Large Inpainting of Face Images with Trainlets

Jeremias Sulam, *Student Member, IEEE,* and Michael Elad *Fellow, IEEE*

***Abstract*—Image inpainting is concerned with the completion of missing data in an image. When the area to inpaint is relatively large, this problem becomes challenging. In these cases, traditional methods based on patch models and image propagation are limited, since they fail to consider a global perspective of the problem. In this work, we employ a recently proposed dictionary learning framework, coined Trainlets, to design large adaptable atoms from a corpus of various datasets of face images by leveraging the Online Sparse Dictionary Learning algorithm. We therefore formulate the inpainting task as an inverse problem with a sparse-promoting prior based on the learned global model. Our results show the effectiveness of our scheme, obtaining much more plausible results than competitive methods.**

Fig. 1. Example of a inpainted image - left: Face image with missing eyes. Right: inpainted result obtained with the proposed approach.

I. Introduction

Image inpainting is a data completion problem that aims to recover – or fill in – missing information in an degraded image. These areas can be given by individual missing pixels distributed along the image, or by continuous regions resulting from scratches, foldings or other forms of degradation of old photographs. In the extreme case where the area to inpaint is relatively large (also known as *hole-filling*), this problem becomes challenging [1].

This ill-posed problem, whose solution is often not even well-defined, has received considerable attention in recent years. Many inpainting approaches rely on Partial Differential Equations (PDF) [2], [3], variational formulations [4], exemplar-based methods [5], sparsity-enforcing priors [6], [7] and combinations of them [8], [9]. Despite their efficient performance, all these works are restricted to either small areas or to the task of object removal, by propagating and filling-in a proper surrounding background.

Some problems, however, require a different approach. We shall focus in the specific problem of inpainting large areas of face images, like the case in Figure 1. As one could foresee, traditional patch-based methods will not be effective in recovering or estimating the missing data. Diffusion based and content propagation approaches will also find this problem too challenging. In fact, any method which seeks to inpaint the missing region by propagating information from the available image data will fail, as all these are oblivious to the fact that they are inpainting a face. This missing information needs to be provided in terms of a global model of the target image.

The task of obtaining an adaptive global model for high dimensional signals is a hard problem. Some attempts include manifold learning techniques, as in [10], where the authors propose to learn an adaptable low-dimensional manifold for images. This work includes inpainting examples on synthetic and texture data, though it is still far from providing a practical solution for real world face images. The recent work in [11], on the other hand, proposes the use of convolutional neural networks to train a global model to inpaint large holes in natural images. This network, however, was trained for general (street) images and it does not apply to our specific problem.

In this work, we propose to build such a global prior employing sparse representations modeling and dictionary learning. The problem of dictionary learning consists of adaptively learning a set of atoms which are able to represent real signals as sparsely as possible, and has been a popular topic in signal and image processing over the last decade [12], [13]. However, due to the computational constraints that this problem entails, all learning methods are typically applied on small patches from the image and not on the image itself [14], [15]. In other words, attempting to obtain such a global dictionary with traditional dictionary learning algorithms would be infeasible.

A novel work which has circumvented this problem is the recently proposed Trainlets framework [16], where the authors proposed an Online Sparse Dictionary Learning (OSDL) algorithm that is able to obtain large adaptable atoms from natural images. Trainlets are built as linear (sparse) combinations of atoms from a fast and analytical dictionary, that of the novel Cropped Wavelets. This work [16] presented some initial results on sparse approximation of face images - indicating their effectiveness in modeling high dimensional data.

In this work we will formulate the inpainting task as an inverse problem regularized by a sparse prior under a global dictionary trained from publicly available datasets. Our results indicate that the proposed approach is able to synthesize missing information which is in a accordance with the global context of the image, yielding natural reconstructed faces.

II. Learning the Model

Sparse representations has shown to be a powerful prior in several inverse problems in image processing (see [12] for a thorough review). This model assumes that a signal $\mathbf{y} \in \mathbb{R}^n$ can be well approximated by a decomposition of the form $\mathbf{D}\mathbf{x}$, where $\mathbf{D}$ is a matrix of size $n \times m$ containing signal atoms in its columns – termed dictionary –, and a sparse vector $\mathbf{x} \in \mathbb{R}^m$.

