Robust 3D Patch-Based Face Hallucination

Chengchao Qu[1,2] Christian Herrmann[1,2] Eduardo Monari[2] Tobias Schuchert[2] Jürgen Beyerer[2,1]
[1]Vision and Fusion Laboratory (IES), Karlsruhe Institute of Technology (KIT)
[2]Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (Fraunhofer IOSB)
firstname.lastname@iosb.fraunhofer.de

Abstract

Incorporating 3D information has proven to be effective in many computer vision tasks and it is no exception in the context of facial analysis. However, limited application has been witnessed in face hallucination (FH), probably due to the difficulty of fitting 3D models onto low-resolution (LR) images. This paper presents a pure 3D approach to address this problem.

By extending the LR image formation process to the 3D domain, the classic Lucas–Kanade algorithm is exploited to improve the precision of the error-prone 3D model fitting on LR images. The established correspondence between the input image and 3D training textures then facilitates reconstruction of high-resolution (HR) patches directly on the mesh, which can be employed to render realistic frontal faces for recognition.

Extensive evaluation on several publicly available datasets reveals superior qualitative and quantitative results over state-of-the-art methods in fitting, FH and recognition, which shows the advantage of the proposed 3D framework over its 2D rivals, especially for non-frontal head poses and low image quality.

1. Introduction

Analysis of face images captured by surveillance cameras is nowadays prevalent for security and forensic applications. Due to the unconstrained conditions, surge of efforts has been made to combat the accompanied pitfalls, *e.g.*, uncontrolled poses, blurring and poor resolution. Single-image face super-resolution (SR), also known as face hallucination (FH) [3], offers an effective solution to enhance the quality of those deficient facial images. Unlike generic SR which is applicable to universal image categories, FH is constrained to a specific domain, leading to results with higher magnification and finer details [40].

The prerequisite for leveraging the prior knowledge in FH is the well registered high-resolution (HR) training data and low-resolution (LR) query image, where 3D fitting is theoretically the best solution to compensate for complex global motion and local deformation of human faces [33]. Nevertheless, fitting 3D models to 2D images is a challenging task, and the LR input could make things even worse for lack of high-frequency facial details. Moreover, existing 3D approaches [29, 33] utilize 3D models solely as an advanced alignment tool and render the 3D training textures as 2D images for conventional 2D FH in the sequel, which, strictly speaking, can only be regarded as 2.5D methods.

We propose a novel framework to facilitate pure 3D FH, of which the basis is a proper reinterpretation of the observation model from the mesh surface to the LR image plane. Thus, the classic Lucas–Kanade algorithm [4] can be extended to the 3DHR–2DLR scenario for robust fitting refinement in terms of both global motion (rotation, scaling and translation) and local deformation (3D shape). Patch-based FH is then directly conducted on the mesh surface to give complete and dense HR texture (even for self-occluded regions), which can be deployed to render frontal face images to alleviate face recognition (FR) across pose. Comparison with state-of-the-art methods on synthetic and real-world LR data justifies the merit of the 3D framework over 2D and 2.5D ones in both efficiency and effectiveness.

Our main contributions are summarized as follows:

1. To the best of our knowledge, this is the first FH algorithm that integrates the LR image formation model into a robust 3D patch-based facial texture SR method.
2. The 3D extension of the Lucas–Kanade algorithm combined with a statistical texture model greatly improves fitting and FH on the ill-posed LR images.
3. Patch-based 3D FH on the mesh naturally fills the hidden facial texture caused by large head poses.

The remainder of this paper is organized as follows. A brief introduction to the previous work is given in §2. §3 first recalls conventional 3D-aided 2D FH as our motivation before we elaborate on our 3D framework. Following this, quantitative and qualitative results are shown in §4 to validate the robustness of the proposed method. Finally, we conclude our work in §5.

2. Related Work

The pioneer FH work of Baker and Kanade [3] established pixel-by-pixel statistical priors from the training faces using feature pyramids and solved the FH problem in a maximum a posteriori (MAP) manner. Holistic face space approaches are often based on the eigenfaces. Capel and Zisserman [11] proposed to infer the HR face with HR principal component analysis (PCA) prior within the MAP framework. Another alternative is the eigentransformation by Wang and Tang [41], *i.e*., projecting the LR input image onto the LR training subspace by PCA and reusing the same weights for FH. By combining global constraints with a Markov random field (MRF) defined on the homogeneous image lattice [16] for the local texture, Liu *et al*. presented a novel two-step FH procedure [25, 26], which was superseded in [45] by adopting nonnegative matrix factorization (NMF) and sparse representation in the respective stages, and by Jia and Gong in the tensor space to handle multiple modalities [21]. Later, positional restrictions were imposed by Ma *et al*. [28] to circumvent the global face in the first stage. Separate subspaces for local patches were shown to be effective in exploring common facial structure.

Prevailing 2D algorithms employ simple alignment of faces as preprocessing. Transformation by similarity using the eyes [26, 41, 45] or affinity with an extra point at tip of the nose [3] or center of the mouth [28] is a widespread technique. On the other hand, pixel-wise registration [4] is arguably a better option for LR faces, which estimates the transformation by energy minimization w.r.t. the whole image window instead of a few feature points [21, 25].

Latest success of 2D FH is partly attributed to the advanced registration techniques that remedy misalignment caused by out-of-plane rotation and complicated facial geometry. Dedeoğlu *et al*. [13] developed a LR active appearance model (AAM) for non-rigid registration with a reversed image formation process to avoid interpolating the LR input face before applying [3]. Non-frontal poses were mitigated by Tappen and Liu [38] by matching and warping training exemplars close to the LR face with PatchMatch [6] and SIFT flow [27] respectively. A convex optimization scheme for the Bayesian SR method in [38] was proposed in [20]. However, this system struggles when the training set is small or it fails to find faces that can match the input well. Yang *et al*. [42] separately dealt with different image components. On the basis of 2D AAM landmarks, main facial features were extracted and similar exemplars were aligned. In addition, statistical prior and PatchMatch [6] were used to hallucinate contours and smooth regions. In [22], Jin and Bouganis devised a unified MAP framework exploiting holistic PCA prior to model blurring and motion, which was extended with a patch-wise formulation in [23] for in-the-wild settings. Nonetheless, homography in their parametric motion can only cope with near-frontal poses.

