Escaping Plato’s Cave using Adversarial Training: 3D Shape From Unstructured 2D Photo Collections

Philipp Henzler
p.henzler@cs.ucl.ac.uk

Niloy Mitra
n.mitra@cs.ucl.ac.uk

Tobias Ritschel
t.ritschel@ucl.ac.uk

University College London

Abstract

We introduce PLATONIC GAN to discover the 3D structure of an object class from an unstructured collection of 2D images, i.e., neither any relation between the images is available nor additional information about the images is known. The key idea is to train a deep neural network to generate 3D shapes which rendered to images are indistinguishable from ground truth images (for a discriminator) under various camera models (i.e., rendering layers) and camera poses. Discriminating 2D images instead of 3D shapes allows tapping into unstructured 2D photo collections instead of relying on curated (e.g., aligned, annotated, etc.) 3D data sets. To establish constraints between 2D image observation and their 3D interpretation, we suggest a family of rendering layers that are effectively differentiable. This family includes visual hull, absorption-only (akin to x-ray), and emission-absorption. We can successfully reconstruct 3D shapes from unstructured 2D images and extensively evaluate PLATONIC GAN on a range of synthetic and real data sets achieving consistent improvements over baseline methods. We can also show that our method with additional 3D supervision further improves result quality and even surpasses the performance of 3D supervised methods.

1. Introduction

A key limitation to current generative models [37, 36, 12, 24, 32, 31] is the availability of suitable training data (e.g., 3D volumes, feature point annotations, template meshes, deformation prior, structured image sets, etc.) for supervision.

While methods exist to learn the 3D structure of classes of objects, they typically require 3D data as input. Regrettably, such 3D data is difficult to acquire, in particular for the “long tail” of exotic classes: ShapeNet might have chair, but it does not have chanterelle or tree.

Addressing this problem, we suggest a method to learn 3D structure from 2D images only (Fig. 1). Reasoning about the 3D structure from 2D observations without assuming anything about their relation is challenging as illustrated from Plato’s Allegory of the Cave [34]: How can we hope to understand higher dimensions from only seeing projections? If multiple views (maybe only two [40, 13]) of the same object are available, multi-view analysis without 3D supervision has been successful, but regrettably most photo collections do not come in this form but are now and will remain unstructured: they show random instances under random pose, uncalibrated lighting in unknown relations, and multiple views of the same objects are not available.

Our first main contribution (Sec. 3) is to use adversarial training of a 3D generator with a discriminator that operates exclusively on widely available unstructured collections of 2D images, which we call platonic discriminator. Here, during training, the generator produces a 3D shape that is projected (rendered) to 2D and presented to the 2D platonic discriminator. Making a connection between the 3D generator and the 2D discriminator, our second key contribution, is enabled by a family of rendering layers that can account for occlusion and color (Sec. 4). These layers do not need any

Figure 1. PLATONIC GANs allow converting an unstructured collection of 2D images of a rare class (subset shown on top) into a generative 3D model (random samples below).
learnable parameters and allow for backpropagation [26]. From these two key blocks we construct a system that learns the 3D shapes of common classes such as chairs and cars, but also exotic classes from unstructured 2D photo collections. We demonstrate 3D reconstruction from a single 2D image as a key application (Sec. 5). While recent works focus on using as little explicit supervision [17, 19, 8, 29, 28, 11] as possible, they all rely on either annotations, 3D templates, known camera poses, specific views or multi-view images during training. Our approach takes it a step further by receiving no such supervision, see Tbl. 1.

Table 1. Taxonomy of different methods that learn 3D shapes with no explicit 3D supervision. We compare Kanazawa et al. [17], Kato et al. [19], Eslami et al. [8], Tulsiani et al. [29], Tulsiani et al. [28], PrGan [11] with our method in terms of degree of supervision.

