Alpha Matting with KL-Divergence Based Sparse Sampling

Levent Karacan, Aykut Erdem, and Erkut Erdem

***Abstract*—In this paper, we present a new sampling-based alpha matting approach for the accurate estimation of foreground and background layers of an image. Previous sampling-based methods typically rely on certain heuristics in collecting representative samples from known regions, and thus their performance deteriorates if the underlying assumptions are not satisfied. To alleviate this, we take an entirely new approach and formulate sampling as a sparse subset selection problem where we propose to pick a small set of candidate samples that best explains the unknown pixels. Moreover, we describe a new dissimilarity measure for comparing two samples which is based on KL-divergence between the distributions of features extracted in the vicinity of the samples. The proposed framework is general and could be easily extended to video matting by additionally taking temporal information into account in the sampling process. Evaluation on standard benchmark datasets for image and video matting demonstrates that our approach provides more accurate results compared to the state-of-the-art methods.**

***Index Terms*—Image Matting, Video Matting, KL-Divergence.**

I. Introduction

ACCURATE estimation of foreground and background layers of an image or video frames plays an important role for many image and video editing applications. In the computer vision literature, this problem is known as alpha matting, and mathematically, it refers to the problem of decomposing a given image or video frame I into two layers, the foreground F and the background B, defined in accordance with the following linear image composition equation.

$$I = \alpha_p F_p + (1-\alpha_p) B_p \tag{1}$$

where α_p represents the unknown alpha matte which defines the true opacity of each pixel p and whose values lies in $[0,1]$ with $\alpha_p = 1$ denoting a foreground pixel and $\alpha_p = 0$ indicating a background pixel. This is a highly ill-posed problem since for each pixel we have only three inputs but seven unknowns (α and the RGB values of F_p and B_p). The general approach to resolve this issue for image matting is to consider a kind of prior knowledge about the foreground and background in form of user scribbles or a trimap to simplify the problem and use the spatial and photometric relations between these known pixels and the unknown ones. As for the video matting, estimating the alpha mattes of each frame is a more challenging task than single image matting since it requires both temporally coherent and spatially accurate maps.

Image matting methods can be mainly categorized into two groups: propagation-based methods [5]–[11] and sampling-based methods [2]–[4], [12]–[16]. The first group defines an affinity matrix representing the similarity between pixels and propagate the alpha values of known pixels to the unknown ones. These approaches mostly differ from each other in their propagation strategies or affinity definitions. The latter group, on the other hand, collects color samples from known foreground and background regions to represent the corresponding color distributions and determine the alpha value of an unknown pixel according to its closeness to these distributions. Early examples of sampling-based matting methods [12], [13] fit parametric models to color distributions of foreground and background regions. Difficulties arise, however, when an image contains highly textured areas. Thus, virtually all recent sampling-based approaches [2]–[4], [14]–[16] consider a non-parametric setting and employ a particular selection criteria to collect a subset of known F and B samples. Then, for each unknown pixel p, they search for the best (F_p, B_p) pair within the representative samples, and once the best pair is found, the final alpha matte is computed as

$$\hat{\alpha}_p = \frac{(I_p - B_p).(F_p - B_p)}{\|F_p - B_p\|^2} \,. \tag{2}$$

The recent sampling-based approaches mentioned above also apply local smoothing as a post-processing step to further improve the quality of the estimated alpha matte. Apart from the two main types of approaches, there are also some hybrid methods which consider a combination of propagation and sampling based formulations [17], or some supervised machine learning based methods which learn proper matting functions from a training set of examples [18]. Very recently, the authors of [19] proposed a deep learning based solution as well.

For video matting, several researchers extend the existing image matting methods so that they can extract temporally coherent alpha mattes by using either user-generated or predefined trimaps along the video frames. Some of these approaches [20]–[22] automatically generate trimaps by using user interaction to segment foreground object and morphological dilation operation and then apply image matting methods to compute alpha matte. The methods [20], [23], [24] which do not directly use temporal information suffer from the temporal inconsistency, on the other hand, the ones [21], [25]–[30] utilize the temporal information present more temporally coherent alpha matte results. These methods differ from each other in terms of how they incorporate temporal information to compute alpha matte along the video sequences. For a more comprehensive up-to-date survey of image and video matting methods, we refer the reader to [31], [32].

The proposed matting approach belongs to the group of sampling-based methods which will be reviewed in the next

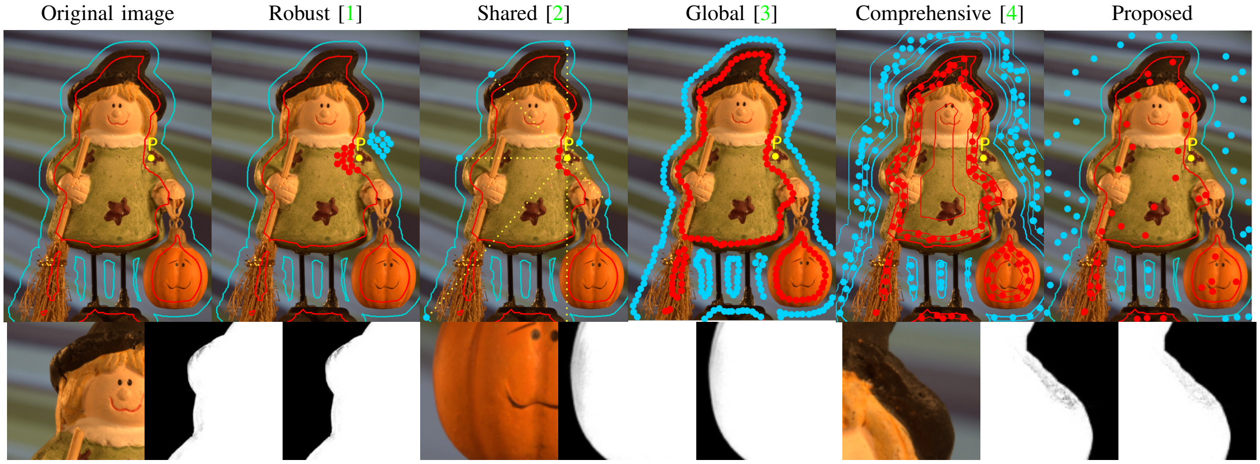

Fig. 1. Non-parametric sampling-based matting approaches. Top row: An input image and the representative samples gathered by the Robust [1], Shared [2], Global [3], Comprehensive [4], and the proposed Sparse Sampling based matting methods. The unknown pixel, the foreground and background samples are shown in yellow, red and blue colors, respectively. Bottom row: Comparison of the estimated alpha mattes by the suggested approach and the state-of-the-art Comprehensive Sampling matting method [4].

subsection. Relying on a non-parametric formulation, these methods typically exploit different strategies to gather the representative foreground and background samples. Our observation is that all these strategies lack a strong theoretical basis, i.e. they require certain assumptions to hold to capture the true foreground and background colors, and moreover, they fail to adequately utilize the relationship between known and unknown regions. In contrast, our approach offers a more principled way to sampling by casting it as a sparse subset selection problem [33], [34], in which the resulting samples refers to a small subset of known foreground and background pixels that best explains the unknown pixels. In particular, sampling is formulated as a row-sparsity regularized trace minimization problem which solely depends on pairwise dissimilarities between known and unknown pixels, and for that, we propose a new KL-divergence based contextual measure as an efficient alternative to chromatic and spatial distances. Besides we extend this sampling approach to video matting by incorporating temporal information together with temporal matting Laplacian to provide temporal coherency. Finally we demonstrate proposed sampling strategy is quite feasible for sparse user input as scribble.

A. Previous work on sampling-based image matting

Sampling-based models differ from each other in (i) how it collects the representative foreground and background samples, and (ii) how it selects the best (F, B) pair for an unknown pixel. Mishima's Blue-screen matting method [35] captures the image of a foreground object in front of a monochrome background. This setup allows efficient estimation of foreground and background distributions via clustering, and then alpha values of unknown pixels are estimated by considering their proximity to the extracted clusters. Another early work, the Knockout system [36], estimates true color values of the foreground and background layers of an unknown pixel by a weighted sum of nearby known pixels with the weights proportional to their spatial distances to the unknown pixel.

Robust matting [1], for an unknown pixel, collects samples from the known nearby foreground and background pixels. Among those samples, it then selects the pair that best fits the linear composting equation defined in Eq. (1). As the selection is carried by considering the color distortion, it provides more robust results than the Knockout system. However, since sampling depends only on the spatial closeness to the unknown pixels, as shown in Fig. 1, the true samples might be missing in the candidate set, decreasing the matting quality. In [37], it has been shown that using geodesic distances improves the results of this model to a certain extent. Shared matting [2] gathers representative samples from the trimap boundary, assuming that, for an unknown pixel, its true foreground and background color can be found at the closest known region boundaries. These pixels are defined as the boundary pixels that lie along the rays which are originated from the unknown pixel and that partition the image plane into disjoint parts of equal planar angles. Then, the best pair among those are used to estimate its alpha value w.r.t. an objective function that depends on spatial and photometric affinity. It falls short, however, when the rays do not reach the true samples. Weighted color and texture (WCT) sampling [14] and its comprehensive version (CWCT) extend Shared matting by combining the local sampling strategy in [2] with a global one that depends on a clustering-based probabilistic model. Moreover, it uses a texture compatibility measure in addition to the color distortion measure to prevent selecting overlapping samples.

Global sampling [3] also collects samples from the trimap boundaries but to avoid the problem of missing true samples, instead of emanating rays from unknown pixels, as in [2], it considers all known boundary samples as a global candidate set. To handle the large number of samples, it employs a simple objective function and an efficient random search algorithm in finding the best sample pair. However, as shown in Fig. 1, the true colors might still be missed in the resulting sample set if they do not lie along the trimap boundaries.