3D-assisted FH has shown its strong potential for the first time in [29], in which Mortazavian *et al*. warped LR images onto a predefined canonical grid by fitting a 3D model [9] before performing [3]. Qu *et al*. [33] argued the negative impact when interpolating LR input and presented a "resolution-aware" approach akin to [13] in conjunction with patch-based FH [28]. The impact of various 2D and 3D alignment techniques was discussed, too. It is worth noting that both methods take advantage of 3D models to merely build accurate correspondence between the LR image and the training data.

It was only recently that the full capability of 3D FH was unleashed. Schumacher *et al*. [36] added the blurring operator into the synthesis stage during 3D Morphable Model (3DMM) fitting [8] to restore degraded facial images. In the example-based approach of Dessein *et al*. [14], the 3D mesh of the mean face was first segmented into uniformly overlapping patches. The LR pixels were then directly back-projected onto the corresponding LR vertices to apply belief propagation (BP) on the HR overlapping patches in a 3D MRF fashion [16]. Despite the positive results on simulated LR data, the inverse image formation model from LR pixels to LR vertices assuming nearest neighbor (NN) interpolation and forward BP with HR vertices within patches of fixed size rules out the flexibility to incorporate viable blurring kernels, which, in comparison, is solved by our direct forward formulation. Additionally, we also allow for a novel LR 3D fitting scheme to boost the performance.

3. Proposed Method

This main section details our 3D framework for super-resolving LR images with a rough initial fitting of the faces. To start with, the LR image formation model is revisited on the 3D mesh space, which makes it possible to generalize the 2D Lucas–Kanade image registration to the non-rigid 3D domain. Finally, we argue the benefit of 3D FH and present a patch-based approach. The proposed method is summarized in Alg. 1.

3.1. Basics: 2.5D Face Hallucination

2.5D FH differs from 2D approaches in that a 3DMM is fitted to create correspondence in lieu of 2D alignment such as geometric transformation or flow methods.

A 3DMM [8] is a statistical face model, which, like 2D AAMs [12], interprets faces in a vector space. Specifically, 3D geometry and albedo are represented with P vertices as $\mathbf{s} = [x_1, y_1, z_1, \ldots, x_P, y_P, z_P]^\top \in \mathbb{R}^{3P}$ and $\mathbf{t} = [r_1, g_1, b_1, \ldots, r_P, g_P, b_P]^\top \in \mathbb{R}^{3P}$ respectively. Applying PCA to the shape and texture data individually yields

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{S}\boldsymbol{\alpha} \quad \text{and} \quad \mathbf{t} = \bar{\mathbf{t}} + \mathbf{T}\boldsymbol{\beta}, \tag{1}$$

where $\bar{\mathbf{s}}$ and $\bar{\mathbf{t}}$ are the mean vectors. $\mathbf{S} \in \mathbb{R}^{3P \times Q_s}$ and $\mathbf{T} \in \mathbb{R}^{3P \times Q_t}$ denote the respective principal modes of variation.

Algorithm 1: Robust 3D patch-based FH

Input: Rough 3D fitting on the LR face
Output: HR 2D and 3D face with refined fitting

// §3.2: 3D image formation model
Determine the sparse LR vertex set $\mathbf{s}^-$ // (3)
Compute the 3D image formation matrix $\mathbf{H}$
// §3.3: 3D fitting enhancement
while *3D motion* $\boldsymbol{\theta}$ *not converged* **do**
    Compute 3D texture $\mathbf{t}'$ by FS–MAP // (5, 6)
    Update $\boldsymbol{\theta}$ by 3D Lucas–Kanade fitting // (9)
    Update $\mathbf{H}$
end
// §3.4: 3D patch-based FH
Divide the LR image into overlapping patches
Compute FH weights for the patches on $\mathbf{s}^-$ // (10)
Reconstruct $\mathbf{t}$ using the same weights

In this way, the coefficients $\boldsymbol{\alpha} \in \mathbb{R}^{Q_s}$ and $\boldsymbol{\beta} \in \mathbb{R}^{Q_t}$ suffice to describe any valid face within the PCA subspaces.

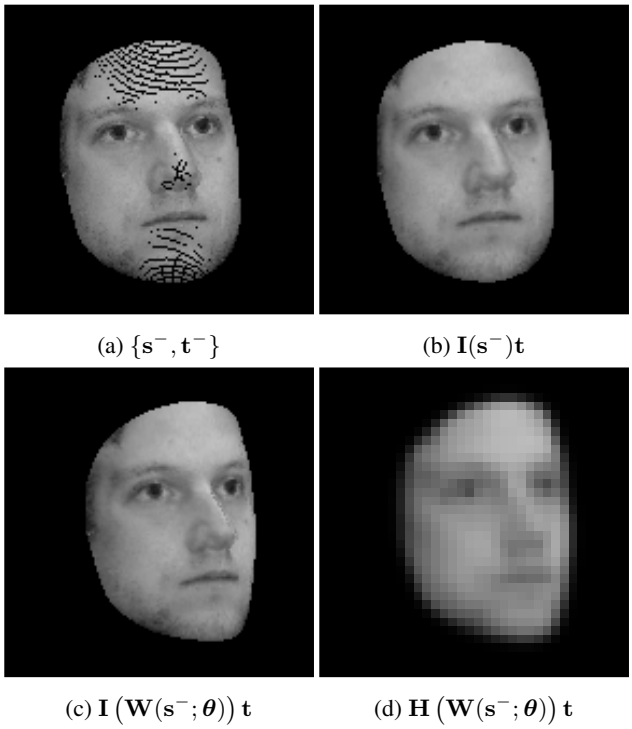
When fitting a 3DMM to an input image, all shape, albedo and photometric parameters, *e.g.*, camera calibration, illumination and shading, are simultaneously estimated [9]. Considering the enormous parameter space and non-convex optimization, this analysis-by-synthesis framework is extremely time-consuming. Alternatively, by leveraging a sparse set of fiducial facial landmarks, the complexity can be dramatically reduced by leaving out the entire appearance parameters [7]. The shape coefficients $\boldsymbol{\alpha}$ are straightforwardly computed using least squares fitting w.r.t. merely dozens of feature points rather than thousands of vertices, which again helps boost the runtime.