<table>
<thead>
<tr>
<th>Supervision at training time</th>
<th>Annotation-free</th>
<th>3D Template-Free</th>
<th>Unknown camera pose</th>
<th>No pre-defined camera poses</th>
<th>Only single view required</th>
<th>Color</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Annotation-free</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>3D Template-Free</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Unknown camera pose</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>No pre-defined camera poses</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Only single view required</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Color</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

2. Related Work

Several papers suggest (adversarial) learning using 3D voxel representations [37, 36, 12, 24, 11, 32, 31, 35, 39, 30, 20] or point cloud input [1, 10]. The general design of such networks is based on an encoder that generates a latent code which is then fed into a generator to produce a 3D representation (i.e., a voxel grid). A 3D discriminator now analyzes samples both from the generator and from the ground truth distribution. Note that this procedure requires 3D supervision, i.e., is limited by the type and size of the 3D data set such as ShapeNet [5].

Girdhar et al. [12] work on a joint embedding of 3D voxels and 2D images, but still require 3D voxelizations as input. Fan et al. [9] produce points from 2D images, but similarly with 3D data as training input. Gadelhan et al. [11] use 2D visual hull images to train a generative 3D model. Cho et al.’s recursive design takes multiple images as input [6] while also being trained on 3D data. Kar et al. [18] propose a simple “unprojection” network component to establish a relation between 2D pixels and 3D voxels but without resolving occlusion and again with 3D supervision.

Cashman and Fitzgibbon [4] and later Carreira et al. [3] or Kanazawa et al. [17] use correspondence to 3D templates across segmentation- or correspondence-labeled 2D image data sets to reconstruct 3D shapes. These present stunning results, for example on animals, but at the opposite end of a spectrum of manual human supervision, in which our approach receives no such supervision.

Closer to our approach is Rezende et al. [25] that also learn 3D representations from single images. However, they make use of a partially differentiable renderer [22] that is limited to surface orientation and shading, while our formulation can resolve both occlusion from the camera and appearance. Also, their representation of the 3D volume is a latent one, that is, it has no immediate physical interpretation that is required in practice, e.g., for measurements, to run simulations such as renderings or 3D printing. This choice of having a deep representation of the 3D world is shared by Eslami et al. [8]. Tulsiani et al. [29] reconstruct 3D shape supervised by multiple 2D images of the same object with known view transformations at learning time. Tulsiani et al. [28] take it a step further and require no knowledge about the camera pose, but still require multiple images of the same object at training time. They have investigated modelling image formation as sums of voxel occupancies to predict termination depth. We use a GAN to train on photo collections which typically only show one view of each instance. Closest to our work is Gadelha et al. [11] which works on an unstructured set of visual hull images and receives three sources of supervision: view information gets explicitly encoded as a dimension in the latent vector; views come from a manually-chosen 1D subspace (circle); and there are only
8 discrete views. We take the image formation a step further to support absorption-only and emission-absorption image formation, allowing to learn from real photos and do so on unstructured collections from-the-wild where no view supervision is available.

While early suggestions how to extend differentiable renderers to polygonal meshs exist, they are limited to deformation of a pre-defined template [19]. We work with voxels, which can express arbitrary topology, e.g., we can generate airplanes with drastically different layout, which are not a mere deformation of a base shape.

Similarly, inter-view constraints can be used to learn depth maps [40, 13] using reprojection constraints: If the depth label is correct, reprojecting one image into the other view has to produce the other image. Our method does not learn a single depth map but a full voxel grid and allows principled handling of occlusions.

A generalization from visual hull maps to full 3D scenes is discussed by Yan et al. [38]. Instead of a 3D loss, they employ a simple projection along major axis allowing to use mere deformation of a base shape.

The two are linked by a fixed-function projection operator, i.e., non-learnable (see Sec. 4).

Let us recall the classic adversarial learning of 3D shapes [36], which is a min-max game

$$\min_{\Theta} \max_{\Psi} \ c_{\text{Dis}}(\Psi) + c_{\text{Gen}}(\Theta)$$  \hspace{1cm} (1)

between the discriminator and the generator cost, respectively $c_{\text{Dis}}$ and $c_{\text{Gen}}$.