Comprehensive sampling matting [4] follows a global strategy and divides the known and unknown regions into a number of segments so that the segment over which the samples are gathered is decided according to the distance of a given unknown pixel to the extracted foreground and background segments. Sample colors are constructed as the means of the color clusters that are obtained via a two-level hierarchical clustering modeled by a parametric Gaussian mixture model. This approach gives better results than the previous non-parametric sampling based approaches. However, there is still a possibility of missing true samples since the sampling strategy depends on spatial closeness. As demonstrated in Fig. 1, the true color samples might be very far away from the unknown pixel.

Sparse coded matting [16] formulates image matting as a sparse coding problem. It computes alpha values from a bunch of sample pairs within a sparse coding framework instead of finding only the best but single pair of foreground and background (F, B) pair. These samples forming the dictionary atoms are collected from the mean color of the superpixels that lie along the boundaries of the trimaps. Thus, it might also suffer from the missing true samples problem. This problem is solved in its extended version [38] by including samples from whole image region. To prevent overlapping color distributions of foreground and background, it adaptively controls the dictionary size according to a confidence value that depends on probabilistic segmentation. A similar sparse coding approach is used in [39] while selecting samples via a two-level hierarchical k-means clustering process.

B. Previous work on video matting

Widely used blue screen matting [40] provides effective video mattes as extracting the foreground objects from a solid color background is easy, but it requires special studio environment. For natural backgrounds, classical video matting approaches [20], [21], [23], [24], [26], [28] first segment the foreground object from the background and construct a trimap, which will accordingly be propagated along the video frames and used as inputs to single image matting methods. The existing models in general differ from each other in terms of either their segmentation and trimap construction strategies or the latter considered matting schemes.

In particular, Chuang et al. [23] propose a video matting method which builds upon Bayesian matting [12] in extracting the foreground layer. Li et al. [20] generate a binary mask via a Graph-Cut based segmentation algorithm, which will be used as input for Coherent matting [41]. Wang et al. [21] employ a Mean-Shift segmentation approach to segment the foreground objects in video sequences. Video SnapCut [22] proposes a new interactive video object extraction system using localized classifiers for local image features such as color, edge and learned shape prior. To impose temporal coherency, it considers the alpha matte computed from previous frame as a prior for the current frame. Tang et al. [26], [28] compute a probability map or an opacity map and again construct a Graph-Cut formulation to segment the video frames into foreground and background layers before applying a 3D Closed-form matting extended from [7]. Bai et al. [24] ask the user to refine the automatically extracted trimaps on some keyframes and then by using optical flow information between frames they propagate these trimaps to all video frames which are later used as inputs to Robust matting [1].

Video matting models which have been recently proposed are particularly focused on the matting part of the pipeline and interested in extracting temporally more coherent alpha mattes. An important direction here is to extend the matting Laplacian [7] by temporal information. For example, [26]–[30], [42] all employ a matting Laplacian extended to 3D by additionally considering the temporal domain. This modification provides extra local smoothness over extracted alpha mattes. Specifically, Choi et al. [27] use Non-local matting [8] approach to define a 3D nonlocal matting Laplacian on a 3D nonlocal neighborhood between video frames to propagate alpha matte values along the video sequences. Li et al. [29] incorporate motion information to KNN Laplacian [9] by using two-frame affinity matrix and propose a closed-form solution. Shahrian et al. [30] propose an improvement over the Comprehensive sampling scheme [4] in which the sampling phase is expanded by considering previous frame samples and by using local texture features to provide temporal and spatial consistency. Finally, they apply a temporal refinement via 3D matting Laplacian and the alpha matte priors computed from the previous frames. Zou et al. consider the non-local principles in [27], [29] and formulate a sparse dictionary learning problem to represent the known foreground and background colors provided from user input. Similarly, a refinement procedure is applied as a final step by employing a two-frame matting Laplacian.

C. Our contributions

As described, all the existing sampling-based image matting methods rely upon different assumptions regarding the selection policy of background and foreground samples. The justification of these assumptions are mostly valid. But still, they are heuristic methods and they all lack a theoretical ground to explain the relationship between known and unknown pixels in all possible situations. As a step towards improving those methods, in this paper we present a new approach for alpha matting. As shown in Fig. 1, the proposed method allows a more effective sampling, and thus provides considerably better alpha mattes especially on the object boundaries. Furthermore, we also show that our sampling scheme can be easily extended to video matting by considering optical flow and temporal refinements schemes, resulting in temporally consistent and spatially accurate alpha mattes.

An earlier version of this work appeared in [44]. Compared to the conference version [44], this paper has the following improvements. First, we present a substantial number of additional experiments on challenging scenarios with sparse user inputs. These experiments demonstrate that the proposed sampling scheme can also effectively deal with these difficult cases as compared to the existing matting models. Second, we extend the proposed approach to video matting by modifying our sampling process in an effectively simple way to additionally take temporal information into account. The proposed temporal sampling method theoretically enables to handle all

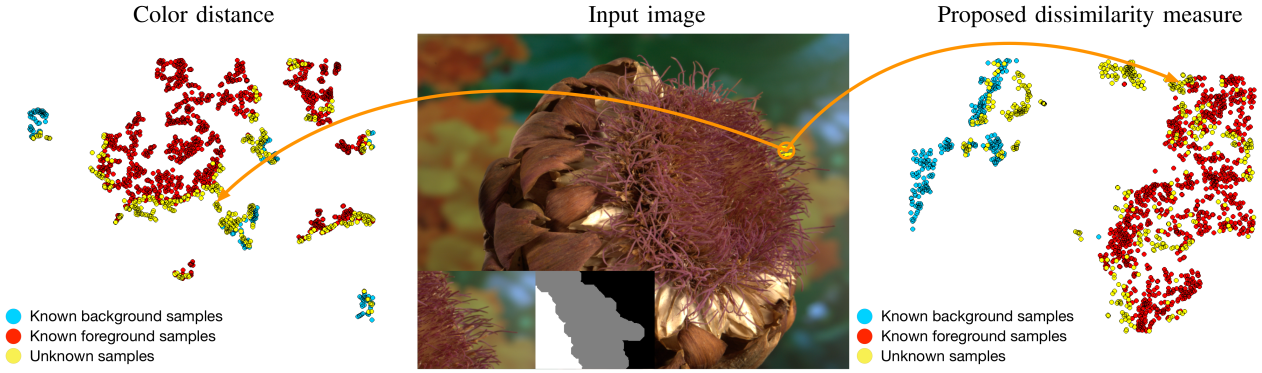

Fig. 2. Distance embedding visualizations using t-SNE method [43] clearly demonstrate that the proposed KL-divergence based dissimilarity measure provides a better discrimination between known foreground and background pixels than using the standard color distance.

video frames on the sampling phase although we present results only for two consecutive frames due to computational complexity. We also analyze the effects of motion information and temporal sampling to alpha matte results. Finally, we add an extra evaluation section about our video matting extension where we extensively test the performance of the approach on a recently introduced benchmark dataset [45]. Our results are highly competitive to the current state-of-the-art, and our method is the best performing approach among the sampling based methods in both image and video matting benchmarks.

To conclude the introduction, the main contributions of this paper can be summarized as follows:

(1) To overcome the limitations of the previous works, we develop a well-founded sampling strategy, which rely on a recently proposed sparse subset selection technique [33], to select a small set of foreground and background samples that best explain the unknown pixels.

(2) We design a new dissimilarity measure between two samples based on KL-divergence between the distributions of the features extracted in the vicinity of the samples. This measure is utilized in both selecting the representative samples and finding the best (F, B) pair for an unknown pixel.

(3) We provide compelling qualitative and quantitative results on a benchmark dataset of images [46] that demonstrate substantial improvements in the estimated alpha mattes upon current state-of-the-art methods.

(4) We extend our sampling strategy to compute alpha mattes for video sequences by incorporating motion information in an effortless way. Besides, we demonstrate the effectiveness of our video matting method on the recently proposed video matting benchmark dataset [45].

(5) We show the feasibility of our method for hard-to-handle case of sparse user inputs on a number of images.

II. Proposed Approach

In this study, we build upon a recent work by Elhamifar et al. [33], and address the sampling process in image and video matting as a sparse subset selection problem. In particular, we find a few representative pixels for the known foreground and background regions solely based on pairwise dissimilarities between the known and unknown pixels. As in other sampling-based approaches, in our formulation, the dissimilarity measure used in comparing two samples is of great importance since it directly affects the quality of selected samples. As we mentioned earlier, another contribution of this study is a new dissimilarity measure which is based on KL-divergence between feature distributions. In the following, we begin with the definition of our dissimilarity measure, and then discuss the details of the proposed algorithm. The steps of the algorithm involves collecting foreground and background color samples from known pixels via sparse subset selection, then we define an objective function to find the best (F, B) pair for an unknown pixel according to linear composition equation. After that, we explain how we extend defined sampling strategy to video matting by utilizing temporal information.

A. Dissimilarity Between Two Samples

Sampling-based approaches generally consider very simple measures which depend on chromatic and/or spatial similarities [1]–[4]. The only exceptions are [14], [15], which also employ some texture similarity measures. Unlike those measures, here, we consider a statistical data representation and propose to use an information-theoretic approach. In particular, our measure depends on a parametric version of the Kullback-Leibler (KL) Divergence [47], a well-known non-symmetric measure of the difference between two probability distributions in information theory, which we describe below. We note that KL-Divergence was used in a different way for video matting previously in [27]

Given an input image, we extract a 9-dimensional feature vector ϕ for each pixel as follows:

$$\phi(x, y) = \begin{bmatrix} x & y & r & g & b & |I_x| & |I_y| & |I_{xx}| & |I_{yy}| \end{bmatrix}^\top \quad (3)$$

with (x, y) denoting the pixel location, $I = [r\ g\ b]$ representing the pixel values of the RGB color space, and I_x, I_y, I_{xx}, I_{yy} respectively corresponding to the first and second-order derivatives of the image intensities, estimated via the filters $[-1\ 0\ 1]$ and $[-1\ 2\ -1]$ in horizontal and vertical directions.