Subsequently, 2D training images are rendered using HR 3D textures either on a flattened grid [29] or directly on the LR input frame [33] to run 2D FH.

Discussion Although existing 2.5D algorithms [29, 33] have so far demonstrated impressive competence in both FH and FR on synthetic LR data, we observe several areas of possible improvements: (1) 3D shape reconstruction is performed either with a standard 3DMM fitting algorithm [9] in [29] or solely based on a few automatically detected facial landmarks in [33], of which the quality is susceptible to image resolution [18, 19]. (2) While rendering novel views of the super-resolved face, the real texture in the self-occluded region remains intractable [33]. These unfavorable aspects can be addressed elegantly with our proposed 3D approach.

3.2. Image Formation Model

The 2D LR observation process [30, 44] models the LR image $\mathbf{z}$ of $N_1 \times N_2$ pixels as a downsampled version of the



(a) $\{\mathbf{s}^-, \mathbf{t}^-\}$ (b) $\mathbf{I}(\mathbf{s}^-)\mathbf{t}$

(c) $\mathbf{I}\left(\mathbf{W}(\mathbf{s}^-;\boldsymbol{\theta})\right)\mathbf{t}$ (d) $\mathbf{H}\left(\mathbf{W}(\mathbf{s}^-;\boldsymbol{\theta})\right)\mathbf{t}$

Figure 1: From 3D shape to LR image: (a) the subsampled vertices $\mathbf{s}^-$, (b) interpolated result of (a), (c) image output with extra rotation applied using the *same* vertex subsampling $\mathbf{s}^-$, (d) final LR output $\mathbf{z}$ of (c).

HR image $\mathbf{x}$ of $mN_1 \times mN_2$ pixels with

$$\mathbf{z} = \left(\mathbf{B}_\mathbf{k} \circ \mathbf{W}(\mathbf{x};\boldsymbol{\theta})\right)\downarrow_m + \mathbf{n}, \tag{2}$$

where $\mathbf{W}$ first warps the original signal via the parametrized motion $\boldsymbol{\theta}$. Then $\mathbf{B}$ imposes the blurring effect with kernel $\mathbf{k}$ and $\downarrow$ denotes decimation with magnification factor m. The imaging noise, often assumed to be white, is reflected in the additive term $\mathbf{n}$. Learning-based SR brings in extra knowledge from internal or external sources to combat the ill-posed problem in recovering $\mathbf{x}$ in Eq. (2). In conventional 2D or 2.5D methods, the motion $\boldsymbol{\theta}$ is compensated for by 2D or 3D alignment techniques and the training data can be warped or rendered to super-resolve the input $\mathbf{z}$.

Elevating the problem setup to the 3D level requires building a direct connection between the 3D shape $\mathbf{s}$ and the LR image $\mathbf{z}$, of which the key challenge is to take into consideration the blurring kernel $\mathbf{k}$. Dessein *et al.* [14] just ignored it and back-projected the LR pixel values to the corresponding LR vertices. This oversimplified NN-like approach turns out to violate the image formation model [15] and struggles with real image data [34] in our evaluation.

Vertex Subsampling To integrate the blurring kernel $\mathbf{k}$, the initial LR 3D shape is first upscaled by factor m onto the

HR image coordinates in accordance with Eq. (2). Since a 3DMM usually has tens of thousands of vertices and involving all of them causes unnecessary computational overhead with hardly any qualitative improvement for FH, we choose to take into account just a small fraction of them. Following this strategy, after the visible vertices are determined, only one of those that fall into each HR pixel grid is selected by

$$\mathbf{s}_i^- = \arg\min_{\mathbf{v}\in\mathcal{V}_i} \|\mathbf{v} - \mathbf{p}_i\|_2, \tag{3}$$

where $\mathcal{V}_i$ denotes all vertices inside the unit square centered at pixel $\mathbf{p}_i$, revealing a reduced set $\mathbf{s}^-$ of roughly the same cardinality as the number of pixels in the HR face. This nearly one-to-one mapping between vertices and pixels ensures fidelity when downscaled to LR.

Fig. 1a depicts the subsampled vertices $\mathbf{s}^-$ of an example 3D face in the associated pixels. Note that holes emerge at regions like forehead or jaw where no vertices are present owing to the non-uniform distribution of the mesh. However, we ignore these less structured places on $\mathbf{s}^-$ as they can be completely eliminated by the subsequent interpolation step (see Fig. 1b). Even after rotating the face by 15°, the *original* subsampling $\mathbf{s}^-$ can still generate natural image output with enough details in Fig. 1c. This is critical for our 3D extension of the Lucas–Kanade algorithm, which alters the face geometry and motion in each iteration.

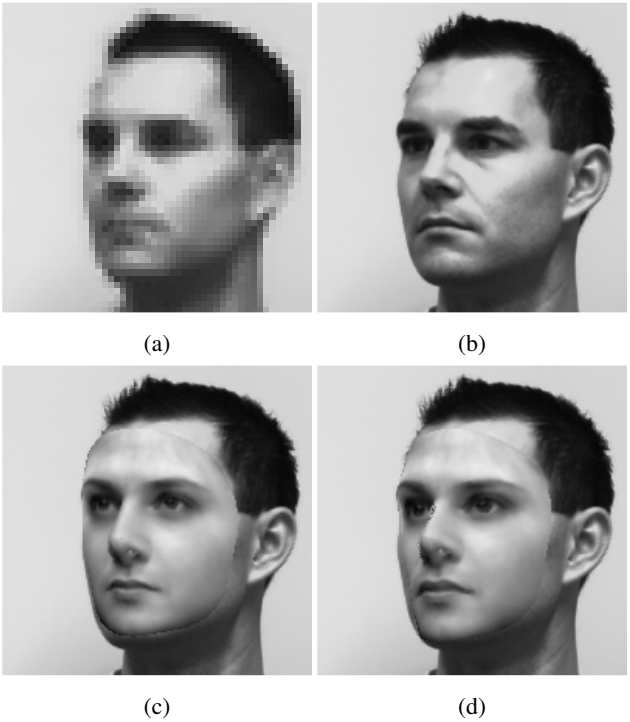
LR Projection for 3D Data After sampling the scattered point cloud $\mathbf{s}^-$, it is interpolated on the HR lattice to obtain a 2D image. Specifically, Delaunay triangulation is carried out on the projected 2D coordinates of $\mathbf{s}^-$ and indices of the triangles in which each pixel is located can be found. At the same time, the barycentric coordinates for the HR pixels w.r.t. the triangulated vertices can also be computed efficiently. Importantly, the whole procedure of linear interpolation is representable as a sparse matrix $\mathbf{I}(\mathbf{s}^-) \in \mathbb{R}^{m^2 N_1 N_2 \times P}$, in which each row has at most three entries indicating the barycentric coordinates of the relevant vertices, multiplied by the grayscale texture $\mathbf{t} \in \mathbb{R}^P$.

