(Modern) Regularization Methods for Inverse Problems (and Machine Learning)

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

Martin Burger

FAU Erlangen-Nürnberg

Joint work with

Many colleagues over the last years

Martin Benning, Leon Bungert, Eva-Maria Brinkmann, Joana Grah, Pia Heins, Meike Kinzel, Jahn Müller, Michael Möller

Guy Gilboa, Tapio Helin, Hanne Kekkonen, Alexandra Koulouri, Elena Resmerita, Stanley Osher, Wotao Yin

Survey reference:

M.Benning, M.Burger, Modern regularisation iver
Acta Numerica 2018.

Deutsche
Forschungsgemeinschaft
DFG

European Research Council
Established by the European Commission

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Inverse Problems and Ill-posedness

Consider operator equation

$$Ku = f$$

with compact operator K acting between Banach spaces

This problem is ill-posed:

- Potential non-existence or non-uniqueness of solutions
- Instability, i.e. discontinuous dependence on data *f*

Model: Forward vs. Inverse

Example image reconstruction

Forward operator maps image *u* to indirect data *f*

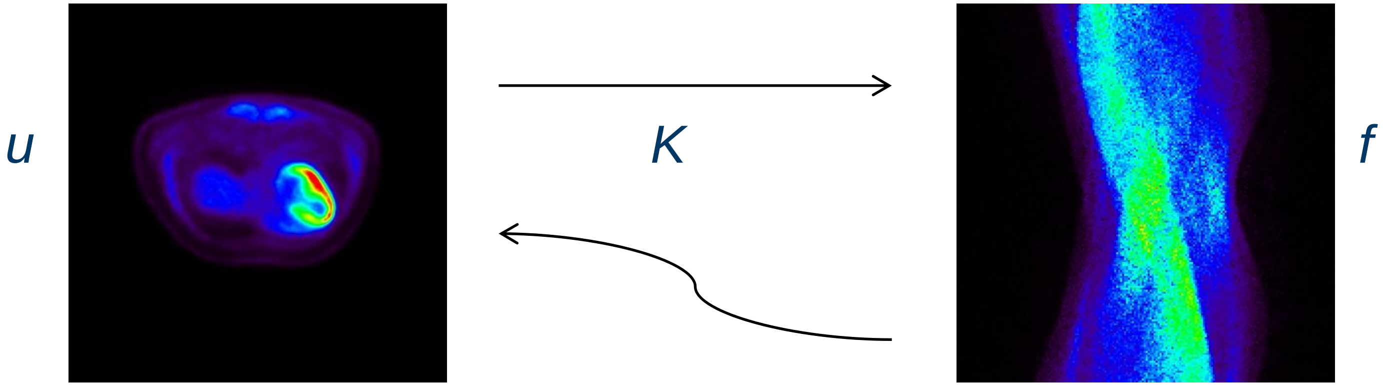


Image reconstruction is the solution of the of the operator equation (application of the inverse operator)

Regularization

The problem needs to be approximated by well-posed one(s)

Topologists answer: restrict domain of *u* to a compact set
[Tikhonov 1943]

Hilbert space theory: approximate least-squares

$$\|Ku - f\|^2 \to \min_u$$

by

$$\|Ku - f\|^2 + \alpha\|u\|^2 \to \min_u$$

[Tikhonov 43/63,Phillips 62,Ridge 70] [Tikhonov, Glasko 64, Morozov 66]

The 70s and 80s

Basic analysis of linear regularization methods in Hilbert spaces, convergence as noise level and regularization parameter tend to zero

First error estimates in dependence of noise level and regularization parameter

[Nashed-Votruba 73/74, Nashed-Wahba 74, Groetsch 80, Groetsch 84, Natterer 80/83]

Application to integral equations of the first kind and Radon inversion (CT)

Projection methods [Natterer 1977]

Iterative regularization methods [Vasilev 83,Groetsch 85,Vainniko 86]

Truncated singular value decomposition [Elden 77, Hansen 86]

The 90s: Nonlinearity

General convergence analysis based singular value decomposition

Quantification of ill-posedness based on decay of singular values

Complete theory of linear regularization finished in the 90s

[Engl-Hanke-Neubauer 96]

Regularization methods for nonlinear inverse problems

- Tikhonov regularization [Tikhonov-Arsenin 77,Seidman-Vogel 87, Engl-Kunisch-Neubauer 89]

- Iterative regularization methods [Hanke-Neubauer-Scherzer 95, Scherzer 95] [Hanke 96, Kaltenbacher-Neubauer-Scherzer 97,Hohage 97] [Hanke, Groetsch 96]

Analysis based on variational techniques, local linearizations of the operator

Modern regularization methods

Paradigms of the 21st century:

- Investigate detailed structure of regularized solutions, non-asymptotic
- Make structured use of available prior information

Methods based on:

- Sparsity and similar priors
- Bayesian prior distributions
- Machine learning and available large data sets

Mostly related to variational methods

Variational Models

Combine fitting term measuring distance between predicted data and f^δ (measured noisy data) with regularization functional *J*

$$\hat{u} \in \arg\min_u \left(F(Ku, f^\delta) + \alpha J(u) \right)$$

Optimality condition, convex *J*

$$K^* \partial_x F(Ku, f) + \alpha p = 0, \qquad p \in \partial J(u).$$

Source condition (range condition): any solution of the variational method satisfies

$$p = K^* w$$

Variational Models

Relation to Bayesian estimation

$$\pi(u|f) = \frac{1}{\pi_*(f)}\pi(f|u)\pi_0(u)$$

Maximum a-posteriori probability estimate satisfies

$$\hat{u} \in \arg\min_u \left(-\log \pi(f|u) - \log \pi_0(u)\right)$$

Compare

$$\hat{u} \in \arg\min_u \Big(F(Ku, f) + \alpha J(u) \Big)$$

Fidelity term

Data fidelity term *F* comes from statistical model of the forward process: negative log-likelihood

Example: additive Gaussian noise leads to quadratic fidelity term

$$\frac{1}{2}\|Ku - f\|_{\Sigma^{-1}}^2$$

Example: Poisson noise (frequent in imaging with photon counts) leads to Kulback-Leibler divergence

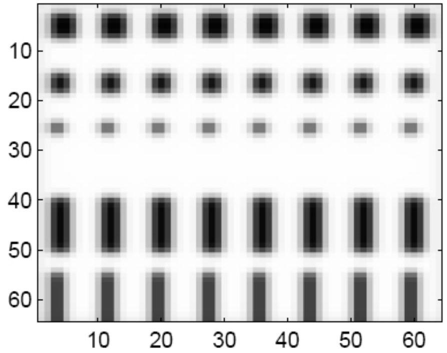
Choice of regularization

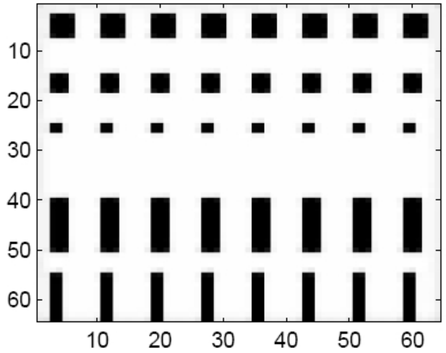
How to choose a suitable regularization functional ?

Simple choice: Gaussian prior = quadratic regularization functional

Example: Sobolev seminorms to enforce smoothness

$$J(u) = \int |\nabla u|^2 \; dx$$





Problem: oversmoothing

Overspmoothing of simple regularizations

Typically forward operator smoothing, i.e. defined on smaller space

Example: $K : L^2 \to Y$

Source condition $p = -\Delta u = K^* w \in L^2$

Elliptic regularity implies at least $u \in H^2$

Hence, no discontinuity, i.e. no edges

Choice of regularization

Alternative: p-Laplacian energy

Similar smoothing properties as long as p > 1, hence consider total variation

$$TV(u) = |u|_{BV} := \sup_{g \in C_0^\infty(\Omega)^d, g \in \mathcal{C}} \int_\Omega u \nabla \cdot g \; dx$$

$$\mathcal{C} = \{ g \in L^\infty(\Omega) \mid |g(x)| \leq 1 \text{ a.e. in } \Omega \}$$

Optimality condition $\quad K^* \partial_x F(Ku, f) + \alpha \nabla \cdot g = 0$

$$g \in \mathcal{C} \qquad \int_\Omega g \cdot dDu = |u|_{BV}$$

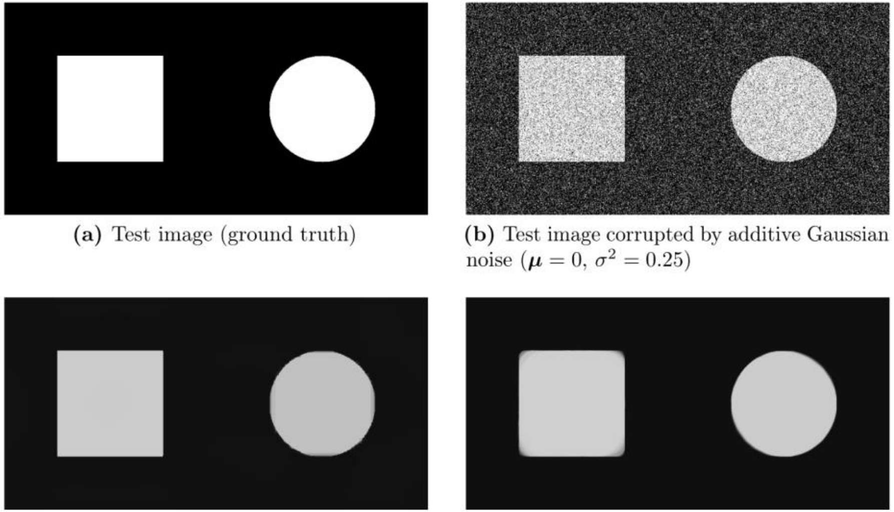
Choice of regularization

Source condition

$$\nabla \cdot g = K^* w$$

Note that g corresponds to (generalized) normal vector field on level sets (discontinuity sets) of u, its divergence equals mean curvature