Next we group the pixels into perceptually meaningful atomic regions using the SLIC algorithm [48]. The motivation behind this step is two folds. First, we use mean color of each foreground or background superpixel to reduce the sample space over which the representative samples are determined.

Second, extracting these superpixels helps us to describe a pixel by means of the characteristics of its neighboring pixels, which provides a source of contextual information.

Let s_p and s_q respectively denote two superpixels. Then, one can use the KL-divergence to measure the distance between s_p and s_q by considering the corresponding feature distributions P and Q as

$$D_{KL}\left(P\|Q\right) = \int_{-\infty}^{\infty} p(x) \ln \frac{p(x)}{q(x)} \mathrm{d}x \tag{4}$$

In our formulation, we assume that each feature distribution can be modeled through a multivariate normal distribution such that $P \sim \mathcal{N}_{s_p} = \mathcal{N}\left(\mu_p, \Sigma_p\right)$. Here, $p(x)$ and $q(x)$ respectively denote these probability density functions of P and Q. Then, the KL-Divergence between two superpixels s_p and s_q is described as follows:

$$\begin{aligned} D_{KL}(\mathcal{N}_{s_p}\|\mathcal{N}_{s_q}) =& \frac{1}{2}\Bigg(\mathrm{tr}\left(\Sigma_q^{-1}\Sigma_p\right) + \ln\left(\frac{\det \Sigma_q}{\det \Sigma_p}\right) \\ &+ (\mu_q - \mu_p)^\top \Sigma_q^{-1}(\mu_q - \mu_p) - k\Bigg) \end{aligned} \tag{5}$$

with $k = 9$ denoting our feature dimension.

Note that the KL-divergence is not symmetric, hence we symmetrize it as follows to obtain a distance metric:

$$dist(s_p, s_q) = D_{KL}(\mathcal{N}_{s_p}\|\mathcal{N}_{s_q}) + D_{KL}(\mathcal{N}_{s_q}\|\mathcal{N}_{s_p}) \tag{6}$$

In measuring the dissimilarity between two superpixels s_p and s_q, we found that, instead of using the metric in Eq. (6), the dissimilarity measure derived below lead to better discrimination:

$$S(s_p, s_q) = \frac{1}{dist(s_p, s_q) + \epsilon} \tag{7}$$

$$d(s_p, s_q) = 1 - min(S(s_p, s_q), 1) \tag{8}$$

where we take $\epsilon = 0.5$ in the experiments.

In Figure 2, we qualitatively verify the effectiveness of our statistical dissimilarity measure over using only the mean color values of the superpixels. For a given input image, we compute the pairwise dissimilarities between the superpixels extracted from the known foreground and background, and unknown regions and then these values are projected to a 2-dimensional space using t-SNE [43]. As can be seen, the proposed KL-divergence based dissimilarity measure provides better discrimination than simply using color distance.

B. Sampling via Sparse Subset Selection

Our strategy to obtain representative samples of known foreground and background regions to encode unknown region is inspired by the recently proposed Dissimilarity-based Sparse Subset Selection (DS3) algorithm [33], which formulate subset selection as a row-sparsity regularized trace minimization problem and presents a convex optimization framework to solve it. Suppose we use $\mathbf{K}$ and $\mathbf{U}$ to represent the set of superpixels extracted from the known foreground (f) and background (b), and unknown (u) regions, with $N = N_f + N_b$ and M elements, respectively:

$$\begin{aligned} \mathbf{K} &= \{s_1^f, ..., s_{N_f}^f, s_1^b, ..., s_{N_b}^b\} \\ \mathbf{U} &= \{s_1^u, ..., s_M^u\} \end{aligned} \tag{9}$$

Assume that the pairwise dissimilarities $\{d_{ij}\}_{i=1,\ldots,N}^{j=1,\ldots,M}$ between superpixels of known region $\mathbf{K}$ and unknown region $\mathbf{U}$ are computed using the dissimilarity measure defined in Eq. (8)[1], and arranged into a matrix form as

$$\mathbf{D} = \begin{bmatrix} \mathbf{d}_1^\top \\ \vdots \\ \mathbf{d}_N^\top \end{bmatrix} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1M} \\ \vdots & \vdots & & \vdots \\ d_{N1} & d_{N2} & \cdots & d_{NM} \end{bmatrix} \in \mathbb{R}^{N \times M} \tag{10}$$

where the entries d_{ij} signifies how well the superpixel i represents the superpixel j, the smaller the value, the higher the degree of representativeness.

According to the method described in [33], in order to find a sparse set of samples of $\mathbf{K}$ that well represents $\mathbf{U}$, one can introduce a matrix of variables $\mathbf{P} \in \mathbb{R}^{N \times M}$ as

$$\mathbf{P} = \begin{bmatrix} \mathbf{p}_1^\top \\ \vdots \\ \mathbf{p}_N^\top \end{bmatrix} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1M} \\ \vdots & \vdots & & \vdots \\ p_{N1} & d_{N2} & \cdots & p_{NM} \end{bmatrix} \tag{11}$$

whose each entry $p_{ij} \in [0, 1]$ is associated to d_{ij} and denote the probability of superpixel i being a representative for superpixel j. Then, the problem can be formulated as the following trace minimization problem regularized by a row-sparsity term:

$$\begin{aligned} \min_{\mathbf{P}} \quad & \gamma \, \|\mathbf{P}\|_{1,\infty} + \mathrm{tr}(\mathbf{D}^\top \mathbf{P}) \\ \mathrm{s.\,t.} \quad & \mathbf{1}^\top \mathbf{P} = \mathbf{1}^\top, \mathbf{P} \geq 0 \end{aligned} \tag{12}$$

where the first term $\|\mathbf{P}\|_{1,\infty} \triangleq \sum_i \|\mathbf{p}_i\|_\infty$ penalizes the size of the representative set, the second term $\mathrm{tr}(\mathbf{D}^\top \mathbf{P}) = \sum_{ij} d_{ij} p_{ij}$ simply measures the total encoding cost, and the parameter γ provides a trade-off between number of samples and encoding quality where smaller values of γ will lead to more number of representative samples. An optimal solution $\mathbf{P}^*$ can be found very efficiently using an Alternating Direction Method of Multipliers (ADMM) approach [33], in which the indices from the nonzero rows of the solution $\mathbf{P}^*$ give us the selected samples of foreground and background superpixels, where we use the mean colors of these superpixels as the candidate set of foreground F and background B colors.

Figure 3 shows the samples obtained with our sparse sampling strategy on an illustrative image. As it can be seen, the proposed approach allows robust selection of a small set samples from the known regions where the selected samples are the samples amongst the ones that best represent the unknown regions. Hence, as compared to the existing sampling based models, we employ less number of samples to determine the alpha matte values of the unknown pixels.

C. Selecting The Best (F, B) Pair

As compared to local sampling methods for image matting, which only collect samples near a given unknown pixel, employing a global scheme, such as ours, has the advantage of not missing any true samples if they are not located in the

[1] We note that the approach is quite general in that it could work with dissimilarities which are asymmetric or violate the triangle inequality.

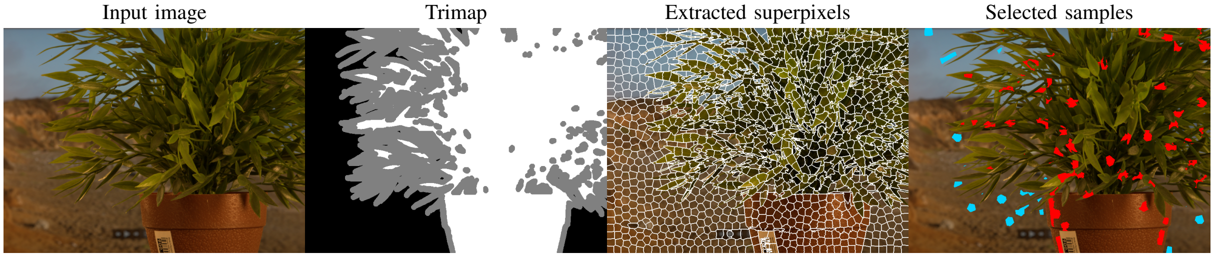

Fig. 3. Sampling via sparse subset selection. Candidate foreground and background samples are shown in red and blue, respectively.

vicinity of the unknown pixel. In some cases, however, there is also a possibility that a local analysis may work better, especially when local samples are more strongly correlated with the unknown pixel. Hence, to get the best of both worlds, we decide to combine our global sparse sampling strategy with a local sampling scheme. Specifically, for a given unknown pixel, we enlarge the global candidate set to include 10 additional foreground and background samples which are selected from the spatially nearest boundary superpixels.

Once candidate foreground and background colors are sampled for an unknown pixel, we select the best foreground and background pair (F, B) and accordingly determine its alpha matte value. In order to identify the best pair, we define a goodness function that depends on four different measures, which are described in detail below. In particular, in our formulation, we adopt the previously suggested chromatic distortion C_u and spatial distance S_u measures [3], [4], [14], [16] and additionally propose two new contextual similarity measures T_u and R_u to better deal with color ambiguity.

For an unknown pixel u and a foreground-background pair (F_i, B_i), the chromatic distortion C_u measures how well the alpha matte $\hat{\alpha}$ estimated via Eq. (2) from (F_i, B_i) fit to the linear composite equation given by Eq. (1), and is defined as

$$C_u(F_i, B_i) = \exp(-\|I_u - (\hat{\alpha}F_i + (1-\hat{\alpha})B_i)\|) \tag{13}$$

where I_u denote the observed color of the unknown pixel u.