All authors are with the Computer Science Department, Technion Israel Institute of Technology.

Algorithm 1: Online Sparse Dictionary Learning

Data: Training samples $\{\mathbf{y}_i\}$, base-dictionary $\mathbf{\Phi}$, initial sparse matrix $\mathbf{A}^0$
Initialization: $\mathbf{G}_{\mathbf{\Phi}} = \mathbf{\Phi}^T\mathbf{\Phi}$; $\mathbf{U} = \mathbf{0}$;
for $j = 1, \ldots, J$ **do**
  Draw a mini-batch $\mathbf{Y}_j$ at random;
  $\mathbf{X}_j \leftarrow$ Sparse Code $(\mathbf{Y}_j, \mathbf{\Phi}, \mathbf{A}^j, \mathbf{G}^j)$;
  $\eta^j = ||\nabla f(\mathbf{A}^j_{\mathcal{S}})||_F / \|\mathbf{\Phi}\nabla f(\mathbf{A}^j_{\mathcal{S}})\mathbf{X}^{\mathcal{S}}_j\|_F$;
  $\mathbf{U}^{j+1}_{\mathcal{S}} = \gamma\mathbf{U}^j_{\mathcal{S}} + \eta^j\ \nabla f(\mathbf{A}^j_{\mathcal{S}})$;
  $\mathbf{A}^{j+1}_{\mathcal{S}} = \mathcal{P}_k\left[\mathbf{A}^j_{\mathcal{S}} - \mathbf{U}^{j+1}_{\mathcal{S}}\right]$;
  Update columns and rows of $\mathbf{G}$ by $\left(\mathbf{A}^{j+1}\right)^T\mathbf{G}_{\mathbf{\Phi}}\mathbf{A}^{j+1}_{\mathcal{S}}$
end
Result: Sparse Dictionary $\mathbf{A}$

The problem of finding such a sparse vector is termed sparse coding, and can formally be expressed as

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{subject to} \quad \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 \leq \epsilon, \tag{1}$$

where ϵ is an allowed deviation in the representation, and the ℓ_0 pseudo-norm is a count on the number of non-zero elements of its argument. When the dictionary is of general content (and overcomplete; i.e, $m > n$), this is an NP-hard problem in general as it is combinatorial in nature. Yet, greedy algorithms and convex relaxation alternatives allow for good approximations of its solution in practice [17], [18].

When combined with the ability to learn the dictionary from real data, and for a specific task, this model has yielded a number of state of the art results [15], [19], [20], [21]. In its general form, the dictionary learning (DL) problem reads as follows

$$\arg\min_{\mathbf{D},\mathbf{X}} \frac{1}{2}\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 \quad \text{subject to} \quad \|\mathbf{x}_i\|_0 \leq p \quad \forall i, \tag{2}$$

where the matrix $\mathbf{Y}$ contains signal examples ordered column-wise. This problem inherits the non-convexity induced by the ℓ_0 pseudo-norm and adds the dictionary as a minimization variable. Though a series of different algorithms have been proposed [22], [14], [15], most methods undertake an alternating minimization approach minimizing over $\mathbf{X}$ and $\mathbf{D}$.

However successful, the dictionary learning problem has traditionally been restricted to the domain of modeling small image patches, thus limiting the kind of problem these methods can address. This limitation arises mainly from computational constraints, but also from the fact that the degrees of freedom of the problem – and the amount data required – become unmanageable as the dimension increases. Some works have attempted to provide more efficient dictionary learning algorithms. The work presented in [23] proposed to lower the complexity of using (and learning) the dictionary by suggesting an adaptable but completely separable structure, yielding an algorithm term SEDIL (Separable Dictionary Learning). Though this is an interesting and effective idea, the complete separability constraint is often too restrictive to represent general images of high dimensions, and its batch-learning algorithm is restricted to relatively *small* training sets.