Subsequently, the convolution and decimation operators in Eq. (2) are converted into matrices $\mathbf{T_k}$ and $\mathbf{S}_m$ respectively, where $\mathbf{T_k}$ denotes a Toeplitz matrix [24] for filter $\mathbf{k}$ and $\mathbf{S}_m \in \mathbb{Z}^{N_1 N_2 \times m^2 N_1 N_2}$ is a sparse shrinkage matrix, *i.e.*, for each LR pixel represented by row i, only column j corresponding to the selected HR pixel is set to one.

Therefore, the complete LR image observation process from the 3D surface is formulated as a matrix multiplication

$$\mathbf{z} = \mathbf{H}\left(\mathbf{W}(\mathbf{s}^-; \boldsymbol{\theta})\right) \cdot \mathbf{t} + \mathbf{n}, \tag{4}$$

where $\mathbf{H} = \mathbf{S}_m \mathbf{T_k} \mathbf{I}$ is a composition matrix of dimension $N_1 N_2 \times P$ integrating all related operations. Both $\mathbf{H}$ and $\mathbf{I}$ are dependent on the (warped) 3D shape. An example LR image generated by this model is illustrated in Fig. 1d.



(a) (b) (c) (d)

Figure 2: Improving 3D fitting: (a) LR input image, (b) HR ground truth of (a), (c) initial fitting on (a) with 3D FS–MAP texture, (d) final result after 10 iterations.

3.3. 3D Fitting Enhancement

LR faces pose a huge challenge on 3D fitting because of a significant amount of information loss and diverse nuisance factors such as blur and noise from various sources. Whereas the accuracy of fitting propels the quality of FH, it can also introduce adverse impact when the precision degrades. Motivated by the classic Lucas–Kanade algorithm [4], we iteratively optimize both global motion and 3DMM deformation tailored for LR images.

Given an approximate fitting, a holistic estimate of the HR texture $\mathbf{t}'$ for the Lucas–Kanade algorithm [4] can be computed using the face space–maximum a posteriori (FS–MAP) approach [11] with the 3D image observation model introduced in §3.2. Assuming we have a FH training dataset of 3D facial textures, applying PCA gives the mean texture $\boldsymbol{\mu}$, a matrix $\mathbf{D}$ composed of eigenvectors and a diagonal matrix $\mathbf{P}$ of eigenvalues resembling Eq. (1). The texture coefficient $\boldsymbol{\beta}'$ on the PCA space and the resulting holistic HR FS–MAP face

$$\mathbf{t}' = \boldsymbol{\mu} + \mathbf{D}\boldsymbol{\beta}' \tag{5}$$

are obtainable in closed form via

$$\hat{\boldsymbol{\beta}}' = \arg\min_{\boldsymbol{\beta}'} \|\mathbf{H}(\mathbf{s}^-)\left(\boldsymbol{\mu} + \mathbf{D}\boldsymbol{\beta}'\right) - \mathbf{z}\|_2^2 + \gamma\|\boldsymbol{\beta}'\|_{\mathbf{P}}^2. \tag{6}$$

The goal of the Lucas–Kanade algorithm [4] is to align an image to a template by minimizing the sum of squared error (SSE) between the two patches. Our 3D FH setup using the 3D FS–MAP texture $\mathbf{t}'$ yields

$$\frac{1}{|\boldsymbol{\Omega}|}\|\mathbf{H}\left(\mathbf{W}(\mathbf{s}^{-};\boldsymbol{\theta})\right)\mathbf{t}'-\mathbf{z}\|_2^2, \tag{7}$$

normalized by the pixel count of the facial region $\boldsymbol{\Omega}$ in the LR template $\mathbf{z}$. $\mathbf{W}(\mathbf{s}^{-};\boldsymbol{\theta})$ transforms the 3D face shape subject to the parametrized warping vector

$$\boldsymbol{\theta}=\left[s,\boldsymbol{\omega}^{\top},t_u,t_v,\boldsymbol{\alpha}^{\top}\right]^{\top}\in\mathbb{R}^{6+Q_s}, \tag{8}$$

which consists of scaling s, 3D rotation vector $\boldsymbol{\omega}$ [37] and translation $[t_u,t_v]^{\top}$ for rigid motion, as well as the 3DMM shape coefficients $\boldsymbol{\alpha}$ in Eq. (1) for local deformation.

Based on a current estimate of $\boldsymbol{\theta}$, the Lucas–Kanade algorithm [4] solves for the incremental update $\boldsymbol{\Delta\theta}$, such that

$$\frac{1}{|\boldsymbol{\Omega}|}\|\mathbf{H}\left(\mathbf{W}(\mathbf{s}^{-};\boldsymbol{\theta}+\boldsymbol{\Delta\theta})\right)\mathbf{t}'-\mathbf{z}\|_2^2 \tag{9}$$

is minimized with respect to $\boldsymbol{\Delta\theta}$. The nonlinear optimization task is addressed by first-order Taylor expansion and solved in an iterative fashion. In each iteration:

1. The 3D face is altered by $\boldsymbol{\theta}$ via $\mathbf{W}(\mathbf{s}^{-};\boldsymbol{\theta})$.
2. The LR projection matrix $\mathbf{H}\left(\mathbf{W}(\mathbf{s}^{-};\boldsymbol{\theta})\right)$ is obtained.
3. The error image $\mathbf{z}-\mathbf{H}\mathbf{t}'$ is computed.
4. The image gradient $\boldsymbol{\nabla}_{\mathbf{H}\mathbf{t}'}$ of $\mathbf{H}\mathbf{t}'$ is computed.
5. The Jacobian of warping $\frac{\partial\mathbf{W}}{\partial\boldsymbol{\theta}}$ at $(\mathbf{s}^{-};\boldsymbol{\theta})$ is evaluated.[1]
6. The steepest descent image $\boldsymbol{\nabla}_{\mathbf{H}\mathbf{t}'}\frac{\partial\mathbf{W}}{\partial\boldsymbol{\theta}}$ is obtained.
7. Nonlinear optimization is performed for $\boldsymbol{\Delta\theta}$.
8. The warping parameter is updated $\boldsymbol{\theta}\leftarrow\boldsymbol{\theta}+\boldsymbol{\Delta\theta}$.