Consequence: solutions of TV regularization can be discontinuous, but have nice discontinuity sets (smooth curvature)



(a) Test image (ground truth)

(b) Test image corrupted by additive Gaussian noise ($\mu = 0$, $\sigma^2 = 0.25$)

(c) Anisotropic TV denoising result ($\alpha = 10$)

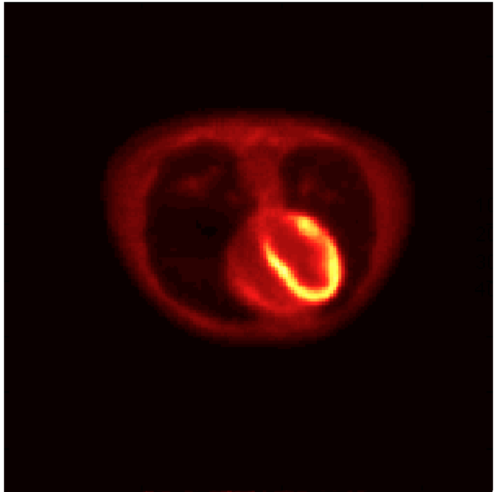
(d) Isotropic TV denoising result ($\alpha = 10$)

Total Variation Regularization

Example: PET reconstruction (inversion of Radon transform with Poisson noise)

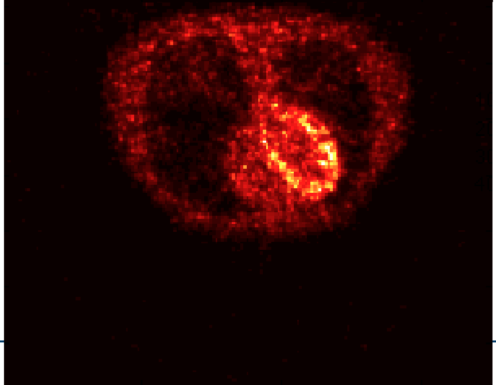
[Müller et al 2013]

20min data
(low noise)

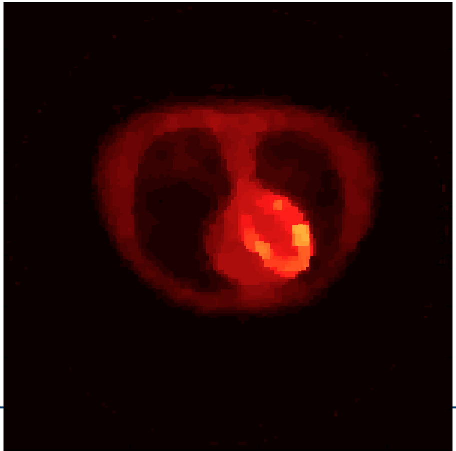


Simple Recon
(EM)

5s data
(high noise)



Simple Recon



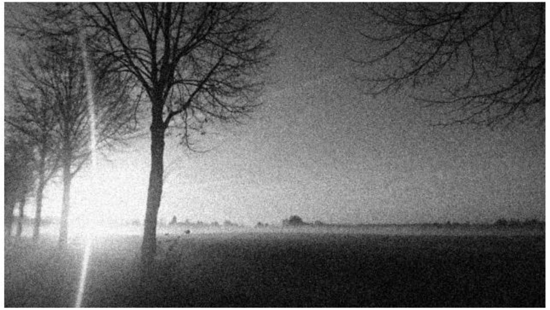
TV

Variants of total variation

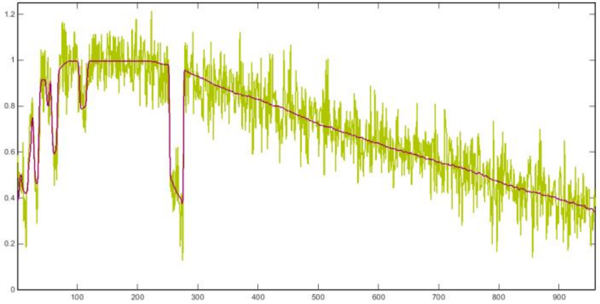
TV regularization suffers from staircasing: piecewise smooth parts often reconstructed by stair-type structure

Example: denoising
K = embedding operator to L^2
[Rudin-Osher-Fatemi 1992]

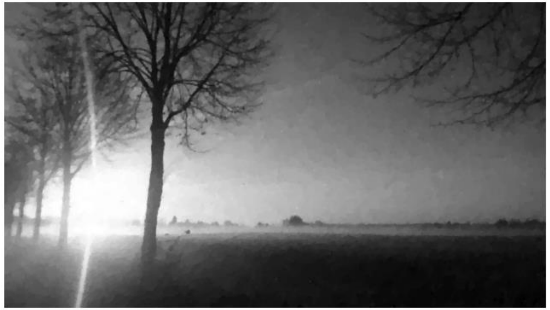
[PhD Brinkmann 2019]



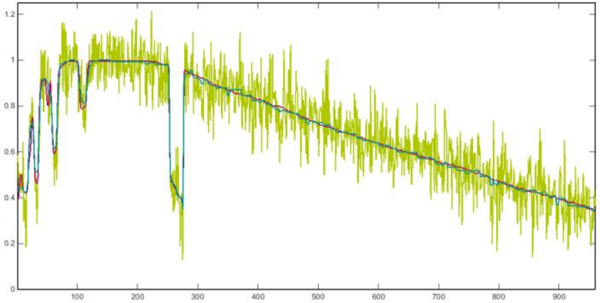
(g) Noisy grayscale photo corrupted by additive Gaussian noise ($\mu = 0, \sigma^2 = 0.01$).



(h) Line profile of the noisy image and the corresponding row in the noiseless image.



(i) TV denoising result with $\alpha = 0.095$.



(j) Line plot of the reconstructed image compared against the noisy and the noiseless image.

Variants of total variation

TV regularization suffers from staircasing: piecewise smooth parts often reconstructed by stair-type structure

Improved versions by infimal convolution [Chambolle-Lions 1997]

$$J(u) = \inf_{u_1+u_2=u} \left(|u_1|_{BV} + |\nabla u_2|_{BV} \right)$$

or total general variation [Bredies-Kunisch-Pock 2010]

$$J(u) = \inf_{Du_1+u_2=Du} \left(|u_1|_{BV} + |u_2|_{BV} \right)$$

Various other generalizations to higher-dimensional (spectral) and time-dependent images

Sparsity

Basic idea in compressed sensing: choose simple solution (minimal combinations) [Donoho 2006, Candes-Tao 2006]

Analysis formulation: for some frame system choose

$$J(u) = \sum_i |\langle u, \phi_i \rangle|$$

Synthesis formulation:

$$J(u) = \sum_i |c_i| \qquad \text{where } u = \sum_i c_i \phi_i$$

Sparsity

Analysis in synthesis formulation

Redefine forward operator

$$\tilde{K} : \ell^2(\mathbb{N}) \to \mathcal{V}, \quad c \mapsto \sum_i c_i K \phi_i$$

Rewritten variational problem

$$\hat{u} = \sum_i \hat{c}_i \phi_i, \quad \hat{c}_i \in \arg\min_c F(\tilde{K}c, f) + \alpha |c|_1$$

Optimality condition

$$(\tilde{K}^* \partial_x F(\tilde{K}\hat{c}, f))_i + \alpha s_i = 0 \qquad s_i \in \operatorname{sign}(\hat{c}_i)$$

Implies sparsity, since only few signs +1 / -1 are possible

One-homogeneous Functionals

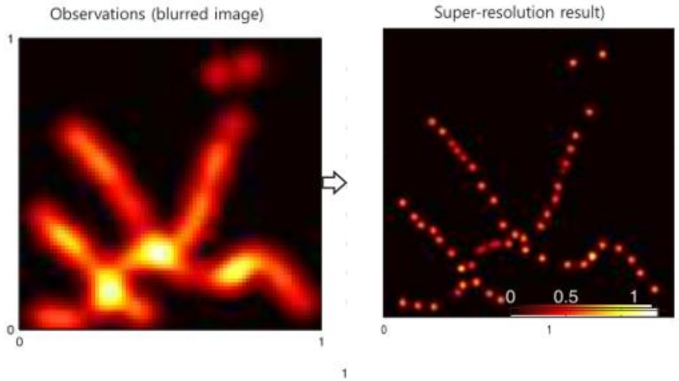
Several other successfull one-homogeneous functionals:

Examples:

- continuum sparsity (total variation of a measure)

[Bredies-Pikkarainien 2013, Duval-Peyre et al 2013-2019]

[mb-Heins-Koulouri 2020]



- Group Sparsity [Eldar-Mishali 2009] [PhD Heins 2014, PhD Kinzel 2021]

- Local Sparsity [mb-Heins-Möller 2014]

- Low rank (nuclear norm of matrix or tensor) [Candes 2010], [Phd Kinzel 2021]