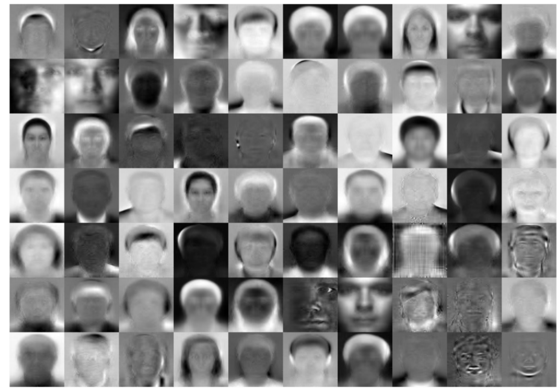

Fig. 2. A subset of the obtained atoms by OSDL.

Recently, the work in [16] proposed the Online Sparse Dictionary Learning (OSDL) algorithm which is able to manage signals of dimensions in the order of the several thousands and beyond. This approach builds on the work of [24], which models the dictionary $\mathbf{D}$ as the product of a fast and efficient *base* dictionary, and an adaptable sparse factor $\mathbf{A}$. This lowers the complexity of both, the degrees of freedom of the problem and the computational cost of applying the dictionary. This way, the dictionary learning problem is formulated as

$$\min_{\mathbf{A},\mathbf{X}} \frac{1}{2}||\mathbf{Y} - \mathbf{\Phi}\mathbf{A}\mathbf{X}||_F^2 \quad \text{subject to} \quad \begin{cases} ||\mathbf{x}_i||_0 \leq p & \forall i \\ ||\mathbf{a}_j||_0 = k & \forall j \end{cases}. \tag{3}$$

In particular, the authors in [16] employ a novel Cropped Wavelets dictionary as the operator $\mathbf{\Phi}$, leveraging the multi-scale analysis properties of wavelets while achieving a completely separable and border-effects free decomposition.

In order to cope with the increase of training data, the work in [16] proposed a dictionary learning algorithm based on ideas from stochastic optimization [25]. In a nutshell, the algorithm performs sparse coding of a mini-batch of training examples with (Sparse) OMP [26], and then updates a subset of the dictionary atoms through a variation of the Normalized Iterative Hard Thresholding algorithm [27]. For completion, we present a summary of this method in Algorithm 1, and we refer the reader to [16] for further details.

Tackling the learning of a global model for face images in particular, we apply OSDL on a compendium of face images taken from different datasets, using the freely available code at the author's website. To increase the variability of the training data – and to obtain a more general model – we employ images taken from the Chinese Passport dataset used in [28] (both in its aligned and non aligned formats), the Chicago Faces Database [29], the AT&T Faces Database[1], and the Cropped Yale Database [30]. All images were rescaled to a size of 100×100 pixels, and employed *as is*; i.e., there was no coherent scaling or alignment involved. All together, these amounted to a training set of roughly 19,000 images. OSDL took approximately 2 days to perform 40 data-sweeps[2]. We employed the Cropped Wavelets as the base dictionary (with Daubechies Wavelets with 4 vanishing moments), which has

[1]Freely available from AT&T Laboratories Cambridge's website.

[2]We run our experiment on a Windows computer with an Intel Xeon E5 CPU, with 64 Gb of RAM running Windows 64 bits. However, no parallel processing was used, and memory consumption did not exceed 16 Gb.

a redundancy of ≈ 1.7. The matrix $\mathbf{A}$ was chosen to be tall (under-complete), having 6,000 atoms in it. The atom sparsity was set to 1000; i.e., these are *only* $\approx 6\%$ sparse. We present some of the obtained atoms in Figure 2, where one can see that not only they resemble faces or face-features, but also the obtained variability between different sizes and configurations.

