STOCHASTIC IMAGE MODELS FROM SIFT-LIKE DESCRIPTORS

A. DESOLNEUX* AND A. LECLAIRE*

Abstract. Extraction of local features constitutes a first step of many algorithms used in computer vision. The choice of keypoints and local features is often driven by the optimization of a performance criterion on a given computer vision task, which sometimes makes the extracted content difficult to apprehend. In this paper we propose to examine the content of local image descriptors from a reconstruction perspective. For that, relying on the keypoints and descriptors provided by the scale-invariant feature transform (SIFT), we propose two stochastic models for exploring the set of images that can be obtained from given SIFT descriptors. The two models are both defined as solutions of generalized Poisson problems that combine gradient information at different scales. The first model consists in sampling an orientation field according to a maximum entropy distribution constrained by local histograms of gradient orientations (at scale 0). The second model consists in simple resampling of the local histogram of gradient orientations at multiple scales. We show that both these models admit convolutive expressions which allow to compute the model statistics (e.g. the mean, the variance). Also, in the experimental section, we show that these models are able to recover many image structures, while not requiring any external database. Finally, we compare several other choices of points of interest in terms of quality of reconstruction, which confirms the optimality of the SIFT keypoints over simpler alternatives.

Key words. Image Synthesis, Random Image Model, Reconstruction from Features, SIFT, Poisson Editing, Maximum Entropy Distributions, Exponential Models

AMS subject classifications. 62M40, 65D18, 68U10, 94A08,

1. Introduction. [1]

A fundamental problem of vision consists in extracting a minimal representation that is sufficient for a human to apprehend the semantic content of an image. Marr and Hildreth [40, 39] proposed a *raw primal sketch* image representation based on the zero-crossings of the Laplacian computed at different scales, which extract spatial positions corresponding to edges, blobs, and terminations. Since this pioneering work, many authors proposed to extract different points of interest (keypoints), or local descriptors (features) based on several differential operators, while being invariant to given image transformations. Extracting keypoints and local features in images is indeed a fundamental step for many imaging tasks [21], like image recognition [63, 33, 9, 10, 26], image matching and rectification [33, 60, 32], object detection and tracking [8, 58, 66, 53], video stabilization [6, 65], image classification [29, 68, 28], etc. In this paper, we propose to discuss the role of such keypoints and descriptors, from a reconstruction point of view.

In the seminal paper [5], Attneave suggests that the most important points for image perception are the ones of maximum curvature. Since then, many techniques have emerged to single out keypoints and build local descriptors around them. Depending on the applicative context, one should use descriptors that are invariant with respect to specific geometric transformations[2] (e.g. image recognition generally needs invariance to homography and illumination change). Here we will only mention a few famous local descriptors, and we refer to [43, 59, 45, 32] for a more comprehensive survey.

Harris and Stephens proposed a combined corner and edge detector based on the

*CMLA, ENS Cachan, CNRS, Université Paris-Saclay, 94235 Cachan, France. (agnes.desolneux@cmla.ens-cachan.fr , arthur.leclaire@cmla.ens-cachan.fr).

[1] A preliminary version of this work was published as a conference paper in [17].

[2] The translation invariance is generally always required, and often trivial.

determinant and trace of the structure tensor of the image [23]. A multiscale variant based on a normalized Laplacian of Gaussian (LoG) scale-space, coined Harris-Laplace was proposed by Milokajczyk and Schmid [42]. The same authors also proposed in [42] the Harris-affine point detector which extends the previous one with a normalization step in order to get invariance to affine transformations. Tuytelaars and Mikolajczyk proposed in [60] two region detectors both starting from anchor points (e.g. Harris points); then the first one selects a region within detected edges around the anchor, and the second one extracts a region by analyzing intensity profiles on rays emanating from the anchor. Rosten and Drummond introduced in [55] the "features from accelerated segment test" (FAST) which is a corner detector accelerated by a machine learning technique. This approach has been further fastened by Mair et al. [37] using optimal decision trees, thus obtaining an "adaptive and generic accelerated segment test" (AGAST). Musé et al. proposed in [48] to extract shapes from the image level lines, and to process them in order to get an affine invariant representation.

In parallel of this research on keypoints, many techniques have been proposed for invariant local descriptions of images. An early descriptor is given by the local binary patterns (LBP) defined by Ojala et al. [51] which extracts signs of differences of image values on pixels located on a circular neighborhood of a keypoint. The LBP were originally designed for texture description but can also be used for face detection [1]. In [33], Lowe introduced the scale invariant feature transform (SIFT) which first extracts the keypoints as local extrema of the "Difference of Gaussian" (DoG) approximation of the LoG, and next computes around each keypoint a local descriptor based on normalized histograms of gradient direction (HOG), see the details in Section 2. Notice that similar HOG descriptors computed on a dense grid were actually used in [14] for person detection; one reference implementation of the HOG descriptors is given in [22]. A fully affine-invariant extension of SIFT, named ASIFT, was proposed by Morel and Yu [45] and consists in applying the SIFT method with the image transformed with several simulated affine maps. The SURF method (Speeded-up robust features) proposed by Bay et al. [7] is closely related in construction to the SIFT method, but allows for a faster implementation. At a higher semantic level, local image behavior can be also represented as visual words [58, 11] which are obtained as cluster points in a feature space. Later, some authors proposed to describe a patch using local binary descriptors (LBD), which extracts the signs of differences between Gaussian measurements taken at different locations. Using different ways of selecting these locations leads to the methods BRISK [30] (binary robust invariant scalable keypoints) or FREAK [2] (fast retina keypoint). All of these descriptors have quite different invariance properties (evaluated either in a theoretical or experimental framework).

Long before the design of these image descriptors, the question of a minimal representation of an image was thoroughly studied, mainly for compression purpose. Through the concept of *raw primal sketch*, Marr [39] suggested that the human visual system processes images by retaining essentially the lines of zero-crossing of the Laplacian at several scales. This leads to the conjecture that an image is uniquely defined by these zero-crossing lines, a conjecture that was later precised by Mallat [38] using wavelet modulus maxima. Both these conjectures were proved wrong by Meyer [41] but still, algorithms for approximate reconstruction were proposed by Hummel and Moniot [24] for zero-crossings and by Mallat and Zhong [38] for the case of wavelet modulus maxima. Besides, unique characterization can be shown to be true under some additional hypotheses [12, 13, 56, 4, 3].

From a more practical point of view, several authors have raised the question of inversion of a feature-based representation. For example, Elder and Zucker [20] proposed an algorithm for image reconstruction from detected contours, based on the heat diffusion. Nielsen and Lillholm [50] consider the problem of variational reconstruction from linear measurements; in addition to the minimum variance reconstruction (given by the pseudo-inverse of the measurements matrix), they propose two variational reconstructions based on either the entropy (of the image seen as a probability distribution on its domain) or the H^1 norm. Interestingly, they discuss the problem of extracting a subset of linear measurements which leads to the best reconstruction and empirically compare three different strategies for that purpose.

Motivated by privacy issues (since the descriptors may be transmitted on an unsecured network), Weinzaepfel et al. [64] addressed image reconstruction from the output of a SIFT transform adapted with elliptic keypoints. One important difference with previous works is that this method exploits a database of image patches: for each keypoint, a patch with similar description is looked for in the database, and all the patches are stitched together with Poisson image editing [52]. Vondrick et al. [62] address reconstruction from dense HOGs by relying on a paired dictionary representation of HOGs and patches. Also, d'Angelo et al. [15] address reconstruction from local binary descriptors by relying on primal-dual optimization techniques; in contrast with [64, 62], this method does not need any external information. Kato and Harada [27] formulate reconstruction from bag of visual words as a problem of quadratic assignment. Finally, Juefei-Xu and Savvides [25] propose to invert the LBP representation with an approach based on paired dictionary learning with an ℓ^0 constraint.

More recently, the success of deep convolutional neural networks in image classification [28, 67] has urged the need of inverting the corresponding representations in order to intuitively understand the kind of information that is extracted at each layer. Even if they do not formulate it as an inverting procedure, Zeiler and Fergus [67] proposed to build a deconvolution network that allows to visualize in image space the stimuli that excite one response at a particular layer of the neural network. Given an image u, Mahendran and Vedaldi [35, 36] proposed to search for a pre-image of an image representation $\varphi(u)$ by minimizing a functional containing a loss term related to the representation φ and a regularizing term (in particular the H^1 norm). Even if the regularizer is convex, the transformation φ is in general highly non-linear so that the resulting optimization problem is not convex; so the output of the inversion may depend on the parameters and initializations of the chosen optimization procedure. On the other hand, Dosovitskiy and Brox [19] suggest to learn an approximate left inverse of the representation (i.e. a mapping φ_L^{-1} such that $\varphi_L^{-1}(\varphi(u)) \approx u$ for every u) in the form of an up-convolutional network. These methods are generic in the sense that they can be applied to any image representation that can be approximated by the output of a convolutional neural network; in particular, the authors of [19] display inversion results for both HOG, SIFT and AlexNet [28] representations. Notice that the inversion/visualization techniques of [67, 19] exploit an external database while the one of [35, 36] does not.

Instead of building a uniquely defined inversion technique (using regularization), another way to perform reconstruction from the image representation φ is to sample from a stochastic model that explores the set of pre-images of $\varphi(u)$. This is particularly relevant if one uses an image representation that is not invertible: for example, the SIFT cells of an image may not cover its whole domain and thus many images could have the same SIFT descriptors. Besides, the HOG descriptors are inherently of a statistical nature: each HOG extracts the distribution of gradient orientations in one small area. Thus they only provide a locally pooled information and thus do not precisely constrain each gradient value. For this reason, the inversion by direct (regularized) optimization proposed in [35, 36] is not adapted to the usual SIFT representation (sometimes called sparse SIFT as opposed to SIFT descriptors computed on a dense grid).

One way to address this problem is to sample from the maximum entropy model that complies with these statistical constraints. Such maximal entropy models were considered by Zhu, Wu and Mumford in [69, 47] for texture modelling based on responses to an automatically selected subset of filters chosen in a filter bank. This approach has been recently extended by Lu, Zhu and Wu to responses to a pre-trained neural network [34]. Maximum entropy models were also used to question the noise models used in the *a contrario* framework for feature detections in images [18]: in [16], for two types of given detections (cluster of points, or line segments), Desolneux proposes explicit computations of maximal entropy image models that lead to the same detections (in average). Let us emphasize that one important difference with previous works is that, more than reconstructing the original image, we aim at exploring the set of images with similar HOG description at the keypoints positions, with the least possible *a priori* of what the reconstruction should look like. In contrast, the dependence on an external database in [52, 19] poses a strong *a priori* on the reconstruction.

In the present paper, we propose two stochastic models that complies with statistical features given by a SIFT-like representation. In order to derive explicit computations, we work on a simplified SIFT transform which extracts multiscale HOGs from regions around the (usual) SIFT keypoints. The first model, called MaxEnt, is indeed an instance of maximum entropy model which complies with local statistical constraints on the gradient orientations (at scale 0, i.e. the image scale). Once the parameters of this model are estimated (using a gradient descent), a target gradient orientation can be sampled, and we recover an image by solving a classical Poisson problem. The second model, called MS-Poisson, consists in first independent sampling of multiscale gradient orientations in all the SIFT cells, and next merging all the pieces by solving a global multiscale Poisson problem. Even if this model does not solve an explicit maximum entropy problem, it allows to coherently merge information given at several scales. Several experiments show that both these models are able to recover large image structures and compare well to the results of [64] while not using any external information. Finally, we discuss the definition of the SIFT keypoints in terms of optimality of reconstruction, thus raising the following question related to visual information theory: "Can we measure the optimality (at fixed memory budget) of some image descriptor in terms of reconstruction?"

The paper is organized as follows. In Section 2, we briefly recall the main steps of the SIFT method, and explain the simplified SIFT descriptors that we use for reconstruction. In Section 3, we build and study the maximum entropy model (MaxEnt) used for reconstruction from monoscale HOGs computed in the SIFT subcells. In Section 4, we propose the multiscale Poisson model (MS-Poisson) that allows to comply with multiscale HOGs taken in the SIFT subcells; the corresponding H^1-regularized multiscale Poisson problem is explicitly solved. Finally, in Section 5 we display several reconstruction results obtained with both models (applied with simplified SIFT, or also the true SIFT), study the variability of the reconstruction (in terms of first and second order moments, but also of SIFT keypoints computed on the reconstruction). We also compare with other existing reconstruction techniques and apply the reconstruction models on other keypoint sets, thus confirming (from the synthesis perspective) the efficiency of the SIFT method for global image description. Finally in Section 6 we conclude the discussion proposed in this paper and open some perspectives for future research. A preliminary version of this work was published as a conference paper in [17]. Compared to the conference version, here we explain in more details the derivation of MaxEnt and MS-Poisson models and we provide some more properties of these models and in particular explicit formulae for the first and second order moments of these models. We also propose several new experiments which illustrate the performance and the variability of these models (with qualitative and quantative evaluation) and question the role of the keypoint definition in the quality of reconstruction.

2. A Brief Summary of the SIFT Method.

In this section we briefly recall the construction of keypoints and local descriptors used in the SIFT method, and we explain the simplified descriptors that will be later used for the reconstruction in the next sections.

2.1. Gaussian Scale-Space and Keypoints.

Following [31], we introduce the Gaussian scale-space in a continuous domain. Let $u : \mathbb{R}^2 \to \mathbb{R}$ be an integrable function. For $\sigma > 0$, we introduce the function $g_\sigma : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$g_\sigma(\boldsymbol{x}) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{|\boldsymbol{x}|^2}{2\sigma^2}\right).$$

The Gaussian scale-space associated with u is then defined by the convolution

$$\forall \boldsymbol{x} \in \mathbb{R}^2,\ \forall \sigma > 0, \quad L_u(\boldsymbol{x}, \sigma) = g_\sigma * u(\boldsymbol{x}) = \int_{\mathbb{R}^2} g_\sigma(\boldsymbol{y}) u(\boldsymbol{x} - \boldsymbol{y}) d\boldsymbol{y}.$$

Another way to parameterize the scale-space is to use a time parameter $t = \sigma^2$ and the kernel $k_t = g_{\sqrt{t}}$ which satisfies

$$\frac{\partial}{\partial t}(k_t(\boldsymbol{x})) = \frac{1}{2}\Delta k_t(\boldsymbol{x}).$$

In other words, $(\boldsymbol{x}, t) \mapsto L_u(\boldsymbol{x}, \sqrt{t})$ is the solution of the heat equation on $\mathbb{R}^2$ with initial condition u (in particular, it is a $\mathcal{C}^\infty$ function on $\mathbb{R}^2 \times (0, \infty)$).