Note that in practice, the Levenberg–Marquardt algorithm is preferred by virtue of its robustness against the more complex problem setup to Gauss–Newton for homography [4].

An instance of the complete fitting process is visualized in Fig. 2. At first sight, the 3D FS–MAP texture hallucinated on the initial fit in Fig. 2c seems plausible. But comparison with Fig. 2d demonstrates how much 3D fitting can be improved in all respects. Not only the head pose and the facial contour, but also the local deformation (*e.g.*, the shape of the mouth) better conforms to the HR ground truth in Fig. 2b.

[1] $\frac{\partial\mathbf{W}}{\partial\boldsymbol{\theta}}=\begin{bmatrix}\mathbf{R}_u\,[u,v,z]^{\top} & 0 & s\cos(\omega_2)z & -s\cos(\omega_3)y\\ \mathbf{R}_v\,[u,v,z]^{\top} & -s\cos(\omega_1)z & 0 & s\cos(\omega_3)x\\ 1\;\;0 & s\mathbf{R}_u\mathbf{H}\mathbf{S} & & \\ 0\;\;1 & s\mathbf{R}_v\mathbf{H}\mathbf{S} & & \end{bmatrix}$ w.r.t. $\boldsymbol{\theta}$ in Eq. (8) using the chain rule, where $\mathbf{R}_u$ and $\mathbf{R}_v$ are the respective rows of the matrix representation $\mathbf{R}$ for the rotation vector $\boldsymbol{\omega}$ using the Rodrigues' formula [37] and the last column denotes the shape dictionary in Eq. (1) projected onto the LR coordinates.

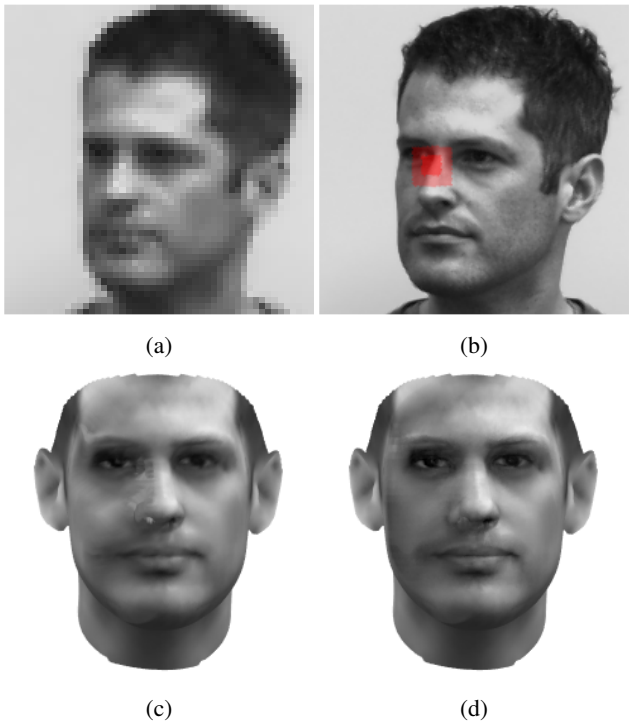


Figure 3: 3D patch-based FH: (a) LR input image, (b) a 3D patch superimposed on the HR image, (c) texture extracted from 2.5D FH, (d) directly super-resolved texture by 3D FH. Best viewed by zooming in the electronic version.

Discussion According to [4], the proposed Lucas–Kanade extension falls under the Forwards Additive variant (*c.f.* [5]). In general, the quadratic expression in Eq. (9) is non-convex and hence not guaranteed to converge globally. However, unlike 3DMM fitting [9] initialized with the mean appearance, we start from relatively good shape and texture estimates so that as few as 10 iterations are found sufficient for convergence in practice. Furthermore, not taking into account the albedo parameters gives rise to considerable benefit in runtime with only a fraction of a second for each iteration. Last but not least, the well-defined LR observation process ensures excellent versatility w.r.t. different image degradation models in real world compared to a fixed number of trained LR 3DMMs [19].

3.4. 3D Patch-Based Face Hallucination

A key difference of our 3D FH to 2.5D systems [29, 33] lies in that we directly obtain the HR 3D texture rather than separately performing 2D FH and texture extraction from the super-resolved image.

Similar to [33], we follow the idea of the 2D patch-based FH [28] to divide the LR image into overlapping patches and enforce local subspaces. Afterwards, the FH procedure is conducted on the face mesh. First, the subset of $\mathbf{s}^{-}$ be-

longing to each patch is determined straightforwardly given the non-empty entries in the sparse matrix $\mathbf{H}$ in Eq. (4). Fig. 3b illustrates the corresponding vertices of a patch in red. On account of convolution with the blurring kernel, the actual vertices in light red for LR patch reconstruction have a larger vicinity. For each of the local patch j, the optimal weights $\hat{\mathbf{w}}_j$ are obtained by

$$\hat{\mathbf{w}}_j = \arg\min_{\mathbf{w}_j} \left\| \sum_{l=1}^{L} w_j^l \mathbf{H} \left(\mathbf{W}(\mathbf{s}_j^-; \boldsymbol{\theta}) \right) \mathbf{t}_j^l - \mathbf{z}_j \right\|_2^2 \quad (10)$$

with the help of our 3D observation model, where the superscript l denotes the index of the L 3D textures in the training data. The complete set $\mathbf{s}$ is subsequently recovered using the same weights as for $\mathbf{s}^-$, where the values of the overlapping vertices are averaged.