Learned Regularizations

Machine learning became attractive in the last years

First idea: learn reconstruction directly

Problems:

- complexity of the inverse problem
- bad generalization (network must have huge Lipschitz constant)
- missing data, hardly pairs of input-output

Alternative: stay close to variational methods and learn regularization only. In image reconstruction this mainly requires favourable (maybe also unfavourable images)

Learned Regularizations

Still solve
$$\hat{u} \in \arg\min_u \Big(F(Ku, f^\delta) + \alpha J(u) \Big)$$

But *J* (and maybe α) obtained from deep learning, given a database of images

Example: adversarial learning [Lunz-Öktem-Schönlieb 18]

Given favourable images $\{u_i\}_{i=1}^n$ and unfavourable ones $\{v_k\}_{k=1}^m$ minimize (with respect to parameters)

$$\frac{1}{n}\sum_{i=1}^n J(u_i) - \frac{1}{m}\sum_{k=1}^m J(v_k) + \lambda \mathbb{E}[(\|\nabla J\| - 1)_+^2]$$

Learned Regularizations

Learned regularization method is itself a random variable in terms of training data

As n and m tend to infinity and under assumption of i.i.d. sampling from appropriate distributions expect convergence to minimizer of

$$\mathbb{E}_u(J) - \mathbb{E}_v(J) + \lambda \mathbb{E}[(\|\nabla J\| - 1)_+^2]$$

Detailed properties of regularizer and subsequent solutions of inverse problem remain unclear

Learned Regularizations

So far, functionals learned based on data sets, but independent of inverse problem (forward operator *K*)

Attractive for computations, but difficult to analyze. Unclear if training data could even be solution of inverse problem

Alternative with guaranteed range condition: minimize

$$\frac{1}{n}\sum_{i=1}^{n}\|(K^*)^{-1}\partial_u J(u_i)\|^2$$

Possibly augmented with other terms, [mb-Mukherjee-Schönlieb, in prep.]

Error estimation

Quantitative estimates in ill-posed problems available only under additional assumptions (conditional stability)

Basic principle: take two elements satisfying source condition

$$p_i = K^* w_i \in \partial J(u_i), \quad i = 1, 2$$

Hölder stability estimate

$$\begin{aligned}\langle p_1 - p_2, u_1 - u_2 \rangle &= \langle K^*(w_1 - w_2), u_1 - u_2 \rangle \\ &\leq (\|w_1\| + \|w_2\|)\|Ku_1 - Ku_2\|\end{aligned}$$

Error estimation

Quantity to be estimated is (symmetric) Bregman distance

$$\begin{aligned} D_J^{\text{symm}}(u_1, u_2) = &\langle p_1 - p_2, u_1 - u_2 \rangle \\ = &(J(u_2) - J(u_1) - \langle p_1, u_2 - u_1 \rangle) + \\ &(J(u_1) - J(u_2) - \langle p_2, u_1 - u_2 \rangle) \end{aligned}$$

Implies directly estimates for both one-sided Bregman distances

Those are limits of scaled Jensen distances for *s=0* and *s=1*

$$D_s(u_1, u_2) = \frac{1}{s(1-s)}(sJ(u_1) + (1-s)J(u_2) - J(su_1 + (1-s)u_2))$$

Error estimation

Estimates between solution of regularized problem

$$\hat{u} \in \arg\min_u \left(F(Ku, f^\delta) + \alpha J(u) \right)$$

and „exact solution“

$$Ku^* = f, \quad p^* = K^* w^* \in \partial J(u^*)$$

Theorem [mb-Osher 2004]

Assumptions as above, J convex. Then

$$\frac{1}{2\alpha} \|K(\hat{u} - u^*)\|^2 + D_J^{\text{symm}}(\hat{u}, u^*) \leq \alpha \|w^*\|^2 + \frac{1}{\alpha} \|f - f^\delta\|^2$$

Error estimation

Corollary 1: a-posteriori estimate

$$J(u^*) - J(\hat{u}) - \langle \hat{p}, u^* - \hat{u} \rangle \leq \alpha \|w^*\|^2 + \frac{1}{\alpha}\|f - f^\delta\|^2$$

Note: single estimate for subgradient appearing in optimality condition

Corollary 2: a-priori estimate

$$J(\hat{u}) - J(u^*) - \langle p^*, \hat{u} - u^* \rangle \leq \alpha \|w^*\|^2 + \frac{1}{\alpha}\|f - f^\delta\|^2$$

Note: possibly multivalued estimate for any subgradient satisfying source condition !

Error estimation

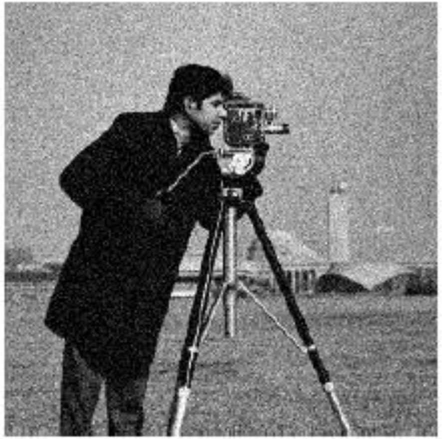
How to make this precise:

- Better characterization of source conditions (too abstract)
- Derive more interpretable quantities from Bregman distances

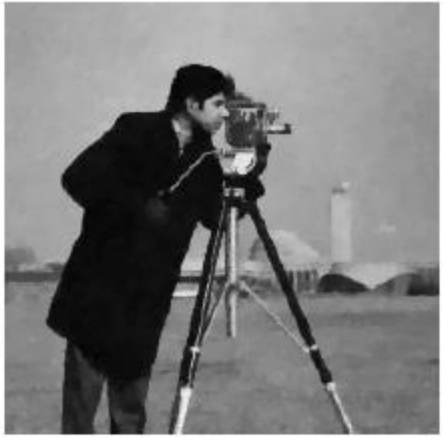
Example: TV denoising of 2D images, K=embedding operator to L^2

clean

noisy



TV result



Source condition

Key issue is the understanding of subgradients

For TV divergence of a generalized normal vector fields

$$p = \nabla \cdot g$$

Assume *u** is piecewise constant with smooth discontinuity set *S* (*C*^{1} and square integrable curvature)

Then we can explicitly construct subgradients respectively g:

Choose *G* with

$$\mathrm{supp}(G) \subset [-1,1], \quad G(0) = 1$$

and

$$g^* = G(\frac{b}{\epsilon})\nabla b$$

Then the source condition is satisfied by $w^* = p^* = \nabla \cdot g^*$

Source condition

We can compute

$$\|w^*\|^2 = \int (\nabla \cdot g)^2 \; dx = \int_{S_\epsilon} (\epsilon^{-1} G' + G \Delta b)^2 \; dx$$

Lemma

There exists a constant *C* such that for ε sufficiently small

$$\|w^*\|^2 \leq C(\epsilon^{-1} \mathrm{Per}(S) + \epsilon \int_S H^2 \; d\sigma)$$

Bregman distance

For one-homogeneous functionals we have

$$J(u^*) = \langle p^*, u^* \rangle$$

Hence Bregman distance becomes

$$J(\hat{u}) - \langle p^*, \hat{u} \rangle = \langle \hat{p} - p^*, \hat{u} \rangle$$

Theorem

For ε sufficiently small

$$TV(\hat{u}|_{\Omega^\epsilon}) \leq C\alpha(\epsilon^{-1}\mathrm{Per}(S) + \epsilon \int_S H^2 \; d\sigma) + \alpha^{-1}\|f - f^\delta\|^2$$

Conclusion: small variation of solution away from *S*

Error estimation

Various generalizations

- Estimates for other schemes (iterative, inverse scales space, gradient flows, discretizations …) [mb-Resmerita-He 2007, Schuster-Kaltenbacher-Hofmann 2012, Grasmair et al, Hofmann et al, mb et al 2008-2021]
- Improved estimates under stronger source conditions [Resmerita 2006, Sprung-Hohage 2019]
- Approximate source conditions / unbounded noise [Hein 2008, Hofmann et all 2008-2021]. [mb-Helin-Kekkonen 2018]

$$d_\rho(\vartheta) := \inf_{w \in Y} \{\|K^* w - \vartheta\|_{X^*} \mid \|w\|_Y \leq \rho\}$$

- Different exponents in fidelity and regularization are rescaling [mb-Bungert 2019]
- Sharpness of estimates by singular vector examples [Benning-mb 2012]

Properties of Variational Regularizations

Possible solutions characterized by range / source conditions

Closer characterization of properties: nonlinear singular vectors

Benning-mb 2012/2018

In the case

$$F(Ku, f^\delta) = G(Ku - f^\delta)$$

we can define a generalized singular system

$$Ku_\sigma = \sigma v_\sigma \quad \text{and} \quad K^* G'(v_\sigma) \in \partial J(\sigma u_\sigma)$$

Singular vectors

For simplicity consider here *J* one-homogeneous and

$$G(f) = \frac{1}{2}\|f\|^2, \quad G' = I$$

Singular vectors satisfy nonlinear eigenvalue problem

$$K^*Ku_\sigma = \sigma p_\sigma, \quad p_\sigma \in \partial J(u_\sigma)$$

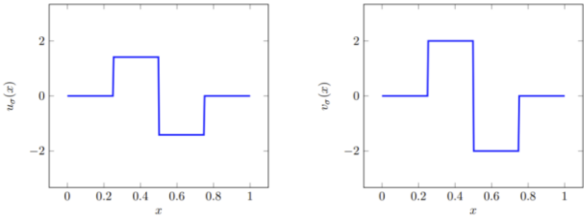
Look at data generated from singular vector

$$f = \lambda K u_\sigma$$

Singular vectors

Systems of nonlinear singular vectors (eigenfunctions) can be of interest for themselves