The spatial distance measure S_u quantifies the spatial closeness of the unknown pixel u to the sample pair (F_i, B_i) according to the distance between the coordinates of these pixels. Therefore, it favors selecting samples that are spatially close to the unknown pixel. It is simply defined as

$$S_u(F_i, B_i) = \exp\left(-\frac{\|u - f_i\|}{Z_F}\right) \cdot \exp\left(-\frac{\|u - b_i\|}{Z_B}\right) \tag{14}$$

where f_i and b_i respectively denote the spatial coordinates of the centers of the superpixels that are associated with the foreground and the background samples F_i and B_i. The scalars $Z_F = (1/n_F)\sum_{k=1}^{n_F}\|u - f_k\|$ and $Z_B = (1/n_B)\sum_{k=1}^{n_B}\|u - b_k\|$ are used as scaling factors, which correspond to the mean spatial distance from the unknown pixel u to all foreground samples F with n_F elements and all background samples B with n_B elements, respectively.

One of the great challenges in image matting is the color ambiguity problem which arises when the foreground and background have similar colors. As most of the matting studies consider pixel based similarities in comparing samples, they generally fail to resolve this ambiguity and incorrectly recognize an unknown foreground pixel as background or vice versa. To account for this, we introduce the following two additional local contextual similarity measures T_u and R_u, which both exploit the similarity function defined in Eq. (7).

The first measure T_u specifies the compatibility of the unknown pixel with the selected foreground and background samples, computed by means of their statistical feature similarities, and it provides a bias towards those pairs (F_i, B_i) that have local contexts similar to that of the unknown pixel, and is formulated as

$$T_u(F_i, B_i) = S(s_{F_i}, s_u) + S(s_{B_i}, s_u) \tag{15}$$

where s_{F_i}, s_{B_i}, and s_u respectively denote the superpixels associated with the corresponding foreground and background samples and the unknown pixel.

The second measure R_u corresponds to a variant of the robustness term in [1], which builds upon the assumption that for any mixed pixel whose color is affected by both the foreground and the background, the true background and foreground colors have similar feature statistics, calculated over the corresponding superpixels. Thus, it favors the selection of the foreground and the background samples that have similar contexts, and is defined as

$$R_u(F_i, B_i) = S(s_{F_i}, s_{B_i}). \tag{16}$$

Putting these four measures together, we arrive at the following objective function to determine the best (F, B) pair:

$$\begin{aligned} O_u(F_i, B_i) = C_u(F_i, B_i)^c \cdot S_u(F_i, B_i)^s \cdot \\ T_u(F_i, B_i)^t \cdot R_u(F_i, B_i)^r, \end{aligned} \tag{17}$$

where c, s, t, r are weighting coefficients, representing the contribution of the corresponding terms to the objective function. Empirically, we observed that that the color distortion C_u and the contextual similarity measure T_u are more distinguishing than others, and thus we set the coefficients as $c = 2, s = 0.5, t = 1, r = 0.5$. Brute-force optimization is done on the objective function in Eq.(17) to select the best background and foreground color samples.

D. Pre- and Post-Processing

Motivated by recent sampling based matting studies [4], [16], we apply some pre- and post-processing steps. First, before selecting the best (F, B) sample pairs, we expand known regions to unknown regions by adopting the preprocessing step used in [4], [16]. Specifically, we consider an unknown pixel u as a foreground pixel if the following condition is satisfied for a foreground pixel $f \in F$:

$$(D(I_u, I_f) < E_{thr}) \wedge (\|I_u - I_f\| \leq (C_{thr} - D(I_u, I_f)), \quad (18)$$

where $D(I_u, I_f)$ and $\|I_u - I_f\|$ are the spatial and the chromatic distances between the pixels u and f, respectively, and f, and E_{thr} and C_{thr} are the corresponding thresholds which are all empirically set to 9. Similarly, an unknown pixel u is taken as a background pixel if a similar condition is met for a background pixel $b \in B$.

Second, as a post-processing, we perform smoothing on the estimated alpha matte by adopting a modified version of the Laplacian matting model [7] as suggested in [2]. That is, we determine the final alpha values α^* by solving the following global minimization problem:

$$\begin{aligned}\alpha^* = \arg\min_{\alpha} \alpha^\top L \alpha + \lambda(\alpha - \hat{\alpha})^\top \Lambda(\alpha - \hat{\alpha}) \\ + \delta(\alpha - \hat{\alpha})^\top \Delta(\alpha - \hat{\alpha})\end{aligned} \quad (19)$$

where the data term imposes the final alpha matte to be close to the estimated alpha matte $\hat{\alpha}$ from Eq. (2), and the matting Laplacian L enforces local smoothing. The diagonal matrix Λ in the first data term is defined using the provided trimap such that it has values 1 for the known pixels and 0 for the unknown ones. The scalar λ is set to 100 so that it ensures no smoothing is applied to the alpha values of the known pixels. The second diagonal matrix Δ, on the other hand, is defined by further considering the estimated confidence scores in a way that it has values 0 for the known pixels and the corresponding confidence values $O_u(F, B)$ from Eq. (17) for the unknown pixels. The scalar δ here is set to 0.1 and determines the relative importance of the smoothness term which considers the correlation between neighboring pixels.

III. Extension To Video Matting

As discussed in the previous section, a successful color sampling method should overcome the color ambiguity problem which occurs when the foreground and the background have similar color distributions. Fortunately, in video matting, the temporal motion information in the video sequences provides extra information to disambiguate this problem in the presence of dissimilar motion patterns. In this regard, we extend proposed sampling strategy to exploit temporal information by addressing both the missing true samples and the color ambiguity problems to obtain more accurate alpha mattes.

A. Motion Aware Temporal Sampling

The similarity metric employed in our sampling approach is quite generic in that we can incorporate any visual feature including motion information in a fairly straightforward way. In our case, we extend the feature vector ϕ in Eq. (3) with the optical flow vectors obtained by [49], as follows:

$$\phi(x, y) = \begin{bmatrix} x & y & r & g & b & |I_x| & |I_y| & |I_{xx}| & |I_{yy}| & v_x & v_y \end{bmatrix}^\top \quad (20)$$

where (v_x, v_y) are optical flow vectors. By doing so, we let the proposed KL-divergence based dissimilarity measure consider contextual motion along with color and orientation.

Such a feature distribution in a local region allows the proposed dissimilarity measure given in Eq. (5) to discriminate color, texture and motion information within our sparse subset selection phase (Eq. (12)). We further use the objective function given in Eq. (17) to select the best (F, B) pair for each unknown pixel. This objective function is also used as a confidence value in the alpha refinement step that will be explained in the following section. As a result, motion information is involved in each step of our video matting method.

Sampling from a single frame can be insufficient due to various types of changes between the video frames such as changes in the illumination, occlusion and changing topology. For this reason, we expand the sampling space in Eq. (9) by including the known foreground and background superpixels from both the previous frame $t-1$ and the current frame t as:

$$\begin{aligned}&\mathbf{K} = \{s_1^{f^{t-1}}, ..., s_{N_f^{t-1}}^{f^{t-1}}, s_1^{b^{t-1}}, ..., s_{N_b^{t-1}}^{b^{t-1}}, s_1^{f^t}, ..., s_{N_f^t}^{f^t}, s_1^{b^t}, ..., s_{N_b^t}^{b^t}\} \\ &\mathbf{U} = \{s_1^{u^t}, ..., s_{M^t}^{u^t}\}\end{aligned} \quad (21)$$

This definition extends the dissimilarity matrix D between the unknown and known superpixels in Eq. (10) to include the elements $\{d_{ij}\}_{i=1,...,N}^{j=1,...,M}$, where M is the number of unknown superpixels and $N = N_f^{t-1} + N_b^{t-1} + N_f^t + N_b^t$ is the number of known foreground and background superpixels extracted from the previous frame $t-1$ and the current frame t. After constructing the dissimilarity matrix and related probability matrix P in Eq. (11), we solve Eq. (12) to pick up the representative superpixels and accordingly the color samples.

Similar to image matting, we enrich global samples with local samples from the current frame. We apply the same procedure that we used in image matting to select the best (F, B) pair but note that the motion information is incorporated into the feature vectors so that the objective function now contains temporal information via our KL-divergence based dissimilarity measure. Figure 4 shows the temporal samples obtained by our modified sampling strategy. The representative samples for the brown background are chosen from the previous frame as the corresponding region becomes occluded by the foreground object in the current frame.

