [4, 38] in 128 × 128 resolution. Since our model assumes a strictly-spherical camera parameterization model, which does not correspond well to the ground truth poses of ShapeNet-SRN [4, 38], we use our encoder to extract poses from the images.

Despite being generative, unsupervised, and not requiring test time optimization in contrast to all other methods, our model's performance does not drop much below the competition. Considering that other models were trained on 128, while our models were trained on 64 × 64 but rendered at 128 × 128 resolution, we observe a super-resolution effect.

Qualitative results. We show some qualitative results of our model's performance on CARLA [8] and CelebA [19] in Fig. 3. We can see that our model can synthesize novel views with good quality while existing few-shot NeRF methods [16, 40, 46] are not able to train on these "in the wild" datasets due to the lack of multi-view supervision. Our model can also produce decent 3D representations even under extreme poses and artifacts (see row 5).

4.2. Ablation studies

We perform a thorough ablation study to verify our design choices by removing the key components one by one and training models under identical settings as the full model. Qualitative results for the following ablations are in Fig. 4; refer to Appendix for the corresponding quantitative results.

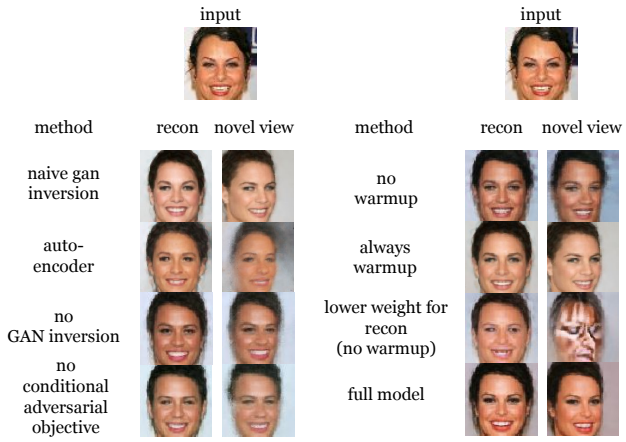

Figure 4. Qualitative results of ablation studies, obtained with an image from the test split of CelebA [19]. λ_{recon} is set to 1 for lower reconstruction weights instead of the warm-up ablation. See Appendix for results obtained by using other λ_{recon} values.

Naive GAN inversion. We compare Pix2NeRF with naive GAN inversion: having a pre-trained GAN, we freeze its weights and train an encoder to map images to their corresponding latent codes. The results show that the encoder can learn an approximate mapping from images to latent code. However, due to the lack of joint distillation, the reconstruction is off from the input image.

Auto-encoder. Another potential approach is to utilize π-GAN's architecture as an auto-encoder, in which the latent space is dropped from the pipeline and training the reconstruction and conditional adversarial objectives only. Under this setting, while the reconstruction achieves decent quality, we can observe visible 3D inconsistency, suggesting difficulty of optimization with the remaining objectives.

No GAN inversion. We proceed with ablations by removing the GAN inversion step from the pipeline. The visual results turn out to be blurry and uncanny compared with full settings. One possible explanation is that this step is a connection between π-GAN training and reconstruction, which significantly affects the overall performance.

No conditional adversarial objective. We further deactivate the conditional adversarial loss and retrain the model. As a result, the renderings become incomplete and have clear visual artifacts. In addition, 3D consistency degrades significantly, which justifies this objective in the given setting.

Warm-up. To verify the effect of the warm-up strategy, we train three separate models and compare their performances: without warm-up, without unfreezing encoder (always warm-up), and assigning a lower weight for reconstruction instead of the warm-up. Without the warm-up strategy, the model tends to overfit the input view and cannot produce meaningful content from novel poses. If we only use the warm-up strategy and never unfreeze the encoder, the distillation is relatively weak, which results in few fine details. With lower reconstruction weight instead of the warm-up, the balance between reconstruction and adversarial objective is missing, resulting in mode collapse for novel view synthesis.

5. Conclusions

In this paper, we introduced Pix2NeRF, a novel unsupervised single-shot framework capable of translating an input image of a scene into a neural radiance field (NeRF), thereby performing single-shot novel view synthesis. The key idea of Pix2NeRF is to utilize generative NeRF models to interpolate missing geometry information. This is accomplished by jointly training an encoder that maps images to a latent space, which disentangles content and pose, and the generative NeRF model while keeping these two parts dependent on each other. Pix2NeRF can go beyond the auto-encoder setting and perform novel scene generation by sampling random content and pose and passing through the generator. Our framework demonstrates high reconstruction quality and 3D consistency, on par and better than previous works.