Then we consider the scale-normalized Laplacian of Gaussian $\sigma^2 \Delta g_\sigma$. The PDE satisfied by k_t gives after change of variables that

$$\sigma \frac{\partial g_\sigma}{\partial \sigma}(\boldsymbol{x}) = \sigma^2 \Delta g_\sigma(\boldsymbol{x}) = \left(\frac{|\boldsymbol{x}|^2 - 2\sigma^2}{2\pi\sigma^4}\right) \exp\left(-\frac{|\boldsymbol{x}|^2}{2\sigma^2}\right).$$

The detection of keypoints will be based on the local extrema of the function

$$D_u(\boldsymbol{x}, \sigma) := \sigma^2 \Delta g_\sigma * u(\boldsymbol{x}) = \sigma^2 \Delta (g_\sigma * u)(\boldsymbol{x}).$$

The following proposition which is recalled without proof shows that these keypoints are covariant to several image transformations.

Proposition 1 ([31]). *We have the following invariance properties.*

1. $\forall a \in \mathbb{R},\ D_{au} = a D_u$.
2. *If v is an affine function of $\boldsymbol{x}$, then $D_{u+v} = D_u$.*
3. *If $\boldsymbol{h} \in \mathbb{R}^2$ and $\tau_{\boldsymbol{h}} u(\boldsymbol{x}) = u(\boldsymbol{x} - \boldsymbol{h})$ is a translated version of u, then*

$$D_{\tau_{\boldsymbol{h}} u}(\boldsymbol{x}, \sigma) = D_u(\boldsymbol{x} - \boldsymbol{h}, \sigma).$$

4. *(Scale invariance) If* $u(\boldsymbol{x}) = v(s\boldsymbol{x})$ *with* $s > 0$*, for all* $\boldsymbol{x} \in \mathbb{R}^2$*, then*

$$D_u(\boldsymbol{x}, \sigma) = D_v(s\boldsymbol{x}, s\sigma).$$

The existence of a keypoint $(\boldsymbol{x}, \sigma)$ indicates the presence of a blob-like structure at position $\boldsymbol{x}$ with scale σ. For example, the Gaussian function g_s ($s > 0$) admits a keypoint $(0, s)$ which corresponds to a strict local minimum of D_{g_s}.

The authors of [46] also discussed the effect of several other image transformations on the SIFT keypoints but left aside the factor σ^2 in the definition of D_u.

2.2. SIFT Summary. In the paper by Lowe [33], the scale-normalized LoG is approximated by a finite difference of Gaussian functions: for a constant scale factor $k > 1$, he considers instead

$$(1) \qquad (\boldsymbol{x}, \sigma) \mapsto (g_{k\sigma} - g_\sigma)(\boldsymbol{x}) \approx (k\sigma - \sigma)\frac{\partial g_\sigma}{\partial \sigma}(\boldsymbol{x}) = (k-1)\sigma^2 \Delta g_\sigma(\boldsymbol{x}).$$

Also, the practical implementation of [33] only works with discretized images, so that the extracted keypoints are actually strict local extrema computed on a discretized scale-space.

Here is a quick summary of the original SIFT method [33]. For technical details we refer the reader to [54]. Here, and in the remaining of the paper, u_0 refers to the original image on which we compute keypoints and local descriptors.

1. Computing SIFT keypoints:
   (a) Extract local extrema of a discrete version of (1).
   (b) Refine the positions of the local extrema in position and scale using a quadratic approximation.
   (c) Discard extrema with low contrast (thresholding low values of (1)) and extrema located on edges (thresholding high values of the ratio between Hessian eigenvalues).
2. Computing SIFT local descriptors associated with the keypoint $(\boldsymbol{x}, \sigma)$:
   (a) Compute one or several principal orientations α. For that, in a square of size $9\sigma \times 9\sigma$ centered at $\boldsymbol{x}$ (and parallel to the image axes), compute a smoothed histogram of orientations of $\nabla g_\sigma * u_0$, and extract its significant local maxima.
   (b) For each detected orientation α, consider a grid of 4×4 square regions around $(\boldsymbol{x}, \sigma)$. These square regions, which we call SIFT subcells, are of size $3\sigma \times 3\sigma$ with one side parallel to α. In each subcell compute the histogram of $\mathrm{Angle}(\nabla g_\sigma * u_0) - \alpha$ quantized on 8 values $(\ell\frac{\pi}{4}, 1 \leq \ell \leq 8)$.
   (c) Normalization: the 16 histograms are concatenated to obtain a feature vector $f \in \mathbb{R}^{128}$, which is thresholded and normalized

$$(2) \qquad f_k \leftarrow \min(f_k, 0.2\|f\|_2) \ , \quad f_k = \min\left(255, \left\lfloor 512\frac{f}{\|f\|_2}\right\rfloor\right)$$

   and finally quantized to 8-bit integers.

When computing orientation histograms in steps 2(a) and 2(b), each pixel votes with a weight that depends on the value of the gradient norm at scale σ and on its distance to the keypoint center $\boldsymbol{x}$. Also in step 2(b), there is a linear splitting of the vote of an angle between the two adjacent quantized angle values.

2.3. Keypoints and Descriptors used in our method. In the reconstruction models proposed in this paper, we work with images defined on a rectangle $\Omega \subset \mathbb{Z}^2$ and

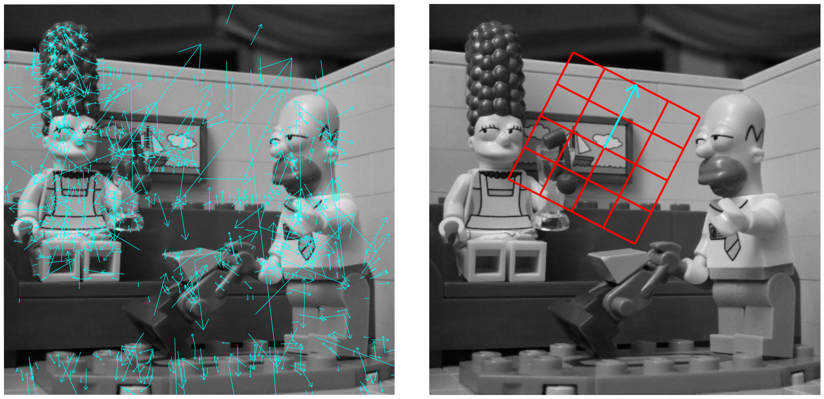

FIG. 1. ***Examples of SIFT keypoints and subcells***. *On the left, one can see an original image (Courtesy of J. Delon) with overimposed SIFT oriented keypoints* $(\boldsymbol{x}, \sigma, \alpha)$ *represented as arrows originating from* $\boldsymbol{x}$*, with orientation* α *and length* 6σ*. On the right, we display the 16 SIFT subcells associated with one particular keypoint. Each subcell is of size* $3\sigma \times 3\sigma$*.*

we consider the oriented keypoints extracted by the original SIFT method. However, we will only work with simplified SIFT descriptors in the sense that we extract hard-binned histograms of gradient orientations at several scales. In other words, we do not include the vote weights nor the normalization step 2(c).

We thus denote by $(s_j)_{j\in\mathcal{J}}$ the collection of SIFT subcells, $s_j \subset \Omega$ (if a $3\sigma \times 3\sigma$ subcell is not entirely contained in Ω, then we replace it with its intersection with Ω). The SIFT *subcells* must not be confounded with the SIFT *cells*: in a SIFT cell, there are 16 SIFT subcells so that different subcells s_j can correspond to the same keypoint. We will denote by $(\boldsymbol{x}_j, \sigma_j, \alpha_j)$ the oriented keypoint associated with s_j. For $\boldsymbol{y} \in \Omega$, we denote by $\mathcal{J}(\boldsymbol{y}) = \{j \in \mathcal{J} \mid \boldsymbol{y} \in s_j\}$ the set of indices of SIFT subcells containing $\boldsymbol{y}$. See Fig. 1 for an illustration.

For technical reasons, the statistics that are used in the two proposed models are slightly different: the MaxEnt model of Section 3 works on orientations at scale 0 whereas the MS-Poisson model of Section 4 works on orientations computed at multiple scales. For that reason, we postpone to the next sections the definition of the extracted statistics.

3. Stochastic Models for Gradient Orientations. In this section, we propose a model for generating random images constrained to have prescribed local HOGs in the SIFT subcells. When designing such a model, the main difficulty arises from the fact that several SIFT subcells can overlap, and thus one has to combine the information available in all corresponding local HOGs in a way that finally complies with all the statistical constraints. In order to cope with this issue, we exploit the framework of exponential distributions to design stochastic orientation models with prescribed statistical features. The obtained distribution is "as uniform (random) as possible" in the sense that it is of maximal entropy among all absolutely continuous distributions which satisfy the desired constraints. We combine this random orientation field with a deterministic magnitude (which is computed with the scales of locally available keypoints) in order to obtain a random objective vector field for the gradient. Finally we solve a Poisson reconstruction problem in order to get back a random image whose gradient is as close as possible as the randomly sampled objective vector field.

3.1. Exponential Models with local HOG. Recall that $\Omega \subset \mathbb{Z}^2$ is a discrete rectangle. We will denote by $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$ the set of angles, and $\mathbb{T}^\Omega$ the set of all possible orientation fields $\theta = (\theta(\boldsymbol{x}))_{\boldsymbol{x}\in\Omega}$ on Ω.

Extracted Statistics. For simplicity, in contrast with the usual SIFT method, in this section we only extract gradient orientations at scale 0 and besides we adopt the same quantization bins for all SIFT subcells

$$B_\ell = \left[(\ell-1)\tfrac{\pi}{4}, \ell\tfrac{\pi}{4}\right), \quad (1 \leqslant \ell \leqslant 8) \tag{3}$$

(i.e. we do not adapt quantization to the principal orientation of the keypoint).

For all $j \in \mathcal{J}$ and $1 \leqslant \ell \leqslant 8$, we thus consider the real-valued function defined on orientation fields by

$$\forall \theta \in \mathbb{T}^\Omega, \quad f_{j,\ell}(\theta) = \frac{1}{|s_j|} \sum_{\boldsymbol{x}\in s_j} \mathbf{1}_{B_\ell}(\theta(\boldsymbol{x})). \tag{4}$$

Thus $f_{j,\ell}(\theta)$ is the proportion of points $\boldsymbol{x} \in s_j$ having their orientation $\theta(\boldsymbol{x})$ in B_ℓ.

Maximum Entropy Distribution. We are then interested in probability distributions P on $\mathbb{T}^\Omega$ such that

$$\forall j \in \mathcal{J}, \forall \ell \in \{1,\dots,8\}, \quad \mathbb{E}_P\big(f_{j,\ell}(\Theta)\big) = \frac{1}{|s_j|} \sum_{\boldsymbol{x}\in s_j} \mathbb{P}(\theta(\boldsymbol{x}) \in B_\ell) = f_{j,\ell}(\theta_0)\ , \tag{5}$$

where $\theta_0 = \mathrm{Angle}(\nabla u_0)$ is the orientation field of the original image u_0, and where Θ is a random orientation field with probability distribution P. In other words, we look for a random model on orientation fields which preserves in average the extracted statistics in the SIFT subcells, see Fig. 2.

Let us emphasize here that we only aim at *average preservation* of the extracted statistics $(f_{j,\ell})$ because of the statistical nature of the SIFT descriptors. As will be clarified with the expression of the MaxEnt model (in particular in the case of non-overlapping SIFT subcells), this average preservation guarantee is sufficient to precisely set the gradient orientation distribution at each point.

There are many probability distributions P on $\mathbb{T}^\Omega$ that satisfy (5), and we will be mainly interested in the ones that are at the same time as "random" as possible, in the sense that they are of maximal entropy. The following theorem shows the existence of such maximal entropy distributions.

THEOREM 2. *There exists a family of numbers $\lambda = (\lambda_{j,\ell})_{j\in\mathcal{J},1\leqslant\ell\leqslant 8}$ such that the probability distribution*

$$dP_\lambda = \frac{1}{Z_\lambda} \exp\left(-\sum_{j,\ell} \lambda_{j,\ell} f_{j,\ell}(\theta) \right) d\theta, \tag{6}$$

where the partition function Z_λ is given by $Z_\lambda = \int_{\mathbb{T}^\Omega} \exp\big(-\sum_{j,\ell} \lambda_{j,\ell} f_{j,\ell}(\theta)\big) d\theta$, satisfies the constraints (5) and is of maximal entropy among all absolutely continuous probability distributions w.r.t. the Lebesgue measure $d\theta$ on $\mathbb{T}^\Omega$ satisfying the constraints (5).

Proof. This result directly follows from the general theorem given in [47]. The only difficulty is to handle the hypothesis of linear independence of the $f_{j,\ell}$. In our framework, the $f_{j,\ell}$ are not independent (in particular because $\sum_{\ell=1}^{8} f_{j,\ell} = 1$, and also

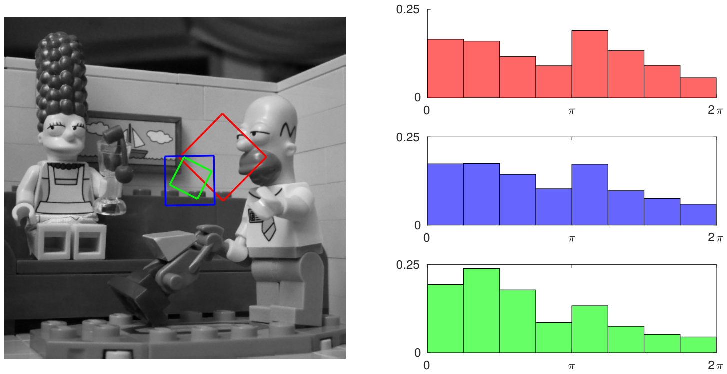

FIG. 2. ***Extracting HOG in SIFT subcells****. On the left, we display an original image (Courtesy of J. Delon) with three overimposed SIFT subcells s_j, and on the right, we display the corresponding HOG $(f_{j,\ell}(\theta_0))_{1\leqslant\ell\leqslant 8}$ extracted in these subcells. The MaxEnt model is a probability distribution on orientation fields that will respect in average the local HOG extracted in the SIFT subcells.*

because there may be other dependencies for instance when one subcell is exactly the union of two smaller subcells). But one can still apply the theorem to an extracted linearly independent subfamily. This gives the existence of the solution for the initial family $(f_{j,\ell})$ (but of course not the unicity). ∎

Remark: We do not repeat here the argument (based on Lagrange multipliers) showing that maximizing entropy under constraints (5) leads to exponential distributions. However, once a solution P_λ has been computed, and if P is an absolutely continuous probability distribution satisfying (5), one can write the Kullback-Leibler divergence using the entropy $H(P)$:

$$D(P||P_\lambda) = \int \log\Big(\frac{P(\theta)}{P_\lambda(\theta)}\Big)P(\theta)d\theta = -H(P) + \log Z_\lambda + \sum \lambda_{j,\ell} f_{j,\ell}(\theta_0), \tag{7}$$

which shows that maximizing $H(P)$ under (5) is equivalent to minimize $D(P||P_\lambda)$. In particular, this shows that the maximal entropy distribution under (5) is unique (because of the strict concavity of the entropy) even if there may be several sets of parameters λ corresponding to that solution.