A favorable byproduct while directly conducting 3D FH is the intrinsic filling of the self-occluded texture. *E.g.*, the hidden side of the nose faithful to the illumination condition is learned from the training data (*c.f.* Figs. 3c and 3d). For large poses as in the last one of Fig. 4i, nearly the entire face half under occlusion is still hallucinated realistically.

Discussion It is worth noting that the newly published 3D MRF approach [14] used irregular 3D patches of fixed size segmented offline, which is cumbersome to further incorporate convolution with different blurring kernels.

4. Experiments

Setup We validate the proposed 3D FH framework on several publicly available datasets, *i.e.*, the Multi-PIE dataset [17] with two sessions of the same 120 subjects under poses from 0° to 45° (05_1, 05_0, 04_1 and 19_0), the real face super-resolution (FSR) dataset [34] of 31 subjects with yaw and pitch head rotation and four images [23] from the PubFig83 dataset [32] for qualitative validation under in-the-wild circumstances. The Multi-PIE images are downsized by 50% and cropped according to the face detector [39] as HR data. LR images are blurred with a Gaussian kernel with $\sigma = 2.4$ and subsampled. No preprocessing is needed for FSR as both ground truth LR and HR faces are simultaneously acquired by a dual-camera imaging system [34]. For the PubFig83 images [32], we follow [23] to set $\sigma = 1.6$. The resize factor is $m = 4$ for all datasets.

A total of 214 Multi-PIE HR shots are collected as the training data for both 2D and 3D methods. A fair out-of-sample evaluation on FH and FR in the Multi-PIE experiments is ensured by temporarily excluding the tested subject from the training data. 3D face fitting is realized with the shape component of the Basel Face Model (BFM) [31].

Following existing studies, FH and quantitative evaluation in PSNR are conducted on the luminance channel due to insensibility of human vision to chrominance channels, which are bicubicly upsampled for color images [23, 38, 42, 45]. SSIM is not reported owing to the irregularly shaped facial masks [23]. Note that frontal, yaw as well as yaw and pitch head poses are abbreviated as F, Y and Y+P respectively in Tabs. 1, 2 and 3.

Table 1: Normalized RMSE ($\times 10^{-2}$) for inner facial landmarks without (✗) / with (✓) fitting enhancement.

	Multi-PIE [17]				FSR [34]		
	0°	15°	30°	45°	F	Y	Y+P
✗	**4.23**	**4.53**	6.69	8.49	3.76	5.12	4.72
✓	4.28	4.75	**6.19**	**7.72**	**3.57**	**4.84**	**4.45**

Fitting As opposed to some 2D and 3D work [14, 28, 33] where alignment is done manually or on HR images, we adopt a more pragmatic setup to carry out face alignment and 3D fitting on LR data. To quantitatively evaluate the benefit of 3D fitting enhancement, we report the normalized RMSE of 2D inner facial landmarks [43] in Tab. 1.

Generally, compared to the initial results, refinement does successfully improve landmarking accuracy. Since the improved landmarks are projected from the 3D shape, discrepancy could have affected our numbers. That means, except for the near-frontal poses on Multi-PIE, this extra stage demonstrates excellent capability to correct the error-prone fitting initialized on LR faces, which will prove crucial in the upcoming FH experiments.

Face Hallucination The performance of our 3D FH is first evaluated against several state-of-the-art algorithms on Multi-PIE and FSR. For 2D methods [20, 28, 38, 42], the authors' code is employed.[2] We further implement the texture normalized 3D MRF using the same parameters as in [14]. Identical 3D fitting engine [35] is used for all 2.5D and 3D systems ([14, 33] and ours), allowing for a fair and convincing comparison. Experimental results are reported in Tab. 2 and Figs. 4 and 5.

Patch-based FH using positional subspaces is sensitive to deficient fitting caused by the average LR face of 6 to 10 pixels of inter-ocular distance (pxIOD) on Multi-PIE. Improved fitting remarkably boosts the final results for all 2D and 3D methods, where our 3D framework tops all situations except for the frontal pose on Multi-PIE. Qualitative comparison shows that with increasing yaw angle, the visual advantage becomes prominent where 2D registration suffers from large out-of-plane rotation. On the noisy FSR data, the frontal case of us also outperforms [28] (*c.f.* the

[2] Following [23], [20] can be regarded as a performance indicator for [38], as it is shown to be as good as [38].

(a) HR (b) LR (c) Bicubic (d) [28] (e) [42] (f) [20] (g) [14] (h) 3D (i) 3D-F

Figure 4: Qualitative FH results on the Multi-PIE dataset [17]. Best viewed by zooming in the electronic version.

(a) HR (b) LR (c) Bicubic (d) [28] (e) [42] (f) [20] (g) [14] (h) 3D (i) 3D-F

Figure 5: Qualitative FH results on the FSR dataset [34]. Best viewed by zooming in the electronic version.

eyes in Figs. 5d and 5h). By contrast, despite the sophisticated alignment mechanism in [20, 38, 42], the output images either are impaired by artifacts and outliers or look blurry. Since faces in FSR with approximately 12 pxIOD are less challenging for most state-of-the-art landmark detectors, fitting refinement is less advantageous than for the smaller Multi-PIE faces.

As discussed in §3, the exemplar-based 3D MRF [14] generates highly detailed faces, however, with neither realistic appearance nor competitive scores using the simplified image formation model. Apart from the ability of natural frontalization, 2.5D [33] and 3D FH yield initially almost identical PSNR values by sharing the core fitting and FH algorithms. The final edge is mainly attributed to the improved fitting accuracy, which reiterates the significance for FH to exploit spatial cues.

Finally, we go beyond controlled scenarios to testify our robustness on the in-the-wild PubFig83 dataset [32] in Fig. 6, where all results except ours are imported from [23]. Under the challenges such as unconstrained poses, expressions and illuminations, especially the faces of as small as 6 to 7 pxIOD, we achieve the most appealing visual quality, surpassing the state-of-the-art [23] in both sharpness and details of facial components with far less training data.