B. Alpha Refinement

After the alpha matte is estimated based on the selected color samples using Eq. 2, we further refine it by post-processing. For this purpose, some video matting methods [26], [28], [30] employ a 3D matting Laplacian defined over a multi-frame neighborhood by warping the neighboring frames to current frame via optical flow. It provides a better temporal coherency as compared to the standard matting Laplacian [7]. However, we observe that obtaining better temporal coherences might worsen the spatial accuracy due to inaccurate estimation of optical flows. Hence, in our experiments, we only consider the standard matting Laplacian that we extend with additional motion confidences.

$$\begin{aligned}\alpha^t = \arg\min_{\alpha^t} {\alpha^t}^\top L^t \alpha^t + \lambda(\alpha^t - \hat{\alpha}^t)^\top \Lambda^t(\alpha^t - \hat{\alpha}^t) \\ + \delta(\alpha^t - \hat{\alpha}^t)^\top \Delta^t(\alpha^t - \hat{\alpha}^t)\end{aligned} \quad (22)$$

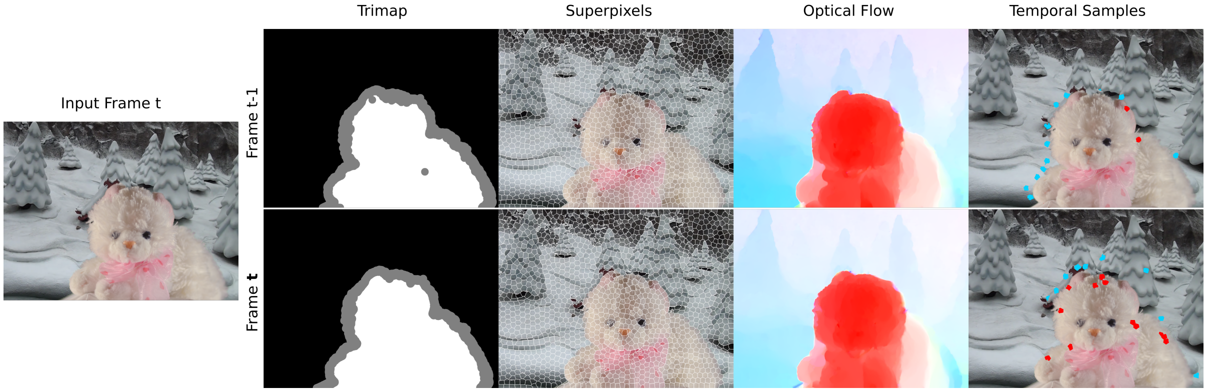

Fig. 4. Temporal sampling using sparse subset selection. Candidate foreground and background samples, which are respectively shown in red and blue color, come from the previous frame $t-1$ and the current frame t.

where L^t is the matting Laplacian, and the other terms are the data fidelity terms. The difference with Eq. (19) lies in the diagonal matrix Δ^t. Specifically, for each unknown pixel, it has the confidence value $O_u^t(F^t, B^t)$ estimated from the modified motion-aware sampling scheme, and 0 for the remaining known pixels at the frame t.

IV. Experiments

We evaluate our alpha matting approach on benchmark datasets used for evaluating image [46] and video [45] matting. First, we conduct an extensive evaluation of the proposed sampling strategy for image matting by providing qualitative and quantitative results and investigate the effects of the parameters in detail. Second, we compare our extension to video matting against state-of-the-art video matting methods both qualitatively and in terms of a set of spatial and temporal quality metrics. Next, we demonstrate that our sampling strategy can also cope with sparse user inputs, beating some recent methods especially proposed for this kind of input. Lastly, we argue the computational complexity of proposed approach.

A. Image Matting

Image matting benchmark dataset [46] contains 35 natural images and each image has a foreground object with different degrees of translucency or transparency. Among those images, 27 of them constitute the training set where the groundtruth alpha mattes are available. On the other hand, the remaining 8 images are used for the actual evaluation, whose ground truth alpha mattes are hidden from the public to prevent parameter tuning. In addition, for each test image, there are three matting difficulty levels that respectively correspond to small, large and user trimaps. To quantitatively evaluate our approach, in the experiments, we consider three different metrics, namely, the mean square error (MSE), the sum of absolute differences (SAD) and the gradient error. We do not report connectivity scores as it is argued in [46] that it is not a robust measure.

Effect of γ parameter. Fig. 5 shows that the average MSE values over all the training images and all trimaps do not vary much for different values of γ. These results seem to be consistent with the theoretical analysis in [33] that for a proper range of values, the DS3 algorithm that we utilize in sampling is guaranteed to find representative samples from all groups when there is a mutual relationship between known and unknown sets. In the remaining experiments, γ is set to 0.025 as it provides the minimum MSE value for the training set.

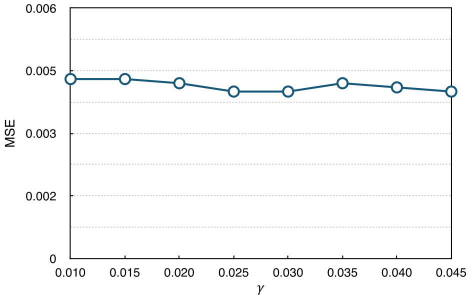

Fig. 5. Effect of γ parameter on the performance. Plot shows average MSE values over all training images and all trimaps.

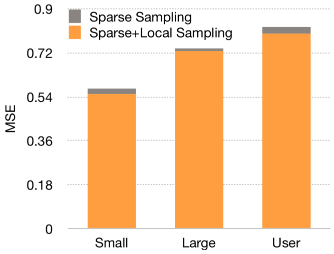

Fig. 6. Effect of including local samples to the representative set obtained with the proposed sparse sampling scheme. Plot shows average MSE values over all test images for three types of trimaps.

Effect of local samples. In Fig. 6, we show the effect of including local samples from boundary to the candidate set found by the proposed sparse sampling scheme. The numbers in the plot refer to the MSE errors averaged over all test images. For each trimap type, adding some closest boundary pixels further improves the performance. The smallest gain is in the large trimaps since having more number of known pixels helps our sparse sampling method to better exploit the associations between the known and unknown regions, eliminating the need for local samples.

Comparison with the state-of-the-art. Table I presents the quantitative comparison of our approach and nine best

TABLE I
EVALUATION OF MATTING METHODS ON THE BENCHMARK DATASET [46] WITH THREE TRIMAPS ACCORDING TO SAD, MSE AND GRADIENT ERROR.

Sum of Absolute Differences					Mean Square Error					Gradient Error				
Method	overall rank	avg. small rank	avg. large rank	avg. user rank	Method	overall rank	avg. small rank	avg. large rank	avg. user rank	Method	overall rank	avg. small rank	avg. large rank	avg. user rank
1. DCNN Matting	**2.4**	3.4	1.1	2.6	1. DCNN Matting	**3**	3.8	1.5	3.9	1. DCNN Matting	**6.3**	8.1	5.3	5.6
2. CSC Matting	**9.1**	13	5.3	9.1	2. LNSP Matting	**8**	5.6	7.4	11.1	2. Graph-based Sparse Matting	**9.1**	7.1	7.8	12.5
3. LNSP Matting	**9.3**	6	8.9	13	3. Proposed Method	**10.2**	9.5	9.1	12	3. Proposed method	**9.3**	7.4	8	12.5
4. Graph-based sparse matting	**9.6**	9.9	10	9	4. CCM	**10.5**	12.8	10.9	7.9	4. LNSP Matting	**9.8**	7.4	9.5	12.6
5. Proposed Method	**9.8**	7.4	9.5	12.6	5. Graph-based sparse matting	**10.9**	11.3	11.1	10.2	5. Comprehensive sampling	**10.5**	10.4	10.3	10.8
6. TSPS-RV Matting	**11.1**	9.6	10.1	13.6	6. TSPS-RV Matting	**11.3**	11.4	8.4	14.1	6. CCM	**12.4**	14.3	12.4	10.6
7. Iterative Transductive Matting	**11.8**	12.8	11.5	11.1	7. Comprehensive sampling	**11.4**	10.1	11.4	12.6	7. SVR Matting	**12.5**	14.1	13.9	9.4
8. Comprehensive sampling	**14.4**	8.1	12	14.3	8. SVR Matting	**11.7**	14.8	10.3	10.1	8. Sparse coded matting	**12.6**	13.6	11.9	12.4
9. SVR Matting	**12.3**	14.8	11.8	10.3	9. CW Color and Texture	**12.5**	12.1	13.5	11.9	9. Segmentation-based matting	**12.9**	15.5	11.4	11.6
10. CW Color and Texture	**12,3**	12.4	13	11.6	10. CSC Matting	**13.5**	16.6	8.1	15.9	10. Global Sampling Matting	**13**	12.4	14.6	12

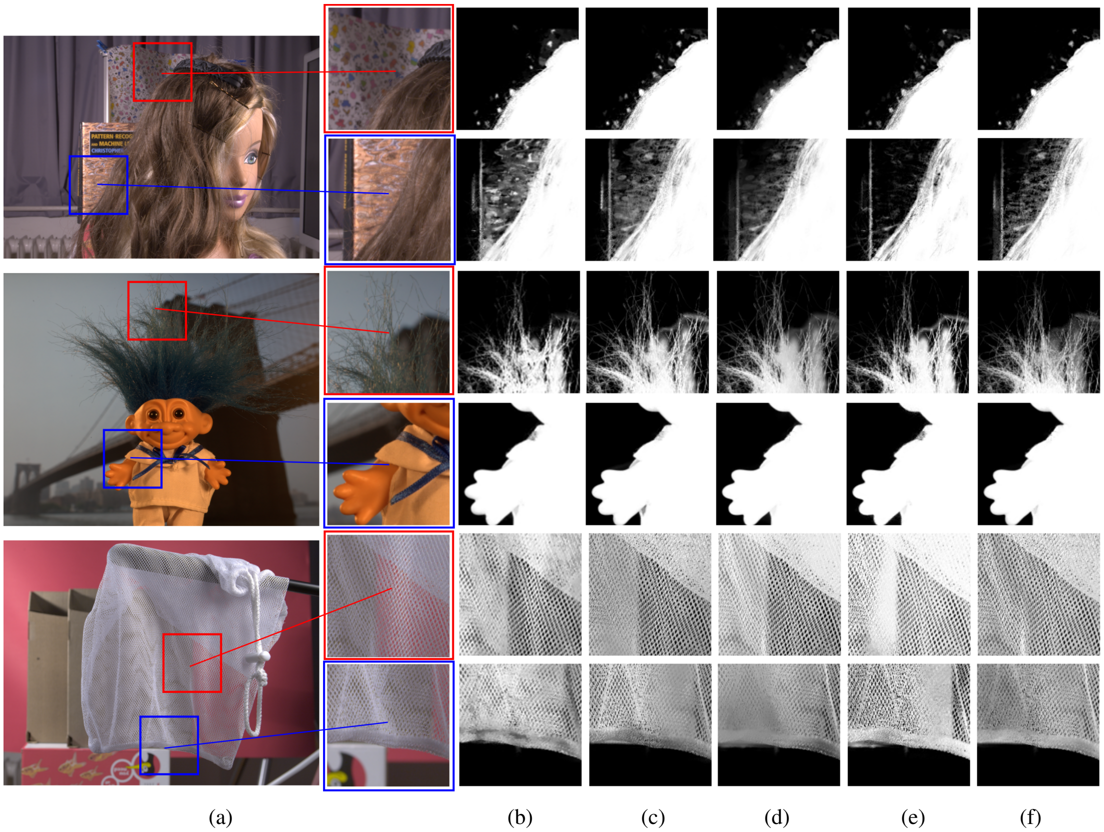

Fig. 7. Visual comparison of our approach with other sampling-based image matting methods. (a) Input image, (b) CWCT sampling [15], (c) Comprehensive sampling [4], (d) LNSP matting [17], (e) Sparse coded matting [16] and (f) Proposed approach.