The discriminator cost is

$$c_{\text{Dis}}(\Psi) = \mathbb{E}_{p_{\text{Data}}(x)}[\log(D_{\Psi}(x))]$$  \hspace{1cm} (2)

where $D_{\Psi}$ is the discriminator with learned parameters $\Psi$ which is presented with samples $x$ from the distribution of real 3D shapes $x \sim p_{\text{Data}}$. Here $\mathbb{E}_p$ denotes the expected value of the distribution $p$.

The generator cost is

$$c_{\text{Gen}}(\Theta) = \mathbb{E}_{p_{\text{Gen}}(z)}[\log(1 - D_{\Psi}(G_{\Theta}(z)))]$$  \hspace{1cm} (3)

where $G_{\Theta}$ is the generator with parameters $\Theta$ that maps the latent code $z \sim p_{\text{Gen}}$ to the data domain.

**PLATONICGAN** The discriminator cost is calculated identical to the common GAN with the only difference that the input samples are rendered 2D images with generation cost

$$c_{\text{Gen}}(\Theta) = \mathbb{E}_{p_{\text{Gen}}(z)}\mathbb{E}_{p_{\text{view}}(\omega)}[\log(1 - D_{\Psi}(R(\omega, G_{\Theta}(z))))]$$  \hspace{1cm} (4)

where $R$ projects the generator result $G_{\Theta}(z)$ from 3D to 2D along the sampled view direction $\omega$. See Sec. 3.1 for details.

While many parameterizations for views are possible, we choose an orthographic camera with fixed upright orientation that points at the origin from an Euclidean position $\omega \in S^2$ on the unit sphere. $\mathbb{E}_{p_{\text{view}}(\omega)}$ is the expected value across the distributions $\omega \sim p_{\text{view}}$ of views.

**PLATONICGAN 3D Reconstruction** Two components in addition to our platonic concept are required to allow for 3D reconstruction, resulting in

$$\min_{\Psi} \max_{\Theta, \Phi} c_{\text{Disc}}(\Psi) + c_{\text{Gen}}(\Theta, \Phi) + \lambda c_{\text{Rec}}(\Theta, \Phi)$$  \hspace{1cm} (5)

where $c_{\text{Gen}}$ includes an encoding step and $c_{\text{Rec}}$ encourages the encoded generated-and-projected result to be similar to the encoder input where $\lambda = 100$. We detail both of these steps in the following paragraphs:

**Generator** The generator $G_{\Theta}$ does not directly work on a latent code $z$, but allows for an encoder $E_{\Phi}$ with parameters $\Psi$ that encodes a 2D input image $I$ to a latent code $z = E_{\Phi}(I)$. The cost becomes,

$$c_{\text{Gen}}(\Theta, \Phi) = \mathbb{E}_{p_{\text{Data}}(I)}\mathbb{E}_{p_{\text{view}}(\omega)}[\log(1 - D_{\Psi}(R(\omega, G_{\Theta}(E_{\Phi}(I))))])$$  \hspace{1cm} (6)

**Reconstruction** We encourage the encoder $E_{\Phi}$ and generator $G_{\Theta}$ to reproduce the input in the $L_2$ sense: by convention the input view is $\omega_0 = (0, 0)$,

$$c_{\text{Rec}}(\Theta, \Phi) = \| y - R(\omega_0, G_{\Theta}(E_{\Phi}(I))) \|^2_2$$  \hspace{1cm} (7)

While this step is not required for generation it is mandatory for reconstruction. Furthermore, it adds stability to the optimization as it is easy to find an initial solution that matches this 2D cost before refining the 3D structure.

### 3.1. Optimization

Two key properties are essential to successfully optimize our PLATONICGAN: First, maximizing the expected
value across the distribution of views $p_{\text{view}}$ and second, back-propagation through the projection operator $R$. We extend the classic GAN optimization procedure in Alg. 1.

**Projection**
We focus on the case of a 3D generator on a regular voxel grid $v_{n_c \times res^2}$ and a 2D discriminator on a regular image $T_{n_c \times res^2}$ where $n_c$ denotes the number of channels and $res = 64$ corresponds to the resolution. In section 4, we discuss three different projection operators. We use $R(\omega, v)$ to map a 3D voxel grid $v$ under a view direction $\omega \in S^2$ to a 2D image $I$.