1D: connection TV regularization – Haar wavelets



(a) $u_{\sigma_1,1/2}$

(b) $v_{\sigma_1,1/2}$

$$u_{\sigma_n,k}(x) := 2^{\frac{n}{2}}\Psi(2^n x - \boldsymbol{k}) \quad \text{with} \quad \Psi(x) := \begin{cases} 1 & x \in [0, \frac{1}{2}) \\ -1 & x \in [\frac{1}{2}, 1) \\ 0 & \text{else} \end{cases}$$

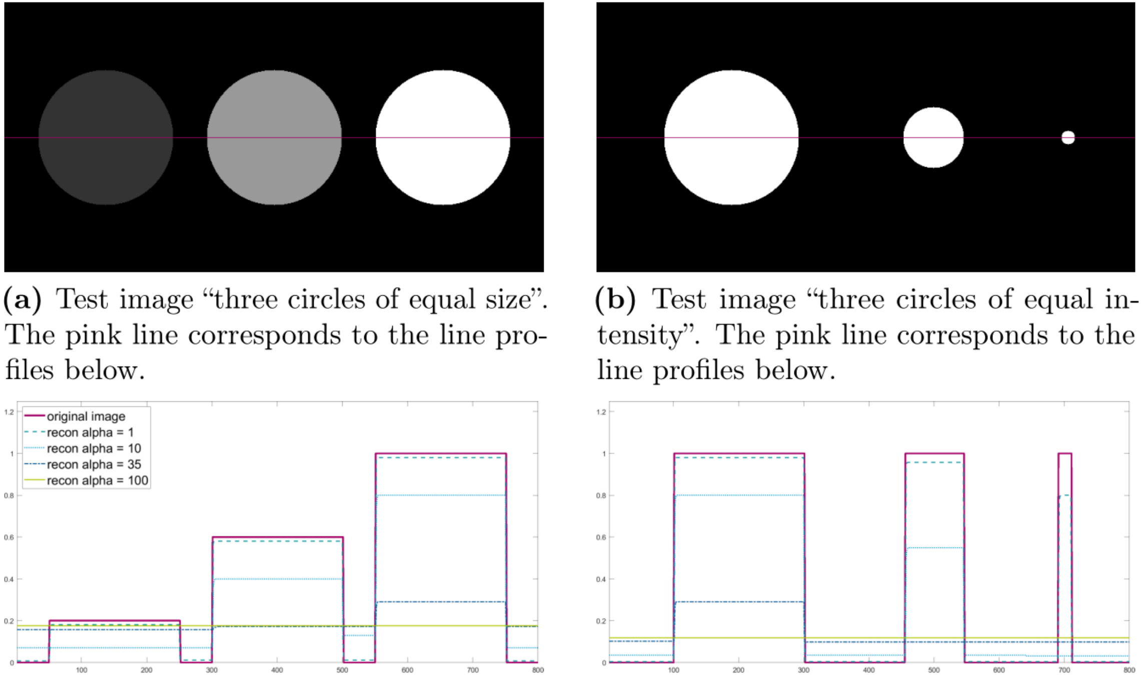
Haar wavelet system = singular vectors of TV with zero Dirichlet values, *K*=embedding operator to *L*^{2}

Bias

In total variation regularization bias = loss of contrast

[Meyer 2002]

[PhD Brinkmann 2019]



(a) Test image “three circles of equal size”. The pink line corresponds to the line profiles below.

(b) Test image “three circles of equal intensity”. The pink line corresponds to the line profiles below.

(c) Line profiles of ROF reconstructions for the above image for several values of α compared against the original (pink).

(d) Line profiles of ROF reconstructions for the above image for several values of α compared against the original (pink).

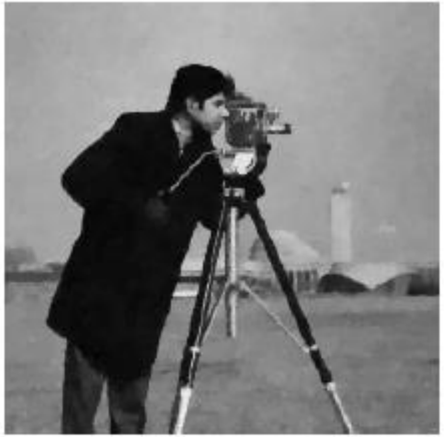
Bias correction

Unfortunately local loss of contrast = missing structures

clean

noisy

u



f-u

Bias correction

Simple debiasing method in l1 synthesis approach:

$$\text{minimize } F(\tilde{K}c, f) \text{ subject to } s_i \in \text{sign}(c_i)$$

Can be written as

$$\text{minimize } F(\tilde{K}c, f) \text{ subject to } \sum_i (|c_i| - |\hat{c}_i| - s_i(c_i - \hat{c}_i)) = 0$$

Generalization

$$\text{minimize } F(Ku, f) \text{ subject to } J(u) - J(\hat{u}) - \langle \hat{p}, u - \hat{u} \rangle = 0$$

Bregman iteration

Approximation with penalty

$$\text{minimize } F(Ku, f) + \frac{1}{\tau}\left(J(u) - J(\hat{u}) - \langle \hat{p}, u - \hat{u} \rangle\right)$$

Can be done in multiple steps: Bregman iteration [Bregman 1967] [Hestenes 1969, Powell 1969] [Osher-mb-Goldfarb-Xu-Yin 2005]

$$u^{k+1} \in \arg\min_u F(Ku, f) + \frac{1}{\tau}\Big(J(u) - J(u^k) - \langle p^k, u - u^k \rangle\Big)$$

Optimality condition = dual update

$$p^{k+1} = p^k + \tau K^* \partial F(Ku^{k+1}, f)$$

Bregman iteration, Inverse Scale Space

Bregman Iteration

$$p^{k+1} = p^k + \tau K^* \partial F(K u^{k+1}, f)$$

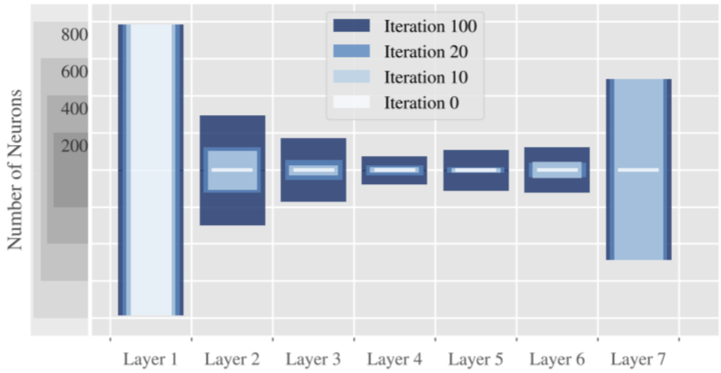
Can also be interpreted as implicit Euler discretization with time step τ

Limit is rather degenerate evolution equation, inverse scale space flow

$$\partial_t p = -K^* \partial F(Ku, f), \quad p \in \partial J(u)$$

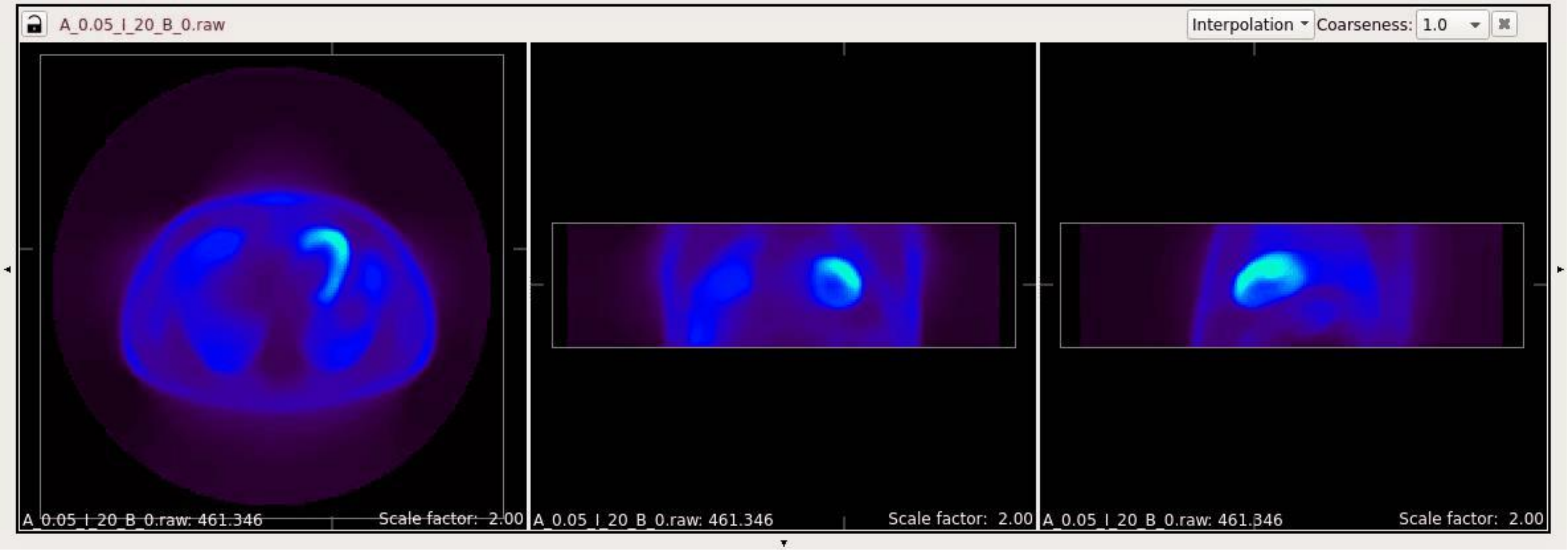
[mb-Gilboa-Osher-Xu 2006, mb-Frick-Osher-Scherzer 2007, Brune-Sawatzky-mb 2011,mb-Möller-Benning-Osher 2012]

