INVERSE PROBLEMS IN COMPUTER VISION USING ADVERSARIAL IMAGINATION PRIORS

Hsiao-Yu Fish Tung & Katerina Fragkiadaki
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{htung,katef}@cs.cmu.edu

ABSTRACT

Given an image, humans effortlessly run the image formation process backwards in their minds: they can tell albedo from shading, foreground from background, and imagine the occluded parts of the scene behind foreground objects. In this work, we propose a weakly supervised inversion machine trained to generate similar imaginations that when rendered using differentiable, graphics-like decoders, produce the original visual input. We constrain the imagination spaces by providing exemplar memory repositories in the form of foreground segmented objects, albedo, shading, background scenes and imposing adversarial losses on the imagination spaces. Our model learns to perform such inversion with weak supervision, without ever having seen paired annotated data, that is, without having seen the image paired with the corresponding ground-truth imaginations. We demonstrate our method by applying it to three Computer Vision tasks: image in-painting, intrinsic decomposition and object segmentation, each task having its own differentiable renderer. Data driven adversarial imagination priors effectively guide inversion, minimize the need for hand designed priors of smoothness or good continuation, or the need for paired annotated data.

Consider Figure 1. We imagine a missing triangle occluding three small black circles rather than three carefully arranged pacman shapes – which is what the pixels depict. In (b), we do not perceive two parts of the sea separated by a standing person, rather a continuous sea landscape. In (c), we explain the input as a "masked 8" rather than two semicircles. Consistent explanations of visual observations in terms of *familiar* concepts and memories we call "imaginations". Imaginations invert the image formation process and propose 3D shape, camera pose, scene layering, spatial layout, albedo, shading, inpainted, un-occluded perceptions of the world, necessary for the understanding of the visual scene and interaction with it. Gestalt philosophers (Smith (1988)) proposed a set or principles to explain formation of such percepts, such as, closure, center surround pop-out, good continuity, smoothness etc, which many works attempt to hand design principles to incorporate those into computational frameworks of e.g., perceptual grouping (Yu (2003)). In this work, we present a learning-based inversion model that uses data-driven priors instead.

We propose a computational model that addresses inverse problems in Computer Vision using adversarial imagination priors. Figure 2 illustrates our model. It is comprised of a generator neural network that given a visual input predicts visual imaginations, such as, in-painted image, un-occluded background scene, object segmentation, albedo and shading etc. Relevant memories, assumed to

(a) (b) (c)

Figure 1: Humans come up with complete and plausible imaginations based on their familiar memories, they imagine, rather than merely labeling pixels.

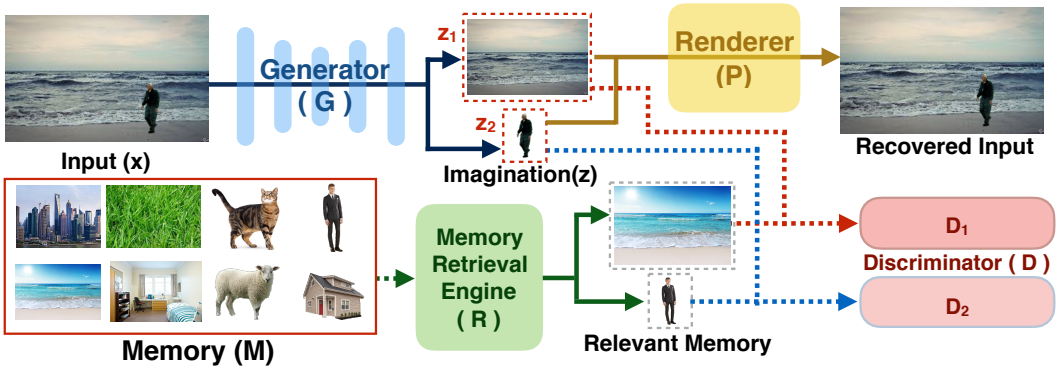

Figure 2: Our model consists of an imagination generator, a graphics-like imagination renderer, a memory retrieval engine and discriminator networks for distribution matching between inferred imaginations and retrieved relevant memories. Here we show our model tailored to the task of figure-ground layer inference, where the imaginations are the segmented foreground object and the completed background scene.

have been acquired from past experience, are retrieved based on coarse attribute matching. Fully-convolutional discriminator networks match statistics of the generated imaginations with retrieved relevant memories. A non-parametric graphics-like differentiable renderer projects such imaginations accordingly and reconstructs the original image. Our model is trained using a combination of adversarial and reconstruction losses.

Architectures we explore ensure the original images can be reconstructed from the inferred imaginations using basic, parameter-free differentiable renderers. This particular choice of decoder function further enforces the imagination spaces to take the particular desired forms, along with the adversarial priors. We are inspired by work of capsules (Tieleman (2014)) that first introduced such domain specific, graphics-like decoders for image generation. We empirically validate the choice of such decoders against standard parametric deconvolutional networks employed by previous works, e.g., inverse graphics network of Kulkarni et al. (2015b).