Table 2: Quantitative FH results in PSNR without / with fitting enhancement.

		Bicubic	[42]	[20]	[14]	2D [28]	2.5D [33]	3D
Multi-PIE [17]	0°	25.60	25.98	26.49	22.10	26.91 / **27.13**	25.92 / 26.25	25.94 / 27.05
	15°	25.49	26.10	26.44	21.52	26.78 / 26.82	26.52 / 26.71	26.56 / **27.15**
	30°	25.27	25.55	26.07	21.38	25.62 / 25.79	25.66 / 26.00	25.69 / **26.35**
	45°	25.27	25.95	26.46	21.34	25.58 / 25.61	26.28 / 26.48	26.31 / **26.62**
FSR [34]	F	26.32	26.02	26.78	18.06	27.11 / 27.04	27.20 / 27.26	27.26 / **27.33**
	Y	25.84	25.71	26.47	16.73	26.16 / 26.35	26.30 / 26.77	26.31 / **26.78**
	Y+P	27.18	26.90	27.82	17.02	27.33 / 27.66	27.50 / 28.03	27.52 / **28.04**

Table 3: FR results in identification rate (%).

		HR	Bicubic	[42]	[20]	[14]	2D [28]	3D	3D-F
Multi-PIE [17]	0°	98.3	72.5	87.5	84.2	17.5	**88.3**	85.0	—
	15°	95.0	58.3	75.0	69.2	13.3	70.8	68.3	**80.8**
	30°	62.5	19.2	35.8	27.5	7.5	30.0	30.0	**49.2**
	45°	38.3	12.5	22.5	16.7	7.5	20.0	21.7	**35.0**
FSR [34]	Y	96.8	77.4	80.6	80.6	38.7	83.9	83.9	**100.0**
	Y+P	77.4	67.7	77.4	74.2	29.0	77.4	67.7	**93.5**

(a) HR (b) Bicubic (c) [42] (d) [20] (e) [23] (f) 3D

Figure 6: Qualitative FH results on the PubFig83 dataset [32]. Best viewed by zooming in the electronic version.

Application to Face Recognition We also test FR with the previous FH results as probe images to verify the practical application of FH, while frontal images of the so far unused second Multi-PIE subset serve as gallery. Since only one session is present in the FSR data, the frontal faces are selected instead. In overall, the identification rate in Tab. 3 using the classic LBP histogram [1] is in accordance with the promising FH outcome. By frontalizing the 3D FH faces (referred to as 3D-F in Tab. 3, see Figs. 4i and 5i), significant improvements are observed and nearly perfect matching scores are achieved on the FSR data, even outperforming HR images by a large margin for faces with moderate yaw and pitch rotation.

Discussion The proposed algorithm runs fully automatically. The sole parameter to be manually set is the blurring kernel. In case of unknown kernels, some class-specific face deblurring methods [2] can be deployed. Currently, we utilize a 3DMM [31] with exclusively neutral facial expression. Hence, hallucination of novel expressions like in the last example of Fig. 6f could lead to artifacts in the mouth. This can be bypassed by adding expression variations [10].

In terms of runtime, the whole workflow in unoptimized MATLAB code including initial 3D reconstruction [35] takes around 10 to 30 seconds depending on the face size on a PC with 3.4 GHz CPU, markedly below that of the competing algorithms [14, 20, 38, 42] with 2 to 5 minutes and [23] with over 15 minutes.

5. Conclusions

This work revisits the fundamental aspects of LR 3D face fitting and presents a novel 3D framework for the problem of LR facial texture hallucination. An effective formulation of the LR image formation process on the 3D mesh is proposed, which opens up the possibility to improve fitting accuracy in the LR scenario and to directly super-resolve the 3D texture with automatic occlusion handling. Superior FH and FR results over state-of-the-art approaches on simulated, real and in-the-wild data justify the theoretical and practical advantage in real-world LR applications.

Acknowledgment

This study was partially supported by the MobilePass project, co-funded by the EU under FP7 grant 608016. We thank the anonymous reviewers for the insightful feedback.

References

[1] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with Local Binary Patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, 2006 (cit. on p. 8).

[2] S. Anwar, C. P. Huynh, and F. Porikli, "Class-specific image deblurring," in *ICCV*, 2015, pp. 495–503 (cit. on p. 8).

[3] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1167–1183, 2002 (cit. on pp. 1, 2).

[4] S. Baker and I. Matthews, "Lucas–Kanade 20 years on: A unifying framework," *Int. J. Comput. Vis.*, vol. 56, no. 3, pp. 221–255, 2004 (cit. on pp. 1, 2, 4, 5).

[5] S. Baker, R. Patil, K. M. Cheung, and I. Matthews, "Lucas–Kanade 20 years on: Part 5," Robotics Institute, Carnegie Mellon University, Tech. Rep. CMU-RI-TR-04-64, 2004 (cit. on p. 5).

[6] C. Barnes, E. Shechtman, A. Finkelstein, and D. B. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, 24:1–24:11, 2009 (cit. on p. 2).

[7] V. Blanz, A. Mehl, T. Vetter, and H.-P. Seidel, "A statistical method for robust 3D surface reconstruction from sparse data," in *3DPVT*, 2004, pp. 293–300 (cit. on p. 3).

[8] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *SIGGRAPH*, 1999, pp. 187–194 (cit. on p. 2).

[9] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1063–1074, 2003 (cit. on pp. 2, 3, 5).

[10] C. Cao, Y. Weng, K. Zhou, Y. Tong, and K. Zhou, "FaceWarehouse: A 3D facial expression database for visual computing," *IEEE Trans. Vis. Comput. Graphics*, vol. 20, no. 3, pp. 413–425, 2014 (cit. on p. 8).

[11] D. Capel and A. Zisserman, "Super-resolution from multiple views using learnt image models," in *CVPR*, vol. 2, 2001, pp. 627–634 (cit. on pp. 2, 4).

[12] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *ECCV*, vol. 1407, 1998, pp. 484–498 (cit. on p. 2).

[13] G. Dedeoğlu, S. Baker, and T. Kanade, "Resolution-aware fitting of active appearance models to low-resolution images," in *ECCV*, vol. 2, 2006, pp. 83–97 (cit. on p. 2).