Independence Property of the MaxEnt Model.

PROPOSITION 3. *Under P_λ the values $\Theta(\boldsymbol{x})$ are independent. Besides, the probability density function of $\Theta(\boldsymbol{x})$ is given by*

$$\frac{1}{Z_{\lambda,\boldsymbol{x}}} e^{-\varphi_{\lambda,\boldsymbol{x}}} = \frac{1}{Z_{\lambda,\boldsymbol{x}}} \sum_{\ell=1}^{8} \exp\Big(-\sum_{j\in\mathcal{J}(\boldsymbol{x})} \frac{\lambda_{j,\ell}}{|s_j|}\Big)\mathbf{1}_{B_\ell} \tag{8}$$

$$\textit{where} \quad Z_{\lambda,\boldsymbol{x}} = \sum_{\ell=1}^{8} \exp\Big(-\sum_{j\in\mathcal{J}(\boldsymbol{x})} \frac{\lambda_{j,\ell}}{|s_j|}\Big)|B_\ell|. \tag{9}$$

Proof. Taking the logarithm of (6), one can group the terms corresponding to the same pixel $\boldsymbol{x}$ so that

$$-\log\frac{dP_\lambda}{d\theta} - \log Z_\lambda = \sum_{j\in\mathcal{J},1\leqslant\ell\leqslant 8} \lambda_{j,\ell} f_{j,\ell}(\theta) = \sum_{\boldsymbol{x}\in\Omega} \varphi_{\lambda,\boldsymbol{x}}(\theta(\boldsymbol{x})), \tag{10}$$

$$\text{where} \quad \varphi_{\lambda,\boldsymbol{x}} = \sum_{\ell=1}^{8} \left(\sum_{j\in\mathcal{J}(\boldsymbol{x})} \frac{\lambda_{j,\ell}}{|s_j|} \right) \mathbf{1}_{B_\ell}. \tag{11}$$

We thus obtain that P_λ can be written in a separable form. ◻

On the one hand, this proposition shows that for a given λ, one can easily sample from the model P_λ. On the other hand, it also allows to compute several statistics associated with this model. In particular, we can compute for any bounded measurable function $\psi : \mathbb{T} \to \mathbb{C}$

$$\mathbb{E}_{P_\lambda}[\psi(\Theta(\boldsymbol{x}))] = \frac{\sum_{\ell=1}^{8} \exp\left(-\sum_{j\in\mathcal{J}(\boldsymbol{x})} \frac{\lambda_{j,\ell}}{|s_j|}\right) \int_{B_\ell} \psi(t)dt}{\sum_{\ell=1}^{8} \exp\left(-\sum_{j\in\mathcal{J}(\boldsymbol{x})} \frac{\lambda_{j,\ell}}{|s_j|}\right) |B_\ell|} \tag{12}$$

It also allows to compute the expected value of the statistics $f(\Theta)$ in the model P_λ (which will be useful in Section 3.3)

$$\mathbb{E}_{P_\lambda}[f_{j,\ell}(\Theta)] = \frac{1}{|s_j|} \sum_{\boldsymbol{x}\in s_j} \mathbb{P}(\Theta(\boldsymbol{x}) \in B_\ell) = \frac{1}{|s_j|} \sum_{\boldsymbol{x}\in s_j} \frac{\exp\left(-\sum_{k\in\mathcal{J}(\boldsymbol{x})} \frac{\lambda_{k,\ell}}{|s_k|}\right)|B_\ell|}{\sum_{1\leqslant \ell' \leqslant 8} \exp\left(-\sum_{k\in\mathcal{J}(\boldsymbol{x})} \frac{\lambda_{k,\ell'}}{|s_k|}\right)|B_{\ell'}|}. \tag{13}$$

But it remains to show how to estimate λ in order to satisfy the constraints (5). These constraints can be rewritten as

$$\forall j, \ell, \quad \sum_{\boldsymbol{x}\in s_j} \frac{1}{Z_{\lambda,\boldsymbol{x}}} \exp\left(- \sum_{k\in\mathcal{J}(\boldsymbol{x})} \frac{\lambda_{k,\ell}}{|s_k|}\right)|B_\ell| = |\{\boldsymbol{x} \in s_j \ ; \ \theta_0(\boldsymbol{x}) \in B_\ell \}| \ . \tag{14}$$

Notice that this system is highly non-linear and is in general difficult to solve.

A simple case: non-overlapped SIFT subcells. When a SIFT subcell s_j is not overlapped, then we have for any $\boldsymbol{x} \in s_j$, $|\mathcal{J}(\boldsymbol{x})| = 1$ and therefore

$$Z_{\lambda,\boldsymbol{x}} = \sum_{\ell=1}^{8} \exp\left(-\frac{\lambda_{j,\ell}}{|s_j|}\right)|B_\ell|. \tag{15}$$

Then (14) gives

$$\forall \ell, \quad \frac{1}{Z_{\lambda,\boldsymbol{x}}} \exp\left(-\frac{\lambda_{j,\ell}}{|s_j|}\right) = \frac{|\{\boldsymbol{x} \in s_j \ ; \ \theta_0(\boldsymbol{x}) \in B_\ell \}|}{|s_j||B_\ell|} = f_{j,\ell}(\theta_0), \tag{16}$$

which gives the marginal distribution on any $\boldsymbol{x} \in s_j$:

$$\frac{1}{Z_{\lambda,\boldsymbol{x}}} e^{-\varphi_{\lambda,\boldsymbol{x}}} = \sum_{\ell=1}^{8} \frac{|\{\boldsymbol{x} \in s_j \ ; \ \theta_0(\boldsymbol{x}) \in B_\ell \}|}{|s_j||B_\ell|} \mathbf{1}_{B_\ell} = \sum_{\ell=1}^{8} f_{j,\ell}(\theta_0) \frac{1}{|B_\ell|} \mathbf{1}_{B_\ell}. \tag{17}$$

So when the subcells do not overlap, the maximum entropy distribution only amounts to independent resampling of the local HOGs, as expected. Notice that we indeed obtain a unique maximal entropy distribution. However, the solutions λ are only unique up to the addition of a constant: indeed the last calculation shows that for a non-overlapped subcell s_j, there exists a constant $c_j > 0$ such that

$$\forall \ell, \quad \lambda_{j,\ell} = -|s_j|(\log f_{j,\ell}(\theta_0) + \log c_j). \tag{18}$$

Maximum-likelihood estimation. If the SIFT subcells intersect, there is no explicit solution anymore. To cope with that, as in [69] we use a numerical scheme to find the maximum entropy distribution P_λ. The solution can be obtained with a traditional maximum likelihood estimation technique, as will be detailed here. Indeed, the minus-log-likelihood function can be written as

$$\Phi(\lambda) = \log Z_\lambda + \sum_{j,\ell} \lambda_{j,\ell} f_{j,\ell}(\theta_0). \tag{19}$$

The gradient of Φ can be obtained by differentiating the partition function

$$\frac{\partial \log Z_\lambda}{\partial \lambda_{j,\ell}} = \frac{1}{Z_\lambda}\frac{\partial Z_\lambda}{\partial \lambda_{j,\ell}} = -\mathbb{E}_{P_\lambda}\big[f_{j,\ell}(\Theta)\big], \tag{20}$$

which gives

$$\frac{\partial \Phi}{\partial \lambda_{j,\ell}} = f_{j,\ell}(\theta_0) - \mathbb{E}_{P_\lambda}\big[f_{j,\ell}(\Theta)\big]. \tag{21}$$

Notice that $\nabla\Phi(\lambda) = 0$ if and only if P_λ satisfies the constraints (5).

Similarly, we can also obtain the second order derivatives

$$\frac{\partial^2 \Phi}{\partial \lambda_{j,\ell}\partial \lambda_{j',\ell'}} = \mathbb{E}_{P_\lambda}\Big[\big(f_{j,\ell}(\Theta) - \mathbb{E}_{P_\lambda}[f_{j,\ell}(\Theta)]\big)\big(f_{j',\ell'}(\Theta) - \mathbb{E}_{P_\lambda}[f_{j',\ell'}(\Theta)]\big)\Big]. \tag{22}$$

One can observe that this Hessian matrix $\nabla^2\Phi(\lambda)$ is actually the covariance of the vector $f(\Theta)$ when Θ has distribution P_λ. In particular it is a semi-positive definite matrix, which shows that Φ is a convex function. The global minima of Φ are exactly the points λ where $\nabla\Phi$ vanishes, which is equivalent to have the constraints (5) on P_λ.

Therefore, we can compute the solution P_λ by a gradient descent algorithm in order to minimize Φ. The complete algorithm is summarized in Section 3.3. Since Φ is not strictly convex, we will not have a guarantee of convergence on the iterates, but on the function values. Since $|f_{j,\ell}(\theta)| \leqslant 1$, it is straightforward to see that all coefficients of the Hessian $\nabla^2\Phi(\lambda)$ have modulus $\leqslant 1$. Therefore, the ℓ^2 operator norm of $\nabla^2\Phi$ is bounded by $8|\mathcal{J}|$, which implies that $\nabla\Phi$ is L-Lipschitz with $L = 8|\mathcal{J}|$. Writing λ^k the iterates of the gradient descent with constant step size $h < \frac{2}{L}$, [49, Th 2.1.14] gives

$$\Phi(\lambda^k) - \min \Phi = \mathcal{O}\Big(\frac{1}{k}\Big). \tag{23}$$

Let us also mention that since Φ is convex smooth, it would be possible to use higher-order optimization schemes to minimize Φ. However, Newton's method will be in general too costly because of the dimension of the system and because the Hessian may be ill-conditioned.

3.2. Monoscale Poisson Reconstruction. Now that we have built a random orientation field Θ with maximum entropy distribution P_λ, we will use it to propose a target vector field V for the image gradient. More precisely, we set the gradient magnitude at $\boldsymbol{x}$ in a deterministic manner, as the inverse scale of the smallest subcell that covers $\boldsymbol{x}$. For pixels $\boldsymbol{x}$ which lie outside the SIFT subcells, we set $V(\boldsymbol{x}) = 0$. This choice allows to give more weight to the locations for which we have information at finer scale. It is also motivated by the following homogeneity argument. Assume that $u : \mathbb{R}^2 \to \mathbb{R}$ has a keypoint $(\boldsymbol{x}, \sigma)$ and for $a > 0$ let $v(\boldsymbol{y}) = u(\frac{\boldsymbol{y}}{a})$. Then, thanks to Proposition 1, v has a keypoint $(a\boldsymbol{x}, a\sigma)$. Let us compare the mean gradient magnitude at scale σ in the corresponding subcell s to the analogous quantity for v. A simple computation shows that

$$\frac{1}{|as|} \int_{\lambda s} |\nabla g_{a\sigma} * v(\boldsymbol{y})| d\boldsymbol{y} = \frac{1}{a} \frac{1}{|s|} \int_{s} |\nabla g_{\sigma} * u(\boldsymbol{y})| d\boldsymbol{y},$$

so that the mean gradient magnitude in the subcell is multiplied by $\frac{1}{a}$ with the change of scale. From this calculation we get the following remark: if two very similar shapes (with similar graylevels) are seen in the image at two different scales with ratio a, then we can obtain a pairwise matching of their SIFT keypoints, and the ratio between the mean gradient magnitude of the two matched subcells is $1/a$. Of course this remark does not extend to the comparison of two SIFT subcells with very different geometric content, but it still provides a general rule for fixing the gradient magnitude as the inverse of the scale. Therefore, we get the random objective vector field

$$\forall \boldsymbol{x} \in \Omega, \quad V(\boldsymbol{x}) = \Big(\max_{j \in \mathcal{J}(\boldsymbol{x})} \frac{1}{\sigma_j} \Big) e^{i\Theta(\boldsymbol{x})} \mathbf{1}_{\mathcal{J}(\boldsymbol{x}) \neq \emptyset}. \tag{24}$$

The aim of the Poisson reconstruction is to compute an image whose gradient is as close as possible to the vector field $V = (V_1, V_2)$. In the case of image editing, this technique has been proposed by Pérez et al. [52] in order to copy pieces of an image into another one in a seamless way. More precisely, the goal is to minimize the functional

$$F(u) = \sum_{\boldsymbol{x} \in \Omega} \| \nabla u(\boldsymbol{x}) - V(\boldsymbol{x}) \|_2^2. \tag{25}$$

Since $F(c + u) = F(u)$ for any constant c, we can impose $\sum_{\boldsymbol{x} \in \Omega} u(\boldsymbol{x}) = 0$. Thus we set

$$U = \mathrm{Argmin}\{ F(u)\, ;\; u : \Omega \to \mathbb{R} \text{ and such that } \sum_{\boldsymbol{x} \in \Omega} u(\boldsymbol{x}) = 0 \}. \tag{26}$$

If we use periodic boundary conditions for the gradient, we can solve this problem with the Discrete Fourier Transform [44]. Indeed, if we use the simple derivation scheme based on periodic convolutions

$$\nabla u(\boldsymbol{x}) = \begin{pmatrix} \partial_1 * u(\boldsymbol{x}) \\ \partial_2 * u(\boldsymbol{x}) \end{pmatrix} \qquad \text{where} \quad \left\{ \begin{array}{ll} \partial_1 & = \delta_{(0,0)} - \delta_{(1,0)} \\ \partial_2 & = \delta_{(0,0)} - \delta_{(0,1)} \end{array} \right. , \tag{27}$$

the problem can be expressed in the Fourier domain with Parseval formula since

$$F(u) = \frac{1}{|\Omega|} \sum_{\boldsymbol{\xi}} |\widehat{\partial}_1(\boldsymbol{\xi}) \widehat{u}(\boldsymbol{\xi}) - \widehat{V}_1(\boldsymbol{\xi})|_2^2 + |\widehat{\partial}_2(\boldsymbol{\xi}) \widehat{u}(\boldsymbol{\xi}) - \widehat{V}_2(\boldsymbol{\xi})|_2^2. \tag{28}$$

Thus, for each $\boldsymbol{\xi}$ we have a barycenter problem which is simply solved by

$$\forall \boldsymbol{\xi} \neq 0, \quad \widehat{U}(\boldsymbol{\xi}) = \frac{\overline{\widehat{\partial}_1}(\boldsymbol{\xi}) \widehat{V}_1(\boldsymbol{\xi}) + \overline{\widehat{\partial}_2}(\boldsymbol{\xi}) \widehat{V}_2(\boldsymbol{\xi})}{|\widehat{\partial}_1(\boldsymbol{\xi})|^2 + |\widehat{\partial}_2(\boldsymbol{\xi})|^2} \quad \text{and} \quad \widehat{U}(0) = 0. \tag{29}$$

Let us emphasize (with the capital letter U) that the solution of this problem is random because the target field V is random.