Our model can infers visual imaginations without having seen *paired annotations*, that is, each input image paired with the corresponding ground-truth. Instead, repositories of relevant memories suffice, in the form of collections of albedo, shading, segmented objects, complete background scenes etc. This distinguishes it from previous works that rely on supervision for decomposing an image into imaginations (Kulkarni et al. (2015b)) or that train image conditioned generator networks using a combination of adversarial and L2 reconstruction loss on imaginations -such as works of Pathak et al. (2016) for in-painting, work of Jiajun Wu (2016) for 3D object reconstruction, work of Dong et al. (2015) for super-resolution. In these works, L2 loss is used to condition the imagination on the input image, e.g., in Jiajun Wu (2016), the adversarial loss ensures the generated voxel grid looks like a 3D object and the L2 loss ensures it is the 3D object corresponding to that particular input image instead of an arbitrary one. Instead, we employ a different method of conditioning: we add a reconstruction loss after our graphics-like decoder, when imaginations are projected (rendered back) to image pixels, reconstructing the original image; our model is like an autoencoder in that sense. In this way, we do not need paired supervision, we can take advantage of unlabelled data and we do not discriminate between training and test phases: adversarial priors useful for inversion can be employed at any time, and the relevant memory repositories may be updated. However, any available annotated pairs can always be used to pretrain the generator network. We did consider such small amount of pretraining in one of the considered tasks and this in this paper to emphasize the power of adversarial priors.

Our model enables feedback from the input image directly to its memory priors: the relevant memory engine retrieves memories based on matching of attribute/ feature descriptors. In this way, priors are tailored to the visual input which alleviates the rare sample problem of traditional training methods, which suffer from the imbalance of training data samples: some examples are way more typical than others, and thus more represented on the neural network weights. Hard negative mining has been used to fight such skewness of training distributions. We empirically show that fixing the prior

distribution instead of adapting it results in undesirable, wrong imaginations, and makes the balance of adversarial and reconstruction losses dependent on example by example basis.

In our model, distribution matching of predicted imaginations and retrieved memories concerns local image statistics. In contrast to most previous use cases of adversarial networks, our work (1) conditions imaginations to the input image -does not generate from random noise- and (2) has a feedback loop by projecting imaginations back to the image through rendering and L2 reconstruction loss. Both (1) and (2) constrain the imagination space and thus our adversarial distribution matching cares mostly about local statistics, rather than global structure, texture matching rather than semantic content. For example, for grass in-painting, we do not care whether the imagination looks similar or exactly the same as a retrieve grass image, we only want to make sure each part in the imagination follows a grass-like texture. We propose fully-convolutional discriminator networks, that predict real versus fake binary tests densely across the feature grid, rather than once for the whole image. This accelerates training, makes our model robust to the size of the network input size and across different field-of-views.

In summary, our contributions are as follows:

- A weakly supervised model for inverse problems given visual input based on adversarial imagination priors and graphics-like decoders.
- Relevant memory retrieval for informative adversarial priors.
- Fully-convolutional discriminator networks for matching local image statistic distributions robust to network input size and image field-of-view.

We demonstrate our model in the tasks of image in-painting, figure-ground layer extraction and intrinsic image decomposition. We show successful imagination prediction without using paired ground-truth annotations. We are working towards updating the draft with inverse problems in videos to convey the generality of the proposed model.

1 Related Work

Vision as an inference problem Both Computer and Human Vision fields have worked towards models that given visual observations attempt to infer hidden properties about the visual scene by *inverting* the image formation process, "un"-doing camera projection, occlusions, motion blur, down-sampling, image masking. Examples are inferring 3D shape and camera pose in videos or images in Tomasi & Kanade (1992), decoupling 3D shape, lighting and albedo interactions in Barron & Malik (2013); Kong et al. (2014), inferring scene depth segmentation layering in Yang et al. (2012), super-resolving low resolution input in Yang et al. (2010), filling in pixels in masked ("hole") images (in-painting) Efros & Leung (1999) etc.

Multimodality Inverse problems are ill-posed: There are many imagination solutions whose projection or rendering would result in the same visual image. Multi-modality of the desired hidden representation causes methods that rely on maximum likelihood to suffer from regression-to-the-mean problem. Despite this fact, direct feed-forward neural networks regressors or classifiers have been trained in a supervised way to achieve such inversion, e.g., depth estimation in Eigen et al. (2014), albedo estimation in Narihira et al. (2015b), volumetric inference in Firman et al. (2016), super-resolution in Dong et al. (2015) etc. The argument is that with large enough receptive fields, ambiguities of inversion are diminished. However, such approaches require human supervision, may not generalize well enough to handle different inputs, adapt to the example at hand effectively, or achieve global consistency of the solution (Narihira et al. (2015a)).

Generative adversarial networks (Goodfellow et al. (2014); Sebastian Nowozin (2016); Radford et al. (2015)) instead have shown to minimize Jensen-Shannon divergence between the matched distributions and exhibit a mode seeking behaviour (Theis et al. (2016)) desirable for inversion. Our adversarial priors can be thought as a surrogate to true perceptual losses, which would involve humans in the loop and would be very expensive to obtain in practise.

Concurrent work of Sønderby et al. (2016) proposes a model for super-resolution that does not required paired supervision, similar to our model. They have an adversarial loss in the high resolution

image and their decoder is a downsampler. Their model is a special case of our model, in which we consider general graphics-like decoders, tailored to each task. Further, they do not consider relevant memory retrieval and do not consider fully-convolutional decoders.