performing matting algorithms on the benchmark hosted at *www.alphamatting.com* [46] where we report the average rankings over the test images according to SAD, MSE and gradient metrics for all three different types of trimap, and the overall ranks, computed as the average over all the images and for all the trimaps. Overall, our approach provides highly competitive results against the state-of-the-art methods. It ranked the third best with respect to the gradient error and mean square error and the fifth best for the sum of absolute differences. Especially, it outperforms all the existing sampling-based matting methods. Fig. 7 provides qualitative comparisons of our approach and the recent matting studies [4], [15]–[17] on the *doll*, *troll* and *net* images from the benchmark dataset.

Textured background. In the first row of Fig. 7, we show the ability of our approach to naturally handle textured backgrounds via the proposed KL-divergence based contextual measure. For the *doll* placed in front of a highly textured background, while other matting methods, including CWCT sampling [15] which employs an additional texture compatibility measure, tend to interpret some of the colored blobs in the background as foreground, our model produces a much more accurate alpha map.

Color ambiguity. When the foreground object and the background have similar color distributions, most matting studies suffer from the so-called color ambiguity and fail to provide reliable alpha values for the unknown pixels. Both second and third rows of Fig. 7 illustrate this issue where the colors of the book and the bridge in the background is very similar to those of the hairs of the *doll* and *troll*, respectively. For these examples, CWCT [15] and Comprehensive sampling [4] give inaccurate estimations whereas LNSP matting [17] oversmooths the foreground matte. Sparse coded matting [16] provides better results but misses some of the foreground details in the hairs. On the other hand, our method is able to achieve significantly better results, providing a more robust discrimination between the background and the foreground.

Missing samples. Previously proposed sampling-based matting methods typically employ certain assumptions while collecting samples from known regions but these assumptions might sometimes lead to missing true foreground and background colors for some unknown pixels. In the fourth row of Fig. 7, we demonstrate the effectiveness of our sparse sampling strategy on the *troll* image. While the other sampling based methods [4], [15], [16] incorrectly recognize the blue ribbon as mixed pixels, our algorithm successfully interprets it as a part of the foreground object. Likewise, the LNSP matting [17] produces an alpha map similar to ours as it uses a non-local smoothness prior in their formulation. If known regions do not include some color samples representing the colors from the unknown region, our method might not give accurate alpha matte results as seen for the *plastic bag* image.

Translucent foreground. Transparent or translucent objects pose another great challenge for matting as they make collecting true foreground color samples difficult. The last two rows of Fig. 7 show the results of two different regions from the *net* image in detail where such a foreground object exists. Due to the characteristics of the test image, all of the competing matting methods fail to differentiate background pixels from the foreground although the distributions of the background and foreground colors are well separated. In contrast, our approach produces a remarkably superior alpha matte.

B. Video Matting

For our video matting experiments, we use the very recently proposed video matting benchmark dataset [45] (*www.videomatting.com*) to evaluate our video matting results. This dataset includes 3 training sequences with available ground-truth maps, and 10 testing sequences with hidden ground-truth maps. Furthermore, for each video frame, 3 different trimaps are generated according to the size of the unknown region as narrow, medium and wide.

Effect of temporal sampling and motion features. Almost all video matting methods [22]–[24], [27]–[30] employ optical flow fields as motion features. However, as discussed in the related papers, optical flow estimation is not always perfect, which may deterioate the quality of extracted alpha mattes. We analyze the effect of our optical-flow based motion features on the training sequences of [45] (*Alex*, *Castle*, *Dmitriv*, see Fig. 8), where the ground truth alpha mattes are available. In Table II, we present the effects of motion and temporal sampling in terms of SSD (sum of squared distances) metric. As can be seen, motion and temporal sampling improve the alpha matte results for *Alex* and *Dmitriv* sequences, however for castle sequence worsen the results. The *Castle* sequence has a complex foreground object and a highly textured background region as compared to other sequences, and we think that these factors negatively affect the extracted optical flow vectors, and consequently the quality of our temporal sampling. In the remaining experiments, on the other hand, we report our results with both using motion features and temporal sampling since this setup provides the best results on the overall training sequences.

TABLE II
EFFECT OF SPATIAL SAMPLING (SINGLE FRAME) VS. TEMPORAL SAMPLING (TWO FRAMES) AND OPTICAL FLOW INFORMATION (OF) TO ALPHA MATTE RESULTS ON SSD SCORES

Video	Trimap	Spatial	Spatial+OF	Temporal+OF
Alex	Narrow	2.134	2.129	2.125
	Medium	1.740	1.707	1.689
	Wide	1.852	1.753	1.729
Castle	Narrow	6.736	7.098	7.158
	Medium	7.160	7.591	7.624
	Wide	7.671	8.150	8.208
Dmitriv	Narrow	1.918	1.916	1.917
	Medium	2.409	2.336	2.329
	Wide	2.734	2.662	2.652

Alex *Castle* *Dmitriv*

Fig. 8. Training sequences from the video matting benchmark dataset [45].

Comparison with other methods. Table III shows the quantitative evaluation of different matting methods on the video matting benchmark dataset [45]. Evaluation is carried out on the test sequences according to the quality metrics which highlight spatial accuracy and temporal coherency of the estimated alpha mattes. Specifically, SSDA (Sum of Squared Distances) error measure is used to evaluate the accuracy of the estimated alpha matte for each pixel. Two additional temporal-coherency metrics, which measure deterioration ratio of the alpha mattes over consequent frames, are used to test the temporal coherency. SSDdt measures the overall variation in the sum squared distances between the estimated alpha matte and the ground-truth for each pixel over the consecutive frames. MESSDdt measure is the generalized version of SSDdt, which additionally considers the optical flow information over video frames. Thus, it provides a more robust comparison of specifically the motion-aware matting methods.

Matting methods are sorted by the average ranking scores according to the accuracy metric SSDA and the temporal-coherency metrics SSDdt and MESSDdt. As can be seen from Table III, our method gives the best results for the spatial accuracy metric and highly competitive results for the temporal-coherency metrics. Here, it is worth noting that the motion-aware MESSDdt metric generally provides a more robust comparison of the methods than the SSDdt metric since the inaccurate optical flow estimations could worsen the quality of alpha mattes as we analyzed above.

In Fig. 9, we also give the qualitative comparisons of our approach against the other best-performing matting methods [7], [18] and the Refine Edge Tool of Adobe on the *Juneau* sequence, for the frames 45, 46 and 47. The zoomed regions demonstrate that our method is affected from the missing true color and the color ambiguity far less than the other methods since it inherently employs motion information in sampling and uses a sampling strategy extended to consider temporal information as well.

TABLE III
EVALUATION OF MATTING METHODS ON THE BENCHMARK DATASET [45] WITH THREE TRIMAPS ACCORDING TO SSDA, SSDDT AND MESSDDT.

SSDA					SSDdt					MESSDdt				
Method	overall rank	avg. narrow rank	avg. medium rank	avg. wide rank	Method	overall rank	avg. narrow rank	avg. medium rank	avg. wide rank	Method	overall rank	avg. narrow rank	avg. medium rank	avg. wide rank
1. Proposed Method	**2.3**	3.2	2.1	1.7	1.Learning Based	**2.8**	2.8	2.8	2.7	1.Learning Based	**2.3**	2.3	2.3	2.4
2. Learning Based	**2.7**	2.5	2.7	3	2. Closed Form	**3.1**	2.8	3.3	3.1	2. Closed Form	**2.7**	2.4	2.6	3
3. Comprehensive Sampling	**3.6**	4	3.3	3.4	3. Refine Edge	**3.4**	3.9	3.4	2.9	3. Proposed Method	**3.6**	3.9	3.5	3.4
4. Closed Form	**3.9**	3.9	3.7	4.1	4. Proposed Method	**4**	4.9	3.7	3.4	4. Refine Edge	**3.6**	4.3	3.7	2.8
5. Shared Matting	**5.3**	4.9	5.4	5.5	5. Comprehensive Sampling	**5**	4.8	4.9	5.2	5. Comprehensive Sampling	**5**	4.9	4.9	5.3
6. Robust Matting	**5.5**	4.5	5.6	6.4	6. Robust sampling	**5.2**	4.3	5.4	6	6. Shared Matting	**6.2**	5.7	6.3	6.6
7. Refine Edge	**5.5**	5.8	5.5	5.2	7. Shared Matting	**5.8**	5.6	5.9	6	7. Robust Matting	**6.4**	5.4	6.4	7.3
8. KNN Matting	**8.4**	8.9	8.6	7.8	8. KNN matting	**8.5**	8.8	8.5	8.2	8. KNN Matting	**7.6**	8.4	7.6	6.9
9. Bayesian Matting	**8.7**	7.9	8.9	9.3	9. Spectral Matting	**9.2**	9.4	9.2	9.1	9. Nonlocal Matting	**9.3**	9.9	9.4	8.5
10. Nonlocal Matting	**9.4**	9.7	9.6	9	10. Nonlocal Matting	**9.3**	9.6	9.3	9.1	10. Spectral Matting	**9.5**	9.7	9.4	9.3