III. Inpainting Formulation

Once the global model has been obtained, we move to describe in detail the inpainting formulation. Consider the original image $\mathbf{y}_0 \in \mathbb{R}^n$ ($n = 10,000$), and a mask $\mathbf{M}$, given by a binary matrix of size $l \times n$, where $l = c \cdot n$. This way, c denotes the fraction of the pixels that have not been removed (and remain) from the degraded image given by $\mathbf{y} = \mathbf{M}\mathbf{y}_0$. Given this degradation model, and leveraging the obtained dictionary $\mathbf{D}$, the inpainting inverse problem can be cast in terms of a pursuit by adding a sparse regularization term. Formally,

$$\min_{\mathbf{x}} \ \|\mathbf{x}\|_0 \text{ subject to } \|\mathbf{y} - \mathbf{M}\mathbf{D}\mathbf{x}\|_2 \leq \epsilon. \tag{4}$$

This is nothing but the sparse coding problem presented in Equation (1), with the incorporation of a degradation mask. Unlike the sparse coding stage in Algorithm 1, we now turn to a relaxation of this formulation moving from the ℓ_0 to the ℓ_1 norm. This way, we replace the problem above with the unconstrained optimization problem given by

$$\min_{\mathbf{x}} \ \|\mathbf{y} - \mathbf{M}\mathbf{D}\mathbf{x}\|_2 + \lambda\|\mathbf{x}\|_1, \tag{5}$$

where λ is a the penalty parameter, compromising between the desired sparsity and the (masked) fidelity term. The shift from the ℓ_0 to the ℓ_1 norm is motivated by a practical aspect: in the inpainting problem, where one does not known a priori the number of non-zero elements needed to obtain a *good* reconstruction (or the equivalent ϵ threshold), it is easier to tune a penalty parameter λ. The number of non-zeros in $\mathbf{x}$ might be larger than those employed during the training, therefore making a greedy pursuit time consuming. In addition, we have found this ℓ_1 approach to yield solutions that are smoother, resulting in more naturally-looking inpainted areas.

Due to the convexity of the problem in Equation (5), a variety of algorithms can be employed to find its solution. Iterative shrinkage algorithms are particularly well-suited for this kind of problems, and we employ FISTA as the specific solver [31][3]. Our implementation of this method benefits from the relatively low-complexity of applying $\mathbf{D}$. Indeed, multiplying a vector by the dictionary (or its transpose) is never done explicitly. Instead, this is computed in terms of the product with the (very) sparse matrix $\mathbf{A}$ and the 1-dimensional wavelet dictionaries, which represent the separable operator $\mathbf{\Phi}$.

IV. Results

For our experiments, we applied the method described in the previous section on a set of testing images, not included in the training set. In order to demonstrate the benefits of the proposed approach based on Trainlets, we compare with a number of other methods; namely: 1) the patch-propagation method of [6], which employs a sparse (patch) prior to inpaint the image, 2) a PCA (global) learned basis, and 3) the Separable Dictionary Learning (SEDIL) algorithm [23], which also trains a global but separable dictionary. For this last method, we trained two 1-dimensional dictionaries of size 100×200 on the same training set, employing the code provided by the authors[4]. Note that both PCA and SEDIL obtain a set of global adaptive atoms by enforcing some constraints: orthogonality and separability, respectively.

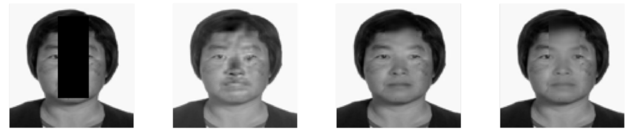

Fig. 3. Inpainting results for increasing values of λ, from left to right. The reader is referred to the supplementary material for a discussion on this point.

The inpainting algorithm resulting from the minimization of Equation (5) depends on the parameter λ, and its value influences the quality of the final reconstruction. An example is presented in Figure 3, where we inpaint the image on the left with the proposed approach for increasing values of this parameter. We expand on this point in the supplementary material accompanying this paper, and provide further demonstrations of this effect. In our experiments, and for a legitimate comparison, we run each method for a series of values of this parameter and then selected the most plausible results for each method separately[5]. The comparison with [6], on the other hand, is not entirely fair: inpaiting methods based on patch propagation are not expected to perform well in this challenging problem, as they cannot inpaint elements (mouth, eyes, etc) that do not appear in the available image region. Yet, we include them for completion and in order to demonstrate the intrinsic need of a global model.