Priors Other research approaches on inverse problems do not employ learning but rather rely on hand designed priors, such as sparsity in Yang et al. (2010), spatial smoothness (for optical flow, depth, albedo etc), temporal smoothness (for shading in Kong et al. (2014)), low-rank 3D shape or trajectory priors in Akhter et al. (2008); Wu et al. (2016), deformable 3D scene models in Kulkarni et al. (2015a). Such hand designed priors, though do not suffer from generalization issues, cannot exploit data available effectively.

Our work proposes data driven priors implemented through adversarial distribution matching between inferred imaginations and retrieval memories. Such priors exploit unlabelled data available in the form of imagination repositories, do not suffer from training and test discrepancies, do not need paired supervision and alleviate the engineering burden of designing good prior models.

Feedback Feedback is visual processing has been incorporated in recent computational models through iterative processing, where each step produces a better estimate of the relevant memory, let is be image reconstruction Raiko et al. (2014), body pose estimation Carreira et al. (2015) etc. Such feedback is incorporated in our adversarial prior model through a memory retrieval mechanism which uses coarse feature extraction and attributes on unoccluded parts of the visual input to retrieve relevant examples, and thus influence the reconstruction in an example by example case, alleviating the problem of data imbalance and finetuning, catastrophic forgetting, hard negative mining of traditional training paradigms.

Domain specific non-parametric decoders Model architectures we explore are based on the fact that the inferred imaginations are such that the original image can be reconstructed using basic, parameter-free operations, such as, camera projection, that project inferred 3D and camera pose to 2D scene Handa et al. (2016), pointwise multiplication for image decomposition, layering that assembles different imaginations based on their depth and segmentation masks. Our work is inspired by work of Tieleman (2014) which proposes capsules, a model for image generation by assembling 2D image pieces and their poses predicted from the encoder into one canvas.

2 MODEL

Our model is illustrated in Figure 2. Given a set of images $X = \{x_i, x_2, \cdots, x_n\}$, and a memory database M, a generator network inverts each image x into a set of imaginations z_1,$\mathrm{z}_2, \cdots \mathrm{z}_K$, which, (1) when rendered back to pixels, the projection should match the corresponding input image; and (2) the imagination statistics should match the distribution of relevant memories retrieved from M through a memory retrieval engine. Our model is trained to minimize the combination of (1) an image reconstruction loss and (2) an adversarial imagination loss that constraints the imagination space(s). The imagination spaces and renderer architecture depend on the inversion task. We consider thress tasks in this work: 1) image in-painting, 2) intrinsic image decomposition, 3) figure-ground layer extraction.

We denote the generator as a mapping function from input to imaginations $G(x)$, the renderer as a mapping function from the imagination to the input image $P(z)$, the image retrieval engine as a mapping function from memories M and input image x to relevant memories $R(M, x)$, and the discriminator for imposing distribution matching between imaginations and retrieved memories as D. In case of multiple imagination spaces (e.g., shading and albedo, in-painted background and foreground object mask etc.) we will use $G_i(x)$ to denote the i-th imagination proposed by the generator. Suppose there are K imagination spaces, the memory retrieval engine will need to retrieved K relevant memory that corresponds to each of the imaginations. Besides, we also need K discriminators to look after each generated imagination. Here we use $R_i(M, x)$ and D_i to denote