Using the notation $\nabla = (\partial_1, \partial_2)^T$, $\widehat{\nabla} = (\widehat{\partial}_1, \widehat{\partial}_2)^T$, $z^* = \bar{z}^T$, we can write

$$\widehat{U}(\boldsymbol{\xi}) = \widehat{\nu}(\boldsymbol{\xi})\widehat{V}(\boldsymbol{\xi}) \quad \text{where} \quad \widehat{\nu}(\boldsymbol{\xi}) = \begin{cases} \frac{\widehat{\nabla}(\boldsymbol{\xi})^*}{|\nabla(\boldsymbol{\xi})|^2} \text{ if } \boldsymbol{\xi} \neq 0 \\ 0 \text{ if } \boldsymbol{\xi} = 0 \end{cases} . \tag{30}$$

Notice that $\widehat{\nu}(\boldsymbol{\xi}) \in \mathbb{C}^{1\times 2}$ and $\widehat{V}(\boldsymbol{\xi}) \in \mathbb{C}^{2\times 1}$ so that (30) is equivalent to

$$U = \nu * V = \nu_1 * V_1 + \nu_2 * V_2. \tag{31}$$

In other words, ν is the (vector-valued) convolution kernel associated to the Poisson reconstruction. This expression allows to compute the moments of the random field U (see also Section 4.3 for a detailed more general calculation).

3.3. Algorithm. Here we summarize the algorithm for estimating and sampling the MaxEnt model proposed in this section. In Fig. 3 we display an example of reconstruction with the MaxEnt model.

Algorithm: Estimating and Sampling the MaxEnt Model

- Maximum-likelihood estimation of λ
  - Compute the observed statistics $f(\theta_0) = (f_{j,\ell}(\theta_0))_{j,\ell}$.
  - Initialization $\lambda \leftarrow 0$. Choose a step size $h < \frac{4}{|\mathcal{J}|}$.
  - For $N(= 10000)$ iterations, compute $\bar{f} = \mathbb{E}_{P_\lambda}[f]$ using (13) and set

$$\lambda \leftarrow \lambda - h(f(\theta_0) - \bar{f}).$$

- Draw a sample θ according to the distribution P_λ.
- Compute the corresponding target vector field

$$V(\boldsymbol{x}) = \Big(\max_{j\in\mathcal{J}(\boldsymbol{x})} \frac{1}{\sigma_j}\Big) e^{i\theta(\boldsymbol{x})} \mathbf{1}_{\mathcal{J}(\boldsymbol{x})\neq\emptyset} \tag{32}$$

- Compute a sample u of MaxEnt via the Poisson reconstruction (29).

For images having many SIFT keypoints in overlapping positions, this algorithm may be slow to converge as can be observed on the case of Fig. 3. This case is relatively simple because it has only 187 keypoints but this corresponds already to $8 \times 16 \times 187 \approx 24000$ $\lambda_{j,\ell}$ parameters to estimate. This is why we use a stopping criterion based on a maximal number of iterations.

3.4. Discussion on MaxEnt Model. One drawback of MaxEnt is that the guarantee on the local distributions of orientations is lost after the Poisson reconstruction step. One way to cope with that would be to consider a model that operates directly on the image values, and not on the orientation field. Theorem 2 could be extended to statistics like

$$\tilde{f}_{j,\ell}(u) = \frac{1}{|s_j|} \sum_{\boldsymbol{x}\in s_j} \mathbf{1}_{B_\ell}(\mathrm{Angle}(\nabla u(\boldsymbol{x}))). \tag{33}$$

It is even possible to consider multiscale statistics using $\nabla g_{\sigma_j} * u$ instead of ∇u (as it will be the case in Section 4). But the analog of Proposition 3 would not hold for these models, so that sampling should rely on a Gibbs strategy. Its cost would be

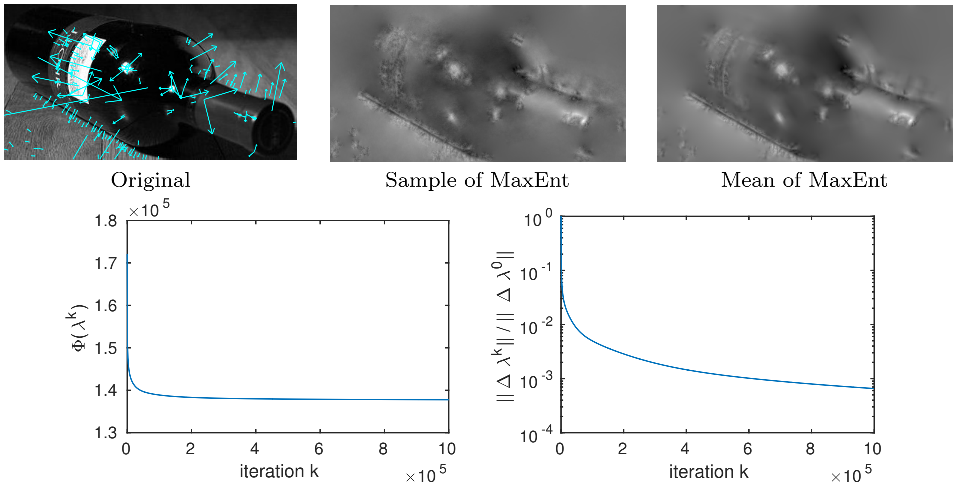

Fig. 3. ***Reconstruction with the MaxEnt model.*** *In the first row from left to right, we display an original image with overimposed 187 oriented keypoints, a sample of the associated MaxEnt model, and the expectation of the MaxEnt model. In the second row we display the evolution of Φ along the iterates, and also the behavior of the difference between iterates $\Delta\lambda^k = \lambda^k - \lambda^{k-1}$. The value of Φ stabilizes in about 10^5 iterations. One can remark that both reconstructions show several important structures of the original image. The mean reconstruction is of course smoother than a sample of the model (because pixels are sampled independently, see Proposition 3).*

clearly prohibitive in the multiscale case due to the large Markov neighborhood size. Even in the monoscale case the convergence of this Gibbs sampler may be very long depending on the parameters λ; and since we would need one sample per iteration of gradient descent to estimate λ, we chose to leave it aside and concentrate on models with reasonably fast sampling.

Also, one can consider another orientation model in which the local HOGs are computed with a quantization that depends on the keypoint orientation. The independence property still holds for this model, and the marginal orientations still have a piecewise constant density, but the number of parameters would be much larger (there would be as many ℓ's as bins of a subdivision that is adapted to all keypoints orientations). Therefore this model is practically untractable, and also only of minor interest. Indeed, in view of the results of Fig. 3, it is likely that the used quantization has only a minor impact on the visual results (provided that we still have a minimal number of bins).

4. Multiscale Poisson Model.

In this section, we propose a stochastic model, called MS-Poisson, for reconstruction using multiscale local HOGs computed in SIFT subcells. This new model is based on a heuristic algorithm for orientation resampling in all SIFT subcells. Therefore, in contrast to the MaxEnt model, the MS-Poisson model can be straightforwardly sampled using the multiscale local HOGs, and does not require an iterative estimation procedure. Another difference is that MS-Poisson is designed to combine information at multiple scales, whereas MaxEnt only operates with the gradient at scale 0.

4.1. Construction of MS-Poisson Model.

Extracted Statistics. The MS-Poisson model is based on local statistics on multiscale gradient orientations. More precisely, in s_j we extract the quantized HOG at scale σ_j

$$H_{j,\ell} = \frac{1}{|s_j|}\left|\left\{\boldsymbol{x} \in s_j \;;\; \text{Angle}(\nabla g_{\sigma_j} * u_0)(\boldsymbol{x}) - \alpha_j \in [(\ell-1)\tfrac{\pi}{4}, \ell\tfrac{\pi}{4})\right\}\right|. \tag{34}$$

In view of resampling, this local HOG can be identified to a piecewise constant density function

$$h_j = \frac{4}{\pi}\sum_{\ell=1}^{8} H_{j,\ell}\mathbf{1}_{[\alpha_j+(\ell-1)\frac{\pi}{4}, \alpha_j+\ell\frac{\pi}{4})}. \tag{35}$$

Notice that, in contrast to the statistics (4) used in the MaxEnt model, the quantization here depends on the local orientation α_j.

Target Vector Fields at Multiple Scales. Using the local orientation distributions h_j, we define vector fields $V_j : \Omega \to \mathbb{R}^2$ that will serve as objective gradients at scale σ_j in the SIFT subcell s_j. We propose to set

$$\forall \boldsymbol{x} \in \Omega, \quad V_j(\boldsymbol{x}) = \frac{1}{\sigma_j} e^{i\gamma_j(\boldsymbol{x})}\mathbf{1}_{s_j}(\boldsymbol{x}), \tag{36}$$

where the orientations $\gamma_j(\boldsymbol{x})$ are independently sampled according to the distribution h_j. Again, as justified in Section 3.2, we set the gradient magnitude in a deterministic way using the inverse of the scale σ_j. Once these vector fields V_j have been sampled, we obtain an image U by solving a multiscale Poisson problem as explained in the next paragraph.

4.2. Multiscale Poisson Reconstruction. In order to simultaneously constrain the gradient at several scales $(\sigma_j)_{j\in\mathcal{J}}$, we propose to consider the following multiscale Poisson energy

$$G(u) = \sum_{j\in\mathcal{J}} w(\sigma_j) \sum_{\boldsymbol{x}\in\Omega} \|\nabla(g_{\sigma_j} * u)(\boldsymbol{x}) - V_j(\boldsymbol{x})\|_2^2, \tag{37}$$

where g_σ is the Gaussian kernel of standard deviation σ, $V_j = (V_{j,1}, V_{j,2})^T$ is the objective gradient at scale σ_j, and $\{w(\sigma_j), j \in \mathcal{J}\}$ is a set of weights. In our application, since there are more keypoints in the fine scales (i.e. with small σ_j), and since the keypoints at fine scales are generally more informative, a reasonable choice is to take all weights $w(\sigma_j) = 1$. But we keep these weights in the formula for the sake of generality. We thus set

$$U = \text{Argmin}\{G(u)\,;\, u : \Omega \to \mathbb{R} \text{ and such that } \sum_{\boldsymbol{x}\in\Omega} u(\boldsymbol{x}) = 0\}. \tag{38}$$

Again, with periodic boundary conditions, this problem can be expressed in Fourier domain as

$$G(u) = \frac{1}{|\Omega|}\sum_{j\in\mathcal{J}}\sum_{\boldsymbol{\xi}} w(\sigma_j)\Big(|\widehat{g}_{\sigma_j}(\boldsymbol{\xi})\widehat{\partial}_1(\boldsymbol{\xi})\widehat{u}(\boldsymbol{\xi}) - \widehat{V}_{j,1}(\boldsymbol{\xi})|_2^2 + |\widehat{g}_{\sigma_j}(\boldsymbol{\xi})\widehat{\partial}_2(\boldsymbol{\xi})\widehat{u}(\boldsymbol{\xi}) - \widehat{V}_{j,2}(\boldsymbol{\xi})|_2^2\Big). \tag{39}$$

As for the monoscale Poisson problem, the solution U is still a barycenter given by $\widehat{U}(0) = 0$ and

$$\forall \boldsymbol{\xi} \neq 0, \quad \widehat{U}(\boldsymbol{\xi}) = \frac{\displaystyle\sum_{j\in\mathcal{J}} w(\sigma_j)\widehat{g}_{\sigma_j}(\boldsymbol{\xi})\Big(\overline{\widehat{\partial}_1(\boldsymbol{\xi})}\widehat{V}_{j,1}(\boldsymbol{\xi}) + \overline{\widehat{\partial}_2(\boldsymbol{\xi})}\widehat{V}_{j,2}(\boldsymbol{\xi})\Big)}{\displaystyle\sum_{j\in\mathcal{J}} w(\sigma_j)|\widehat{g}_{\sigma_j}(\boldsymbol{\xi})|^2\Big(|\widehat{\partial}_1(\boldsymbol{\xi})|^2 + |\widehat{\partial}_2(\boldsymbol{\xi})|^2\Big)}. \tag{40}$$

Let us remark that in the above formula, we have $\widehat{g}_{\sigma_j}(\boldsymbol{\xi}) \in \mathbb{R}$ since g_{σ_j} is even.

Regularization. Notice that, depending on the finest scale, the denominator may numerically vanish in the high frequencies because of the term $\widehat{g}_{\sigma_j}(\boldsymbol{\xi})$ (as it is the case in a deconvolution problem). Therefore, it may be useful to add a regularization term controlled by a parameter $\mu > 0$. Then, if we set

$$U = \operatorname{Argmin}\{G(u) + \mu\|\nabla u\|_2^2\,;\, u : \Omega \to \mathbb{R} \text{ and such that } \sum_{\boldsymbol{x}\in\Omega} u(\boldsymbol{x}) = 0\}, \tag{41}$$

then we get the well-defined solution U given by $\widehat{U}(0) = 0$ and

$$\forall \boldsymbol{\xi} \neq 0, \quad \widehat{U}(\boldsymbol{\xi}) = \frac{\displaystyle\sum_{j\in\mathcal{J}} w(\sigma_j)\widehat{g}_{\sigma_j}(\boldsymbol{\xi})\Big(\,\overline{\widehat{\partial}_1(\boldsymbol{\xi})}\widehat{V}_{j,1}(\boldsymbol{\xi}) + \overline{\widehat{\partial}_2(\boldsymbol{\xi})}\widehat{V}_{j,2}(\boldsymbol{\xi})\Big)}{\left(\mu + \displaystyle\sum_{j\in\mathcal{J}} w(\sigma_j)|\widehat{g}_{\sigma_j}(\boldsymbol{\xi})|^2\right)\Big(|\widehat{\partial}_1(\boldsymbol{\xi})|^2 + |\widehat{\partial}_2(\boldsymbol{\xi})|^2\Big)}. \tag{42}$$

As we will see in Section 5.1, the parameter μ allows to attenuate the noise generated by the randomly sampled gradient fields in the fine scale SIFT subcells. We will see (empirically) that the value $\mu = 50$ realizes a good compromise between recovered details and smoothness.

We end this paragraph by summarizing the MS-Poisson sampling algorithm.

Algorithm: Sampling the MS-Poisson Model

- In each subcell s_j, draw independent orientations $\gamma_j(\boldsymbol{x}), \boldsymbol{x} \in s_j$ according to the p.d.f. h_j.
- Set $V_j = \frac{1}{\sigma_j}\mathbf{1}_{s_j} e^{i\gamma_j}$.
- Compute U by solving the MS-Poisson problem (41) with targets V_j, with $w(\sigma_j) = 1$ and $\mu = 50$.