We present a subset of our results in Figure 4, and more examples can be found in the supplementary material. As expected, the local method of [6] provides results that are not in agreement with the global context. On the other hand, the performance of SEDIL is limited, while PCA sometimes manages to recover somewhat of a natural result. Still, the constraints imposed by both of these two methods appear to be too restrictive for this problem. As can be seen, Trainlets provide the best results – often making it hard to distinguish between the original and the synthetic inpainted image. Some cases are particularly interesting: in the third image, where the glare in the glasses occlude the left eye, our approach manages to restore it; in the fourth image, we inpaint an eye which was not originally there due to lighting conditions, still in a plausible manner. More interesting examples can be found in the supplementary material.

V. Conclusion

We have presented a simple inpainting algorithm which exploits the representation power of Trainlets and the learning capabilities of the OSDL algorithm, obtaining a global model

[3]While we employ FISTA for the minimization of Equation (5), the learning algorithm (OSDL) still employs OMP for the Sparse Coding stage.

[4]Note that this is a batch method, and we employed 2,000 iterations. Training with SEDIL took approximately 2.5 days, resulting in both dictionary learning algorithms running for about the same time.

[5]Note that the selection of the best (most plausible) result is somewhat subjective, for which we have used our most fair judgment.

Fig. 4. Inpainting results. From left to right: masked image, patch propagation [6], PCA, SEDIL [23], Trainlets [16], and the original image.

for a diverse collection of face images. When this model is deployed with a sparse prior, we obtain very plausible reconstructions, outperforming competing methods. While the comparison with CNN-based models, such as that in [11], exceeds the scope of this work, we will be excited to see other groups undertake this very interesting study. To facilitate such comparisons, we give access to our Trainlets software package, along with our trained model (dictionary) and inpainting code.

An interesting observation is that once a good global model is at our disposal, there is no need for any extra algorithmic manipulation of the data: there is no symmetry, exemplar-based copying or other form of external regularization enforced in the reconstruction; this is naturally captured by the learning process. Exploring the ability of a similar approach in tackling other inverse problems is an interesting direction of research and part of ongoing work. Finally, while the our method is very effective in modeling images from a similar class, employing this approach for the inpainting of large areas in natural images is unlikely to succeed, as learning a global model for such general cases is a significantly more challenging task. In this case, improvements on the learning algorithm (and the model) would be needed before attempting to solve this kind of inverse problem.

Acknowledgments

The research leading to these results has received funding from the European Research Council under European Unions Seventh Framework Programme, ERC Grant agreement no. 320649. The authors thank Michael Zibulevsky for the insightful discussions, and Elad Richardson for his helpful assistance in the experimental section.

References

[1] C. Guillemot and O. Le Meur, "Image inpainting: Overview and recent advances," *Signal Processing Magazine, IEEE*, vol. 31, no. 1, pp. 127–144, 2014. 1

[2] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 417–424, ACM Press/Addison-Wesley Publishing Co., 2000. 1

[3] D. Tschumperlé, "Fast anisotropic smoothing of multi-valued images using curvature-preserving pde's," *International Journal of Computer Vision*, vol. 68, no. 1, pp. 65–82, 2006. 1

[4] C. Ballester, M. Bertalmio, V. Caselles, G. Sapiro, and J. Verdera, "Filling-in by joint interpolation of vector fields and gray levels," *Image Processing, IEEE Transactions on*, vol. 10, no. 8, pp. 1200–1211, 2001. 1

[5] A. Criminisi, P. Pérez, and K. Toyama, "Region filling and object removal by exemplar-based image inpainting," *Image Processing, IEEE Transactions on*, vol. 13, no. 9, pp. 1200–1212, 2004. 1

[6] Z. Xu and J. Sun, "Image inpainting by patch propagation using patch sparsity," *Image Processing, IEEE Transactions on*, vol. 19, no. 5, pp. 1153–1165, 2010. 1, 3, 4

[7] O. G. Guleryuz, "Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising-part ii: adaptive algorithms," *Image Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 555–571, 2006. 1

[8] A. Bugeau, M. Bertalmío, V. Caselles, and G. Sapiro, "A comprehensive framework for image inpainting," *Image Processing, IEEE Transactions on*, vol. 19, no. 10, pp. 2634–2645, 2010. 1

[9] O. Le Meur and C. Guillemot, "Super-resolution-based inpainting," in *Computer Vision–ECCV 2012*, pp. 554–567, Springer, 2012. 1