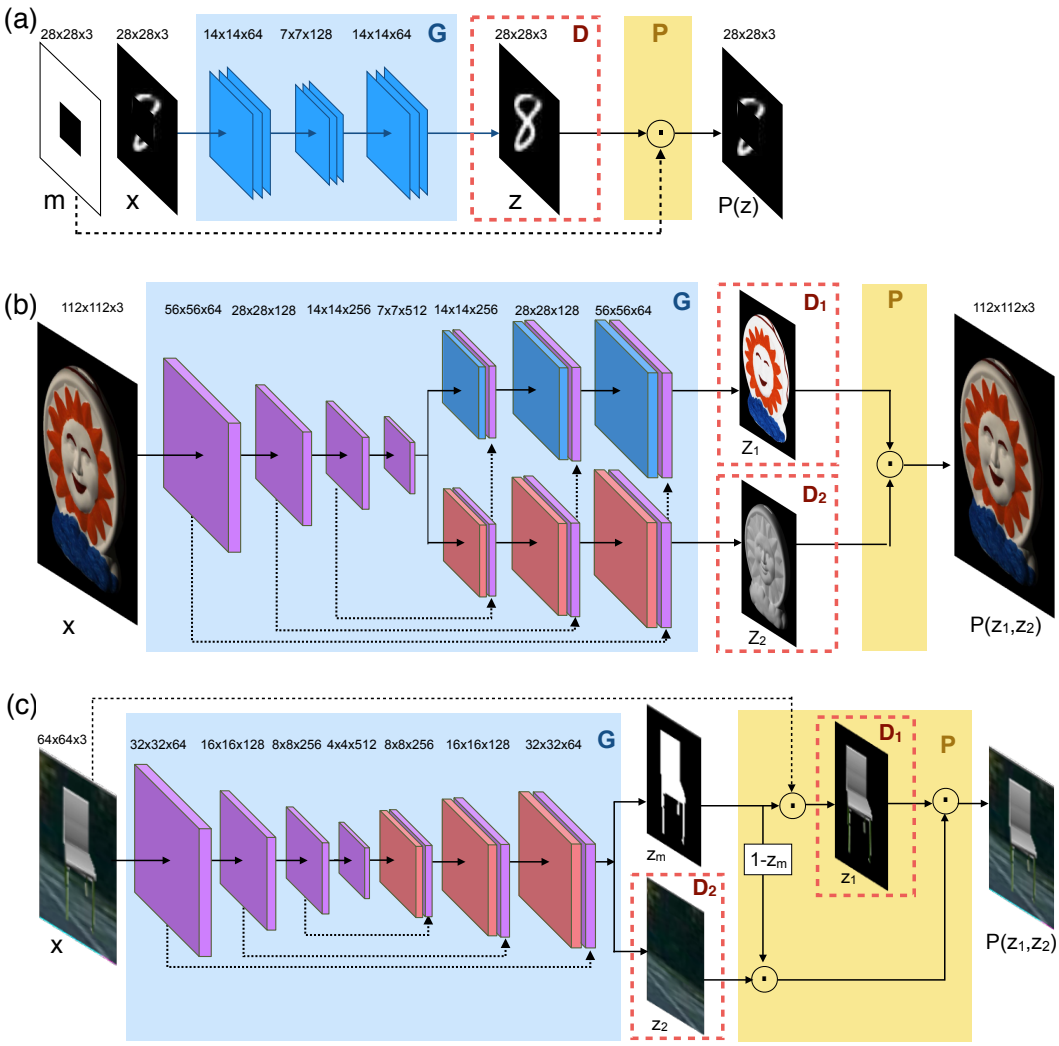

Figure 3: Model architecture for (a) image in-painting, (b) intrinsic image decomposition, and (c) figure-ground layer extraction.

the corresponding retrieved relevant memory and discriminator for the i-th imagination space. Our loss reads as follows:

$$\min_D \max_G \mathbb{E}_{x \in X} \underbrace{||P(G(x)) - x||^2}_{\text{reconstruction loss}} + \beta \sum_{i=1}^{n} \underbrace{\log D_i(R_i(M, x)) + \log(1 - D_i(G_i(x)))}_{\text{adversarial loss}}, \quad (1)$$

where β the relative weight of reconstruction and adversarial losses.

2.1 Imagination Generator G

Given an image, the generator outputs one or more imaginations. In the tasks we consider, imaginations have a retinotopic representation, that is, they have the same size as the input image. Our generators are convolutional/deconvolutional neural networks with skip-layer connections from the encoding to the decoding layers. Skip-layer connections much improve the precision of the produced imaginations. We share weights of the first convolutional layers across multiple imagination spaces. Figure 3 shows the generator architectures we used for the different inversion tasks.

2.2 IMAGINATION RENDERER P

Below we present our domain-specific renderers for three Computer Vision tasks: image in-painting, intrinsic image decomposition and figure-ground layer extraction.

Image in-painting The input is a masked image x, an image whose content is covered by a black contiguent mask m. The task is to invert such masking and produce an imagination that corresponds to the complete (in-painted) image before the masking operation, as shown in Figure 3 (a). The rendering function P in this case is defined as $P(\mathrm{z}) = m \odot \mathrm{z}$, where $\odot$ denotes pointwise multiplication.

Intrinsic image decomposition Given an image x, the generator generates albedo z_1 and shading z_2, as shown in Figure 3 (b). For Lambertian surfaces that the product of albedo and shading should recover the original image, we thus define our renderer to be: $P(\mathrm{z}_1,\mathrm{z}_2) = \mathrm{z}_1 \odot \mathrm{z}_2$. Note that we need two discriminator networks, one that controls the statistics distribution of generated albedos and one that controls the statistics distribution of generated shading imaginations. In practise, instead of pointwise multiplication, we used addition in the log space.

Figure-ground layer extraction In this task, given an image, we want to invert the layering superimposition caused by the objects against their background and produce imaginations of the segmented objects and in-painted background scene. Given an image x, the generator outputs a foreground segmentation mask z_m, corresponding image foreground $\mathrm{z}_1 = x \odot \mathrm{z}_m$ and an in-painted background z_2 such that the in-painted background matches the relevant background memories and the image foreground matches memories of segmented relevant objects with clean (black) background, as shown in Figure 3 (c). Our renderer in this case is defined to be: $P(\mathrm{z}_1,\mathrm{z}_2,\mathrm{z}_m) = (1-\mathrm{z}_m) \odot \mathrm{z}_2 + \mathrm{z}_1 \odot x$, that is, it overlays the object on the in-painted background.

2.3 FULLY-CONVOLUTIONAL DISCRIMINATOR D

We propose fully-convolutional discriminator architectures for matching local image statistics between inferred imaginations and retrieved relevant memories. Fully-convolutional discriminators employ many -instead of one- classifiers centered at grid points of the feature maps, that calculate the confidence scores of being real of fake pattern for each of the local receptive fields, in different layers of the network. Fully-convolutional discriminators allow better generalization from relevant memories to generated imaginations as they match only local statistics and not global patterns. Further, they are much faster and more stable to train as the number of examples fed into the discriminator increases. In our experimental section, we show empirically that fully-convolutional adversarial loss accelerates and stabilizes training.