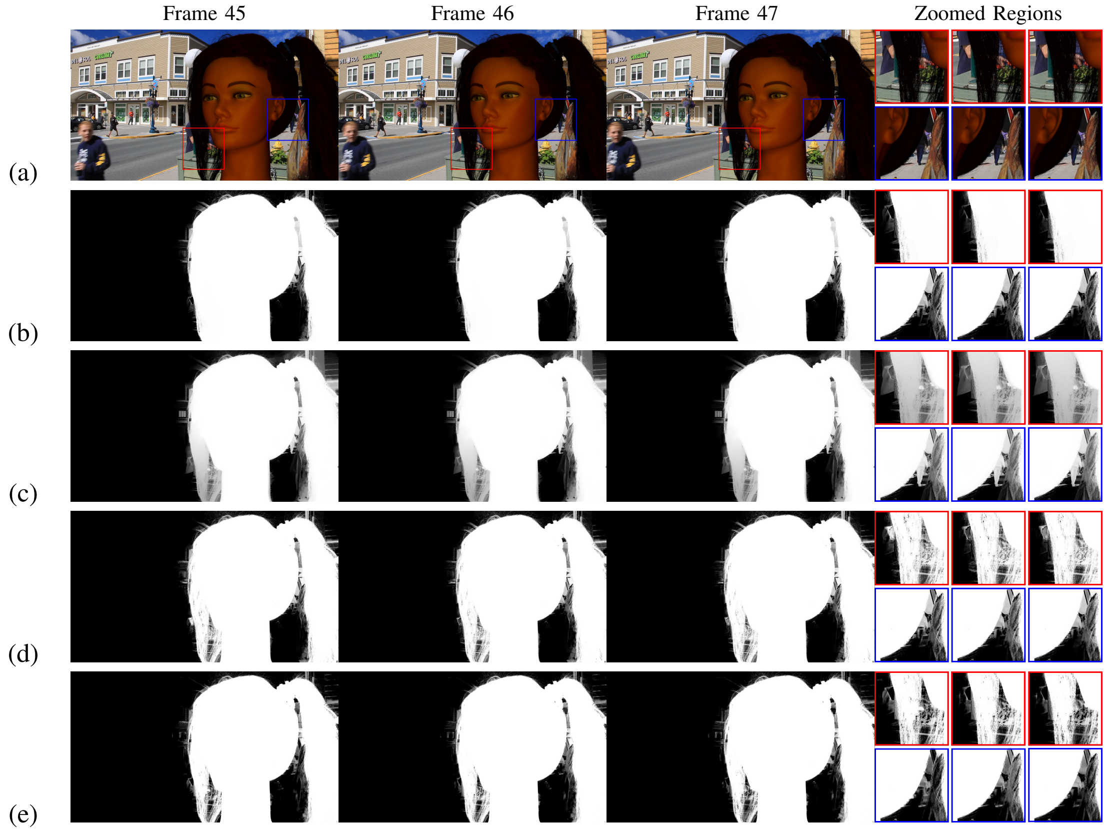

Fig. 9. Visual comparison between the proposed and the other best-performing video matting methods on the video benchmark dataset [45]. (a) Input video frames, (b) Closed Form [7], (c) Refine Edge Tool, (d) Learning Based [18] and (e) Proposed approach.

C. Sparse Scribbles

Our approach can also work with sparse user inputs since we do not make any spatial assumption while collecting color samples, which is the case for many previous sampling-based image matting methods [1]–[4]. More specifically, we apply our sampling strategy on the sparse user inputs by considering the superpixels that contain any user scribbled image pixel as the known superpixels and the others as the unknown superpixels. Consequently, we construct the dissimilarity matrix between known and unknown superpixels via our proposed KL-Divergence based dissimilarity measure and select the representative color samples from the known user scribbles using the sparse subset selection strategy described in Eq. (12).

We compare the performance of our method with KNN Matting [9] and Nonlocal Matting [8] methods which are both tailor-fit to work with sparse user inputs and with Closed Form [7] and Comprehensive sampling [4] methods. In Fig. 10, some image matting results along with the estimated MSE (Mean Square Error) scores are given. Our method in general produces better results than all these methods. This also demonstrates that the proposed sampling scheme for image matting is a generic and theoretically well-grounded sampling strategy for alpha matting problem.

D. Runtime Performance

In our work, we used the ADMM-based serial implementation of the DS3 method, but it is indeed highly parallelizable [33]. Overall, the runtime performance of our current implementation is better than Comprehensive sampling (CS) as our algorithm selects much less and more representative samples from the known regions, which significantly reduces runtime costs of the subsequent steps. For example, for the

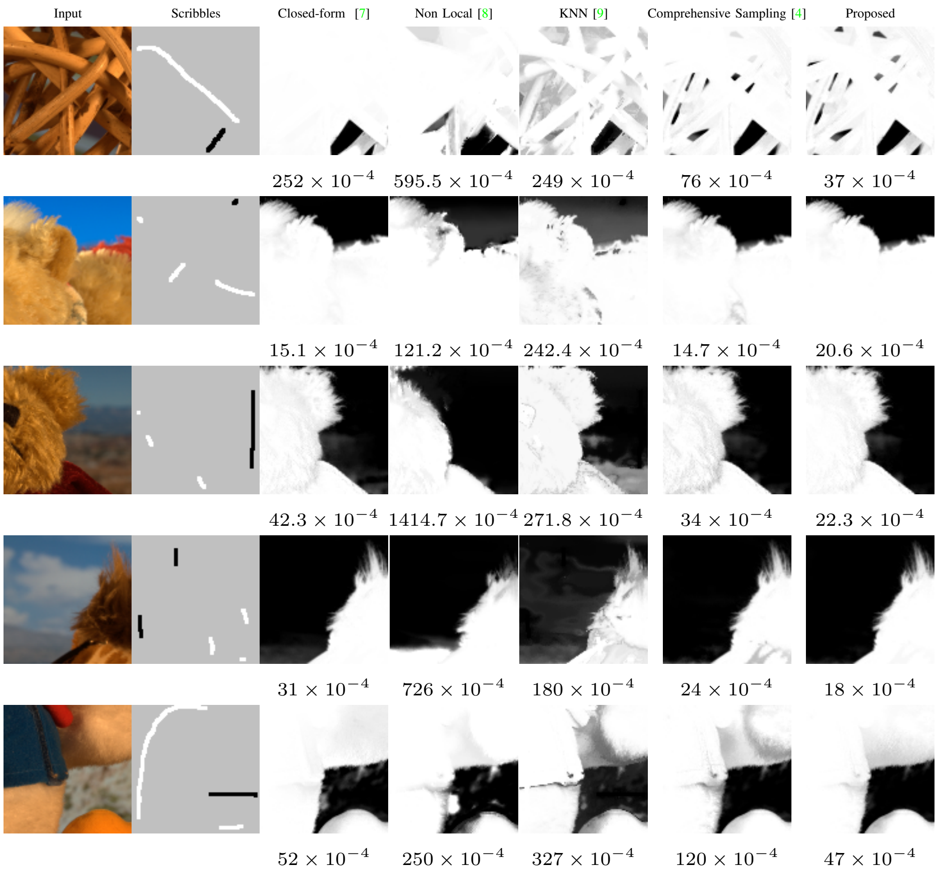

Fig. 10. Alpha matting results with scribble based user input and the corresponding MSE scores estimated over the input images.

doll, *donkey* and *elephant* images in [46], the average running times over all trimaps are 341 secs for our method, and 414 secs for CS, on a PC with an Intel Xeon 2GHz CPU.

V. Conclusion

In this paper, we developed a new and theoretically well-grounded sampling strategy for image matting and extended it to video matting. Rather than making assumptions about the possible locations of true color samples, or performing a direct clustering of all known pixels, our sampling scheme solves a sparse subset selection problem over known pixels to obtain a small set of representative samples that best explain the unknown pixels. This property also makes our sampling method directly applicable to sparse user inputs provided to estimate alpha matte. Moreover, it employs a novel KL-divergence based contextual measure in both collecting the candidate sample set and finding the best (F, B) pair for an unknown pixel. Our experiments on both image and video benchmark datasets clearly demonstrate that our approach is superior to existing sampling-based image and video matting methods and achieves state-of-the-art results.

Acknowledgment

This research was supported in part by The Scientific and Technological Research Council of Turkey (TUBITAK), Career Development Award 112E146.