We further define $R(\omega, v) := \rho(T(\omega)v)$ with rotation matrix $T(\omega)$ according to the view direction $\omega$ and an image formation function $\rho(v)$ that is view-independent. The same transformation is shared by all implementations of the rendering layer, so we will only discuss the key differences of $\rho$ in the following. Note that a rotation and a linear resampling is back-propagatable and typically provided in a deep learning framework, e.g., as torch.nn.functional.grid_sample in PyTorch [23]. While we work in orthographic space, $\rho$ could also model a perspective transformation.

**View sampling** We assume uniform view sampling.

**4. Rendering Layers**
Rendering layers (Fig. 3) map 3D information to 2D images so they can be presented to a discriminator. We first rotate the 3D volume (Fig. 3, a) into camera space from view direction $\omega$ (Fig. 3, b), such that the pixel value $p$ is to be computed from all voxel values $v_i$ and only those (Fig. 3, c). The rendering layer maps a sequence of $n_v$ voxels to a pixel value $\rho(v) \in \mathbb{R}^{n_c \times res^2} \rightarrow \mathbb{R}^{n_c \times res^2}$. Composing the full image $I$ just amounts to executing $\rho$ for every pixel $p$ resp. all voxels $v = v_1, \ldots, v_{n_v}$ at that pixel.

Note, that the rendering layer does not have any learnable parameters. We will now discuss several variants of $\rho$, implementing different forms of volume rendering [7].

**Visual hull (VH)** Visual hull [21] is the simplest variant (Fig. 4). It converts scalar density voxels into binary opacity images. A voxel value of 0 means empty space and a value of 1 means fully occupied, i.e., $v_i \in [0, 1]$. Output is a binary value indicating if any voxel blocked the ray. It is approximated as

$$\rho_{\text{VH}}(v) = 1 - \sum_i (-v_i).$$

Note that the sum operator can both be back-propagated and is efficiently computable on a GPU using a parallel scan. We can apply this to learn 3D structure from binary 2D data such as segmented 2D images.

**Absorption-only (AO)** The absorption-only model is the gradual variant of visual hull. This allows for “softer” attenuation of rays. It is designed as:

$$\rho_{\text{AO}}(v) = 1 - \prod_i (1 - v_i).$$

If $v_i$ are fractional the result is similar to an x-ray, i.e., $v_i \in [0, 1]$. This image formation allows learning from x-rays or other transparent 2D images. Typically, these are single-channel images, but a colored variant (e.g., x-ray at different wavelength or RGB images of colored transparent objects) could technically be done.

**Emission-absorption (EA)** Emission-absorption allows the voxels not only to absorb light coming towards the observer but also to emit new light at any position. This interplay of emission and absorption can model occlusion, which we will see is useful to make 3D sense of a 3D world. Fig. 3 uses emission-absorption with high absorption, effectively realizing an opaque surface with visibility.

A typical choice is to have the absorption $v_a$ monochromatic and the emission $v_e$ chromatic.

---

**Algorithm 1** PLATONIC-GAN Reconstruction Update Step

1. $I \leftarrow \text{SAMPLEIMAGE}(p_{\text{Dat}})$
2. $\omega \leftarrow \text{SAMPLEVIEW}(p_{\text{view}})$
3. $z \leftarrow \text{E}(I)$
4. $v \leftarrow \text{G}(z)$
5. $\text{rendering} \leftarrow R(\omega, v)$
6. $\text{renderingFront} \leftarrow R(\omega_0, v)$
7. $c_{\text{dis}} \leftarrow \log(D(I) + \log(1 - D(\text{rendering})))$
8. $c_{\text{gen}} \leftarrow \log(1 - D(\text{rendering}))$
9. $c_{\text{rec}} \leftarrow \text{L2}(I - \text{rendering})$
10. $\Psi \leftarrow \text{MAXIMIZE}(c_{\text{dis}})$
11. $\Theta, \Phi \leftarrow \text{MINIMIZE}(c_{\text{gen}} + \lambda c_{\text{rec}})$