Remark: In (42), one can observe that the solution to MS-Poisson actually solves a monoscale Poisson problem with objective vector field V whose Fourier transform is given by

$$\widehat{V}(\boldsymbol{\xi}) = \frac{\displaystyle\sum_{j\in\mathcal{J}} w(\sigma_j)\widehat{g}_{\sigma_j}(\boldsymbol{\xi})\widehat{V}_j(\boldsymbol{\xi})}{\mu + \displaystyle\sum_{j\in\mathcal{J}} w(\sigma_j)|\widehat{g}_{\sigma_j}(\boldsymbol{\xi})|^2}. \tag{43}$$

4.3. First and Second Order Moments. In order to compute the statistics of the MS-Poisson model, we remark that the multiscale Poisson reconstruction is actually a linear process. Indeed, for each j, let $\nu_j : \Omega \to \mathbb{R}^{1\times 2}$ be the vector-valued kernel defined by its discrete Fourier transform

$$\forall \boldsymbol{\xi} \neq 0, \quad \widehat{\nu}_j(\boldsymbol{\xi}) = \frac{w(\sigma_j)\widehat{g}_{\sigma_j}(\boldsymbol{\xi})\widehat{\nabla}(\boldsymbol{\xi})^*}{\left(\mu + \displaystyle\sum_{j'\in\mathcal{J}} w(\sigma_{j'})|\widehat{g}_{\sigma_{j'}}(\boldsymbol{\xi})|^2\right)|\widehat{\nabla}(\boldsymbol{\xi})|^2} \quad \text{and } \widehat{\nu}_j(0) = 0. \tag{44}$$

Then, as in Section 3.2 we get the convolutive expression

$$U = \sum_{j\in\mathcal{J}} \nu_j * V_j = \sum_{j\in\mathcal{J}} \Big(\nu_{j,1} * V_{j,1} + \nu_{j,2} * V_{j,2}\Big). \tag{45}$$

From this expression we can compute the moments of U. By linearity

$$\mathbb{E}(U) = \sum_{j\in\mathcal{J}} \nu_j * \mathbb{E}(V_j), \tag{46}$$

so that computing this expectation only amounts to compute $\mathbb{E}(V_j) = \frac{1}{\sigma_j}\mathbf{1}_{s_j}\mathbb{E}(e^{i\gamma_j})$.

We can also compute the variance. Since the objective fields $(V_j)_{j\in\mathcal{J}}$ are independent, we have

$$\mathrm{Var}(U(\boldsymbol{x})) = \sum_{j\in\mathcal{J}} \mathrm{Var}(\nu_j * V_j(\boldsymbol{x})). \tag{47}$$

Also, the $V_j(\boldsymbol{y})$ for different pixels $\boldsymbol{y}$ are independent so that

$$\begin{aligned}
\mathrm{Var}(\nu_j * V_j(\boldsymbol{x})) &= \mathrm{Var}\Big(\sum_{\boldsymbol{y}\in\Omega} \nu_j(\boldsymbol{x}-\boldsymbol{y})V_j(\boldsymbol{y})\Big) = \sum_{\boldsymbol{y}\in\Omega} \mathrm{Var}(\nu_j(\boldsymbol{x}-\boldsymbol{y})V_j(\boldsymbol{y})) && (48)\\
&= \sum_{\boldsymbol{y}\in\Omega} \nu_j(\boldsymbol{x}-\boldsymbol{y})\mathrm{Cov}(V_j(\boldsymbol{y}))\nu_j^T(\boldsymbol{x}-\boldsymbol{y}) && (49)\\
&= \sum_{\boldsymbol{y}\in\Omega} \nu_{j,1}^2(\boldsymbol{x}-\boldsymbol{y})\mathrm{Var}(V_{j,1}(\boldsymbol{y})) + \nu_{j,2}^2(\boldsymbol{x}-\boldsymbol{y})\mathrm{Var}(V_{j,2}(\boldsymbol{y})) && (50)\\
&\qquad + 2\nu_{j,1}(\boldsymbol{x}-\boldsymbol{y})\nu_{j,2}(\boldsymbol{x}-\boldsymbol{y})\mathrm{Cov}(V_{j,1}(\boldsymbol{y}), V_{j,2}(\boldsymbol{y})). && (51)
\end{aligned}$$

Therefore the variance of this model can be obtained by summing convolutions of the kernels ν_j with the covariances of V_j. Since $V_j(\boldsymbol{y}) = \frac{1}{\sigma_j}e^{i\gamma_j(\boldsymbol{y})}\mathbf{1}_{s_j}$ where $\gamma_j(\boldsymbol{y})$ has p.d.f. h_j given by (34), we can explicitly compute its covariance.

More generally, we can compute the covariance between two pixel values of U in a similar way, which gives

$$\mathrm{Cov}\big(U(\boldsymbol{x}), U(\boldsymbol{y})\big) = \sum_{j\in\mathcal{J}}\sum_{\boldsymbol{z}\in\Omega} \nu_j(\boldsymbol{x}-\boldsymbol{z})\mathrm{Cov}(V_j(\boldsymbol{z}))\nu_j^T(\boldsymbol{y}-\boldsymbol{z}). \tag{52}$$

5. Results and Discussion.

In this section, we give empirical evidence that both models MS-Poisson and MaxEnt are able to generate images that are similar to the original image in many aspects, which is confirmed by several quantitative results (in particular based on normalized correlations). We discuss the impact of the regularization parameter μ of the MS-Poisson model on the quality of the sampled images. We also compare MaxEnt and MS-Poisson in terms of local variance of the sampled images, and also in terms of resulting SIFT keypoints computed in the sampled images. After explaining how to adapt the MS-Poisson model to operate on true SIFT descriptors we compare with previous approaches of [64, 19]. Finally we discuss the impact of the keypoints definition on the quality of the reconstruction.

5.1. Results with MaxEnt and MS-Poisson model.

5.1.1. Comparison between MaxEnt and MS-Poisson. Let us first compare the reconstruction results obtained with MaxEnt and with MS-Poisson. On Fig. 4, using an original image with 386 keypoints, we display a sample of MaxEnt and a sample of MS-Poisson, together with the expected images of these models. One first remark is that both models are able to retrieve several geometric structures of the original image, so that much semantic content of the image can still be understood. For both models, one can observe that the samples are very close to the expected image, which will be later confirmed by the variance analysis on Fig. 8.

One crucial difference between MaxEnt and MS-Poisson is that they do not rely on the same gradient information. Indeed, MS-Poisson exploits gradients extracted at multiple scales while MaxEnt only operates with gradients at scale $\sigma = 0$ (i.e. the same scale as the image). This is why the results obtained with MS-Poisson will generally look blurrier than the ones obtained with MaxEnt. Besides, because of the multiscale nature of the input of MS-Poisson, the corresponding optimization problem had to be regularized; and the adopted H^1-regularization term is also a source of blur in the result. This is confirmed by Fig. 5 where we display several MS-Poisson reconstructions with varying regularization parameter μ. In Fig. 5 and in many other experiments, we observed that the parameter $\mu = 50$ realizes a good compromise between preserving geometric structures and removing spurious oscillations.

In the last row of Fig. 4, we also compare with the reconstructions obtained with the true gradient orientations (resp. multiscale gradient orientations) computed in the SIFT subcells and the gradient magnitude computed as in MaxEnt (resp. MS-Poisson). So the difference with MaxEnt (or MS-Poisson) is that local (multiscale) gradient orientations are not pooled in histograms but directly extracted pixelwise; in other words, there is no local resampling of the orientations. Thus, in some sense, these images are the best ones we could hope using Poisson reconstruction. Comparing these images with samples of MS-Poisson and MaxEnt precisely shows the effect of local resampling of the (multiscale) orientations; observe in particular the man's face and also the folds of his t-shirt. These images thus correspond to much more precise reconstructions, but it is interesting to notice that in certain regions where attention will be focused (near the face e.g.), there are enough keypoints at fine scales in order to get back satisfying pieces of images even after local resampling. Also, one must keep in mind that the loss of the gradient magnitude information is in practice difficult to cope with and may force us to erroneously amplify the noise in the reconstruction. As one can see in the bottom left of Fig. 4, it is obvious if one tries to set the gradient magnitude to 1 in the global Poisson reconstruction.

5.1.2. Quantitative evaluation. As mentioned in [15], there is no reliable criterion to quantitatively evaluate the quality of the result for such reconstruction problems. In our context where only gradient orientations are extracted, it is reasonable to evaluate the reconstruction quality based on the normalized correlation to the input image (which is invariant under affine contrast change). If $u, v : \Omega \to \mathbb{R}$ are two images, the normalized correlation is defined as

$$r(u,v) = \frac{1}{|\Omega|} \sum_{\boldsymbol{x} \in \Omega} \left(\frac{u(\boldsymbol{x}) - \bar{u}}{\sigma_u} \right) \left(\frac{v(\boldsymbol{x}) - \bar{v}}{\sigma_v} \right) \in [-1, 1], \tag{53}$$

where $\bar{u} = \frac{1}{|\Omega|} \sum_{\boldsymbol{x} \in \Omega} u(\boldsymbol{x})$ and $\sigma_u^2 = \frac{1}{|\Omega|} \sum_{\boldsymbol{x} \in \Omega} (u(\boldsymbol{x}) - \bar{u})^2$. On Fig. 4, for each result we have indicated the normalized correlation value r. Surprisingly, the higher correlation values are attained with results linked to the MS-Poisson model (even if it only have access to HOG computed on a blurred gradient). Besides, the value

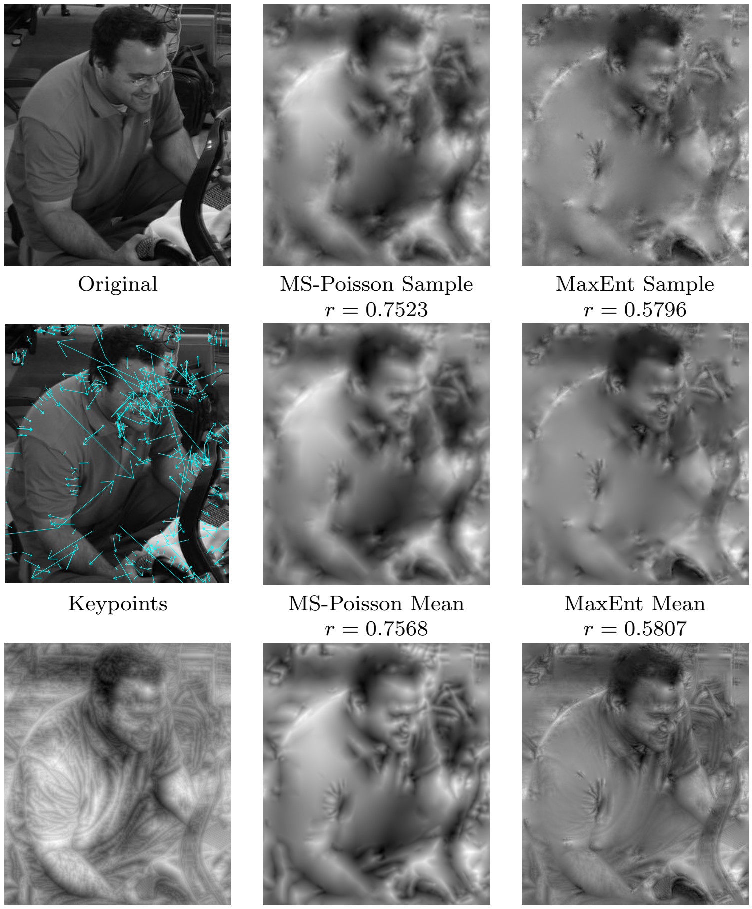

Fig. 4. ***Reconstruction results with MaxEnt and MS-Poisson models.*** *In the first column we display an original image, the corresponding oriented keypoints, and the Poisson reconstruction with true gradient orientations of the whole image and magnitude set to* 1*. In the second column we display a sample of the MS-Poisson model, the expectation of this model, and the multiscale Poisson reconstruction using the true multiscale gradient orientations in the SIFT subcells. In the third column, we display a sample of the MaxEnt model, the expectation of this model, and the Poisson reconstruction using the true gradient orientations in the SIFT subcells. For each result we indicate the value of the normalized correlation r with respect to the original image. See the text for comments on these results. (Images are better seen on the electronic version)*

attained by the samples (or mean) of MS-Poisson is close to the one obtained with the true multiscale HOGs. In contrast, the correlations obtained with the MaxEnt model are lower. This is better explained by the results of Fig. 6 in which we display values of local normalized correlations obtained with both models: for each pixel $\boldsymbol{x}$ we extract patches $p_{\boldsymbol{x}}(u), p_{\boldsymbol{x}}(v)$ of compared images u, v and we compute the normalized correlation $r(p_{\boldsymbol{x}}(u), p_{\boldsymbol{x}}(v))$. On the one hand, MaxEnt result is everywhere much noisier (because gradient orientations are sampled independently). On the other hand, the regularization involved in MS-Poisson helps to propagate good correlations values in regions located near SIFT subcells. Also, this criterion based on normalized

FIG. 5. ***Influence of the regularization parameter μ in MS-Poisson.*** *As expected, increasing μ penalizes more the L^2-norm of the gradient and thus makes the image blurrier. Here again we indicate the value of the normalized correlation r with respect to the original image. We empirically observed that a good compromise between recovered details and smoothness is often attained around $\mu = 50$. (Images are better seen on the electronic version)*

correlation confirms the choice for the regularization parameter $\mu = 50$, see Fig. 5.

Another interesting way of performing quantitative evaluation in our context is to compare the HOG computed in the SIFT subcells to the ones of the original image. For each subcell, we can compute histograms H_u, H_v of gradient orientations (with 8 bins) for the images u, v and then compute the total variation distance between these histograms, defined as $\frac{1}{2}\sum_{\ell=1}^{8} |H_u(\ell) - H_v(\ell)| \in [0, 1]$. Again, we use gradients at scale 0 when considering the MaxEnt model, and scaled gradients when considering the MS-Poisson model. We can then average the HOG distances obtained for all SIFT subcells, weighted by the number of pixels in each subcell. With this methodology, for the image of Fig. 4, we obtain a mean distance around 0.27 for MS-Poisson and 0.16 for MaxEnt. This value is lower for MaxEnt because the model is inherently made to satisfy the HOG constraint. One can better understand these results by examining the orientation fields of both models, as proposed in Fig. 7, in particular the effect of the final Poisson reconstruction (keeping in mind that MS-Poisson can also be written with a single objective vector field given by Eq. (43)). In this figure, one clearly observes that the objective vector field for MS-Poisson is already very smooth (and certainly too smooth to account for fine local variations in orientation). In constrast the objective vector field for MaxEnt better accounts for the fine variations, but is much noisier, even after the Poisson reconstruction step.