References

[1] J. Wang and M. F. Cohen, “Optimized color sampling for robust matting,” in *CVPR*, 2007, pp. 1–8. 2, 3, 4, 6, 11

[2] E. S. Gastal and M. M. Oliveira, “Shared sampling for real-time alpha matting,” *Computer Graphics Forum*, vol. 29, no. 2, pp. 575–584, 2010. 1, 2, 4, 7, 11

[3] K. He, C. Rhemann, C. Rother, X. Tang, and J. Sun, “A global sampling method for alpha matting,” in *CVPR*, 2011, pp. 2049–2056. 1, 2, 4, 6, 11

[4] E. Shahrian, D. Rajan, B. Price, and S. Cohen, “Improving image matting using comprehensive sampling sets,” in *CVPR*, 2013, pp. 636–643. 1, 2, 3, 4, 6, 9, 10, 11, 12

[5] J. Sun, J. Jia, C.-K. Tang, and H.-Y. Shum, “Poisson matting,” *ACM Trans. Graph.*, vol. 23, no. 3, pp. 315–321, 2004. 1

[6] L. Grady, T. Schiwietz, S. Aharon, and R. Westermann, “Random walks for interactive alpha-matting,” in *VIIP*, 2005, pp. 423–429. 1

[7] A. Levin, D. Lischinski, and Y. Weiss, “A closed-form solution to natural image matting,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 228–242, 2008. 1, 3, 7, 10, 11, 12

[8] P. Lee and Y. Wu, “Nonlocal matting,” in *CVPR*, 2011, pp. 2193–2200. 1, 3, 11, 12

[9] Q. Chen, D. Li, and C.-K. Tang, “KNN matting,” in *CVPR*, 2012, pp. 869–876. 1, 3, 11, 12

[10] Y. Shi, O. C. Au, J. Pang, K. Tang, W. Sun, H. Zhang, W. Zhu, and L. Jia, “Color clustering matting,” in *ICME*, 2013, pp. 1–6. 1

[11] B. He, G. Wang, C. Shi, X. Yin, B. Liu, and X. Lin, “Iterative transductive learning for alpha matting,” in *ICIP*, 2013, pp. 4282–4286. 1

[12] Y.-Y. Chuang, B. Curless, D. H. Salesin, and R. Szeliski, “A Bayesian approach to digital matting,” in *CVPR*, vol. 2, 2001, pp. II–264. 1, 3

[13] J. Wang and M. F. Cohen, “An iterative optimization approach for unified image segmentation and matting,” in *ICCV*, vol. 2, 2005, pp. 936–943. 1

[14] E. Shahrian and D. Rajan, “Weighted color and texture sample selection for image matting,” in *CVPR*, 2012, pp. 718–725. 1, 2, 4, 6

[15] E. Varnousfaderani and D. Rajan, "Weighted color and texture sample selection for image matting," *IEEE Trans. Image Processing*, vol. 22, no. 11, pp. 4260–4270, Nov 2013. 1, 4, 9, 10

[16] J. Johnson, D. Rajan, and H. Cholakkal, "Sparse codes as alpha matte," in *BMVC*, 2014. 1, 3, 6, 9, 10

[17] X. Chen, D. Zou, S. Z. Zhou, Q. Zhao, and P. Tan, "Image matting with local and nonlocal smooth priors," in *CVPR*, 2013, pp. 1902–1907. 1, 9, 10

[18] Z. Zhang, Q. Zhu, and Y. Xie, "Learning based alpha matting using support vector regression," in *ICIP*, 2012, pp. 2109–2112. 1, 10, 11

[19] D. Cho, Y.-W. Tai, and I. Kweon, "Natural image matting using deep convolutional neural networks," in *ECCV*, 2016, pp. 626–643. 1

[20] Y. Li, J. Sun, and H.-Y. Shum, "Video object cut and paste," *ACM Trans. Graphics*, vol. 24, no. 3, pp. 595–600, 2005. 1, 3

[21] J. Wang, P. Bhat, R. A. Colburn, M. Agrawala, and M. F. Cohen, "Interactive video cutout," *ACM Trans. Graphics*, vol. 24, no. 3, pp. 585–594, 2005. 1, 3

[22] X. Bai, J. Wang, D. Simons, and G. Sapiro, "Video snapcut: robust video object cutout using localized classifiers," *ACM Trans. Graphics*, vol. 28, no. 3, p. 70, 2009. 1, 3, 10

[23] Y.-Y. Chuang, A. Agarwala, B. Curless, D. H. Salesin, and R. Szeliski, "Video matting of complex scenes," in *ACM Trans. Graphics*, vol. 21, no. 3, 2002, pp. 243–248. 1, 3, 10

[24] X. Bai, J. Wang, and D. Simons, "Towards temporally-coherent video matting," *Computer Vision/Computer Graphics Collaboration Techniques*, pp. 63–74, 2011. 1, 3, 10

[25] S.-Y. Lee, J.-C. Yoon, and I.-K. Lee, "Temporally coherent video matting," *Graphical Models*, vol. 72, no. 3, pp. 25–33, 2010. 1

[26] Z. Tang, Z. Miao, and Y. Wan, "Temporally consistent video matting based on bilayer segmentation," in *ICME*, 2010, pp. 370–375. 1, 3, 7

[27] I. Choi, M. Lee, and Y.-W. Tai, "Video matting using multi-frame nonlocal matting Laplacian," in *ECCV*, 2012, pp. 540–553. 1, 3, 4, 10

[28] Z. Tang, Z. Miao, Y. Wan, and D. Zhang, "Video matting via opacity propagation," *The Visual Computer*, vol. 28, no. 1, pp. 47–61, 2012. 1, 3, 7, 10

[29] D. Li, Q. Chen, and C.-K. Tang, "Motion-aware knn laplacian for video matting," in *ICCV*, 2013, pp. 3599–3606. 1, 3, 10

[30] E. Shahrian, B. Price, S. Cohen, and D. Rajan, "Temporally coherent and spatially accurate video matting," in *Computer Graphics Forum*, vol. 33, no. 2. Wiley-Blackwell, 2014, pp. 381–390. 1, 3, 7, 10

[31] Q. Zhu, L. Shao, X. Li, and L. Wang, "Targeting accurate object extraction from an image: A comprehensive study of natural image matting," *IEEE Trans. Neural Networks and Learning Systems*, vol. 26, no. 2, pp. 185–207, 2015. 1

[32] J. Wang and M. Cohen, "Image and video matting: a survey," *Foundations and Trends in Computer Graphics and Vision*, vol. 3, no. 2, pp. 97–175, 2007. 1

[33] E. Elhamifar, G. Sapiro, and S. S. Sastry, "Dissimilarity-based sparse subset selection," *arXiv preprint arXiv:1407.6810*, 2014. 2, 4, 5, 8, 11

[34] E. Elhamifar, G. Sapiro, and R. Vidal, "Finding exemplars from pairwise dissimilarities via simultaneous sparse recovery," in *NIPS*, 2012, pp. 19–27. 2

[35] Y. Mishima, "Soft edge chroma-key generation based upon hexoctahedral color space," Oct. 11 1994, US Patent 5,355,174. 2

[36] A. Berman, A. Dadourian, and P. Vlahos, "Method for removing from an image the background surrounding a selected object," Oct. 17 2000, US Patent 6,134,346. 2

[37] C. Rhemann, C. Rother, and M. Gelautz, "Improving color modeling for alpha matting," in *BMVC*, 2008. 2

[38] J. Johnson, E. S. Varnousfaderani, H. Cholakkal, and D. Rajan, "Sparse coding for alpha matting," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3032–3043, 2016. 3

[39] X. Feng, X. Liang, and Z. Zhang, "A cluster sampling method for image matting via sparse coding," in *ECCV*, 2016, pp. 204–219. 3

[40] A. R. Smith and J. F. Blinn, "Blue screen matting," in *SIGGRAPH*, 1996, pp. 259–268. 3

[41] H.-Y. Shum, J. Sun, S. Yamazaki, Y. Li, and C.-K. Tang, "Pop-up light field: An interactive image-based modeling and rendering system," *ACM Trans. Graphics*, vol. 23, no. 2, pp. 143–162, 2004. 3

[42] D. Zou, X. Chen, G. Cao, and X. Wang, "Video matting via sparse and low-rank representation," in *ICCV*, 2015, pp. 1564–1572. 3

[43] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-SNE," *J. Machine Learning Research*, vol. 9, pp. 2579–2605, Nov 2008. 4, 5

[44] L. Karacan, A. Erdem, and E. Erdem, "Image matting with KL-divergence based sparse sampling," in *ICCV*, 2015, pp. 424–432. 3

[45] M. Erofeev, Y. Gitman, D. Vatolin, A. Fedorov, and J. Wang, "Perceptually motivated benchmark for video matting," in *BMVC*, 2015, pp. 99.1–99.12. 4, 8, 10, 11

[46] C. Rhemann, C. Rother, J. Wang, M. Gelautz, P. Kohli, and P. Rott, "A perceptually motivated online benchmark for image matting," in *CVPR*, 2009, pp. 1826–1833. 4, 8, 9, 12

[47] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, pp. 79–86, 1951. 4

[48] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012. 4

[49] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, "High accuracy optical flow estimation based on a theory for warping," in *ECCV*, 2004, pp. 25–36. 7

Levent Karacan received his B.Sc. degree in Computer Engineering from Erciyes University, Turkey in 2011, his M.Sc. degree in Computer Engineering from Hacettepe University, Ankara, Turkey in 2013, where he is currently pursuing the Ph.D degree in Computer Engineering. His current research interests include image and video processing, image editing and computational photography.

Aykut Erdem received his B.Sc. and M.Sc. degrees in Computer Engineering in 2001 and 2003 from Middle East Technical University, Ankara, Turkey. Upon receiving his Ph.D. degree in 2008, he worked as a post-doctoral researcher in the Computer Science Department of CaFoscari University of Venice, Italy from 2008-2010. In 2010, he joined Hacettepe University, Turkey, where he is now an Assistant Professor at the Department of Computer Engineering. His research interests currently focused on image matting, summarization of videos and large image collections, and integrating language and vision.

Erkut Erdem received his B.Sc. and M.Sc. degrees respectively in 2001 and 2003 from the Dept. of Computer Engineering, Middle East Technical University (METU), Turkey. After completing his Ph.D. at the METU in 2008, he continued his postdoctoral studies at Télécom ParisTech, Ecole Nationale Supérieure des Télécommunications, France between 2009 and 2010. He is an Assistant Professor at the Department of Computer Engineering, Hacettepe University, Turkey since 2014. His research interests include image editing and smoothing, visual saliency prediction and language and vision.