Incorporating Priors within Learning Algorithms

ShahRukh Athar
Department of Computer Science
Stony Brook University
Stony Brook, NY-11794, U.S.A

I. Overview

Over the past few years significant progress has been made within computer vision due to the effectiveness of convolutional neural networks (CNNs) as the base on which many computer vision algorithms are built on. The effectiveness of CNNs is primarily attributed to the use of the convolutional operator that encodes translation equivariance and the locality of image features as an explicit prior in an otherwise poorly constrained learning algorithm (i.e fully connected networks). This has allowed CNNs to not only be great image feature detectors but has also allowed them to model images [13], [19], [1], [2], [12], [6]. The central focus of my research so far has been to figure out ways to go beyond the simple convolution to incorporate principled and relevant task-specific priors within learning algorithms with the aim to improve their performance and potentially solve previously unsolved problems. My work on Latent Convolutional Models [1] aimed at developing a universal prior for images that could be used for any image restoration task. We developed such a prior by imposing a convolutional prior on the latent space of images through a parametrization using convolutional networks. Such a parametrization allowed image restoration to be performed on a constrained high-dimensional convolutional manifold that restorted images pretty well from even the most severe image degradations [1]. Along similar lines, I worked on the development of DefGAN [2] that significantly improved the results of facial expression editing over the prior state-of-the-art by imposing physical priors of human facial expressions within the learning algorithm. DefGAN carried out facial expression editing through a disentangled mechanism that explicitly modeled *facial muscle movement* in the *motion editing* phase followed by a texture editing phase. This explicit disentanglement allowed us to vastly improve both the generalization ability of facial expression editing methods and the quality of results they produced. Following this paradigm of incorporating task specific knowledge as priors within the algorithms we develop, the most logical next step is to work towards incorporating, either explicitly or implicitly, knowledge about the three dimensional nature of our world. I strongly believe that a 3D world prior can at the very least improve generalization of current methods (e.g editing faces with extreme pose variations) and potentially lead to exciting new avenues of research.

II. Incorporating Priors within Learning Algorithms

A. Imposing Convolutional Priors on the Latent Space of Images

1) Latent models of images and the Deep Image Prior: The efficacy of a convolutional prior was strongly bolstered by results from the Deep Image Prior [19] which showed that *untrained* convolutional networks can act as a really good prior for image restoration problems. More specifically, parametrizing an image using a convolutional neural network and optimizing over its weights while carrying out the desired image restoration instead of directly doing it in the pixel space gives great results. Given a degraded image x^* recovering the restored image $\hat{x}$ can be modelled as the following optimization problem

$$\hat{x} \leftarrow \underset{x}{\operatorname{argmin}} - [\log p(x^*|x) + \log p(x)] \tag{1}$$

The Deep Image prior replaces the prior term with an implicit prior using a convolutional neural network f_θ as follows

$$\hat{\theta} \leftarrow \underset{\theta}{\operatorname{argmin}} - [\log p(x^*|f_\theta(s))]\,; \quad \hat{x} = f_{\hat{\theta}}(s) \tag{2}$$

where s is a constant input to the network. A major drawback of the deep image prior is that, since these networks are untrained, the 'convolutional prior' is too general and *cannot capture higher level semantic structure within images*. For example, DIP cannot inpaint an image of a human face with a centered hole because it has no a-priori knowledge of where a nose must be placed relative to the eyes. In fact, the priors within an untrained convolutional network are only able to bias an optimization towards low-level image statistics and are, predictably, unable to capture higher level semantics such as a the structure of a human face. Capturing such structure necessarily requires learning.