Recent development: stochastic linearized Bregman methods for training sparse deep neural networks [Bungert-Roith-Tenbrinck-mb 2021]



PET Reconstruction

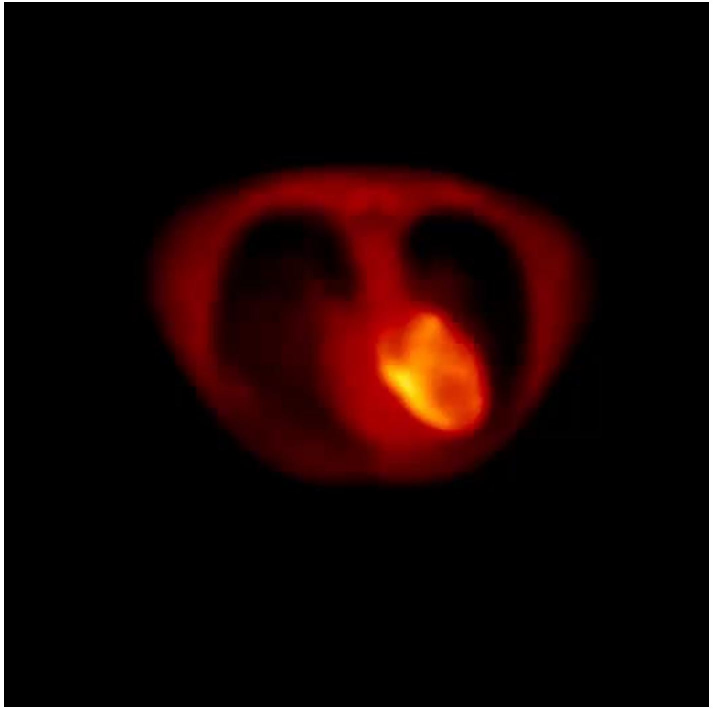
Increasing Bregman iterations

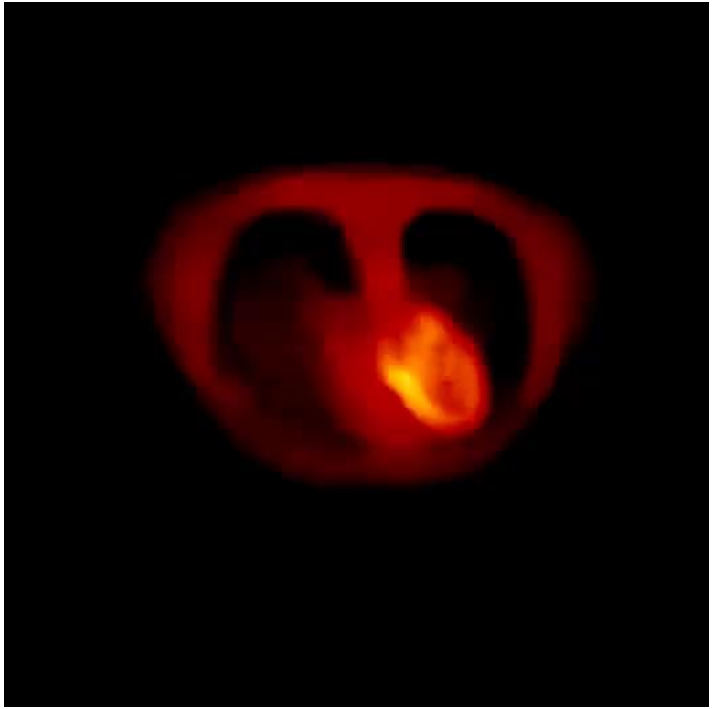


Cardiac PET Reconstruction

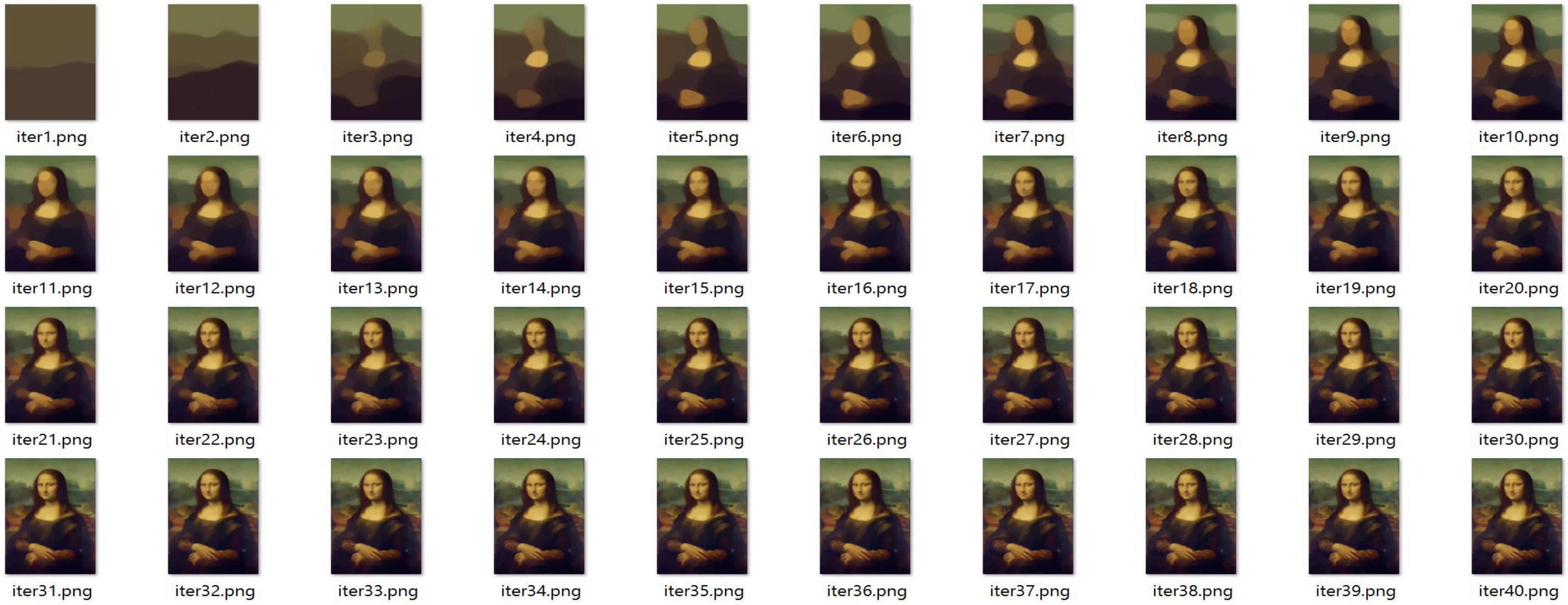
20 min data, simple

5s data Bregman TGV





Multiscale Decomposition

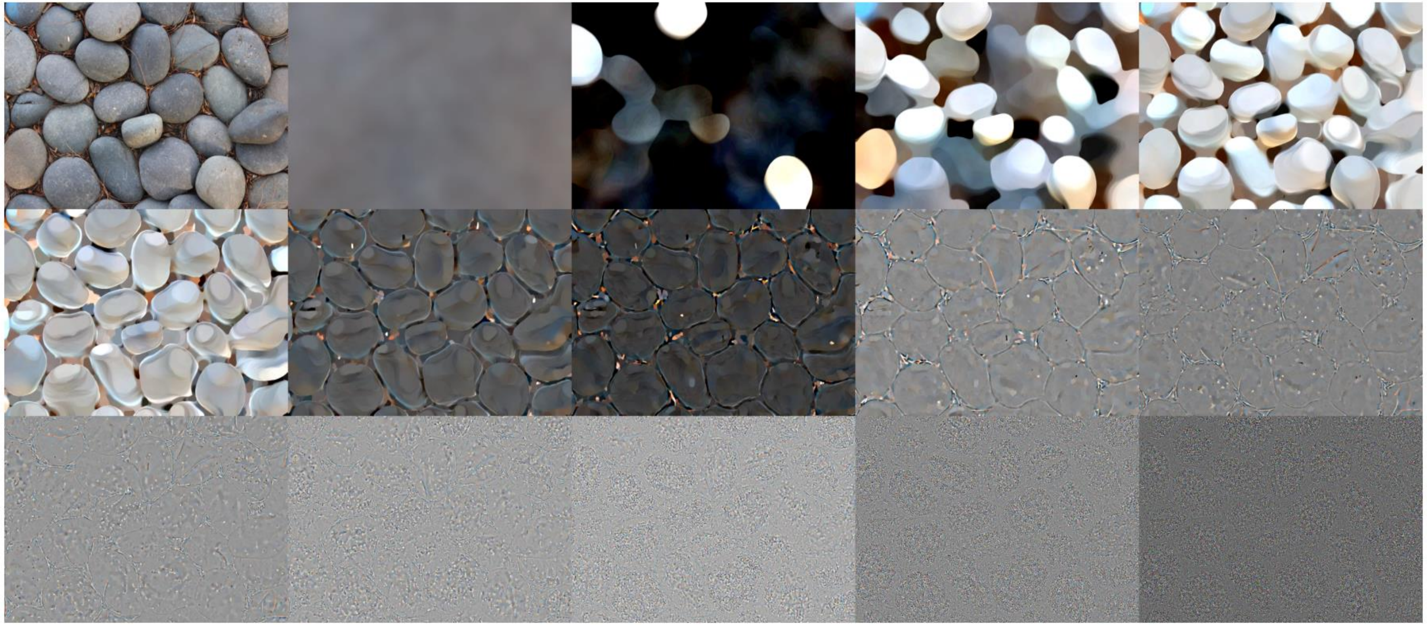


Multiscale Decomposition

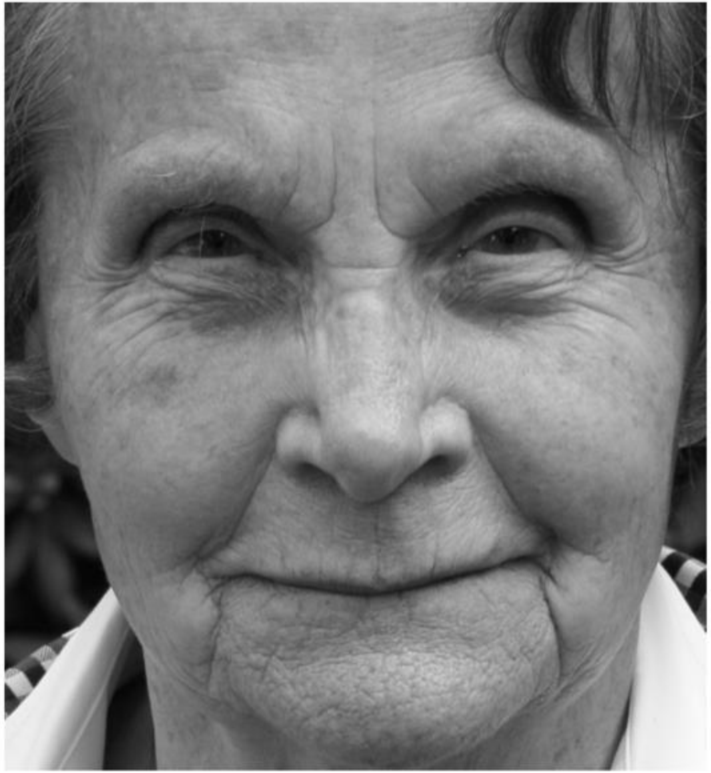
Inverse scale space method

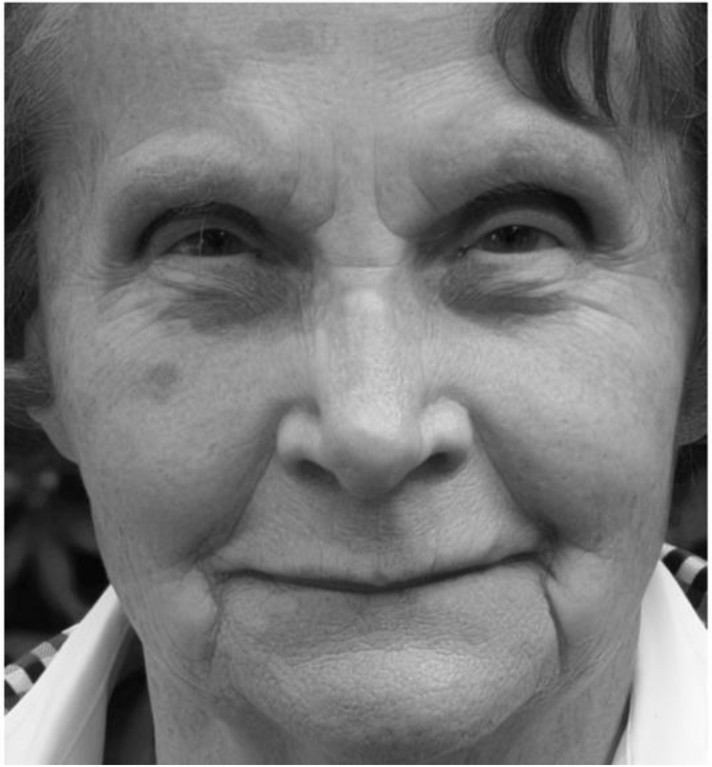
$$\partial_t p = -K^* \partial F(Ku, f), \quad p \in \partial J(u)$$