[14] A. Dessein, W. A. P. Smith, R. C. Wilson, and E. R. Hancock, "Example-based modeling of facial texture from deficient data," in *ICCV*, 2015, pp. 3898–3906 (cit. on pp. 2, 3, 6–8).

[15] N. Efrat, D. Glasner, A. Apartsin, B. Nadler, and A. Levin, "Accurate blur models vs. image priors in single image super-resolution," in *ICCV*, 2013, pp. 2832–2839 (cit. on p. 3).

[16] W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Comput. Graph. Appl. Mag.*, vol. 22, no. 2, pp. 56–65, 2002 (cit. on p. 2).

[17] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010 (cit. on pp. 6–8).

[18] C. Herrmann, C. Qu, D. Willersinn, and J. Beyerer, "Impact of resolution and image quality on video face analysis," in *AVSS*, 2015, pp. 1–6 (cit. on p. 3).

[19] G. Hu, C. H. Chan, J. Kittler, and W. Christmas, "Resolution-aware 3D morphable model," in *BMVC*, 2012, pp. 109.1–109.10 (cit. on pp. 3, 5).

[20] P. Innerhofer and T. Pock, "A convex approach for image hallucination," in *ÖAGM–AAPR*, 2013 (cit. on pp. 2, 6–8).

[21] K. Jia and S. Gong, "Generalized face super-resolution," *IEEE Trans. Image Process.*, vol. 17, no. 6, pp. 873–886, 2008 (cit. on p. 2).

[22] Y. Jin and C. Bouganis, "Face hallucination revisited: A joint framework," in *ICIP*, 2013, pp. 981–985 (cit. on p. 2).

[23] Y. Jin and C.-S. Bouganis, "Robust multi-image based blind face hallucination," in *CVPR*, 2015, pp. 5252–5260 (cit. on pp. 2, 6–8).

[24] A. Levin, Y. Weiss, F. Durand, and W. T. Freeman, "Efficient marginal likelihood optimization in blind deconvolution," in *CVPR*, 2011, pp. 2657–2664 (cit. on p. 4).

[25] C. Liu, H.-Y. Shum, and W. T. Freeman, "Face hallucination: Theory and practice," *Int. J. Comput. Vis.*, vol. 75, no. 1, pp. 115–134, 2007 (cit. on p. 2).

[26] C. Liu, H.-Y. Shum, and C.-S. Zhang, "A two-step approach to hallucinating faces: Global parametric model and local nonparametric model," in *CVPR*, vol. 1, 2001, pp. 192–198 (cit. on p. 2).

[27] C. Liu, J. Yuen, and A. Torralba, "SIFT flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, 2011 (cit. on p. 2).

[28] X. Ma, J. Zhang, and C. Qi, "Hallucinating face by position-patch," *Pattern Recogn.*, vol. 43, no. 6, pp. 2224–2236, 2010 (cit. on pp. 2, 5–8).

[29] P. Mortazavian, J. Kittler, and W. Christmas, "3D-assisted facial texture super-resolution," in *BMVC*, 2009, pp. 119.1–119.11 (cit. on pp. 1–3, 5).

[30] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: A technical overview," *IEEE Signal Process. Mag.*, vol. 20, no. 3, pp. 21–36, 2003 (cit. on p. 3).

[31] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3D face model for pose and illumination invariant face recognition," in *AVSS*, 2009, pp. 296–301 (cit. on pp. 6, 8).

[32] N. Pinto, Z. Stone, T. Zickler, and D. Cox, "Scaling up biologically-inspired computer vision: A case study in unconstrained face recognition on facebook," in *CVPRW*, 2011, pp. 35–42 (cit. on pp. 6–8).

[33] C. Qu, C. Herrmann, E. Monari, T. Schuchert, and J. Beyerer, "3D vs. 2D: On the importance of registration for hallucinating faces under unconstrained poses," in *CRV*, 2015, pp. 139–146 (cit. on pp. 1–3, 5–8).

[34] C. Qu, D. Luo, E. Monari, T. Schuchert, and J. Beyerer, "Capturing ground truth super-resolution data," in *ICIP*, 2016, pp. 2812–2816 (cit. on pp. 3, 6–8).

[35] C. Qu, E. Monari, T. Schuchert, and J. Beyerer, "Adaptive contour fitting for pose-invariant 3D face shape reconstruction," in *BMVC*, 2015, pp. 87.1–87.12 (cit. on pp. 6, 8).

[36] M. Schumacher, M. Piotraschke, and V. Blanz, "Hallucination of facial details from degraded images using 3D face models," *Image Vis. Comput.*, vol. 40, pp. 49–64, 2015 (cit. on p. 2).

[37] R. Szeliski, "Image formation," in *Computer vision: Algorithms and applications*, D. Gries and F. B. Schneider, Eds. Springer London, 2011, ch. 2, pp. 27–86 (cit. on p. 5).

[38] M. F. Tappen and C. Liu, "A bayesian approach to alignment-based image hallucination," in *ECCV*, 2012, pp. 236–249 (cit. on pp. 2, 6–8).

[39] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004 (cit. on p. 6).

[40] N. Wang, D. Tao, X. Gao, X. Li, and J. Li, "A comprehensive survey to face hallucination," *Int. J. Comput. Vis.*, vol. 106, no. 1, pp. 9–30, 2014 (cit. on p. 1).

[41] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Trans. Syst. Man Cybern. C, Appl. Rev.*, vol. 35, no. 3, pp. 425–434, 2005 (cit. on p. 2).

[42] C.-Y. Yang, S. Liu, and M.-H. Yang, "Structured face hallucination," in *CVPR*, 2013, pp. 1099–1106 (cit. on pp. 2, 6–8).

[43] H. Yang, X. Jia, C. C. Loy, and P. Robinson. (2015). An empirical study of recent face alignment methods. arXiv: 1511.05049 [cs.CV] (cit. on p. 6).

[44] J. Yang and T. Huang, "Image super-resolution: Historical overview and future challenges," in *Super-resoluton imaging*, P. Milanfar, Ed. CRC Press, 2010, ch. 1, pp. 1–33 (cit. on p. 3).

[45] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, 2010 (cit. on pp. 2, 6).