2) A Convolutional Latent Space: One possible way learn high level image semantics is by learning a latent variable model $g_\phi(z) = x$ that maps a latent variable z deterministically to an image x. Image restoration then reduces to an optimization in the latent space z

$$\hat{z} \leftarrow \underset{z}{\operatorname{argmin}} - \left[\log p(x^*|g_\phi(z)) + p(z)\right]\,; \quad \hat{x} = g_\phi(\hat{z}) \tag{3}$$

Generally, the priors over the latent space are quite simple, such as a gaussian distribution in the case of GANs, which leads to underfitting the likelihood. Better likelihoods can be obtained by increasing the dimensionality of the latent space but this comes at a cost of restoration quality. An ideal

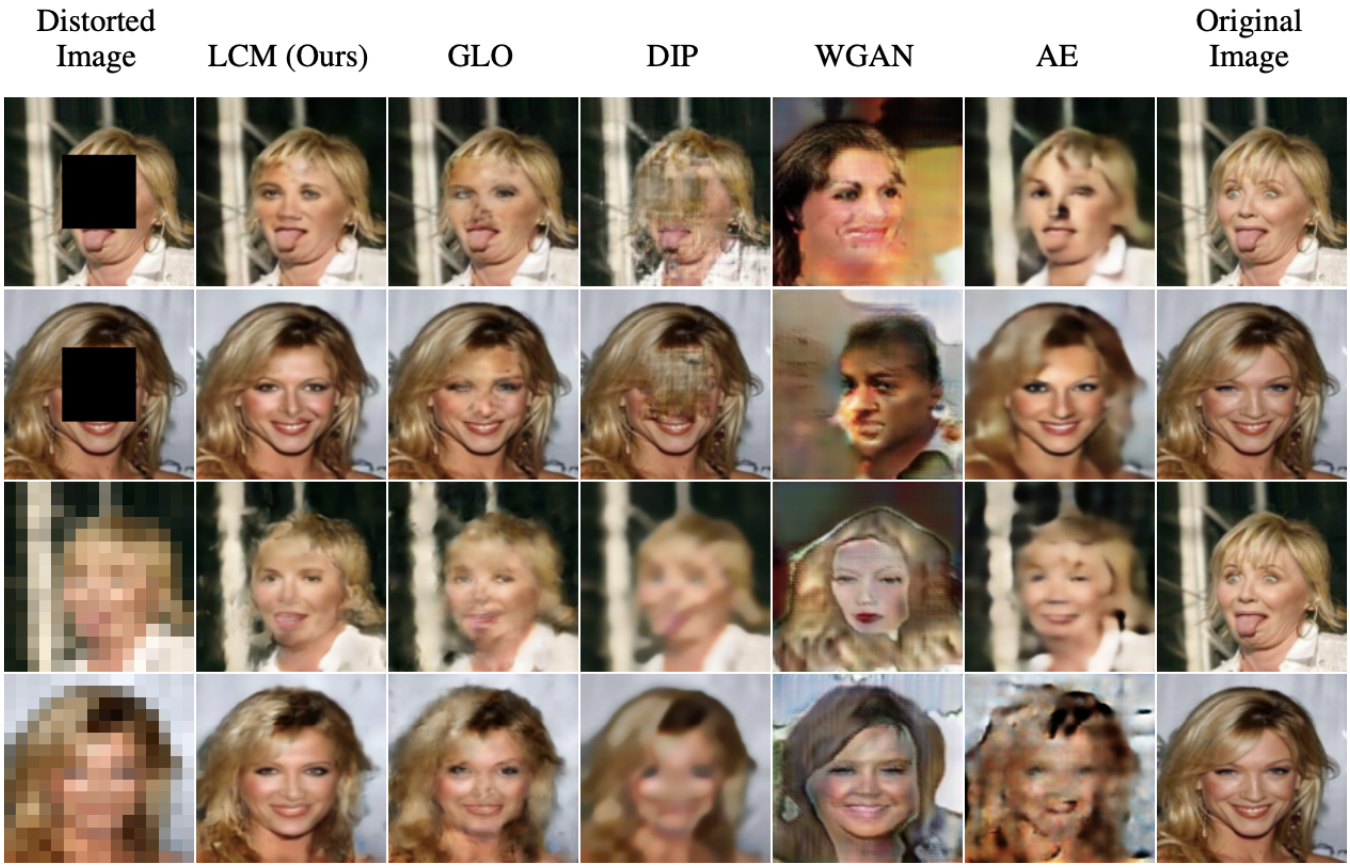

Fig. 1. Image Restoration using Latent Convolutional Models (LCM): In this figure we can see that an LCM not only fits the image likelihood better but also give incredibly realistic restorations as compared to GLO [4], DIP[19], WGANs [8], and Autoencoders (AE)