Take (scaled) time derivatives of *u*

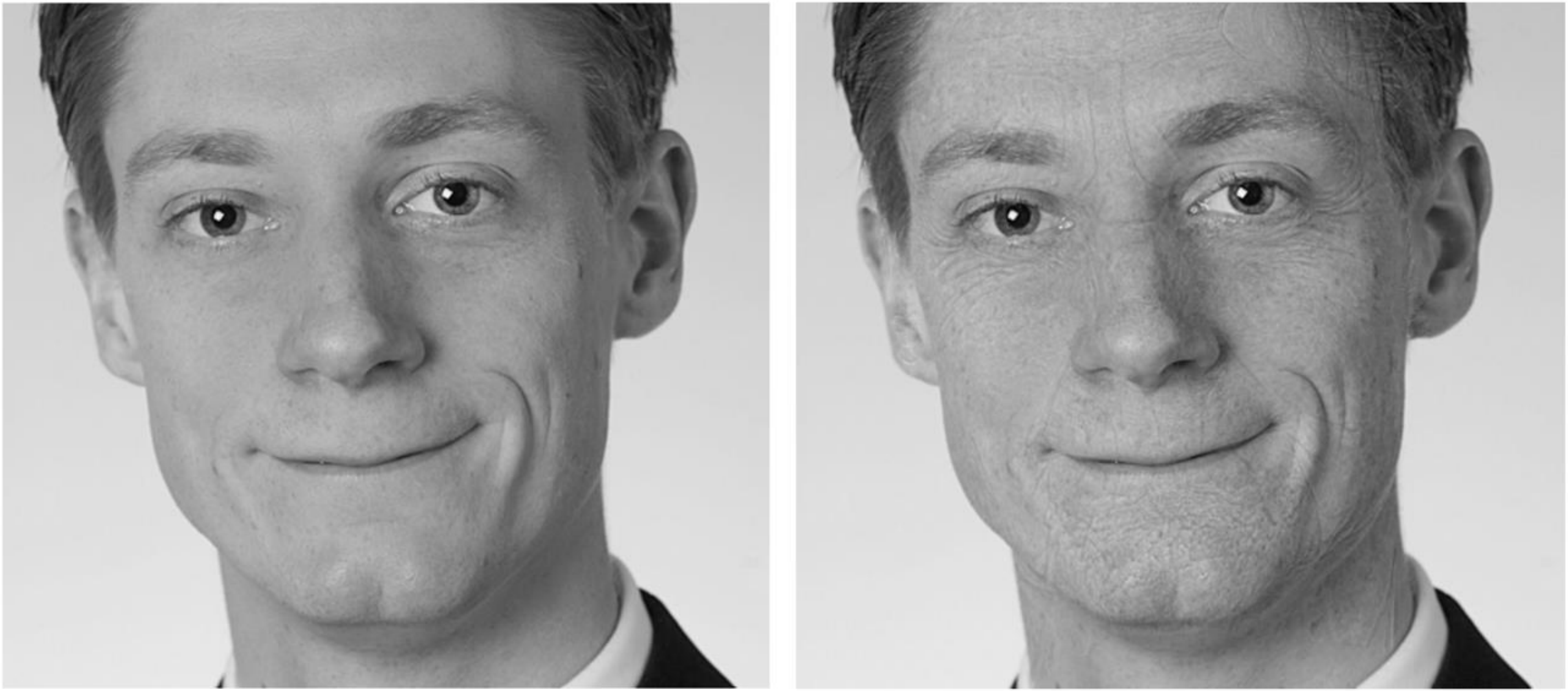


Filtering: Juvenation

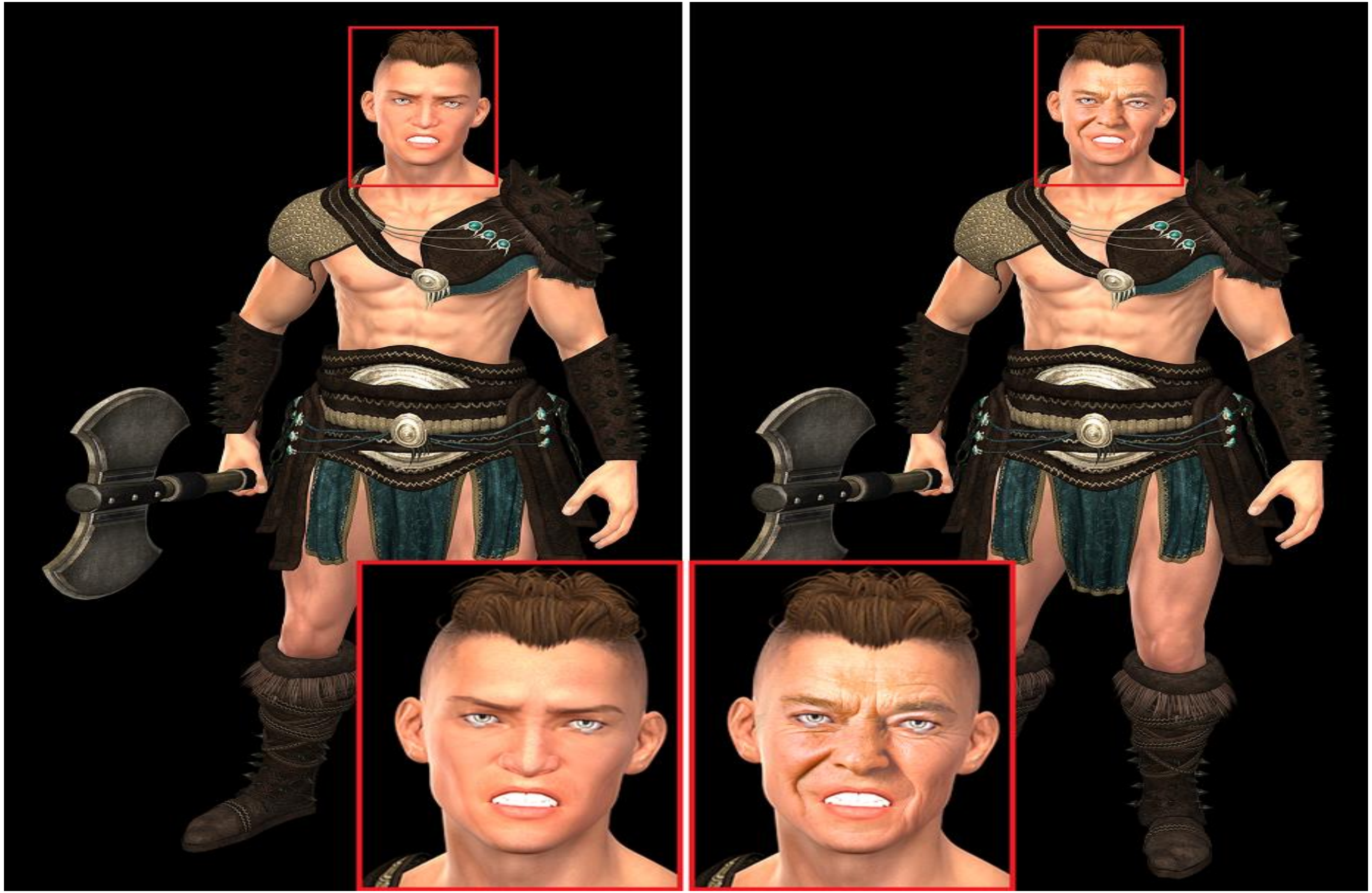




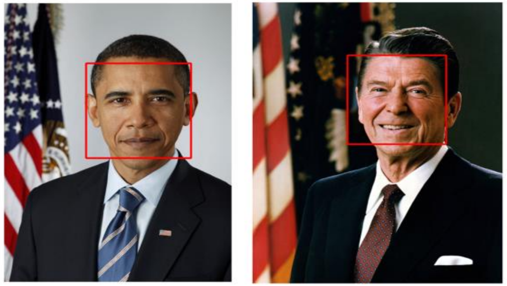
Filtering: Ageing



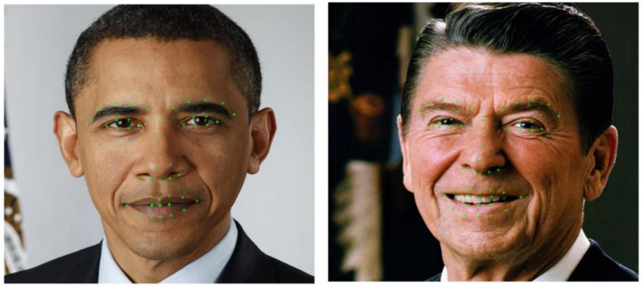
Personalized Avatar



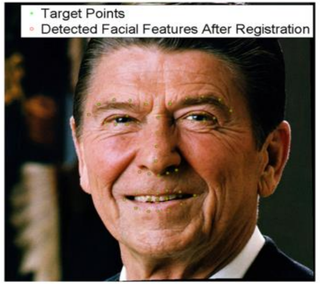
Advanced: Automated Image Fusion



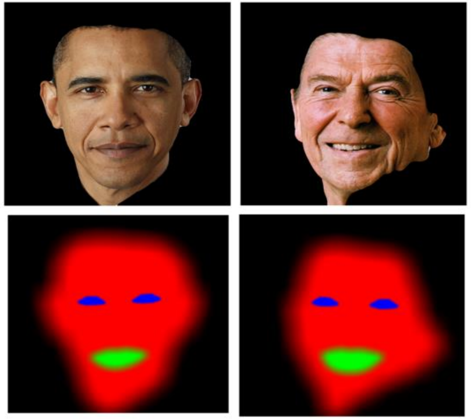