5.1.3. Second order statistics. As we have seen in Section 4.3, it is possible to compute the second order statistics of the reconstructed image in each model. In Fig. 8 we display the standard deviations of all pixels values in each model. One first remark is that MaxEnt has in general much larger variance than MS-Poisson which can be explained by the fact that the output of MS-Poisson is in some sense a weighted

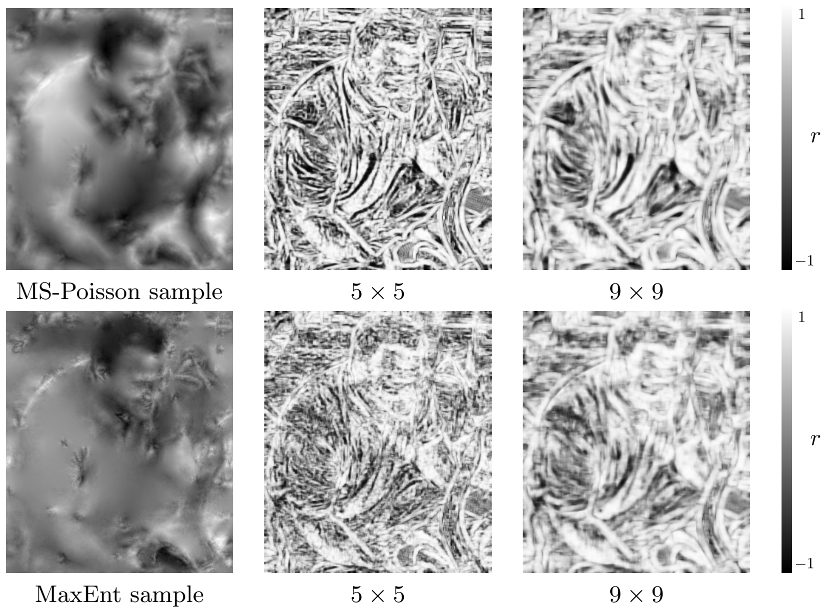

FIG. 6. ***Comparison between MS-Poisson and MaxEnt with local normalized correlations.*** *On the left column we display samples of the models MS-Poisson and MaxEnt. On the other columns, we display the local normalized correlation of the sample (first column) with respect to the original image. The local normalized correlations are computed on patches of size* 5×5 *and* 9×9*, with values in* $[-1, 1]$*. See the text for comments. (Images are better seen on the electronic version)*

average of many local reconstructions. Also it is interesting to see that the image regions with larger variance are located in the SIFT subcells which contain sharp geometric details. That being said, the variance of both these models is relatively small compared to the global range of the mean image, which indicates that both these models have quite small variations around the mean.

5.1.4. Discard boundary keypoints. Let us emphasize that in our experiments, we used all the keypoints computed by the SIFT methods and we did not discard keypoints located near the image boundaries. The positions of the corresponding local extrema in the normalized scale-space are indeed highly dependent on the boundary conditions used to compute the scale-space. This explains why SIFT keypoints near the image boundaries are often discarded for particular applications, e.g. image matching. In our reconstruction problem, there is no reason to discard such keypoints, and we use the information available in SIFT subcells as soon as they intersect the image domain (if the SIFT subcell is not entirely contained in the domain, we consider only the pixels in the intersection of the subcell and the domain). But still, it is clear that for some images, the reconstruction will be quite different when discarding those keypoints. For example in the case of Fig. 9, if boundary keypoints are discarded, then several parts of the man's body are not as properly retrieved in the reconstruction, thus affecting the semantic understanding of the image.

5.1.5. Matching keypoints between the original and reconstructed images. Finally, it is interesting to compare the keypoints computed on the original image and the ones computed on several samples of the models. As one can see on Fig. 10, we get back similar keypoints in many regions of the image, but still with some

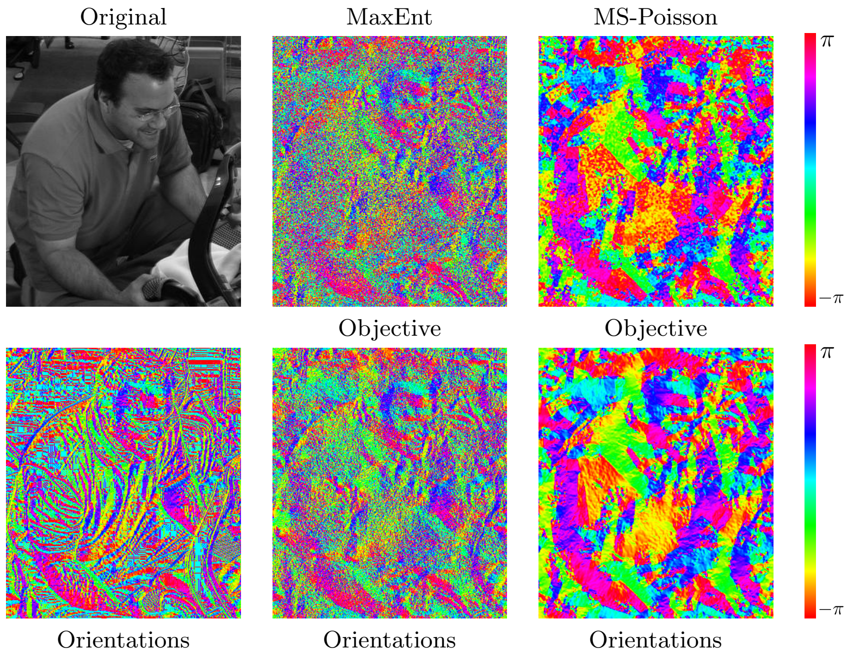

FIG. 7. ***Orientation fields of the original and reconstructions.*** *On the left column we display the original image (top) and the corresponding gradient orientations (bottom). On column 2 (MaxEnt) and column 3 (MS-Poisson), we display the orientation of the objective vector field (before Poisson reconstruction, top) and the orientation of the resulting gradient field (after Poisson reconstruction, bottom). In the regions that are covered by several SIFT subcells, one can see that the local HOGs are quite well preserved (especially for MaxEnt) even if the orientations are locally shuffled. One can also observe that the Poisson reconstruction step smoothes slightly the orientation field. (Images are better seen on the electronic version)*

variations in positions, scales and orientations. In particular, we observe variations when taking different samples of the model (sometimes, some keypoints associated with low contrast regions may even disappear). Notice also that we get back less keypoints in the MS-Poisson model: indeed, since it is more regular we loose some extrema in the scale-space. Besides, the regularization tends to change the scale of the structures, thus the scales of the keypoints is often larger than in the original image.

In order to give a more quantitative evaluation of the variations of the keypoints over different samples of the model, it is possible to use the matching algorithm available with the online implementation [54] (we used the proposed default parameters). This algorithm follows the matching method proposed in [33] which essentially pairs SIFT keypoints by thresholding the ratio between the distances to the first and second nearest neighbors (computed with the ℓ^2-distance between SIFT descriptors). First we can comment on what happens when matching two different samples of the same model. For the MS-Poisson model, when matching the two samples shown in Fig. 10, among the 206 keypoints found on the first image (resp. 211 on the second image), 150 keypoints are matched. The mean spatial distance (resp. mean scale variation, mean angle variation) between matched keypoints is about 0.54 (resp. 0.15, 0.050). Similar numbers can be given for the MaxEnt model, but in this case much less keypoints are correctly matched: over the 452 keypoints found on the first image (resp. 458 on the

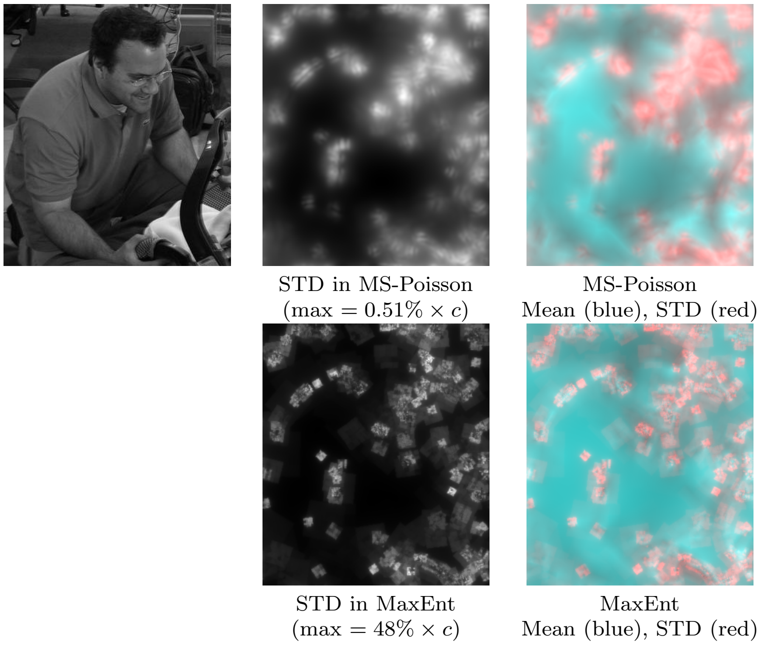

FIG. 8. ***Standard deviations of MS-Poisson and MaxEnt models.*** *On the top left we display the original image. On the rest of the figure we display the images formed with the standard deviations (STD) of the models MS-Poisson (first row) and MaxEnt (second row). On the second column we display the raw STD values. On the third column, the red component corresponds to the raw STD values (same as in the second column) and the blue component corresponds to the mean image* $m = \mathbb{E}(U)$ *of the model (MaxEnt or MS-Poisson). Let us emphasize that for better visualization the images of the second column are renormalized so that the white color corresponds to the indicated maximum value (expressed as a percentage of the empirical standard deviation* $c = \sqrt{|\Omega|^{-1}\sum m(\boldsymbol{x})^2 - (|\Omega|^{-1}\sum m(\boldsymbol{x}))^2}$ *of the mean image* m*). These results clearly indicate that the MS-Poisson model is much more concentrated around its expectation than MaxEnt. (Images are better seen on the electronic version)*

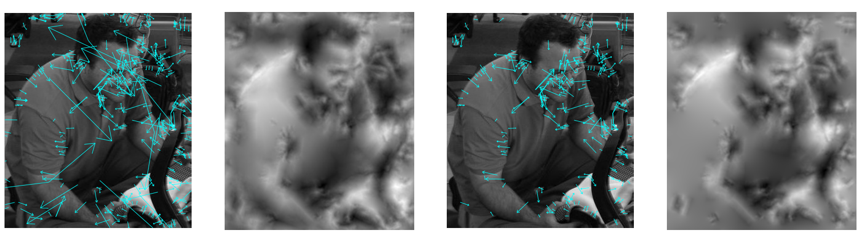

FIG. 9. ***Discard keypoints near image boundary.*** *In this figure, we examine the effect of discarding keypoints whose associated SIFT cell is not entirely contained in the image domain. The displayed reconstructions are samples of the MS-Poisson model.*

second image), only 184 are matched. This reflects again the larger variance of the MaxEnt model.

More interestingly, we can try to match the SIFT keypoints between the original image and the reconstructions. Unfortunately, only a few SIFT points are properly matched this way: among the 477 keypoints found in the original image, around 10 keypoints are properly matched in samples of the MS-Poisson model, and no keypoints are matched when comparing to a sample of MaxEnt. This shows that even if these

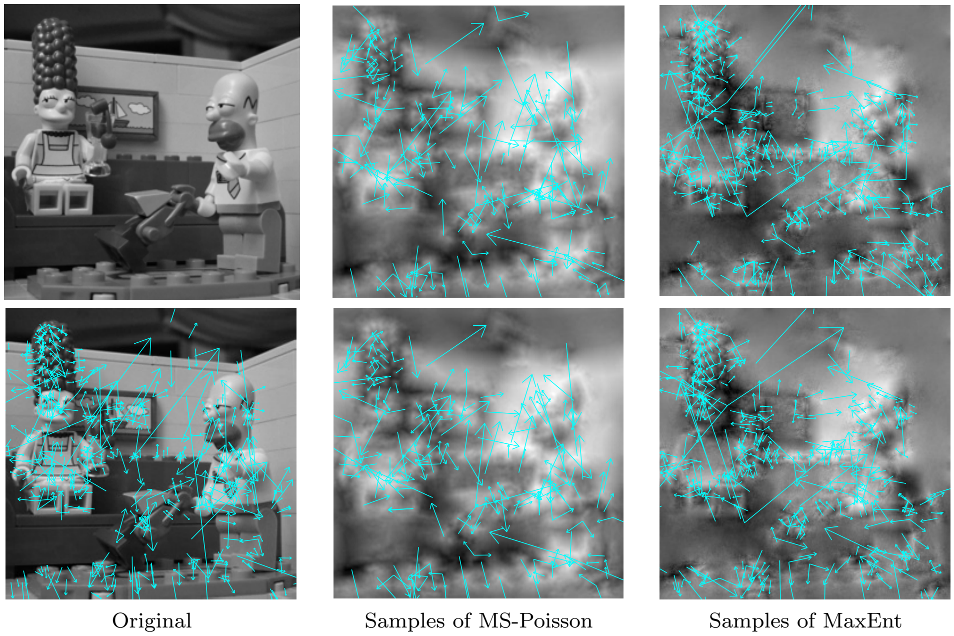

FIG. 10. ***Keypoints after reconstruction.*** *In the first column we display an original image and the same image with its SIFT keypoints. In the second column we display two samples of the MS-Poisson model. In the third column we display two samples of the MaxEnt model. We display the keypoints associated to these images as overimposed blue arrows. Notice that several keypoints are retrieved after reconstruction, with still some variations in positions and orientations. Notice also that we observe some variations in the keypoints associated to different samples of these models. See the text for additional comments. (Images are better seen on the electronic version)*

models are able to recover gradient orientations in a somehow blurry manner, this is not sufficient to precisely get back the content of SIFT descriptors. By the way, the fact that only 75% (resp. 50%) of the keypoints are matched between two samples of MS-Poisson (resp. MaxEnt) illustrates the sensitivity of the SIFT descriptors to small random perturbations.

5.2. Reconstruction from true SIFT descriptors. The two models MS-Poisson and MaxEnt are designed to propose stochastic reconstructions of an image based on simplified SIFT descriptors, that is, multiscale HOGs extracted around the SIFT keypoints. But it is also possible to test these reconstruction models with the true SIFT descriptors. For that, for each keypoint, we still consider the location, scale and principal orientation, but, following the discussion of Section 2.2, starting from the normalized feature vector $(f_k) \in \mathbb{R}^{128}$, we improperly build target histograms for the 16 corresponding SIFT subcells: for each $p \in \{1, \dots, 16\}$, to the corresponding p-th subcell s_j we associate the discrete histogram

$$\widetilde{H}_{j,\ell} = \frac{f_{16(p-1)+\ell}}{\sum_{\ell'=1}^{8} f_{16(p-1)+\ell'}} \quad (1 \leqslant \ell \leqslant 8). \tag{54}$$

We can thus sample the MS-Poisson model using the $(\widetilde{H}_{j,\ell})$ values as a substitute for the extracted multiscale HOG $(H_{j,\ell})$.