Limitations and future work. The current setting in consideration is limited to one category per dataset and cannot directly generalize beyond the chosen category. Alternative research directions include local conditional fields similar to PixelNeRF [46] and GRF [40], which can generalize to unseen categories, multi-instance, and even real-world scenes. Being a general framework, Pix2NeRF is not limited to using π-GAN as its backbone. Newer generative NeRF models, e.g. EG3D [3] could potentially achieve better visual quality. Additionally, architecture search, especially with respect to the encoder remains a challenging problem. Utilizing more mature encoder architectures from 2D GAN feed-forward inversion literature, e.g. pixel2style2pixel [32], could potentially improve the performance of Pix2NeRF significantly.

Ethical consideration. As with most modern conditional generative models, Pix2NeRF can be misused by generating content to spread misinformation or perform targeted attacks. The growing popularity of deepfake celebrity accounts in social media suggests that new use cases, markets, and novel ways of monetizing this kind of data will follow.

Acknowledgement. We thank Eric R. Chan for generously sharing the implementation of π-GAN, giving helpful suggestions and clarification throughout the project. Anton Obukhov is funded by Toyota Motor Europe via TRACE Zürich.

References

[1] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans, 2021. 6, 1

[2] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *arXiv*, 2020. 2, 3, 4, 6, 7

[3] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *arXiv*, 2021. 8

[4] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015. 6, 7, 1, 4

[5] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo, 2021. 2

[6] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[7] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7911–7920, June 2021. 2

[8] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 5, 6, 7, 1

[9] Yilun Du, Katherine M. Collins, Joshua B. Tenenbaum, and Vincent Sitzmann. Learning signal-agnostic manifolds of neural fields, 2021. 7

[10] Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio. Feature-wise transformations. *Distill*, 2018. https://distill.pub/2018/feature-wise-transformations. 3

[11] Emilien Dupont, Miguel Bautista Martin, Alex Colburn, Aditya Sankar, Josh Susskind, and Qi Shan. Equivariant neural rendering. In *International Conference on Machine Learning*, pages 2761–2770. PMLR, 2020. 7

[12] SMA Eslami, DJ Rezende, F Besse, F Viola, AS Morcos, M Garnelo, A Ruderman, AA Rusu, I Danihelka, K Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–+, 2018. 7

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Y. Bengio. Generative adversarial networks. *Advances in Neural Information Processing Systems*, 3, 06 2014. 2

[14] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis, 2021. 3

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500, 2017. 6, 1

[16] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories, 2021. 2, 3, 7

[17] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014. 6

[18] Chen-Hsuan Lin, Chaoyang Wang, and Simon Lucey. Sdf-srn: Learning signed distance 3d object reconstruction from static images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 1

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 4, 5, 6, 7, 8, 1, 2, 3

[20] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *CVPR*, 2021. 2

[21] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. *arXiv preprint arXiv:2103.15606*, 2021. 2

[22] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[23] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2

[24] Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *The IEEE International Conference on Computer Vision (ICCV)*, Nov 2019. 2, 3, 7

[25] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[26] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[27] Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. High-fidelity performance metrics for generative models in pytorch, 2020. Version: 0.3.0, DOI: 10.5281/zenodo.4957738. 6

[28] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703, 2019. 6

[30] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. *CoRR*, abs/1709.07871, 2017. 3

[31] Pierluigi Zama Ramirez, Alessio Tonioni, and Federico Tombari. Unsupervised novel view synthesis from a single image. *CoRR*, abs/2102.03285, 2021. 2, 3, 4, 5, 1

[32] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 8

[33] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[34] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *CoRR*, abs/1606.03498, 2016. 6, 1

[35] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3, 6, 7

[36] Vincent Sitzmann, Eric R. Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. Metasdf: Meta-learning signed distance functions. In *arXiv*, 2020. 2

[37] Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Proc. NeurIPS*, 2020. 3

[38] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, 2019. 2, 5, 6, 7, 1, 4

[39] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3d models from single images with a convolutional network, 2016. 7

[40] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d scene representation and rendering. In *arXiv:2010.04595*, 2020. 2, 3, 7, 8

[41] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 2

[42] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5, 6, 1

[43] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF–: Neural radiance fields without known camera parameters. *https://arxiv.org/abs/2102.07064*, 2021. 2

[44] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *CVPR*, 2020. 2, 5

[45] Xudong Xu, Xingang Pan, Dahua Lin, and Bo Dai. Generative occupancy fields for 3d surface-aware image synthesis. In *Advances in Neural Information Processing Systems(NeurIPS)*, 2021. 3

[46] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 2, 3, 7, 8

[47] Yuxuan Zhang, Wenzheng Chen, Huan Ling, Jun Gao, Yinan Zhang, Antonio Torralba, and Sanja Fidler. Image gans meet differentiable rendering for inverse graphics and interpretable 3d neural rendering. In *International Conference on Learning Representations*, 2021. 2

[48] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis, 2021. 3