1. Face detection



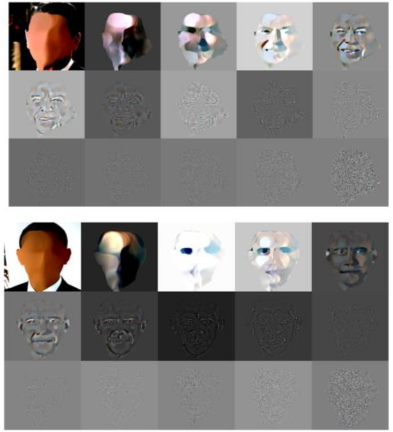
2. Landmark detection



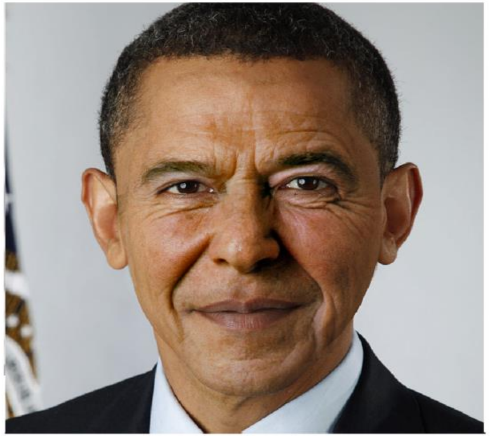
3. Registration



4. Face segmentation

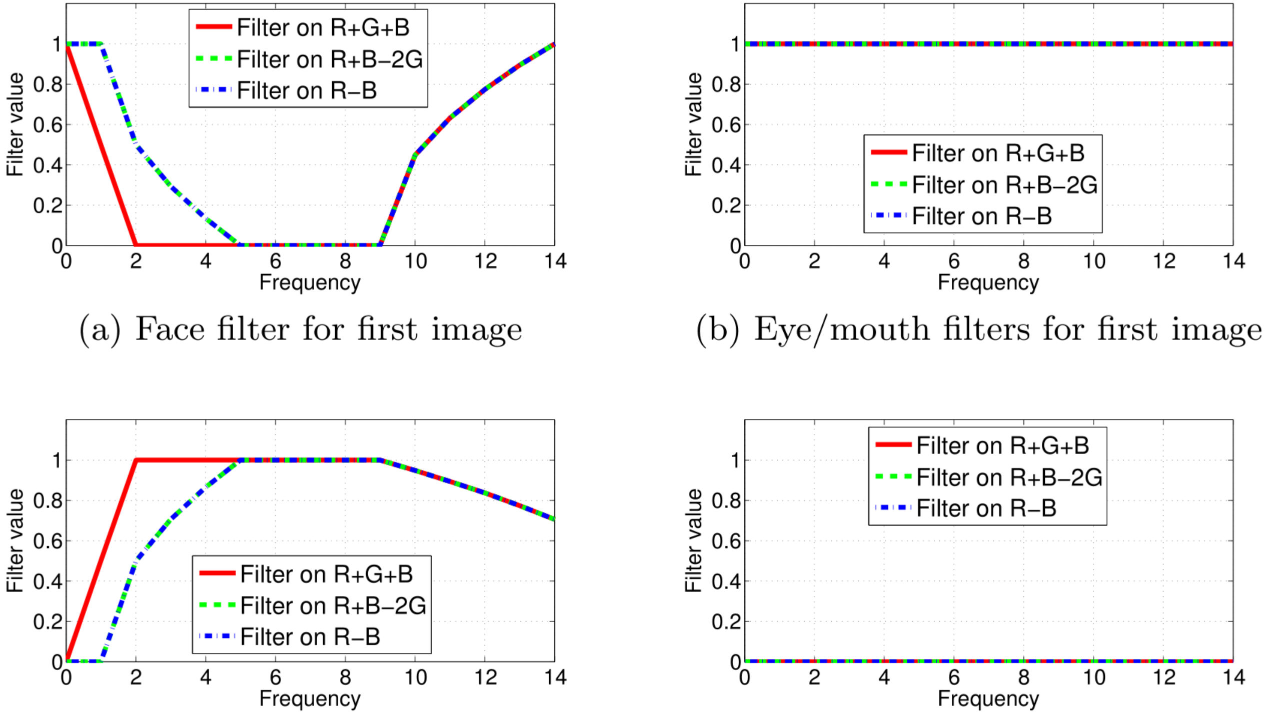


5. Spectral decomposition



6. Image fusion

Note: Spatially varying filters



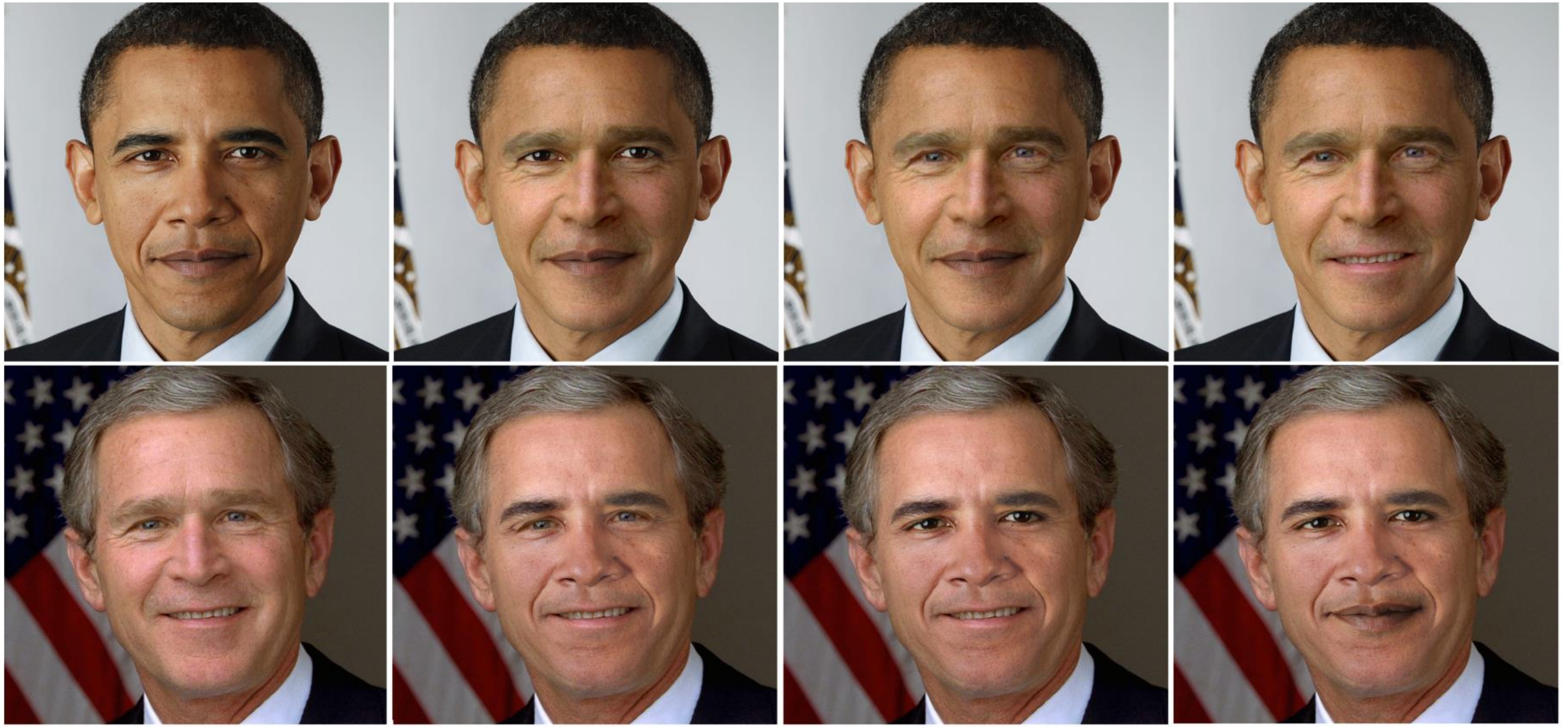
(a) Face filter for first image