On Fig. 11, we display several reconstruction results obtained with the model MS-Poisson based on the multiscale HOGs or the true SIFT descriptors. As could be expected, the reconstruction results obtained with the true SIFT descriptors are not as good as the ones obtained from multiscale HOGs, in particular many fine scale structures are lost, and the shape of small objects is not recovered in a coherent way (see for example the wings in the butterfly image). However, large-scale structures of the image are still retrieved quite properly which often suffices to understand the semantic content of the image.

In order to get sharper results, we should adapt the reconstruction models to account for the normalizations applied in the original SIFT method. It appears quite straightforward to adapt the models to histograms computed with linear votes (instead of binary votes). However, it seems much more difficult to cope with the final normalization and thresholding (see Equation (2)), which dramatically reduce the quantity of information. Also, in the true SIFT descriptors, the pixels vote for orientations values with a weight that is proportional to the gradient magnitude. This explains why it is difficult to retrieve the local HOG from the SIFT descriptors in the absence of any information about the local gradient magnitude.

5.3. Comparison with previous works. In this paragraph, we propose to compare our reconstruction models with the ones obtained by the methods by Weinzaepfel et al. [64] and Dosovitskiy & Brox [19]. One important difference between these two other approaches and ours is that our method relies only on the content provided in the SIFT subcells while these methods exploit an external database either to copy local information from patches with similar SIFT descriptors (as in [64]) or to build an up-convolutional neural network for reconstruction (as in [19]). Thus our work has no intention to outperform these methods in terms of visual quality of reconstruction (in particular, our method has absolutely no possibility of recovering the color information). Notice that we cannot compare to the method of [36] which is adapted to "dense SIFT" (i.e. SIFT descriptors computed on a dense set of patches) and not "sparse SIFT" (i.e. SIFT descriptors computed around the keypoints).

They are also minor differences in the extracted information because both these works do not rely on the original implementation of the SIFT method. The method of [64] actually uses "elliptic" interest regions (extracted using the Hessian-affine method by [42]) in which normalized multiscale HOG are computed (in the same way as in the original SIFT method). In contrast, Dosovitskiy and Brox use circular keypoints and descriptors that are computed with the VLFeat library [61]. But in order to apply an up-convolutional neural network to these features, they need to derive a grid-based representation of these features: the image is divided in 4×4 cells and each cell containing a keypoint is being associated with the corresponding oriented keypoint and feature vector. If there is no keypoint, then they associate the zero vector, and if there are several keypoints they randomly choose one of them (see the details in [19, Section III]).

One advantage of the MS-Poisson model, compared to the result of [64], is that it is defined through the minimization of the global MS-Poisson energy (37). Therefore, it produces images that are globally coherent while respecting as much as possible the local constraints given by the multiscale HOGs. In contrast, the result of [64] is clearly affected by stitching artifacts which are inherent to their reconstruction method. On the other hand, their method is able to copy pieces of clean patches so that their reconstruction looks locally sharper (but also noisier).

However, the reconstructed images obtained in [19] are both globally coherent and quite sharp. Indeed, our method does not rely on an external database so it cannot compete with the one of [19], and in particular it cannot get back information which

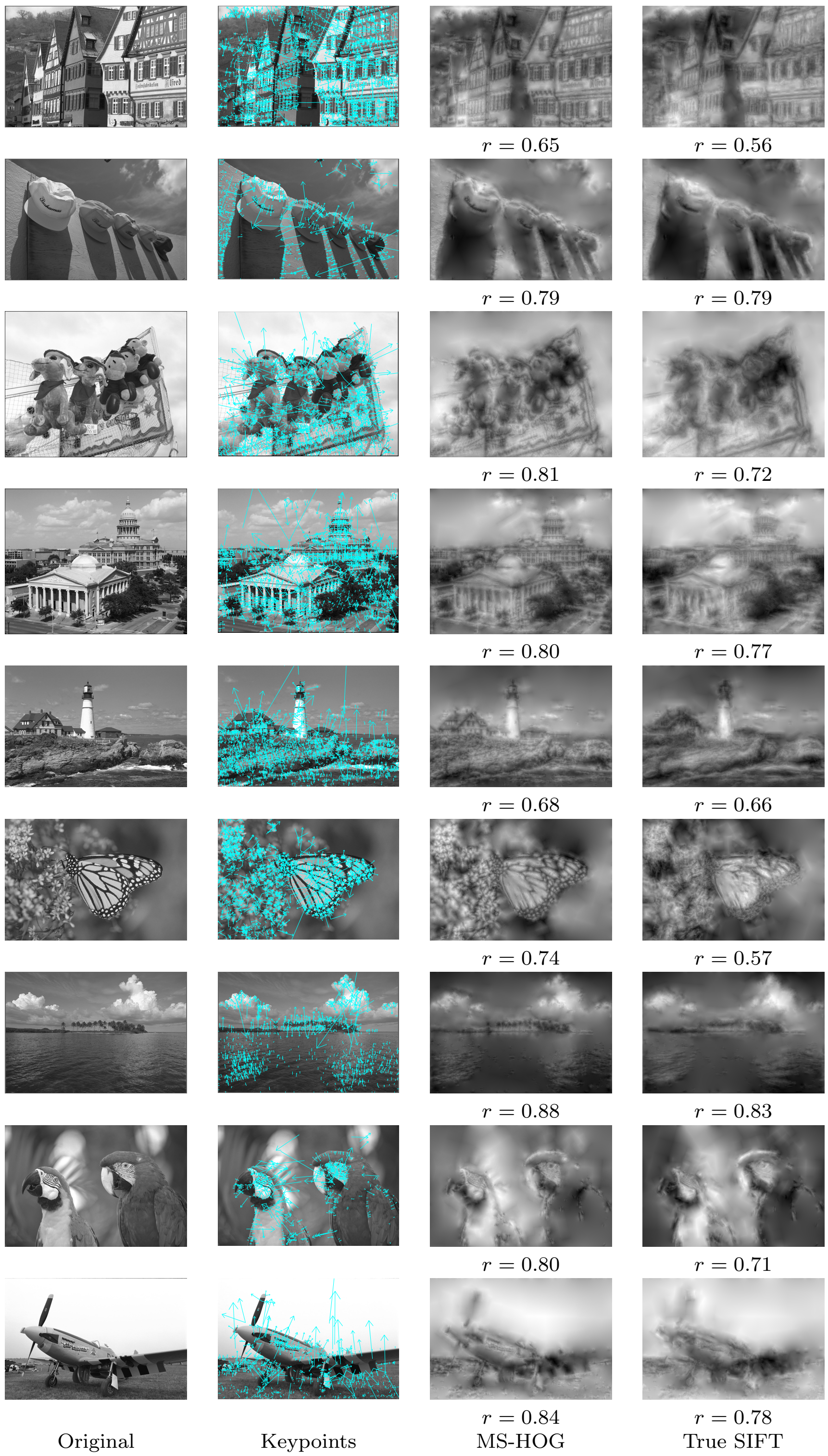

FIG. 11. ***Reconstruction results from multiscale HOG or SIFT descriptors*** *with images of the Live database [57]. For each row, from left to right, we display an original image, the same image with overimposed SIFT keypoints, a sample of the MS-Poisson model obtained from multiscale HOG, and a sample of the MS-Poisson model obtained from the true SIFT descriptors. Notice that the reconstruction from true SIFT descriptors is less sharp but still recovers many geometric structures of the initial image.*

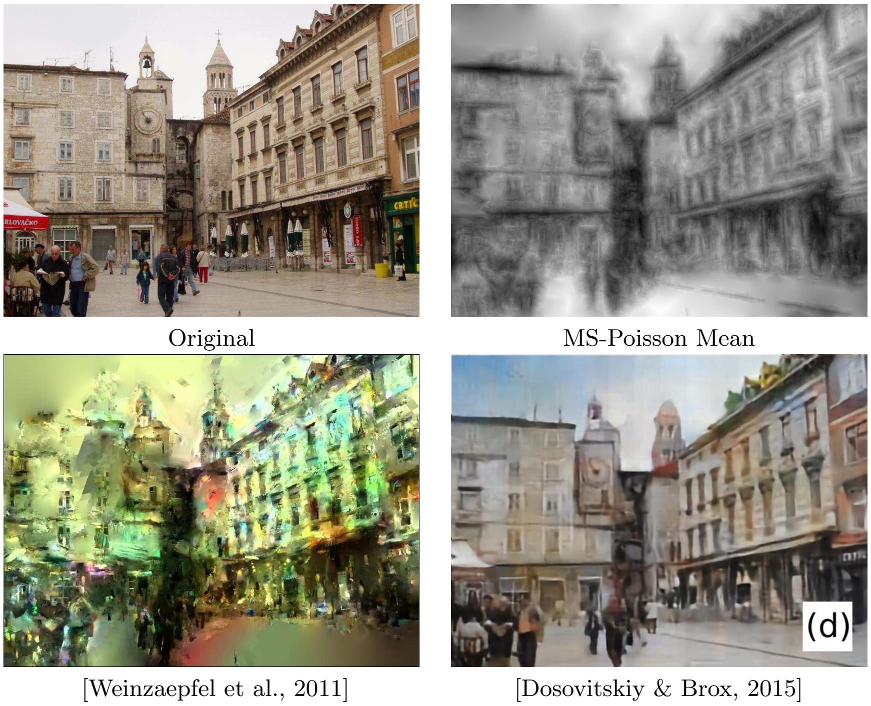

FIG. 12. ***Comparison for SIFT reconstruction.*** *In the first row we display the original image and the reconstruction results obtained as the expectation of the MS-Poisson model computed on the true SIFT descriptors (see Section 5.2). In the second row we display the results obtained with the methods of [64] and [19]. Notice that the MS-Poisson model provides images that are blurrier but also more globally coherent than the ones obtained by the method of [64]. However, this model does not compete with [19] in terms of restitution and visual quality since it does not rely on any external information.*

are completely lost in the SIFT descriptors (global contrast, or also color information).

5.4. Reconstruction with other keypoints. In this paragraph we question the very definition of the SIFT keypoints in terms of synthesis, in a similar way that what was done in [50]. Indeed, one can wonder if selecting the local extrema of $(\boldsymbol{x}, \sigma) \mapsto \sigma^2 \Delta g_\sigma * u(\boldsymbol{x})$ is the best possible choice for points of interest in order to extract relevant information for synthesis.

For that, we propose to compare with two other sets of keypoints extracted in a very different way. The first choice ("Min-Rec-Error") is driven by the following intuition: using Taylor formula around a point $\boldsymbol{x}$, one can write when $\sigma \to 0$ that

$$\int u(\boldsymbol{x}+\boldsymbol{z}) g_\sigma(\boldsymbol{z}) d\boldsymbol{z} - u(\boldsymbol{x}) = \sigma^2 \Delta u(\boldsymbol{x}) + o(\sigma^2). \tag{55}$$

Therefore, nearby the positions $\boldsymbol{x}$ where $\Delta u(\boldsymbol{x})$ is close to zero, one can approximately recover $u(\boldsymbol{x})$ by averaging neighboring values. In this sense, it seems relevant to extract more information at the points where the average reconstruction fails, and in particular at the maxima of $|\Delta u|$.

But one could also directly work with the reconstruction error: we thus propose to extract local maxima of the function

$$(\boldsymbol{x}, \sigma) \longmapsto |g_\sigma * u(\boldsymbol{x}) - u(\boldsymbol{x})|. \tag{56}$$

In our implementation, we detect these maxima on a discretized scale-space with 30 scales $s = 2^{r/6}$, $0 \leqslant r < 30$. Besides, in order to draw a comparison with a fixed number of keypoints, we only keep the points having an "edgeness" value below a threshold. As in the original SIFT method, the edgeness measure is obtained as the ratio $\frac{\mathrm{Tr}(H)^2}{\det H}$ of the principal curvatures, where H is the Hessian of the smoothed image $g_2 * u$. The threshold is adapted in order to get the same number n_{kp} of keypoints than the ones provided by the SIFT method.

The second and third choices ("Random-unif" and "Random-grad") consists in selecting keypoints in a random manner. More precisely, for the choice "Random-unif", we independently sample n_{kp} keypoints by choosing uniformly a position $\boldsymbol{x}$ in the image domain, a uniform orientation $\alpha \in \mathbb{T}$, and a scale by sampling an exponential distribution whose parameter is adjusted so that the expectation is the same as the mean scale of the usual SIFT keypoints. Modelling by the exponential distribution is empirically justified by the fact that the distribution of scales of SIFT keypoints is concentrated in the fine scales. For the choice "Random-grad", we do the same except that the positions are randomly drawn using a probability distribution which is proportional to the gradient magnitude of the smoothed image $g_2 * u$.

For these new sets of keypoints, we computed the average image of the MS-Poisson model. The results are displayed on Fig. 13. They clearly indicate that the usual SIFT keypoints lead to a reconstruction that is visually better than the others. The main problem of the "Min-Rec-Error" keypoints is that they do not extract enough small scale information: for the examples shown in Fig. 13 the average scale of these keypoints is approximately twice larger than the one of the SIFT keypoints. Besides, for both "Min-Rec-Error" and random keypoints, the spatial locations are not concentrated around geometric details as can be the case with the SIFT keypoints. The comparison with "Random-grad" is particularly interesting: indeed the reconstruction with "Random-grad" keypoints is slightly better than the one with "Random-unif" keypoints, but still it fails to recover fine details. The main problem of the "Random-grad" approach is that it is not contrast invariant and thus it favors points with strong gradients in uniform regions over points in salient regions with low contrast. Thus, the usual definition of SIFT keypoints (and in particular the thresholding steps) is confirmed to be a relevant choice for extracting visual information near salient structures, both from the analysis or the synthesis perspective.

6. Conclusion.

In this paper we proposed two stochastic models (MaxEnt, respectively MS-Poisson) for reconstructing an image based only on the information contained in the (monoscale, respectively multiscale) local HOGs computed in the SIFT subcells. With both models we get back images which are close to the original in terms of semantic content. This is still true if we compute the reconstructions based on the true SIFT descriptors. One benefit of these models over competing approaches is that they do not rely on any external image database, and besides the convolutive expressions found in this paper allow to compute statistics of the corresponding output random fields (e.g. local variance).