prior would give a great likelihood without compromising restoration quality. We developed such a prior in [1] by constraining the latent space through a parametrization using convolutional networks. Effectively imposing a *convolutional prior on the latent space*. This allows us to optimize over a very high dimensional latent space which is constrained in manner that allowed us to *achieve very good image restorations* ***without*** *compromising the likelihood fit.* Concretely, the optimization is now carried out as follows

$$\hat{\theta} \leftarrow \underset{\theta}{\operatorname{argmin}} - \left[\log p(x^* | g_\phi(f_\theta(s)))\right]; \quad \hat{x} = g_\phi(f_{\hat{\theta}}(s)) \quad (4)$$

where, s is a constant input, $f_\theta(s) : s \rightarrow z$ maps a constant input s to a latent vector z and $g_\phi(z) : z \rightarrow x$ maps that latent vector $z = f_\theta(s)$ to an image x. Such a convolutional prior on the latent space significantly improves image restoration over a large variety of image degradations. A sample of inpainting and super-resolution results can be found in Figure 1.

B. Priors for Facial Expression Editing

1) StarGAN and GANimation: Facial expression editing is traditionally cast as an image-to-image translation problem where an input face has some expression (for example happy) and one would like to take this image to some target expression (for example sad). Such a translation however must be tightly constrained so that attributes of the image that are invariant to expression change (such as identity, background, environment etc.) do not change. StarGAN [6] was one of the first works to show remarkable results on this task. Given an input image I_x with an input expression x and target expression label y StarGAN directly translates the expression $I_x \rightarrow I_y$ through a single feed forward generator

$$I_y = G(I_x, y) \quad (5)$$

Invariance with respect to the input identity, environment and so on were enforced through cycle consistency losses. A major drawback of StarGAN was that it required a dataset with discrete expression labels to train on. Such datasets are typically not available in in-the-wild settings thus it has limited generalization ability [2]. Further, expression editing is only limited to the expression labels available. StarGAN cannot edit an input face to a 'smirk' or a 'half-smile' simply because they cannot be represented in the discreet label space of expressions it has been trained on. GANimation [12] eliminated those shortcomings of StarGAN by using Action Units [7] as the representation of expressions instead of discrete labels. Action Units can encode any anatomically possible human expression thus allowing GANimation, in principle, to edit any input image to any target expression as long as one has the expression's AU representation. Further, it is possible to label a large scale in-the-wild dataset with AUs using off the shelf algorithms [3]. GANimation goes a step further in enforcing identity and environment invariance