(b) Eye/mouth filters for first image

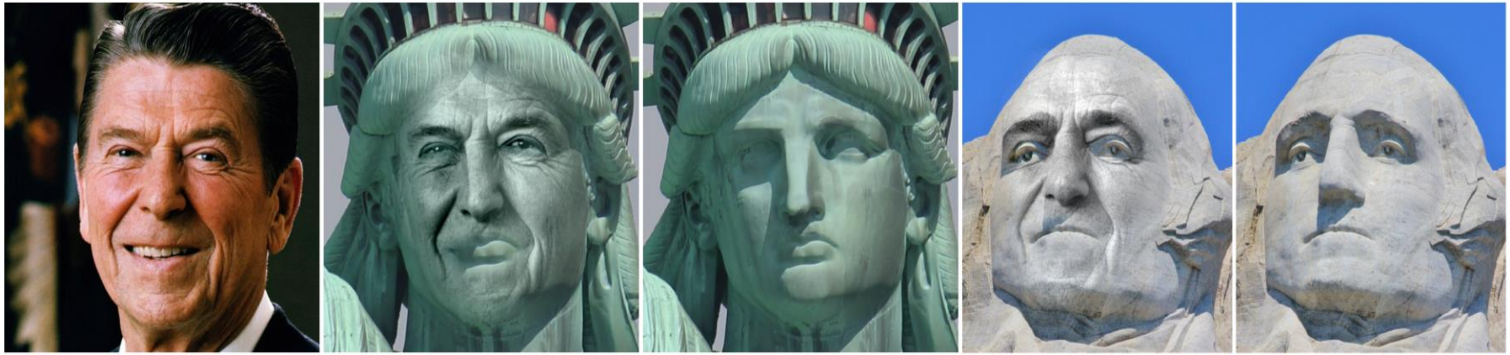
(c) Face filter for second image

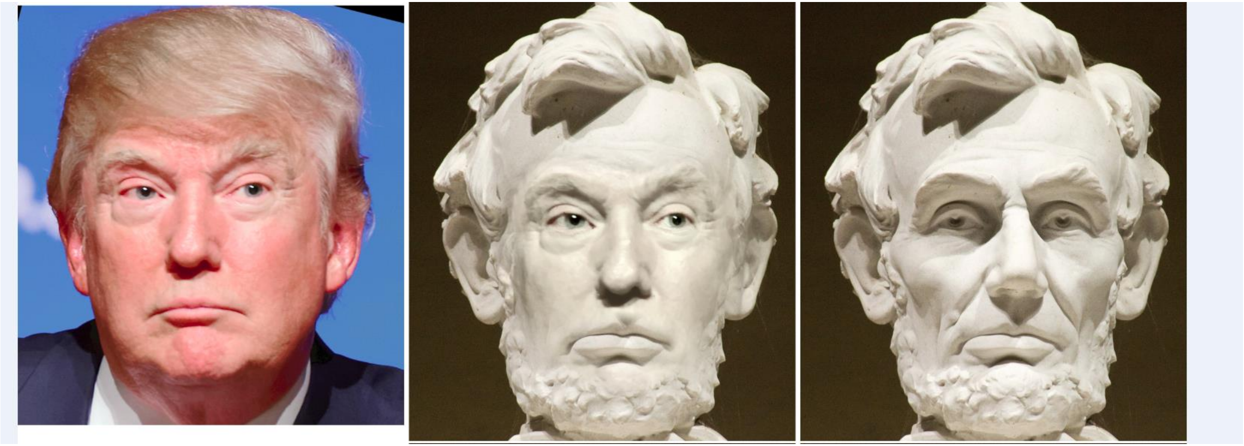
(d) Eye/mouth filters for second image

Local fusion

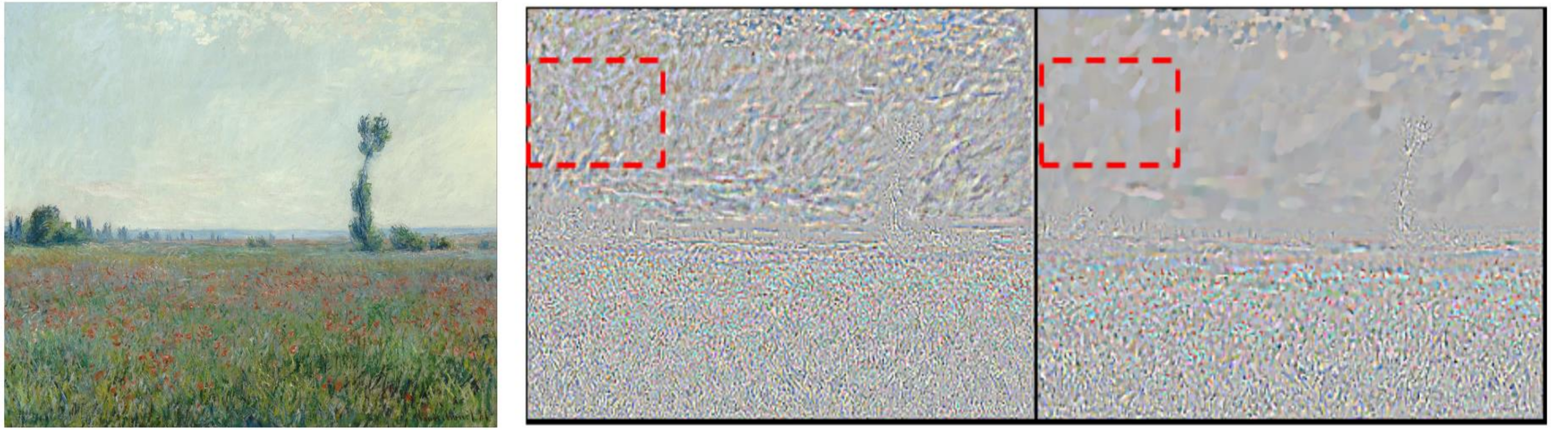


Nonstandard: Image / Sculptures





Paint it like Monet



Brush stroke patterns from Poppy field (1881), locally extracted from higher frequencies