---

**Figure 3. Rendering layers (Please see text).**

**Figure 4. Different image formation models visual hull (VH), absorption-only (AO) and emission-absorption (EA).**
The complete emission-absorption equation is

\[
\rho_{EA}(v) = \sum_{i=1}^{n_v} (1 - \prod_{j=1}^{n} (1 - v_{a,j})) v_{e,i} \tag{10}
\]

While such equations are typically solved using ray-marching [7], they can be rewritten to become differentiable in practice: First, we note that the transmission \( t_i \) from voxel \( i \) is one minus a product of one minus the density of all voxels before \( i \). Similar to a sum such a cumulative product can be backpropagated and computed efficiently using parallel scans, e.g., using \( \text{torch.cumprod} \). A numerical alternative, that performed similar in our experiments, is to work in the log domain and use \( \text{torch.cumsum} \).

5. Evaluation

Our evaluation comprises of a quantitative (Sec. 5.4) and a qualitative analysis (Sec. 5.5) that compares different techniques (Sec. 5.2).

5.1. Data sets

Synthetic We evaluate on two synthetic data sets: (a) ShapeNet [5] and (b) mammalian skulls [16]. For our quantitative analysis, we use ShapeNet models as 3D ground truth is required. 2D images of 3D shapes are rendered for the three image formation models VH, AO, EA. Each shape is rendered from a random view (50 per object), with random natural illumination. ShapeNet only provides 3D density volumes which is not sufficient for EA analysis. To this end, we use volumetric projective texturing to propagate the appearance information from thin 3D surface crust as defined by ShapeNet’s textures into the 3D voxelization in order to retrieve RGBA volumes where \( A \) corresponds to density. We use shapes from following classes: airplane, car, chair, rifle and lamp. The same train / val / test split as proposed by [5] is adopted.

We also train on a synthetic x-ray data set that consists of 466,200 mammalian skull x-rays [16]. We used the monkey skulls subset of that data set (~30k x-rays).

Real We use two data sets of rare classes: (a) chanterelle (60 images) and (b) tree (37 images) (not strictly rare, but difficult to 3D-model). These images are RGBA, masked, on white background. Note, that results on these input data has to remain qualitative, as we lack the 3D information to compare to and do not even have a second view of the same object to even perform an image comparison.

5.2. Baselines and comparison

2D supervision First, we compare the publicly available implementation of PrGAN [11] with our platonic method. PrGAN is trained on an explicitly created data set adhering to their view restrictions (8 views along a single axis). Compared to our method it is only trained on visual hull images, however for evaluation purposes absorption-only and emission-absorption (in form of luminance) images are used as input images at test time. Note that PrGAN allows for object-space view reconstruction due to view information in the latent space whereas our method performs reconstruction in view-space. Due to the possible ambiguities in the input images (multiple images can belong to the same 3D volume) we transform our results into object-space using a Monte-Carlo approach to find the best transformation according to the ground truth volume.

3D supervision The first baseline with 3D supervision is MULTI-VIEW, that has multiple images of the same object [38] in a known spatial relation. Note, that this is a stronger requirement to PLATONIC-GAN that does not require any structure in the adversarial examples: geometry, view, light, all change, while in this method only the view changes in a prescribed way. The second competitor is a classic 3DGAN [36] trained with Wasserstein GAN loss [2] and a gradient penalty [15]. To compare PLATONIC-GAN against methods having access to 3D information, we also propose a variant PLATONIC3D by adding the PLATONIC-GAN loss terms (for all the available images) to the 3DGAN framework.