2.4 MEMORY RETRIEVAL ENGINE R

Given an input image and a memory database of the same imagination types we want to generate, e.g., albedos of natural images, the memory retrieval engine retrieves the most relevant memories. The details of memory retrieval depend on the inversion task.

Image in-painting In this case, we measure the L2 pixel distance between the *visible part* of the input image and images in our memory database, and retrieve the top nearest neighbors.

Intrinsic image decomposition We retrieve relevant shadings by L2 pixel matching between the grayscale version of the input image and albedo memories. We retrieve relevant albedos by computing pixel matching between the input image and albedo memories.

Figure-ground layer extraction Our foreground object memories are segmented objects, as shown in Figure 3 (c) z_1. We retrieve relevant segmented objects according to an object detector output, that makes sure our segmented object imaginations agree on the object category with retrieved memories (here the object category of interest is "chair"). For the in-painted background memories, we use L2 pixel distance between the current image and images from the SUN scene dataset.

In any of the aforementioned inversion tasks, after the model starts to generate reasonable initial imaginations, we can use those to retrieve more relevant memories. Such iterative feedback between memory and visual processing though very reasonable to do, we did not consider it in this work to keep the framework simple.

3 EXPERIMENTS

We show results of our model for (1) image in-painting, (2) intrinsic image decomposition and (3) figure-ground layer extraction. The corresponding model architectures are shown in Figure 3 and further training details are provided in the Appendix.

3.1 IMAGE IN-PAINTING

We used the MNIST dataset and masked parts of its digit images. Specifically, we randomly selected 2500 samples of digits $0, 1, 2, 3$ from the dataset and overlayed a squared mask at the center over them to create our input images. Our memory database M contains 1000 samples for each of the ten digits. We purposefully designed such distribution mismatch between the input image dataset and memory database to study the usefulness of retrieved memories under a controlled setup. The set of digit images contained in M does not intersect the set of images we used to create our input images. In other words, the groundtruth imaginations for our input images are not contained in our memory database.

Figure 4 shows the results of four in-painting models: (1) a baseline with L2 pixel loss between imaginations and retrieved relevant memories ($\mathrm{BmemL2}$), (2) our model with memories retrieved uniformly at random from M rather than conditioned on input images (we supress our memory retrieval engine R) ($\mathrm{Bmemrand}$), (3) our model with memories retrieved uniformly at random from M and with larger weight on the reconstruction loss ($\mathrm{BmemrandHR}$), (4) our model.

Treating retrieved relevant memories as the golden ground-truth produces blurry images, as shown in Figure 4 Row 2. L2 matching optimizes the wrong objective, aside of the fact that it suffers from regression-to-the-mean error even with perfectly correct paired ground-truth, as noted in previous works (e.g., Sønderby et al. (2016)). $\mathrm{Bmemrand}$ produces imaginations that look like reasonable digits but do not match the corresponding input image, as shown in Figure 4 Row 3, rightmost column. Such discrepancy between memories and desired imagination distributions cannot be corrected by increasing the reconstruction loss over the adversarial loss ($\mathrm{BmemrandHR}$), shown in Figure 4 Row 4. Then, the resulting imaginations do not look like correct digits anymore. Our model correctly in-paints the masked digits, as shown in Figure 4 Row 5.

For each input digit image we show in Figure 4 Row 6 the closest retrieved memory from our engine R. By comparing the output of our model with the closest memory, we see that we learn to

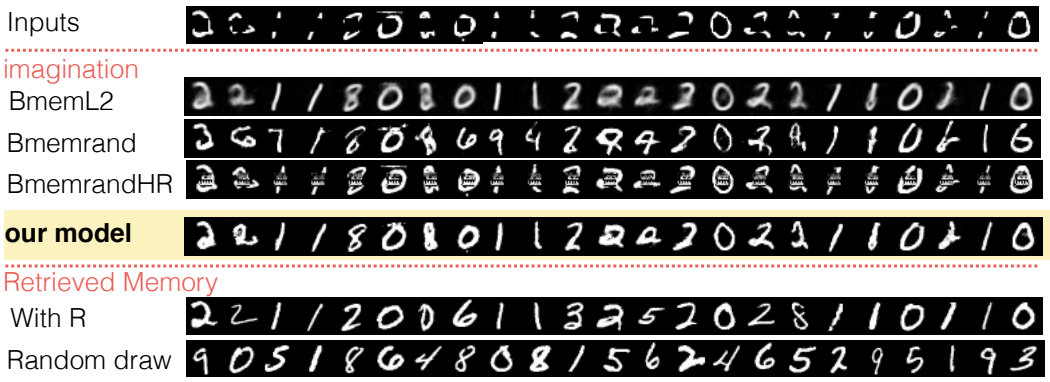

Figure 4: Results of image in-painting on MNIST dataset. *Row 2:* $\mathrm{BmemL2}$ treats retrieved relevant memories as ground-truth imagination and penalizes the l2 loss between them. *Row 3:* $\mathrm{Bmemrand}$: our model without memory retrieval but rather uniform at random memory access. Imaginations do not respect the input image conditioning. *Row 4:* $\mathrm{BmemrandHR}$: our model without memory retrieval but increased weight on the reconstruction loss. Imaginations do not look like correct digits. *Row 5:* Our model. It produces correct in shape and texture digit imaginations in contrast to the baselines above. *Row 6:* Top closest relevant memories retrieved by our engine. *Row 7:* Random memory retrieval.

interpolate on the memory space and form an imagination that fits the current input image, without copy pasting, as a nearest neighbor memory engine alone would do.