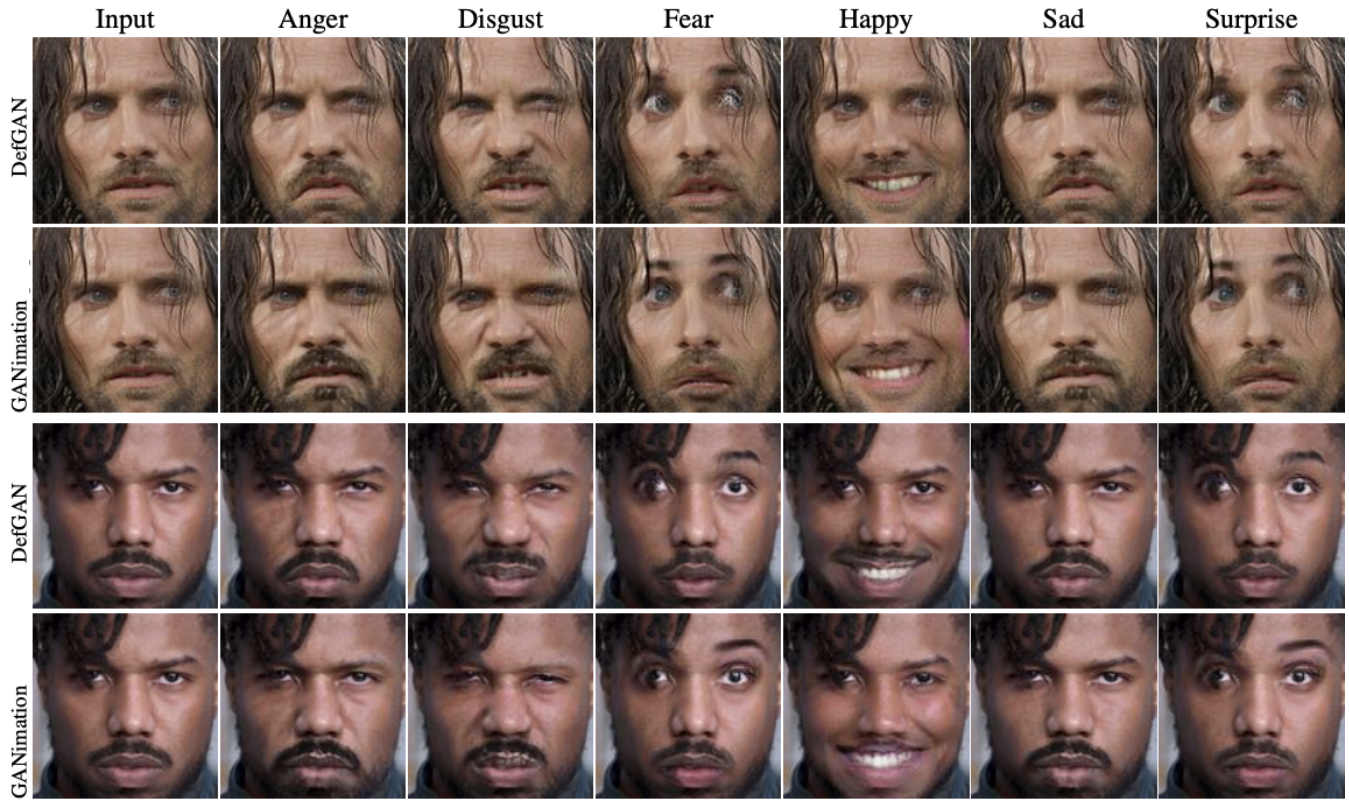

Fig. 2. Editing Facial Expressions: Here we show images edited by DefGAN and GANimation [12] to various target expressions. GANimation [12] tends to produce artifacts (results of 'Anger' and 'Disgust' in row 2) or ends up hallucinating inaccurate textures (results of 'Happy' in row 2 and results of 'Anger' in row 4). In contrast, the editing results of DefGAN are more consistent with fewer artifacts and more accurate textures.

by an explicit masking mechanism as follows

$$\begin{aligned} T, M &= G(I_x, y) \\ I_y &= M \odot I_x + (1 - M) \odot T \end{aligned} \tag{6}$$

Where, T is a Texture Map and M is a binary Mask. This masking mechanism allows the network to copy the invariant pixels (such as the background and parts of the face that do not change over the expression change) only generate the required changes within T. The invariance is further imposed by cycle consistency losses. One can view the cycle consistency losses in StarGAN and GANimation, and the masking mechanism in GANimation as incorporating explicit priors within network architectures that heavily bias networks to preserving the invariants of the target task.

2) Explicit disentanglement of motion and texture as a prior for editing expressions: While GANimation significantly improves the generalization ability of StarGAN and allows expression editing to any anatomically possible target expression it does not take into account an important prior of expression editing, *facial movement*. In our experiments we found that, due to GANimation's complete reliance on hallucinating a Texture Map (T) to model changes in expression, it ends up over-editing images and changes parts of the image that, while not strictly invariant, must not change significantly. We remedy this in DefGAN [2] by *explicitly modelling muscle movement though a deformation in the pixel space*. In other words, we encode an important physical prior of expression editing within the learning architecture. The full editing process is now as follows

$$\begin{aligned} I_y^* &= G_{\text{Def}}(I_x, x, y) \\ I_y &= G_{\text{Texture}}(I_y^*, y) \end{aligned} \tag{7}$$

Where, G_{Def} is a 'deforming' generator [14] that models facial muscle movement as a deformation field over the pixel space giving the deformed image I_y^* and G_{Texture} takes as input this deformed image (that has all the necessary movements) and outputs the final edited image I_y.