However, several questions raised by this work remain open. First it would be interesting to consider generalizations of the MS-Poisson model with different image priors, i.e. adopt other regularization terms in the functional. It is likely that solving the corresponding optimization problem may require an iterative procedure, but on the other hand the solutions may exhibit cleaner geometric structures which are better extrapolated outside the SIFT subcells. Also, there is more to discuss about the optimality of keypoints with respect to the quality of reconstructed images. In

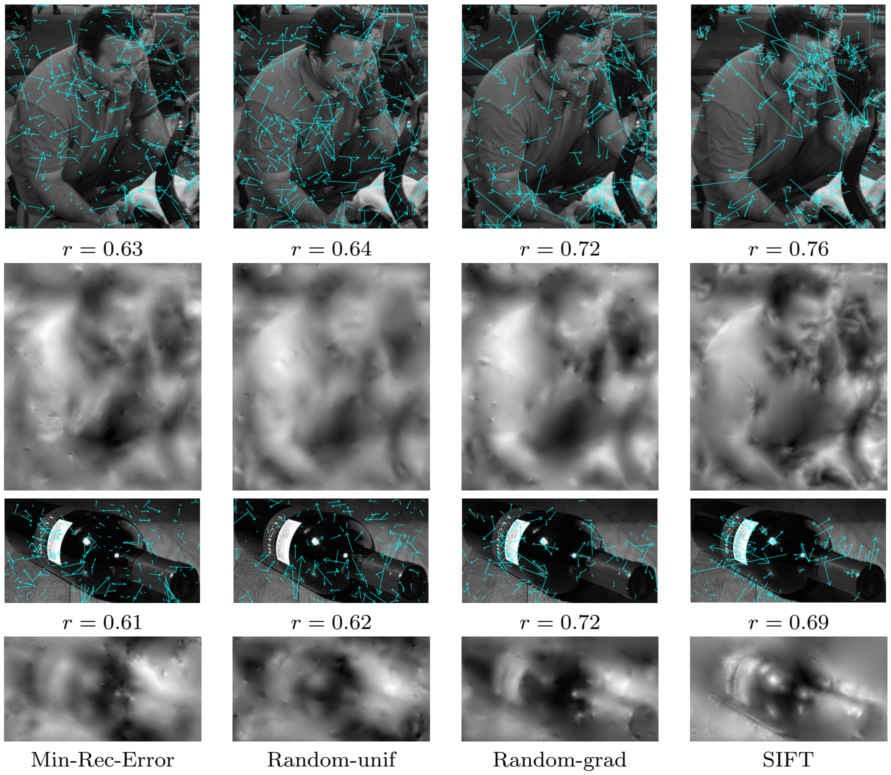

FIG. 13. ***Reconstruction with other keypoints.*** *The first column ("Minimum reconstruction error") corresponds to the keypoints obtained as local minima of* (56). *The second ("Random-unif") and third column ("Random-grad") corresponds to the randomly selected keypoints. The last column corresponds to the standard SIFT keypoints. The original images are displayed on Fig. 3 and Fig. 4. Above each reconstruction we indicate the value of the normalized correlation toe the original image. See the text in Section 5.4 for the precise definition of these sets of keypoints, and additional comments.*

particular, here we adopted one unique reconstruction strategy in order to compare different sets of keypoints. But it seems possible to optimize both the sets of keypoints and the reconstruction strategy in order to maximize a criterion linked to the proximity of the reconstruction to the input original image. This could be thought of as a kind of auto-encoding procedure in which the encoder is constrained to have a very particular form (that is, keypoint extractor).

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikainen, *Face description with local binary patterns: Application to face recognition*, IEEE transactions on pattern analysis and machine intelligence, 28 (2006), pp. 2037–2041.

[2] A. Alahi, R. Ortiz, and P. Vandergheynst, *Freak: Fast retina keypoint*, in IEEE Conference on Computer vision and pattern recognition (CVPR), IEEE, 2012, pp. 510–517.

[3] B. Allen and M. Kon, *The Marr Conjecture and Uniqueness of Wavelet Transforms*, arXiv preprint arXiv:1401.0542, (2015).

[4] B. Allen and M. Kon, *Unique recovery from edge information*, in Sampling Theory and Applications (SampTA), 2015 International Conference on, IEEE, 2015, pp. 312–316.

[5] F. Attneave, *Some informational aspects of visual perception.*, Psychological review, 61 (1954), p. 183.

[6] S. Battiato, G. Gallo, G. Puglisi, and S. Scellato, *SIFT features tracking for video stabilization*, in Image Analysis and Processing, 2007. ICIAP 2007. 14th International Conference on, IEEE, 2007, pp. 825–830.

[7] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, *Speeded-up robust features (SURF)*, Computer vision and image understanding, 110 (2008), pp. 346–359.

[8] M. Black and A. Jepson, *Eigentracking: Robust matching and tracking of articulated objects using a view-based representation*, International Journal of Computer Vision, 26 (1998), pp. 63–84.

[9] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, *Learning mid-level features for recognition*, in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2010, pp. 2559–2566.

[10] Y.-L. Boureau, N. Le Roux, F. Bach, J. Ponce, and Y. LeCun, *Ask the locals: multi-way local pooling for image recognition*, in Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 2651–2658.

[11] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, *Visual categorization with bags of keypoints*, in Workshop on statistical learning in computer vision, ECCV, 2004.

[12] S. Curtis and A. Oppenheim, *Reconstruction of multidimensional signals from zero crossings*, J. Opt. Soc. Am. A, 4 (1987), pp. 221–231, doi:10.1364/JOSAA.4.000221.

[13] S. Curtis, S. Shitz, and A. Oppenheim, *Reconstruction of nonperiodic two-dimensional signals from zero crossings*, IEEE Transactions on Acoustics, Speech, and Signal Processing, 35 (1987), pp. 890–893.

[14] N. Dalal and B. Triggs, *Histograms of oriented gradients for human detection*, in Proceedings of the IEEE CVPR, vol. 1, 2005, pp. 886–893.

[15] E. d'Angelo, L. Jacques, A. Alahi, and P. Vandergheynst, *From bits to images: Inversion of local binary descriptors*, IEEE Transactions on PAMI, 36 (2014), pp. 874–887.

[16] A. Desolneux, *When the a contrario approach becomes generative*, International Journal of Computer Vision, 116 (2016), pp. 46–65.

[17] A. Desolneux and A. Leclaire, *Stochastic image reconstruction from local histograms of gradient orientation*, in Proceedings of the sixth International Conference on Scale Space and Variational Methods in Computer Vision (SSVM), Springer, Lecture Notes in Computer Science, 2017, pp. 133–145.

[18] A. Desolneux, L. Moisan, and J. Morel, *From Gestalt theory to image analysis: a probabilistic approach*, vol. 34, Springer Science & Business Media, 2007.

[19] A. Dosovitskiy and T. Brox, *Inverting Visual Representations with Convolutional Networks*, arXiv:1506.02753 [cs], (2015).

[20] J. H. Elder and S. W. Zucker, *Scale space localization, blur, and contour-based image coding*, in Computer Vision and Pattern Recognition, 1996. Proceedings CVPR'96, 1996 IEEE Computer Society Conference on, IEEE, 1996, pp. 27–34.

[21] O. Faugeras, *Three-dimensional computer vision: a geometric viewpoint*, MIT press, 1993.

[22] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, *Object detection with discriminatively trained part-based models*, IEEE Transactions on PAMI, 32 (2010), pp. 1627–1645.

[23] C. Harris and M. Stephens, *A combined corner and edge detector.*, in Alvey vision conference, vol. 15, Citeseer, 1988, p. 50.

[24] R. Hummel and R. Moniot, *Reconstructions from zero crossings in scale space*, IEEE Transactions on Acoustics, Speech, and Signal Processing, 37 (1989), pp. 2111–2130.

[25] F. Juefei-Xu and M. Savvides, *Learning to invert local binary patterns.*, in BMVC, 2016.

[26] S. K. and Z. A., *Very deep convolutional networks for large-scale image recognition*, in Proceedings of the International Conference on Learning Representations, 2014.

[27] H. Kato and T. Harada, *Image reconstruction from bag-of-visual-words*, in Proceedings of the IEEE CVPR, 2014, pp. 955–962.

[28] A. Krizhevsky, I. Sutskever, and G. Hinton, *Imagenet classification with deep convolutional neural networks*, in Advances in neural information processing systems, 2012, pp. 1097–1105.

[29] S. Lazebnik, C. Schmid, and J. Ponce, *Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories*, in IEEE computer society conference on Computer vision and pattern recognition, vol. 2, IEEE, 2006, pp. 2169–2178.

[30] S. Leutenegger, M. Chli, and R. Y. Siegwart, *Brisk: Binary robust invariant scalable keypoints*, in IEEE International Conference on Computer Vision (ICCV), IEEE, 2011, pp. 2548–2555.

[31] T. Lindeberg, *Feature detection with automatic scale selection*, International journal of computer vision, 30 (1998), pp. 79–116.

[32] T. Lindeberg, *Image matching using generalized scale-space interest points*, Journal of Mathematical Imaging and Vision, 52 (2015), pp. 3–36.
[33] D. Lowe, *Distinctive image features from scale-invariant keypoints*, International Journal of Computer Vision, 60 (2004), pp. 91–110.
[34] Y. Lu, S. Zhu, and Y. N. Wu, *Learning FRAME models using CNN filters for knowledge visualization*, CoRR, abs/1509.08379 (2015), http://arxiv.org/abs/1509.08379, arXiv:1509.08379.
[35] A. Mahendran and A. Vedaldi, *Understanding deep image representations by inverting them*, in IEEE CVPR, 2015, pp. 5188–5196.
[36] A. Mahendran and A. Vedaldi, *Visualizing deep convolutional neural networks using natural pre-images*, International Journal of Computer Vision, 120 (2016), pp. 233–255.
[37] E. Mair, G. D. Hager, D. Burschka, M. Suppa, and G. Hirzinger, *Adaptive and generic corner detection based on the accelerated segment test*, in European conference on Computer vision, Springer, 2010, pp. 183–196.
[38] S. Mallat and S. Zhong, *Characterization of signals from multiscale edges*, IEEE Transactions on PAMI, 14 (1992), pp. 710–732.
[39] D. Marr, *Vision: A computational investigation into the human representation and processing of visual information*, W.H. Freeman and Company, 1982.
[40] D. Marr and E. Hildreth, *Theory of edge detection*, Proceedings of the Royal Society of London B: Biological Sciences, 207 (1980), pp. 187–217.
[41] Y. Meyer, *Wavelets-algorithms and applications*, vol. 1, Society for Industrial and Applied Mathematics Translation, 1993.
[42] K. Mikolajczyk and C. Schmid, *Scale & affine invariant interest point detectors*, International journal of computer vision, 60 (2004), pp. 63–86.
[43] K. Mikolajczyk and C. Schmid, *A performance evaluation of local descriptors*, IEEE Transactions on PAMI, 27 (2005), pp. 1615–1630.
[44] J.-M. Morel, A. Petro, and C. Sbert, *Fourier implementation of Poisson image editing*, Pattern Recognition Letters, 33 (2012), pp. 342–348.
[45] J.-M. Morel and G. Yu, *ASIFT: A new framework for fully affine invariant image comparison*, SIAM Journal on Imaging Sciences, 2 (2009), pp. 438–469.
[46] J.-M. Morel and G. Yu, *Is SIFT scale invariant?*, Inverse Problems and Imaging, 5 (2011), pp. 115–136.
[47] D. Mumford and A. Desolneux, *Pattern Theory: The Stochastic Analysis of Real-World Signals*, A K Peters/CRC Press, Natick, Mass, 2010.
[48] P. Musé, F. Sur, F. Cao, Y. Gousseau, and J.-M. Morel, *An a contrario decision method for shape element recognition*, International Journal of Computer Vision, 69 (2006), pp. 295–315.
[49] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer, 2004.
[50] M. Nielsen and M. Lillholm, *What do features tell about images?*, in Scale-Space, vol. 1, Springer, 2001, pp. 39–50.
[51] T. Ojala, M. Pietikäinen, and T. Mäenpää, *Multiresolution gray-scale and rotation invariant texture classification with local binary patterns*, IEEE Transactions on PAMI, 24 (2002), pp. 971–987.
[52] P. Pérez, M. Gangnet, and A. Blake, *Poisson Image Editing*, in ACM SIGGRAPH 2003 Papers, SIGGRAPH '03, 2003, pp. 313–318, doi:10.1145/1201775.882269.
[53] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, *Object retrieval with large vocabularies and fast spatial matching*, in IEEE Conference on Computer Vision and Pattern Recognition, 2007., IEEE, 2007, pp. 1–8.
[54] I. Rey Otero and M. Delbracio, *Anatomy of the SIFT Method*, Image Processing On Line, 4 (2014), pp. 370–396, doi:10.5201/ipol.2014.82.
[55] E. Rosten and T. Drummond, *Machine learning for high-speed corner detection*, in Computer Vision–ECCV 2006, Springer, 2006, pp. 430–443.
[56] J. Sanz and T. Huang, *Theorems and experiments on image reconstruction from zero crossings*, 1987. IBM Almaden Research Center.
[57] H. Sheikh, Z. Wang, L. Cormack, and A. Bovik, *Live image quality assessment database release 2 (2005)*, 2005.
[58] J. Sivic and A. Zisserman, *Video Google: A text retrieval approach to object matching in videos*, in Proceedings of the IEEE ICCV, 2003, pp. 1470–1477.
[59] T. Tuytelaars and K. Mikolajczyk, *Local invariant feature detectors: a survey*, Foundations and trends in computer graphics and vision, 3 (2008), pp. 177–280.
[60] T. Tuytelaars and L. Van Gool, *Matching widely separated views based on affine invariant*

regions, International journal of computer vision, 59 (2004), pp. 61–85.

[61] A. Vedaldi and B. Fulkerson, *Vlfeat: An open and portable library of computer vision algorithms*, in Proceedings of the 18th ACM international conference on Multimedia, ACM, 2010, pp. 1469–1472.

[62] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba, *Hoggles: Visualizing object detection features*, in Proceedings of the IEEE ICCV, 2013, pp. 1–8.

[63] C. Wallraven, B. Caputo, and A. Graf, *Recognition with local features: the kernel recipe*, in Proceedings of the IEEE ICCV, 2003, pp. 257–264.

[64] P. Weinzaepfel, H. Jégou, and P. Pérez, *Reconstructing an image from its local descriptors*, in Proceedings of the IEEE CVPR, 2011, pp. 337–344.

[65] J. Yang, D. Schonfeld, and M. Mohamed, *Robust video stabilization based on particle filter tracking of projected camera motion*, IEEE Transactions on Circuits and Systems for Video Technology, 19 (2009), pp. 945–954.

[66] A. Yilmaz, O. Javed, and M. Shah, *Object tracking: A survey*, ACM computing surveys (CSUR), 38 (2006), p. 13.

[67] M. Zeiler and R. Fergus, *Visualizing and understanding convolutional networks*, in Proceedings of ECCV, Springer, 2014, pp. 818–833.

[68] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, *Local features and kernels for classification of texture and object categories: A comprehensive study*, International journal of computer vision, 73 (2007), pp. 213–238.

[69] S. Zhu, Y. Wu, and D. Mumford, *Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling*, International Journal of Computer Vision, 27 (1998), pp. 107–126.