3.2 INTRINSIC IMAGE DECOMPOSITION

We use the MIT intrinsic image dataset of Grosse et al. (2009). We use ten objects for training and ten objects for testing. During training, our inputs are images of the training objects and our memory database contains albedos and shadings for the training objects. At test time, we just evaluate our generator on images of the test objects, without finetuning our model. We used random image cropping for data augmentation as described in the Appendix.

Figure 5 Left shows results of our model which never uses paired annotations, that is, does not have access to pairing of each RGB image with its ground-truth albedo and shading. The results are comparable to an oracle model that has access to such paired supervision and optimizes a regression loss, similar to previous work of Narihira et al. (2015b), shown in Figure 5 Right. Our model effectively generalizes to unseen objects (Figure 5 Bottom Right). Figure 6 shows how using fully-convolutional discriminators on albedo and shading stabilize training and converge faster.

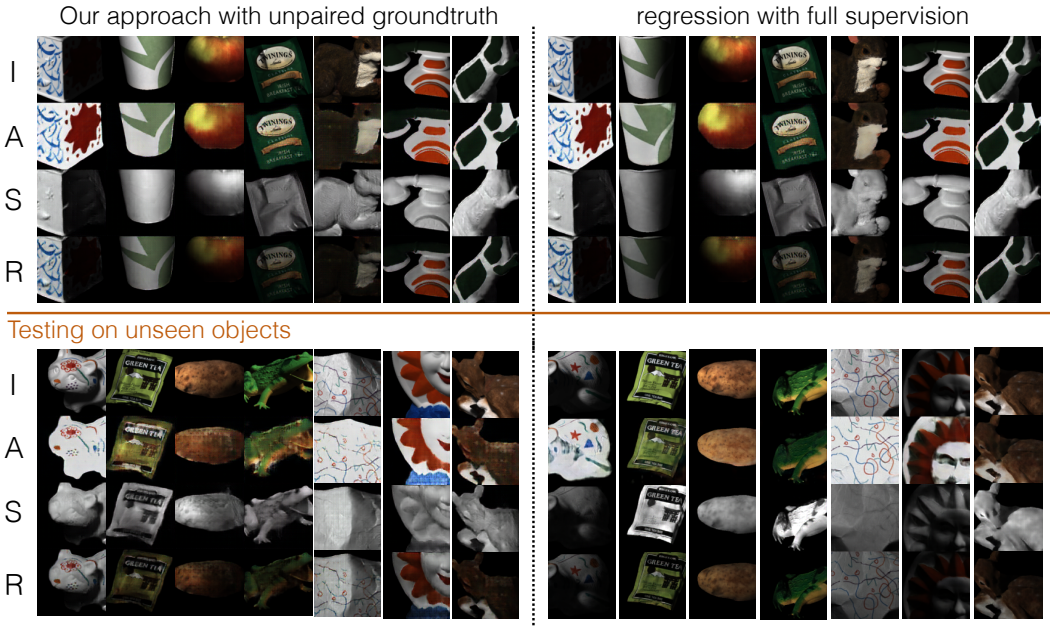

Figure 5: MIT intrinsic decomposition with unpaired shading and albedo. I: Input Image, A: Inferred albedo, S: Inferred shading, R: Reconstructed Image using A and S. *Left:* inferred albedo and shading using our weakly supervised method. *Right:* inferred albedo and shading using a fully supervised model that minimizes regression loss. The bottom part shows the results of decompositions on unseen objects.

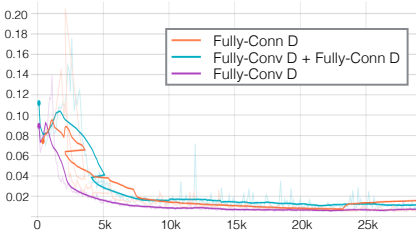

Figure 6: L2 distance between inferred albedo imagination and the ground truth. Fully-convolutional discriminators (purple and green lines) converge faster than fully connected ones, that employ only one fake/real classifier per image.

3.3 FIGURE-GROUND LAYER EXTRACTION

We use the seeing 3D chairs dataset of Aubry et al. (2014) as object memory database which contains 1200 different chairs and the SUN scene dataset Xiao et al. (2010) as the background memory database which contains 131000 images. Our input images are generated by randomly selecting an SUN image and cropping it to 64×64. After that, we overlaid an chair image on top of it as our input image.

We use 200 background images and 200 chair images to generate a small subset of "labelled data", where we provide the network with ground-truth of mask and background and train the network using regression loss, as described in the Appendix. Such small scale supervised pretraining suffices for stability of our model in this task; it is very realistic to assume the existence of such strong sparse supervision. Figure 7 shows the inferred mask and background we obtain.