5.3. Evaluation Metrics

2D evaluation measures Since lifting 2D information to 3D can be ambiguous, absolute 3D measures might not be the best suitable measures for evaluation on our task. For instance, a shift in depth of an object under an orthographic camera assumption will result in a higher error for metrics in 3D, but the shift would not have any effect on a rendered image. Thus, we render both the reconstructed and the reference image from the same 10 random views and compare their images using SSIM / DSSIM [33] and VGG16 [27] features. For this re-rendering, we further employ four different rendering methods: the original (i.e., \( \rho \)) image formation (IF), volume rendering (VOL), iso-surface rendering with an instance, a shift in depth of an object under an orthographic camera assumption will result in a higher error for metrics in 3D, but the shift would not have any effect on a rendered image. Thus, we render both the reconstructed and the reference image from the same 10 random views and compare their images using SSIM / DSSIM [33] and VGG16 [27] features. For this re-rendering, we further employ four different rendering methods: the original (i.e., \( \rho \)) image formation (IF), volume rendering (VOL), iso-surface rendering with an iso-value of .1 (ISO) and a voxel rendering (VOX), all under random natural illumination.

3D evaluation measures We report root-mean-squared-error (RMSE), intersection-over-union (IoU) and chamfer distance (CD). For the chamfer distance we compute a weighted directional distance:

\[
d_{CD}(T, O) = \frac{1}{N} \sum_{p_i \in T} \min_{p_j \in O} w_j \|p_i - p_j\|_2^2,
\]

where \( T \) and \( O \) correspond to output and target volumes respectively, and \( w_j \) denotes the density value of the voxel at location \( p_j \). The weighting makes intuitive sense as our results have scalar values rather than binary values, i.e., higher
Table 2. Performance of different methods with varying degrees of supervision (superv.) (rows) on different metrics (columns) for the class airplane (200 shapes from the test set are used). Evaluation is performed on all three image formations (IF): visual hull (VH), absorption-only (AO) and emission-absorption (EA). Note, DSSIM and VGG values are multiplied by 10, RMSE by $10^2$ and CD by $10^3$. Lower is better except for IoU where higher is better.

<table>
<thead>
<tr>
<th>Method</th>
<th>IF Superv.</th>
<th>2D Image Re-synthesis</th>
<th>3D Volume</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>VH DSSIM VGG</td>
<td>AO DSSIM VGG</td>
</tr>
<tr>
<td>PrGAN [11]</td>
<td>vis. hull</td>
<td>2D 1.55 6.57</td>
<td>1.37 4.85</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td>2D 1.14 5.37</td>
<td>1.16 4.93</td>
</tr>
<tr>
<td>Multi-view [38]</td>
<td></td>
<td>3D 0.87 4.89</td>
<td>0.80 4.31</td>
</tr>
<tr>
<td>3DGAN [36]</td>
<td>vis. hull</td>
<td>3D 0.83 5.01</td>
<td>0.75 4.02</td>
</tr>
<tr>
<td>Ours 3D</td>
<td>vis. hull</td>
<td>2D+3D 0.81 4.82</td>
<td>0.77 3.98</td>
</tr>
<tr>
<td>PrGAN [11]</td>
<td>abs.-only</td>
<td>2D 1.41 6.40</td>
<td>1.27 4.80</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td>2D 0.94 5.35</td>
<td>0.93 4.46</td>
</tr>
<tr>
<td>Multi-view [38]</td>
<td></td>
<td>3D 0.95 4.99</td>
<td>0.78 4.23</td>
</tr>
<tr>
<td>3DGAN [36]</td>
<td>abs.-only</td>
<td>3D 0.67 4.37</td>
<td>0.69 3.77</td>
</tr>
<tr>
<td>Ours 3D</td>
<td>abs.-only</td>
<td>2D+3D 0.66 4.36</td>
<td>0.66 3.73</td>
</tr>
<tr>
<td>PrGAN [11]</td>
<td>em.-abs.</td>
<td>2D 1.31 6.22</td>
<td>1.15 4.77</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td>2D 2.18 6.35</td>
<td>1.99 5.38</td>
</tr>
<tr>
<td>Multi-view [38]</td>
<td></td>
<td>3D 1.62 6.21</td>
<td>1.53 4.58</td>
</tr>
<tr>
<td>3DGAN [36]</td>
<td>em.-abs.</td>
<td>3D 0.89 5.28</td>
<td>0.78 3.93</td>
</tr>
<tr>
<td>Ours 3D</td>
<td>em.-abs.</td>
<td>2D+3D 0.82 4.71</td>
<td>0.82 3.96</td>
</tr>
</tbody>
</table>