[10] G. Peyré, "Manifold models for signals and images," *Computer Vision and Image Understanding*, vol. 113, no. 2, pp. 249–260, 2009. 1

[11] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros, "Context encoders: Feature learning by inpainting," 2016. 1, 4

[12] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images," *SIAM Review.*, vol. 51, pp. 34–81, Feb. 2009. 1

[13] J. Mairal, F. Bach, and J. Ponce, "Sparse modeling for image and vision processing," *Foundations and Trends® in Computer Graphics and Vision*, vol. 8, no. 2-3, pp. 85–283, 2014. 1

[14] M. Aharon, M. Elad, and A. M. Bruckstein, "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Trans. on Signal Process.*, vol. 54, no. 11, pp. 4311–4322, 2006. 1, 2

[15] J. Mairal, F. Bach, and G. Sapiro, "Non-local Sparse Models for Image Restoration," *IEEE International Conference on Computer Vision.*, vol. 2, pp. 2272–2279, 2009. 1, 2

[16] J. Sulam, B. Ophir, M. Zibulevsky, and M. Elad, "Trainlets: Dictionary learning in high dimensions," *IEEE Transactions on Signal Processing*, vol. 64, no. 12, pp. 3180–3193, 2016. 1, 2, 4

[17] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal Matching Pursuit: Recursive Function Approximat ion with Applications to Wavelet Decomposition," *Asilomar Conf. Signals, Syst. Comput. IEEE.*, pp. 40–44, 1993. 2

[18] J. A. Tropp, "Just Relax : Convex Programming Methods for Identifying Sparse Signals in Noise," *IEEE Transactions on In*, vol. 52, no. 3, pp. 1030–1051, 2006. 2

[19] Y. Romano, M. Protter, and M. Elad, "Single image interpolation via adaptive nonlocal sparsity-based modeling," *IEEE Trans. on Image Process.*, vol. 23, no. 7, pp. 3085–3098, 2014. 2

[20] J. Sulam and M. Elad, "Expected patch log likelihood with a sparse prior," in *Energy Minimization Methods in Computer Vision and Pattern Recognition*, Lecture Notes in Computer Science, pp. 99–111, Springer International Publishing, 2015. 2

[21] R. Giryes and M. Elad, "Sparsity-based poisson denoising with dictionary learning," *Image Processing, IEEE Transactions on*, vol. 23, no. 12, pp. 5057–5069, 2014. 2

[22] K. Engan, S. O. Aase, and J. H. Husoy, "Method of Optimal Directions for Frame Design," in *IEEE Int. Conf. Acoust. Speech, Signal Process.*, pp. 2443–2446, 1999. 2

[23] S. Hawe, M. Seibert, and M. Kleinsteuber, "Separable dictionary learning," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 438–445, 2013. 2, 3, 4

[24] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double Sparsity : Learning Sparse Dictionaries for Sparse Signal Approximation," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1553–1564, 2010. 2

[25] L. Bottou, "Online algorithms and stochastic approximations," in *Online Learning and Neural Networks*, Cambridge University Press, 1998. revised, Oct 2012. 2

[26] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient Implementation of the K-SVD Algorithm using Batch Orthogonal Matching Pursuit," *Technion - Computer Science Department - Technical Report.*, pp. 1–15, 2008. 2

[27] T. Blumensath and M. E. Davies, "Normalized iterative hard thresholding: Guaranteed stability and performance," *IEEE Journal on Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 298–309, 2010. 2

[28] O. Bryt and M. Elad, "Compression of facial images using the K-SVD algorithm," *J. Vis. Commun. Image Represent.*, vol. 19, pp. 270–282, May 2008. 2

[29] D. S. Ma, J. Correll, and B. Wittenbrink, "The chicago face database: A free stimulus set of faces and norming data," *Behavior research methods*, vol. 47, no. 4, pp. 1122–1135, 2015. 2

[30] A. Georghiades, P. Belhumeur, and D. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 23, no. 6, pp. 643–660, 2001. 2

[31] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM journal on imaging sciences*, vol. 2, no. 1, pp. 183–202, 2009. 3