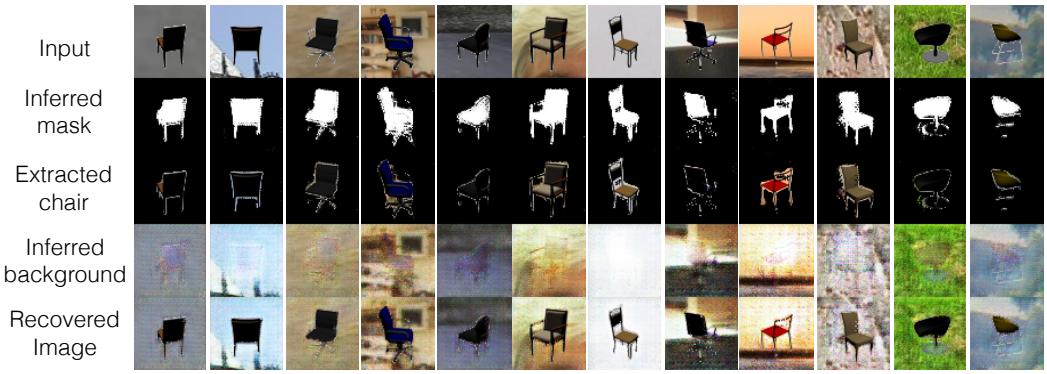

Figure 7: Results for Figure-ground layer extraction. *Row 1:* input images. *Rows 2,4:* The segmentation mask and in-painted background proposed by the generator. *Row 5:* By superimposing the inferred mask on the in-painted background, the network outputs the recovered image, which should match the input.

4 CONCLUSION

We have presented a weakly supervised inverse model of images that predicts imaginations of hidden representations which then renders through image formation or layering to reconstruct the original image. It regularizes the inferred hidden representations using convolutional adversarial priors by distribution matching against retrieved relevant memories. It does not assume paired supervision and can handle multimodal imagination spaces. We have empirically validated are design choices of fully-convolutional adversarial discriminator networks and relevant memory retrieval. We believe the proposed learning paradigm better exploits unlabelled data in the form of images, depth maps, albedo, shading or segmentation maps and complements well human paired annotations.

We are working towards updating the paper with two inversion problems in videos, visual odometry and motion object segmentation. Videos allow for much stronger observation module, with imagination projections from frame to frame, as well as temporal constraining of the imaginations in time. Further, we are working towards quantifying generalization of our imaginations from training to test images, specifically measuring how well our model can do with increasing dissimilarity between memories in the database and input images.

REFERENCES

Martn Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Man, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Vigas, Oriol Vinyals, Pete Warden, Martin Wattenberg,

Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems, 2015. URL http://download.tensorflow.org/paper/whitepaper2015.pdf.

Ijaz Akhter, Yaser Ajmal Sheikh, Sohaib Khan, and Takeo Kanade. Nonrigid structure from motion in trajectory space. In *Neural Information Processing Systems*, December 2008.

Mathieu Aubry, Daniel Maturana, Alexei Efros, Bryan Russell, and Josef Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*, 2014.

Jonathan Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. Technical Report UCB/EECS-2013-117, EECS Department, University of California, Berkeley, May 2013. URL http://www2.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-117.html.

Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. 2015.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *CoRR*, abs/1501.00092, 2015. URL http://arxiv.org/abs/1501.00092.

Alexei A. Efros and Thomas K. Leung. Texture synthesis by non-parametric sampling. In *ICCV*, pp. 1033–1038, 1999. doi: 10.1109/ICCV.1999.790383. URL http://dx.doi.org/10.1109/ICCV.1999.790383.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *CoRR*, abs/1406.2283, 2014. URL http://arxiv.org/abs/1406.2283.

Michael Firman, Oisin Mac Aodha, Simon Julier, and Gabriel J. Brostow. Structured Prediction of Unobserved Voxels From a Single Depth Image. In *CVPR*, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Roger Grosse, Micah K. Johnson, Edward H. Adelson, and William T. Freeman. Ground-truth dataset and baseline evaluations for intrinsic image algorithms. In *International Conference on Computer Vision*, pp. 2335–2342, 2009. doi: http://dx.doi.org/10.1109/ICCV.2009.5459428.

Ankur Handa, Michael Blösch, Viorica Patraucean, Simon Stent, John McCormac, and Andrew J. Davison. gvnn: Neural network library for geometric computer vision. *CoRR*, abs/1607.07405, 2016. URL http://arxiv.org/abs/1607.07405.

Tianfan Xue William T. Freeman Joshua B. Tenenbaum Jiajun Wu, Chengkai Zhang. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. 2016.

Naejin Kong, Peter V. Gehler, and Michael J. Black. Intrinsic video. In *Computer Vision – ECCV 2014*, volume 8690 of *Lecture Notes in Computer Science*, pp. 360–375. Springer International Publishing, September 2014.

Tejas D. Kulkarni, Pushmeet Kohli, Joshua B. Tenenbaum, and Vikash K. Mansinghka. Picture: A probabilistic programming language for scene perception. In *CVPR*, pp. 4390–4399. IEEE Computer Society, 2015a. ISBN 978-1-4673-6964-0. URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2015.html#KulkarniKTM15.

Tejas D. Kulkarni, Will Whitney, Pushmeet Kohli, and Joshua B. Tenenbaum. Deep convolutional inverse graphics network. *CoRR*, abs/1503.03167, 2015b. URL http://arxiv.org/abs/1503.03167.