In Tbl. 3 we compare the mean VGG error of a vanilla 3D GAN trained only on 3D shapes, a platonic approach accessing only 2D images, and PLATONIC3D that has access to both. We keep the number of 2D images fixed, and

**3D supervision** Comparing our PLATONIC3D variant to the 3D baselines we observe our method to mostly outperform them for 2D metrics. Not surprisingly our method performs worse for 3D metrics as our platonic approach only operates in 2D.

In Tbl. 3 we look into the performance across different classes. rifle performs best: the approach learns quickly from 2D that a gun has an outer 3D shape that is a revolute structure. chair performs worst likely due to its high intra-class variation.

Table 3. Reconstruction performance of our method for different image formation models (columns) on different classes (rows). The error metric is SSIM (higher is better).

<table>
<thead>
<tr>
<th>Class</th>
<th>VH VOL</th>
<th>AO ISO</th>
<th>VOX VOL</th>
<th>ISO VOX</th>
<th>VOX ISO VOX</th>
</tr>
</thead>
<tbody>
<tr>
<td>plane</td>
<td>0.93 0.92</td>
<td>0.93 0.94</td>
<td>0.93 0.93</td>
<td>0.93 0.85</td>
<td>0.76 0.77</td>
</tr>
<tr>
<td>rifle</td>
<td>0.95 0.94</td>
<td>0.95 0.94</td>
<td>0.95 0.94</td>
<td>0.95 0.90</td>
<td>0.78 0.80</td>
</tr>
<tr>
<td>chair</td>
<td>0.86 0.85</td>
<td>0.85 0.86</td>
<td>0.85 0.86</td>
<td>0.86 0.80</td>
<td>0.61 0.63</td>
</tr>
<tr>
<td>car</td>
<td>0.841 0.846</td>
<td>0.851 0.844</td>
<td>0.846 0.850</td>
<td>0.850 0.800</td>
<td>0.731 0.743</td>
</tr>
<tr>
<td>lamp</td>
<td>0.920 0.915</td>
<td>0.920 0.926</td>
<td>0.914 0.920</td>
<td>0.920 0.883</td>
<td>0.790 0.803</td>
</tr>
</tbody>
</table>

3D supervision

The error metric is SSIM (higher is better).

5.4. Quantitative evaluation

Tbl. 2 summarizes our main results for the airplane class. Concerning the image formation models, we see that the overall values are best for AO, which is expected: VH asks for scalar density but has only a binary image; AO provides internal structures but only needs to produce scalar density; EA is hardest, as it needs to resolve both density and color. Nonetheless the differences between us and competitors are similar across the image formation models.

**2D supervision** We see that overall, our 2D supervised method outperforms PrGAN for VH and AO. Even though PrGAN was not trained on EA it wins for all metrics against our 2D supervised method. However, it even outperforms the 3D supervised methods 3DGAN and MULTI-VIEW which demonstrates the complexity of the task itself. However, PrGAN for EA only produces density voxel volumes unlike all other methods that produce RGBA volumes. Comparing our 2D supervised method against the 3D supervised methods we see that overall our method produces competitive results. In the case of MULTI-VIEW we sometimes even perform better.

**densities get penalized more, and \( N \) is the total number of voxels in the volume. We give preference to such a weighting opposed to finding a threshold value for binarization.**
Table 4. Effect of number of 3D shapes and 2D images on learning different GANs methods in terms of mean DSSIM error. Lower is better.