A great example of how such explicit modelling of facial muscle movement benefits expression editing can be seen when editing an image of a bearded man from neutral to a smile. A one can see in row 4 of Figure 2 GANimation hallucinates incorrect textures which makes the beard disappear. This happens because GANimation's editing process does not account for the fact that many expression transformations happen though the movement of facial muscles and not through the appearance and disappearance of texture. In fact, Action Units themselves encode muscle movement [7], GANimation ignores this, DefGAN **encodes** it. In the paper, we also show that expression transformations that only involve texture hallucination (such as blinking) are mostly handled by the texture network G_{Texture} and the deformation network G_{Def} produces an identity deformation field and vice-versa in case of expression transformations that mostly involve muscle movement (such as neutral to disgust).

III. Moving beyond the pixel space

All my past works have focused on incorporating relevant priors within network architectures and learning algorithms that act in the pixel space. Perhaps, the most obvious prior that is not explicitly accounted for in these works is the 3D nature of underlying objects within images. More often than not, networks do learn this 3D structure due to the amount of variation in the training data [5], [9], [16], [20], [15], [14] but there are obvious cases of failures [11]. Recently, there has been a significant amount of work in incorporating 3D knowledge for various computer vision tasks. Most direct perhaps is the use of 3DMMs for generating talking heads [10], [18], [17]. These methods perform most transformations (such as pose and expression changes) in three dimensions and then reproject and 'clean-up' the rendered image to give photo-realistic images. Such a clean-up is necessary because 3DMMs are only able to represent a small sample of all possible human faces. As a result, shapes, textures and expressions underfit most optimizations carried out over them and require significant post-processing in the image space once re-projected. I hope to investigate efficient ways to improve the representation capacity of explicit 3D models such as 3DMMs, perhaps by using appropriate reparametrizations. I strongly believe this would further bolster the performance of current learning algorithms on a large variety of computer vision tasks and open up new avenues of research.

References

[1] S. Athar, E. Burnaev, and V. Lempitsky. Latent convolutional models. In *Proc. ICLR*, 2019.

[2] S. Athar, Z. Shu, and D. Samaras. Self-supervised deformation modeling for facial expression editing. *arXiv preprint arXiv:1911.00735*, 2019.

[3] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018.

[4] P. Bojanowski, A. Joulin, D. Lopez-Paz, and A. Szlam. Optimizing the latent space of generative networks. *arXiv preprint arXiv:1707.05776*, 2017.

[5] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[6] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[7] F. W. Ekman, P. Facial action coding system: A technique for the measurement of facial movement. In *Consulting Psychologists Press*, 1978.

[8] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

[9] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *Iclr*, 2(5):6, 2017.

[10] H. Kim, P. Garrido, A. Tewari, W. Xu, J. Thies, M. Nießner, P. Pérez, C. Richardt, M. Zollhöfer, and C. Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018.

[11] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y.-L. Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7588–7597, 2019.

[12] A. Pumarola, A. Agudo, A. Martinez, A. Sanfeliu, and F. Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[13] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[14] Z. Shu, M. Sahasrabudhe, R. Alp Guler, D. Samaras, N. Paragios, and I. Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 650–665, 2018.

[15] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *Computer Vision and Pattern Recognition, 2017. CVPR 2017. IEEE Conference on*, pages –. IEEE, 2017.

[16] A. Shysheya, E. Zakharov, K.-A. Aliev, R. Bashirov, E. Burkov, K. Iskakov, A. Ivakhnenko, Y. Malkov, I. Pasechnik, D. Ulyanov, et al. Textured neural avatars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2397, 2019.

[17] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):1–13, 2017.

[18] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

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

[20] E. Zakharov, A. Shysheya, E. Burkov, and V. Lempitsky. Few-shot adversarial learning of realistic neural talking head models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9459–9468, 2019.