Takuya Narihira, Michael Maire, and Stella X. Yu. Learning lightness from human judgement on relative reflectance. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 2965–2973, 2015a. doi: 10.1109/CVPR.2015.7298915. URL http://dx.doi.org/10.1109/CVPR.2015.7298915.

Takuya Narihira, Michael Maire, and Stella X. Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. *CoRR*, abs/1512.02311, 2015b. URL http://arxiv.org/abs/1512.02311.

Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. 2016.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *CoRR*, abs/1511.06434, 2015. URL http://arxiv.org/abs/1511.06434.

Tapani Raiko, Li Yao, KyungHyun Cho, and Yoshua Bengio. Iterative neural autoregressive distribution estimator (nade-k). In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, NIPS'14, pp. 325–333, Cambridge, MA, USA, 2014. MIT Press. URL http://dl.acm.org/citation.cfm?id=2968826.2968863.

Ryota Tomioka Sebastian Nowozin, Botond Cseke. f-gan: Training generative neural samplers using variational divergence minimization. In *Neural Information Processing Systems 2016*. Neural Information Processing Systems, October 2016.

Barry Smith. Gestalt Theory: An Essay in Philosophy. In Barry Smith (ed.), *Foundations of Gestalt Theory*, pp. 11–81. Philosophia Verlag, December 1988.

Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*, 2016.

L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations*, Apr 2016. URL http://arxiv.org/abs/1511.01844.

Tijmen Tieleman. Optimizing neural networks that generate images. *Ph.D. Thesis*, 2014.

Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: A factorization method. *Int. J. Comput. Vision*, 9(2):137–154, November 1992. ISSN 0920-5691. doi: 10.1007/BF00129684. URL http://dx.doi.org/10.1007/BF00129684.

Jiajun Wu, Tianfan Xue, Joseph J. Lim, Yuandong Tian, Joshua B. Tenenbaum, Antonio Torralba, and William T. Freeman. *Single Image 3D Interpreter Network*, pp. 365–382. Springer International Publishing, Cham, 2016. ISBN 978-3-319-46466-4. doi: 10.1007/978-3-319-46466-4_22. URL http://dx.doi.org/10.1007/978-3-319-46466-4_22.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pp. 3485–3492. IEEE Computer Society, 2010. ISBN 978-1-4244-6984-0. URL http://dblp.uni-trier.de/db/conf/cvpr/cvpr2010.html#XiaoHEOT10.

Jianchao Yang, John Wright, Thomas S. Huang, and Yi Ma. Image super-resolution via sparse representation. *Trans. Img. Proc.*, 19(11):2861–2873, November 2010. ISSN 1057-7149. doi: 10.1109/TIP.2010.2050625. URL http://dx.doi.org/10.1109/TIP.2010.2050625.

Yi Yang, Sam Hallman, Deva Ramanan, and Charless C. Fowlkes. Layered object models for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(9):1731–1743, 2012. doi: 10.1109/TPAMI.2011.208. URL http://dx.doi.org/10.1109/TPAMI.2011.208.

Stella Yu. Computational models of perceptual organization. Technical report, Robotics Institute, Carnegie Mellon University, 2003.

A IMPLEMENTATION DETAILS

All models are implemented in Tensor Flow Abadi et al. (2015).

Image in-painting Details of the generator architecture are illustrated in Figure 3 (a). The discriminator consists of one convolutional layer and one fully-connected layer on top. Both the generator and the discriminator use batch normalization with relu activation for the generator and leaky relu activations for discriminator. We initialize all weights with sampling from normal distribution with standard deviation 0.02. We use the Adam optimizer with a fixed learning rate of 0.0002.

Intrinsic Image Decomposition For each of the images in the MIT dataset, we randomly crop a region of 112×112. For memory retrieval, we find the top nearest-neighbors from 100 crops, using the method described in Section 2.4.

Details of the generator architecture are illustrated in Figure 3 (b). Each convolutional layer will pass though batch normalization layer, leaky relu activation and max pooling layer before sending to the next convolutional layer. Discriminators for both albedo and shading contain four convolutional layers with batch normalization leaky relu activations. The fully-convolutional adversarial loss is built on top of the fourth layer. We initialize all weights with sampling from normal distribution with standard deviation 0.02. We use the Adam optimizer with a fixed learning rate of $10\mathrm{e}{-6}$. We put 0.1 weight on the l2 recovery loss.

Figure-ground layer extraction Details of the generator architecture are illustrated in Figure 3 (c). Each convolutional layer passes though batch normalization layer, leaky relu activation and max pooling layer before sending to the next convolutional layer. Discriminator for both objects and background contains four convolutional layers with batch normalization and leaky relu activations. We initialize all weights with sampling from normal distribution with standard deviation 0.02.. The generator is pre-trained using 200 images annotated with groundtruth in-painted background and foreground object mask using an L2 pixel loss. Then, the model is finetuned using only the described adversarial imagination loss and image reconstruction loss. We use the Adam optimizer with a fixed learning rate of $10\mathrm{e}{-5}$. Such pretraining, though small in scale, much helped stability of our model. We have also experimented with adding noise to retrieved memories to make the task of the discriminator harder at the beginning of the training, as in Sønderby et al. (2016). Small scale supervised pretraining suffices for stability of the model in this task, and it is very realistic to assume the existence of such strong sparse supervision.