<table>
<thead>
<tr>
<th></th>
<th>2D images</th>
<th>3D shapes</th>
<th>2D-3D ratio</th>
<th>DSSIM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>70k</td>
<td>5</td>
<td>14k</td>
<td>.135</td>
</tr>
<tr>
<td></td>
<td>70k</td>
<td>50</td>
<td>1.4k</td>
<td>.108</td>
</tr>
<tr>
<td></td>
<td>70k</td>
<td>250</td>
<td>280</td>
<td>.106</td>
</tr>
<tr>
<td></td>
<td>70k</td>
<td>1.5k</td>
<td>47</td>
<td>.101</td>
</tr>
<tr>
<td>3D</td>
<td>.125</td>
<td>.125</td>
<td>.125</td>
<td>.125</td>
</tr>
<tr>
<td>Ours</td>
<td>.134</td>
<td>.108</td>
<td>.102</td>
<td>.099</td>
</tr>
</tbody>
</table>

Figure 5. Visual results for 3D reconstruction of three classes (airplane, chair, rifle) from multiple views.

vary the number of 3D shapes available. The number of shapes increases along the horizontal axis. We see that a PLATONICGAN (red line) can beat both other approaches in a condition where little 3d data is available (left). When more 3D data is available, the 3D GAN (blue line) and wins over PLATONICGAN. We conclude that adding image information to a 3D corpus helps, and when the corpus is small enough, even performs better than 3D supervised methods. Qualitative results can be seen in Fig. 6.

5.5. Qualitative

**Synthetic** Fig. 5 shows typical results for the reconstruction task. We see that our reconstruction can produce airplane 3D models representative of the input 2D image. Most importantly, these 3D models look plausible for multiple views, not only from the input one. The results on the chair category also show that the model captures the relevant variation, ranging from straight chairs over club chairs to cantilever chairs. For gun, the results turn out almost perfect, in agreement with the numbers reported before. In summary, we think the quality is comparable to supervised GANs where 3D supervision is required.

**Synthetic rare** We explored reconstructing skulls from x-ray (i.e., the AO IF model) images [16] in Fig. 9. We find the method to recover both external and internal structures.

**Real rare** Results for rare classes are seen in Fig. 1 and Fig. 7. We see that our method produces plausible details from multiple views while respecting the input image, even in this difficult case. No metric can be applied to these data as no 3D volume is available to compare in 3D or re-project.

6. Discussion

**Why not having a multi-view discriminator?** It is tempting to suggest a discriminator that does not only look at a single image, but at multiple views at the same time to judge if the generator result is plausible holistically. But while we can generate “fake” images from multiple views $p_{Data}$, the set of “real” natural images does not come in such a form. As a key advantage, our method only expects unstructured data: online repositories hold images with unknown camera, 3D geometry or illumination.

**Failure cases** are depicted in Fig. 8. Our method struggles
to reconstruct the correct pose as lifting 2D images to 3D shapes is ambiguous for view-space reconstruction.

**Supplemental** Our supplemental materials show novel-view videos, more analysis and both data and network definitions will be made publicly available upon publication.

7. Conclusion

In this paper, we have presented **PlatonicGAN**, a new approach to learning 3D shapes from unstructured collections of 2D images. The key to our “escape plan” is to train a 3D generator outside the cave that will fool a discriminator seeing projection inside the cave.

We have shown a family of rendering operators that can be GPU-efficiently back-propagated and account for occlusion and color. These support a range of input modalities, ranging from binary masks, over opacity maps to RGB images with transparency. Our 3D reconstruction application is build on top of this idea to capture varied and detailed 3D shapes, including color, from 2D images. Training is exclusively performed on 2D images, enabling 2D photo collections to contribute to generating 3D shapes.

Future work could include shading that is related to gradients of density [7] into classic volume rendering. Furthermore, any sort of back-propagatable rendering operator $\rho$ can be added. Devising such operators is a key future challenge. Other adversarial applications such as 2D supervised...
completion of 3D shapes seems worth exploring. Enabling object-space as opposed to view-space reconstruction would help to prevent failure cases as shown in Fig. 8.

While we combine 2D observations with 3D interpretations, similar relations might exist in higher dimensions, between 3D observations and 4D (3D shapes in motion) but also in lower dimensions, such as for 1D row scanner in robotics or 2D slices of 3D data such as in tomography.

References


[11] M. Gadelha, S. Maji, and R. Wang. 3d shape induction from robotics or 2D slices of 3D data such as in tomography.


